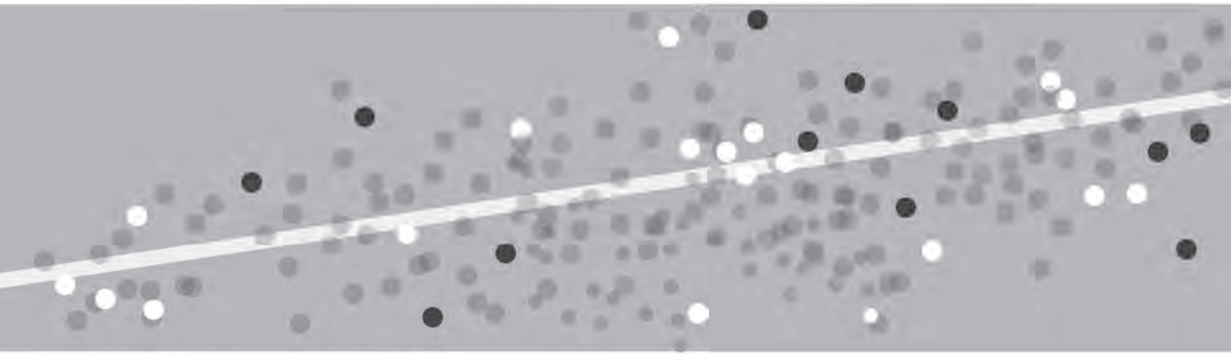
The SAGE Handbook of

Regression Analysis and Causal Inference



Edited by

Henning Best and Christof Wolf

SAGE reference

Los Angeles | London | New Delhi
Singapore | Washington DC

2015

15

Fixed-effects panel regression

Josef Brüderl and Volker Ludwig*

INTRODUCTION

Fixed-effects (FE) regression is a method that is especially useful in the context of causal inference (Gangl, 2010). While standard regression models provide biased estimates of causal effects if there are unobserved confounders, FE regression is a method that can (if certain assumptions are valid) provide unbiased estimates in this situation (other methods are instrumental variables regression and regression discontinuity; see Chapters 13 and 14 in this volume). Since unobserved confounders are ubiquitous in social science applications, FE regression should be standard in the toolkit of modern social research.

FE regression is most often used with panel data, and therefore the focus of this chapter will be on FE regression with panel data. However, before we begin, we want to place FE regression in a wider context. FE regression is not only applicable with panel data, it can be used with any kind of multi-level data. These are data where lower-level units are nested within higher-level groups (sometimes also called 'clusters'). A multi-level regression (see Chapter 7 in this volume) has to assume that there is neither unit-specific nor group-specific unobserved heterogeneity.[1] With non-experimental data, this assumption is often violated due to self-selection on the group level. However, the assumption of no unobserved heterogeneity can be weakened if the researcher uses FE models. A fixed-effects regression is specified on the level of the units and includes group-specific constants (the so-called 'fixed effects'). Because group-specific fixed effects wipe out all group-specific unobserved heterogeneity, FE models only require the assumption of no unit-specific unobserved heterogeneity. Thus, compared with standard regression models, FE models allow a causal effect to be identified under weaker assumptions. Obviously, this makes FE models attractive for social researchers undertaking causal analysis.

There are many examples of multi-level data where FE models have been used:

- Pupils nested within classes nested within schools: School class fixed effects wipe out unobserved heterogeneity on the class level (Legewie, 2012).
- Workers nested within firms (matched employer-employee data): Firm fixed effects control for unobservables on the firm level (Hinz and Gartner, 2005).
- Siblings nested within families: Family fixed effects wipe out unobservables on the family level (Arránz Becker et al., 2013).

- Individuals nested within countries: Country fixed effects control for all country-level heterogeneity.
- Repeated observations within individuals: Individual fixed effects control for all person-level heterogeneity.

The latter case describes an FE model for panel data: here we observe persons repeatedly over several panel waves. Panel data are especially useful for applying FE models because, due to their richness, they allow many relevant social science questions to be investigated. Therefore, the combination of panel data and FE modeling is especially promising.

The basic fixed-effects framework

Panel data are typically set up in long format. That is, the observations of each subject are ordered chronologically, and the time series (panels) of subjects are stacked below each other (pooled data). We refer to a setting where there are short time series, but many cross-sections ($N \to \infty$, T fixed). This is the typical situation when estimating an effect of some causal variable using data from an annual panel survey conducted in years $t = 1, \ldots, T$ for individuals $i = 1, \ldots, N$. In this case a single observation from a subject i is called a person-year. A leading case is estimation of the causal effect of an event that occurs during the life course (e.g. marriage). We assume that the outcome variable Y is continuous, while the K regressors $X_1, \ldots, X_K$ may be measured on any scale.

FE estimation builds on the error components model,

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i + \epsilon_{it}. \tag{15.1}$$

Here, y_{it} denotes the observed outcome of person i at time t, $\mathbf{x}_{it}$ is the $(1 \times K)$ vector of covariates of this person measured contemporaneously, and $\boldsymbol{\beta}$ is the corresponding $(K \times 1)$ vector of parameters to be estimated. The error term of this model is split into two components. The α_i are stable person-specific characteristics which not only are often unobserved by the researcher (e.g. cognitive ability, genetic disposition, personality), but also are very often related to the covariates. Hence, the α_i are unobserved effects capturing time-constant individual heterogeneity. The second component ϵ_{it} is an idiosyncratic error that varies across subjects and over time. The intercept α that is standard in regression models is dropped here due to collinearity with the person-specific errors α_i (in fact, these can be seen as person-specific intercepts).

Formally, such an error term decomposition is always possible. However, the two terms can only be identified if panel data are available (or, more generally, multi-level data). The reason is that we can infer the person-specific characteristics only from repeated observations. With cross-sectional data the error components model is not identified.

The easiest way to estimate the parameters of the model is by pooling the data and running ordinary least squares (pooled OLS, POLS). Consistency of POLS requires exogeneity of time-constant individual heterogeneity and idiosyncratic errors. POLS estimation does not distinguish between the two error components, which are replaced by the composite error $\nu_{it} = \alpha_i + \epsilon_{it}$. The condition for consistency is $E(\mathbf{x}_{it}'\nu_{it}) = \mathbf{0}$ which is equivalent to assuming $E(\mathbf{x}_{it}'\alpha_i) = \mathbf{0}$ and $E(\mathbf{x}_{it}'\epsilon_{it}) = \mathbf{0}$. The second condition requires that idiosyncratic errors are contemporaneously exogenous, an assumption that is often reasonable. But the first condition imposes exogeneity of stable characteristics, which very often is not a reasonable assumption. In case of self-selection into treatment, the assumption is violated and POLS estimates are biased and inconsistent.

On the other hand, the key assumption for consistency of the FE estimator is the strict exogeneity condition imposed on the idiosyncratic errors,

$$E(\epsilon_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \alpha_i) = E(\epsilon_{it}|\mathbf{x}_{it}, \alpha_i) = 0. \tag{15.2}$$

Strict exogeneity implies $E(\mathbf{x}'_{is}\epsilon_{it}) = \mathbf{0}$, for all $s, t = 1, \ldots, T$. This rules out not only contemporaneous correlation of regressors and the idiosyncratic errors, but also correlation of past and future values of covariates and errors. This is why the assumption is called 'strict' exogeneity. Obviously this assumption is somewhat stronger than the contemporaneous exogeneity assumption required for POLS. The key advantage, however, of the FE framework is that no assumption is needed concerning the relation of stable characteristics and regressors. In fact, α_i and $\mathbf{x}_{it}$ can be related in arbitrary ways which will not result in biased estimates of the coefficients.

FE estimation applies POLS to transformed data where the transformation (called 'demeaning' or 'within transformation') extracts the variation within subjects over time, but discards variation across units. Averaging equation (15.1) over time gives

$$\bar{y}_i = \bar{\mathbf{x}}_i\boldsymbol{\beta} + \alpha_i + \bar{\epsilon}_i, \tag{15.3}$$

where $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$, $\bar{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T} \mathbf{x}_{it}$ and $\bar{\epsilon}_i = T^{-1}\sum_{t=1}^{T} \epsilon_{it}$ are person-specific means. Because α_i is time-constant for each person it is identical to the mean for that person. Therefore, subtracting equation (15.3) from equation (15.1) eliminates α_i and any bias that might result from its association with regressors.

The demeaned regression is given by

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + \epsilon_{it} - \bar{\epsilon}_i. \tag{15.4}$$

The conventional FE estimator is the POLS estimator applied to these demeaned data. As mentioned above and shown in the statistical framework section below, the strict exogeneity assumption is sufficient for consistency of the FE estimator. The main point is that the assumption of no person-specific unobserved heterogeneity $E(\mathbf{x}'_{it}\alpha_i) = \mathbf{0}$ is no longer needed. The FE estimator is consistent even if $E(\mathbf{x}'_{it}\alpha_i) \neq \mathbf{0}$. Thus, with panel data and the FE estimator it is possible to identify a causal effect under weaker assumptions (compared to cross-sectional OLS or POLS).

A didactic example

To strengthen the intuition on FE methodology we now demonstrate with stylized data how the FE estimator works. The data are plotted in Figure 15.1.[2] For didactic reasons we imagine that these are the wage careers of four men ($N = 4$) over six panel waves ($T = 6$). The outcome variable is the monthly wage in euros. The treatment variable is a marriage dummy that is zero before marriage (black dots) and unity afterwards (grey triangles). As can be seen, two of the men are low earners who do not marry during the observation window. The other two men are high earners and they marry between panel waves (time) 3 and 4. The wage careers are constructed so that with marriage there is a wage increase of €500. Thus we built a causal marriage effect of plus €500 into these data (marital wage premium). Further, we built self-selection into these data: it is the high earners who marry (perhaps because they are the more attractive marriage partners). This has the consequence that these data are plagued by person-specific heterogeneity.

Next we want to compare the results of three wage regressions that are specified as follows:

$$w_{it} = \beta m_{it} + \alpha_i + \epsilon_{it}. \tag{15.5}$$

Here w_{it} is the monthly wage and m_{it} is the marriage dummy. β is then an estimate for the marital wage premium.

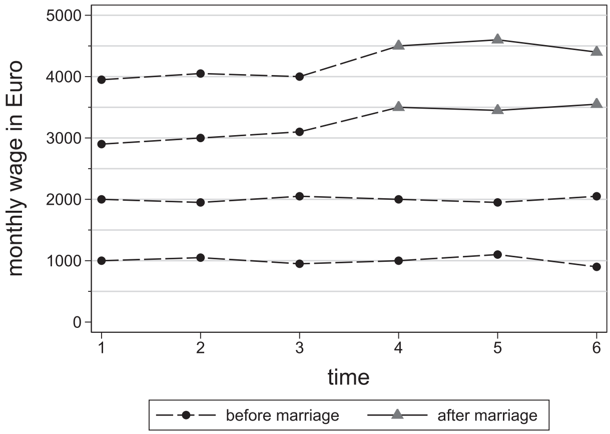


Figure 15.1 Fictional wage careers of four men

First, we estimate a *cross-sectional OLS regression* at $t = 4$ (see Table 15.1, column (1)). Because the regression coefficient of a dummy is simply the difference of the group means, we obtain an estimate of the marriage effect of plus €2500 (it can easily be seen from Figure 15.1 that at $t = 4$ the wage average of those married is €4000 and of those not married is €1500). Obviously, this estimate of the marital wage premium is strongly biased upwards. The statistical reason for this bias is that we have person-specific unobserved heterogeneity in these data. The person-specific error term α_i is correlated with the marriage dummy m_{it}.

More intuitively, the reason is that a cross-sectional regression compares outcomes of different persons: the wage of those married with the wage of those not married. We term this a *between comparison*. A between comparison only works if the assumption of unit (here, person) homogeneity holds: the persons we compare must not differ in anything relevant but the treatment (in other words, there is no unobserved heterogeneity). However, in our data this assumption is violated: men differ not only by marriage but also on other relevant characteristics that affect their wage and marriage propensity. In fact, in most non-experimental social science research the assumption of unit homogeneity will not be tenable: units differ in so many respects that it is impossible to control for all of them.

Would capitalizing on the panel structure of our data remedy the problem? To investigate this, in a next step we estimate an *OLS regression with the pooled data* (POLS, see Table 15.1, column (2)). POLS uses all 24 observations and estimates a marital wage premium of plus €1833. This is the mean of the six wage observations after marriage minus the mean of the 18 wage observations before marriage. Again, the statistical reason for this heavy upward bias is that we have unobserved heterogeneity.

More intuitively, the reason is that POLS, once again, does a between comparison. With panel data, we can distinguish *between* and *within variation*. Between variation is generated by person differences (more technically, it is the variation of the person-specific means $\bar{y}_i$). Within variation is the variation generated by changes over time within the persons (more technically, it is the variation of the demeaned data $y_{it} - \bar{y}_i$). POLS uses both components of variation to estimate the marriage effect. But, as we already know, the between variation is 'contaminated' (Allison, 2009, p. 3) by unobserved heterogeneity due to self-selection. Therefore, the POLS estimate is so strongly biased. Compared with cross-sectional OLS the bias is lower, because POLS also uses the within variation (which is not contaminated).

Table 15.1 Comparing three regression methods: the effect of marriage on wages

	(1) Cross-sectional OLS	(2) POLS	(3) FE
Marriage	2500	1833	500
Constant	1500	2167	–
Number of persons	4	4	2
Number of person-years	4	24	12

Source: own computations from fabricated data.

Obviously, panel data *per se* do not remedy the problem of unobserved heterogeneity. Part of the variation in panel data – namely between variation – is contaminated if there is self-selection on the person level. Therefore, estimation techniques that recur on between variation will be biased. However, if the within variation is exogenous, a solution to the problem of unobserved person heterogeneity would be to base the estimation on within variation alone.

A *fixed-effects regression* does exactly this: it discards the between variation and infers the causal effect from the within variation only. In Table 15.1, column (3), we see that FE regression in fact provides the correct estimate of the marital wage premium: plus €500. By demeaning the data all between variation has been eliminated. Only within variation is left and estimates are based on within variation only. Therefore, person-specific heterogeneity no longer disturbs estimation. FE estimation is not biased by person-specific unobserved heterogeneity.

To get more intuition on how FE estimation 'works', we plot the demeaned data in Figure 15.2 (data are jittered for better visibility). Marriage (demeaned) has always value zero for those who never marry. Their 12 person-years are therefore in the middle of the plot. Marriage (demeaned) is -0.5 for the six person-years before marriage and +0.5 after marriage. Now, what determines the FE regression line? The FE regression line is estimated by OLS applied to the demeaned data. First, because demeaned variables have a mean of zero, the FE regression line passes through $(0,0)$. This makes clear that those persons who never marry contribute nothing to the FE estimation. The reason is that they have no within variation on the treatment variable. Therefore, the FE regression line is based only on the 12 observations of those two men who eventually marry. Second, the slope of the regression line is determined by the mean of the (demeaned) wage of the six person-years after marriage minus the mean wage of the six person-years before marriage. This difference is €500.

Thus we see that the FE estimator does a pure *within comparison*: the wages after a marriage are compared with the wages before a marriage. And this within comparison is not biased by any kind of person heterogeneity. The intuition on within comparison becomes even clearer when one describes FE estimation in a slightly different (but equivalent) way. Compute for each man who married the within wage difference (average wage after marriage minus average wage before). This is the marital wage premium estimated for each man separately. Then the FE estimator is the (weighted) average of the individual wage premia.

The main point is that FE estimation does not infer the causal effect from a comparison of different persons, but by comparing within person change that is induced by a treatment event. Thus, FE estimation no longer needs the strong and in many situations untenable assumption of unit homogeneity. It needs the assumption of 'temporal homogeneity': nothing relevant changes over time, only the treatment. This assumption usually is also strong, because many things change over time. However, the assumption is valid in our stylized data and therefore the FE estimator works well here.

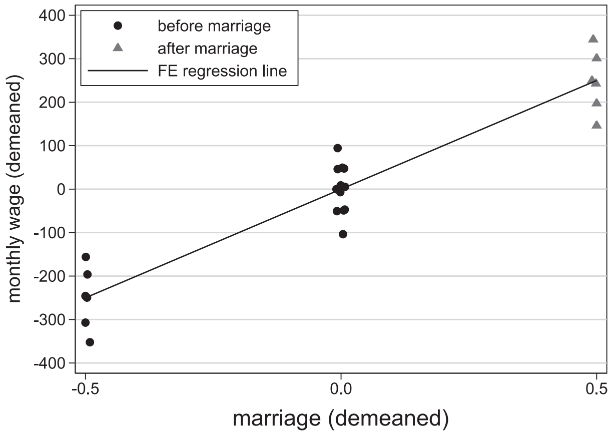


Figure 15.2 The 'mechanics' of FE estimation

More generally, the assumption of temporal homogeneity can be weakened easily by including a control group of non-treated units. In our stylized data these would be the men who never marry. By including time dummies in addition to the marriage dummy they also contribute to FE estimation by providing an estimate of the time trend. This trend is then differenced out from the within comparison in the treatment group. One no longer has to assume that nothing relevant changes over time, but only that the time trends are parallel in the treatment and control group (parallel trends assumption, see below). More formally, this is the strict exogeneity assumption introduced above. In many social science research contexts, this assumption is weaker than a unit homogeneity assumption. This is the reason why panel data and FE estimation allow a causal effect to be identified under weaker assumptions than standard regression with cross-sectional data.

STATISTICAL FRAMEWORK OF FIXED-EFFECTS REGRESSION AND ALTERNATIVE PANEL ESTIMATORS

The basic fixed-effects framework continued

In the previous section we introduced the FE estimator by applying POLS to the demeaned data. We can write the transformed estimation equation (15.4) as

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{\epsilon}_{it}, \tag{15.6}$$

where the dots indicate that respective variables have been demeaned. Consistency of FE estimation is preserved because strict exogeneity is sufficient for consistency of POLS applied to the demeaned data. Assumption $E(\ddot{\mathbf{x}}'_{it}\alpha_i) = \mathbf{0}$ holds because α_i has gone after demeaning. The assumption $E(\ddot{\mathbf{x}}'_{it}\ddot{\epsilon}_{it}) = \mathbf{0}$ holds because $E(\mathbf{x}'_{is}\epsilon_{it}) = \mathbf{0}$, for all $s, t = 1, \ldots, T$. Hence, $E(\ddot{\mathbf{x}}'_{it}\ddot{\epsilon}_{it}) = E[(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\epsilon_{it} - \bar{\epsilon}_i)] = \mathbf{0}$. Neither ϵ_{it} nor $\bar{\epsilon}_i$ is related to $\mathbf{x}_{it}$ or $\bar{\mathbf{x}}_i$ by assumption. Like any assumption, strict exogeneity may fail. Two important violations may arise if time-constant unobservables produce heterogeneous growth in the outcome and if current outcomes are determined by prior outcomes. These problems are discussed later in this section.

The conventional FE estimator can be derived by the analogy principle, replacing expectations by sample moments. Write (15.6) as a system of T OLS equations

$$\ddot{\mathbf{y}}_i = \ddot{\mathbf{X}}_i\boldsymbol{\beta} + \ddot{\boldsymbol{\epsilon}}_i, \tag{15.7}$$

where $\ddot{\mathbf{y}}_i$ and $\ddot{\boldsymbol{\epsilon}}_i$ are $T \times 1$, $\ddot{\mathbf{X}}_i$ is $T \times K$ and $\boldsymbol{\beta}$ is $K \times 1$. Premultiplying (15.7) by $\ddot{\mathbf{X}}_i'$, taking expectations and solving for $\boldsymbol{\beta}$ gives

$$\begin{aligned}
\ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i &= \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\boldsymbol{\beta} + \ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\epsilon}}_i \\
E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i) &= E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\boldsymbol{\beta}) + E(\ddot{\mathbf{X}}_i'\ddot{\boldsymbol{\epsilon}}_i) = \boldsymbol{\beta}E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i) \\
\boldsymbol{\beta} &= \left[E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i)\right]^{-1}E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i).
\end{aligned} \tag{15.8}$$

Note that a rank condition must hold for $E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i)$ to be invertible (i.e. nonsingular). The rank of the matrix must be equal to the number of regressors, rk $\sum_{n=1}^{N} E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i) = K$, which imposes linear independence among the demeaned covariates.

Demeaning not only wipes out unobserved time-constant variables α_i, but along with them all observed time-constant covariates. The effect of variables that do not change for any person over time (e.g. sex, birth cohort) thus cannot be estimated. This is often seen as a shortcoming of FE estimation, but it actually is a major strength of the method because the number of potential confounders that need be measured and included in the model is reduced tremendously. For example, one never has to worry about bias induced by genetic differences between persons, a potential confounder that is virtually impossible to measure in large surveys.

Given that strict exogeneity and the rank condition are satisfied, the FE estimators are identified as

$$\hat{\boldsymbol{\beta}}_{\mathrm{FE}} = \left(\sum_{n=1}^{N}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1}\left(\sum_{n=1}^{N}\ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i\right). \tag{15.9}$$

Inference

For tests of hypotheses concerning the FE coefficients further assumptions are necessary. The idiosyncratic errors must be homoscedastic and serially uncorrelated to obtain consistency of the variance–covariance matrix from which to get the standard errors of coefficients. Both conditions hold if

$$E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i'|\mathbf{x}_i, \alpha_i) = E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i') = \sigma_\epsilon^2\mathbf{I}_T. \tag{15.10}$$

Since FE regression is based on POLS for the demeaned data, however, the above condition must guarantee that POLS conditions of homoscedasticity and no serial correlation hold for the transformed errors. The POLS conditions are (Wooldridge, 2010, p. 192)

$$E(\epsilon_{it}^2\mathbf{x}_{it}'\mathbf{x}_{it}) = \sigma_\epsilon^2E(\mathbf{x}_{it}'\mathbf{x}_{it}), \tag{15.11}$$

$$E(\epsilon_{it}\epsilon_{is}\mathbf{x}_{it}'\mathbf{x}_{is}) = \mathbf{0}, \quad \text{for all } t \neq s. \tag{15.12}$$

Assumption (15.11) is the homoscedasticity condition for each of the T cross-sections. Assumption (15.12) states that the composite errors must be uncorrelated over time.

The conventional asymptotic variance of the FE estimator is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}) \overset{\mathrm{asy}}{\approx} \hat{\sigma}_\epsilon^2\left(\sum_{i=1}^{N}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1}, \tag{15.13}$$

where the standard errors are the square roots of the diagonal elements. After FE regression, the residuals are given by $\hat{\ddot{\epsilon}}_{it} = \ddot{y}_{it} - \ddot{\mathbf{x}}_{it}\hat{\boldsymbol{\beta}}$. These are squared and then summed over t and i to estimate the error variance $\hat{\sigma}_{\epsilon}^{2} = \sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\ddot{\epsilon}}_{it}^{2}/(NT - N - K)$, where $NT - N - K$ are the correct degrees of freedom. (The usual standard errors obtained when running OLS on the demeaned equation are not valid. They have to be adjusted for the fact that N means have to be calculated for demeaning.)

Assumption (15.10) guarantees that (15.11) holds because the variance of the demeaned errors of period t is $E(\ddot{\epsilon}_{it}^{2}) = E[(\epsilon_{it} - \bar{\epsilon}_{i})^{2}] = E(\epsilon_{it}^{2}) + E(\bar{\epsilon}_{i}^{2}) - 2E(\epsilon_{it}\bar{\epsilon}_{i}) = \sigma_{\epsilon}^{2} + \sigma_{\epsilon}^{2}/T - 2\sigma_{\epsilon}^{2}/T = \sigma_{\epsilon}^{2}(1 - 1/T)$, which does not vary over t. However, assumption (15.12) is violated. Even if the idiosyncratic errors are serially uncorrelated in the untransformed data so that condition (15.10) holds, the demeaned errors are negatively correlated because $E(\ddot{\epsilon}_{it}\ddot{\epsilon}_{is}) = E[(\epsilon_{it} - \bar{\epsilon}_{i})(\epsilon_{is} - \bar{\epsilon}_{i})] = E(\epsilon_{it}\epsilon_{is}) - E(\epsilon_{it}\bar{\epsilon}_{i}) - E(\epsilon_{is}\bar{\epsilon}_{i}) + E(\bar{\epsilon}_{i}^{2}) = 0 - \sigma_{\epsilon}^{2}/T - \sigma_{\epsilon}^{2}/T + \sigma_{\epsilon}^{2}/T = -\sigma_{\epsilon}^{2}/T$. Although this term usually is small in practice (especially as T gets larger), in many applications there is substantial serial correlation already before transformation. It is therefore recommended to use panel-robust standard errors by default.

The panel-robust standard errors correct for arbitrary clustering of time series and heteroscedasticity. They are obtained from

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}) \overset{\text{asy}}{\approx} (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\left(\sum_{i=1}^{N}\ddot{\mathbf{X}}_{i}'\hat{\boldsymbol{\epsilon}}_{i}\hat{\boldsymbol{\epsilon}}_{i}'\ddot{\mathbf{X}}_{i}\right)(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}. \tag{15.14}$$

For large N and small T, the degrees of freedom are $NT - K$, which is smaller than the adjustment of the usual standard errors. It is therefore quite often the case that the panel-robust standard errors are actually smaller. Angrist and Pischke (2009) recommend reporting the larger of the two standard errors to be on the conservative side.

Panel-robust standard errors are biased if there are only few clusters. Kézdi (2004) shows for simulated data that the clustered standard errors perform very well if there are 50 or more clusters. Stock and Watson (2008) present simulation results showing little bias of clustered standard errors in the event of heteroscedasticity and 100 or more clusters. This result also holds when errors are at the same time serially correlated (modeled as a first-order moving average). Nevertheless, as Angrist and Pischke (2009, Chapter 8) argue, standard errors will always be biased in finite samples. However, it is not the prime task of social researchers to get the standard errors right. Much more important is to get the coefficient estimates right. Therefore, we recommend following their advice: 'Your standard errors probably won't be quite right, but they rarely are. Avoid embarrassment by being your own best skeptic, and especially, DON'T PANIC!' (Angrist and Pischke, 2009, p. 327).

Other basic within estimators

There are three other within estimators which yield results that are identical or similar to the FE estimator. The first, least squares dummy variable (LSDV) regression, is conceptually simple and makes the intuition of within estimation very clear. Instead of demeaning the data one could include $N - 1$ dummy variables in a POLS regression, one indicator variable for each person (leaving one out as the reference category). That is, one 'fixes' the time-constant differences between subjects by replacing the common constant of the POLS model with individual-specific constants. The term 'fixed-effects' regression actually stems from this approach where the unobserved effects are viewed as fixed quantities and individual constants as parameters that need to be estimated.[3] However, in large samples it is more appropriate to view them as random variables, which is the modern approach to FE regression. The LSDV and FE estimators are identical.

The second alternative to FE regression is first differencing (FD). Instead of demeaning the data, FD builds on differences between ensuing observations from the same subject. This is another within transformation that wipes out time-constant unobservables (since these are the same at t and $t-1$). FD is the POLS regression of

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta\epsilon_{it}, \tag{15.15}$$

where $\Delta y_{it} = y_{it} - y_{it-1}$, $\Delta \mathbf{x}_{it} = \mathbf{x}_{it} - \mathbf{x}_{it-1}$, $\Delta\epsilon_{it} = \epsilon_{it} - \epsilon_{it-1}$. FD and FE yield identical estimates for $\boldsymbol{\beta}$ when $T = 2$. For panels that are longer than two periods, estimates differ in general. They will differ substantially if there is very strong or very weak serial correlation in the untransformed data. Both estimators require strict exogeneity. However, while FE builds on the assumption of no serial correlation prior to demeaning (see condition (15.10)), FD relies on no serial correlation in the differenced errors. The latter assumption is equivalent to very strong correlation in the untransformed errors. FE and FD therefore rely on assumptions that are opposite extremes.

The third well-known alternative is the difference-in-differences (DiD) estimator. The DiD approach is similar to LSDV in that both explicitly model differences in outcome levels. However, DiD is more parsimonious because only mean differences between groups are modeled, not individual differences. For example, if one is interested in the effect of a training program that takes place between years $t = 1$ and $t = 2$ and uses DiD for evaluation of earnings in $t = 2$, one may estimate by POLS the equation

$$y_{it} = \alpha_0 + \delta_1 p_{it} + \alpha_1 d_i + \delta_2 p_{it} d_i + \mathbf{x}_{it}\boldsymbol{\beta} + \epsilon_{it}. \tag{15.16}$$

This regression contains a dummy for the second (post-treatment) year, p_{it}, a time-constant dummy indicating that a person belongs to the treated group, d_i, the interaction of these two, $p_{it}d_i$, and possibly control variables, $\mathbf{x}_{it}$. The overall constant α_0 gives us the mean earnings of the control group in the first year (pre-treatment). α_1 tells us by how much the mean earnings of the treated differed from those of the non-treated before treatment. The coefficient on the year dummy, δ_1, is the change in mean earnings of the non-treated from year 1 to year 2, and δ_2 captures by how much the change in mean earnings across the two years (i.e. from before to after treatment) differed in the treated group. δ_2 is thus an estimate of the treatment effect. Since only group mean differences are modeled, the DiD estimator essentially builds on aggregate data. This is why it does not necessarily require panel data, but can be applied also to a pseudo-panel of pooled independent cross-sections. This property makes DiD a very useful tool for policy evaluation when only pseudo-panel data are available.

Any of the four basic within estimators is appropriate for dealing with time-constant confounders. Nevertheless, FE regression has some important practical advantages over the others. While LSDV is conceptually very simple, it is computationally impracticable with large N because any statistical software will refuse to estimate a model containing several thousand regressors. As noted above, FD might be preferable in the presence of strong serial correlation. However, besides that advantage, it has the disadvantage of being inefficient because the initial period is dropped in any case. Moreover, the inefficiency can be very large in the presence of missing data, because first differences can be built only on ensuing observations. For example, if one person is observed at $t = 1, 3, 5$, then FE would use the three person-years, but in FD the person would be dropped completely. With (balanced) panel data for $T = 2$, the DiD estimator is identical to FE and FD. For longer panels, however, it differs in general. In fact, it can give misleading answers because all variables enter the regression in levels. If there are control variables (which usually is the case) their effect is likely to be biased, which may also induce bias on the treatment effect. It is therefore recommended to use FE (or FD) where all variables are transformed (Wooldridge, 2010, p. 321).

Extending the fixed-effects approach to account for heterogeneous growth

As mentioned above, strict exogeneity may fail if some unobservable produces heterogeneity with regard to individual trajectories of the outcome. More precisely, the assumption must fail if heterogeneous growth is related to the covariate(s) of interest. Strict exogeneity implies that individual outcome trajectories of treated subjects would have developed parallel to trajectories of non-treated subjects *had they not chosen treatment*. In this section, we show how conventional within estimators fail if this assumption of 'parallel trends' is not met, and how failure of FE can be repaired by extending the FE approach to allow for individual-specific slopes.

Any of the four basic within estimators can be extended to allow for heterogeneous growth. Suppose there are $T = 3$ observations per person and consider the FD model with one regressor x_{it} (say, a binary treatment indicator) and individual-specific slope parameters of (process) time. The structural model can be written as a system of three equations:

$$\text{at } t, \quad y_{it} = \alpha_{1i} + \alpha_{2i}t + \beta x_{it} + \epsilon_{it}, \tag{15.17}$$

$$\text{at } t-1, \quad y_{it-1} = \alpha_{1i} + \alpha_{2i}(t-1) + \beta x_{it-1} + \epsilon_{it-1}, \tag{15.18}$$

$$\text{at } t-2, \quad y_{it-2} = \alpha_{1i} + \alpha_{2i}(t-2) + \beta x_{it-2} + \epsilon_{it-2}. \tag{15.19}$$

Here, α_{1i} are individual-specific constants capturing any differences in outcome *levels* between individuals produced by stable characteristics. In addition, α_{2i} are individual-specific slopes that capture individual differences in the *growth* of outcomes over time. Technically, α_{2i} are stable characteristics that interact with time to produce outcome trajectories specific to individuals. These may be observable or unobservable characteristics. For example, the steepness of wage trajectories might differ by cohort, because younger cohorts face better labor market opportunities over the life course. Or they might differ by unobserved career orientation, because the motivated are likely to get promotions at a faster rate.

Individual heterogeneity in the outcome levels, i.e. in α_{1i}, is eliminated by first differencing. Subtract (15.18) from (15.17), and (15.19) from (15.18), to get

$$\text{at } t, \quad (y_{it} - y_{it-1}) = \alpha_{2i} + \beta(x_{it} - x_{it-1}) + (\epsilon_{it} - \epsilon_{it-1}), \tag{15.20}$$

$$\text{at } t-1, \quad (y_{it-1} - y_{it-2}) = \alpha_{2i} + \beta(x_{it-1} - x_{it-2}) + (\epsilon_{it-1} - \epsilon_{it-2}). \tag{15.21}$$

Estimation of the FD model will nevertheless be biased if α_{2i} is related to x_{it}. Strict exogeneity will hold only if the individual trends do not differ in the treated and control group, on average. Hence, it must be true that $E(\alpha_{2i}|x_{it}, \alpha_{1i}) = 0$. In words, the outcome trajectories of the two groups would have developed parallel to each other had there been no treatment. There might have been differences in outcome levels prior to treatment, and there might have been a treatment effect for the treated. But conditioning on α_{1i} and x_{it}, there should be no difference in the mean outcomes.

How can we cure the bias? A simple way to get rid of α_{2i} is to apply differencing once more to get

$$\text{at } t, \quad (y_{it} - 2y_{it-1} + y_{it-2}) = \beta(x_{it} - 2x_{it-1} + x_{it-2}) + (\epsilon_{it} - 2\epsilon_{it-1} + \epsilon_{it-2}). \tag{15.22}$$

This extension of FD, second differencing (SD), provides an unbiased estimate of the treatment effect even if there is heterogeneity with respect to individual growth that is systematically related to treatment, that is, if the parallel trends condition is violated.

The LSDV and DiD estimators can also be extended to eliminate bias that is due to α_{2i}. Instead of just including dummy variables in a POLS regression (LSDV), one could additionally include interactions of these dummies with time. The coefficients on the interaction terms would then give estimates for individual growth. Similarly, when following the DiD approach, one would add the interaction of time and the time-constant indicator of the treatment group.

As explained above, conventional FE estimation is based on within transformation of the data, also known as demeaning. The idea was to subtract the individual mean as a time-constant estimate of individuals' outcome trajectory to eliminate individual differences in outcome levels. The extension of this approach now subtracts a time-varying estimate instead (i.e. an estimate of individual growth curves), to eliminate the individual-specific slopes α_{2i} along with the individual-specific constants α_{1i}. Intuitively, this idea leads to the following estimation procedure:

(1) Estimate for each subject i the individual POLS regression $y_{it} = \alpha_{1i} + \alpha_{2i}t + \xi_{it}$ and get predicted values $\hat{y}_{it} = \hat{\alpha}_{1i} + \hat{\alpha}_{2i}t$.
(2) From the actual outcome values y_{it}, subtract the estimated values $\hat{y}_{it}$ from step (1) to get detrended outcomes $\tilde{y}_{it} = y_{it} - \hat{y}_{it}$.
(3) Repeat steps (1) and (2) for the independent variable(s) to get, for any variable x_j, the detrended variable $\tilde{x}_{jit} = x_{jit} - \hat{x}_{jit}$.
(4) Pool the transformed data from steps (2) and (3) and run a POLS regression.

The predicted values $\hat{y}_{it}$ from step (1) are time-varying predicted values of the expected outcome of individual i. These estimated values are used in step (2) to purge individual i's outcomes from the expected individual-specific trend. Hence, the $\tilde{y}_{it}$ are just the residuals of the individual time-series regression from step (1). Detrending all variables of the model in this way thus makes clear what the extended within transformation actually does to the model. The estimate of the treatment effect is based only on variation within subjects over time that cannot be predicted from the (initial) level and slope of individual trajectories.

The fixed-effects model with individual-specific constants and slopes (FEIS), introduced by Polachek and Kim (1994) and generalized by Wooldridge (2010, pp. 377–381), can be described as follows.[4] If the data are in long format, we can write the structural model for unit i as

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\alpha}_i + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{15.23}$$

where $\mathbf{y}_i$ and $\boldsymbol{\epsilon}_i$ are $T \times 1$, $\mathbf{Z}_i$ is $T \times J$, $\boldsymbol{\alpha}_i$ is $J \times 1$, $\mathbf{X}_i$ is $T \times K$ and $\boldsymbol{\beta}$ is $K \times 1$.

Now define a 'residual maker' matrix for unit i: $\mathbf{M}_i \equiv \mathbf{I}_T - \mathbf{Z}_i\left(\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i'$. Premultiply equation (15.23) through by $\mathbf{M}_i$. This gives $\mathbf{M}_i\mathbf{y}_i = \mathbf{M}_i\mathbf{Z}_i\boldsymbol{\alpha}_i + \mathbf{M}_i\mathbf{X}_i\boldsymbol{\beta} + \mathbf{M}_i\boldsymbol{\epsilon}_i$. Since $\mathbf{M}_i\mathbf{Z}_i = \mathbf{Z}_i - \mathbf{Z}_i\left(\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i'\mathbf{Z}_i = \mathbf{0}$, $\boldsymbol{\alpha}_i$ is eliminated. Conventional FE is a special case of this model where $\mathbf{Z}_i \equiv \mathbf{1}$, that is, the model includes only individual constants. Another special case is the model including additionally individual-specific linear time trends such that $\mathbf{Z}_i \equiv (\mathbf{1}, \mathbf{t})$. As discussed above, this model could be estimated by POLS on second differences. However, FEIS is more efficient, and this advantage can be very important if panels are short and/or if panels have gaps due to missing data (analogous to the pros and cons of FE and FD; see the previous subsection). FEIS is also much more general than SD because individual slopes can be assumed for variables other than calendar time. For example, one could assume that unobserved career orientation produces steeper wage trajectories over labor market experience (and perhaps experience squared). Given there are enough observations per subject (at least $J+1$ are needed to remove $\boldsymbol{\alpha}_i$), the model is easily extended to include further variables that interact with unobservables. These variables do not have to be a function of time. For example, returns to schooling might be greater for very motivated people. If the motivated people are also more likely to participate in further education, the effect of a training program on wages would be biased in an FE regression model, but the bias could be removed using an FEIS model with individual slopes for schooling.

The general detrended model for unit i at time t can be written as

$$\tilde{y}_{it} = \tilde{\mathbf{x}}_{it}\boldsymbol{\beta} + \tilde{\epsilon}_{it}. \tag{15.24}$$

Because detrending eliminated $\boldsymbol{\alpha}_i$ from the model (along with $\tilde{\mathbf{z}}_{it}$) the FEIS estimators are obtained by running POLS on the detrended data. Consistency of $\boldsymbol{\beta}_{\text{FEIS}}$ requires

$$E(\epsilon_{it}|\mathbf{z}_{it}, \mathbf{x}_{it}, \boldsymbol{\alpha}_i) = 0, \tag{15.25}$$

which is a weaker form of the strict exogeneity condition required for conventional FE because idiosyncratic errors are expected to be unrelated to regressors only conditional on individual slopes. Hence, the parallel trends assumption is no longer needed.

Despite this obvious advantage, the FEIS model has not yet been widely used in the social sciences (for an exception, see Ludwig and Brüderl, 2011). Morgan and Winship (2007) are even skeptical about the usefulness of the model, because 'substantial amounts of data are needed to estimate it, and certainly from more than just one pretreatment time period and one posttreatment time period' (pp. 264–265). While it is true that the method is 'data hungry', it is also the case that there are some long-running panel surveys around that provide long enough panels.[5] The main practical reason why researchers do not use FEIS models seems to be that it is not implemented in standard statistical software. Therefore, a Stata ado file (xtfeis.ado) is available for download from the website accompanying this volume.

Random-effects models

A very popular class of panel estimators is based on the random-effects (RE) approach. The conventional RE model starts from the error components model given in (15.1). RE requires the same strict exogeneity assumption as FE. As for POLS, however, an assumption is needed about stable unobserved characteristics. The orthogonality condition $E(\alpha_i|\mathbf{x}_i) = E(\alpha_i) = 0$ states that stable unobserved confounders may not be related to any of the regressors. Because this assumption is so restrictive, RE often will fail to identify the causal effect of interest. If both conditions hold and the conditional mean is modeled correctly, the RE estimator is consistent because $E(\nu_{it}|x_{it}) = 0$.

We can write the model as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\nu}_i, \tag{15.26}$$

where $\boldsymbol{\nu}_i = \boldsymbol{\alpha}_i\mathbf{j}_T + \boldsymbol{\epsilon}_i$ and $\mathbf{j}_T$ is a $(T \times 1)$ vector of ones.

RE estimation exploits the time structure of the composite errors, which is why strict exogeneity is required. Define the $(T \times T)$ matrix $\boldsymbol{\Omega} \equiv E(\boldsymbol{\nu}_i\boldsymbol{\nu}_i')$. This is the variance–covariance matrix of the composite errors. In order to estimate it, further assumptions on the error structure are needed. Standard assumptions of constant variance of the idiosyncratic errors across t, $E(\epsilon_{it}^2|\mathbf{x}_i, \alpha_i) = E(\epsilon_{it}^2) = \sigma_\epsilon^2$ and $E(\epsilon_{it}\epsilon_{is}|\mathbf{x}_i, \alpha_i) = E(\epsilon_{it}\epsilon_{is}) = 0$ for all $t \neq s$, as well as constant variance of the stable components across i, $E(\alpha_i^2|\mathbf{x}_i) = \sigma_\alpha^2$, are sufficient (Wooldridge, 2010, p. 294). The two assumptions together imply equality of the conditional and unconditional variance–covariance matrix, $E(\boldsymbol{\nu}_i\boldsymbol{\nu}_i'|\mathbf{x}_i) = E(\boldsymbol{\nu}_i\boldsymbol{\nu}_i')$. The variance–covariance matrix

$\boldsymbol{\Omega}$ then has a special form. Elements on the diagonal are $\sigma_\alpha^2 + \sigma_\epsilon^2$ and off-diagonal elements are σ_α^2. This special form arises because under strict exogeneity $E(\alpha_i \epsilon_{it}) = 0$, hence, $E(\nu_{it}^2) = E(\alpha_i^2) + E(\epsilon_{it}^2) + 2E(\alpha_i \epsilon_{it}) = \sigma_\alpha^2 + \sigma_\epsilon^2$, and given the additional assumptions on the error terms, $E(\nu_{it}\nu_{is}) = E(\alpha_i^2) = \sigma_\alpha^2$.

The RE estimator is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{RE}} = \left(\sum_{n=1}^{N} \mathbf{X}_i' \hat{\boldsymbol{\Omega}}_i^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{n=1}^{N} \mathbf{X}_i' \hat{\boldsymbol{\Omega}}_i^{-1} \mathbf{y}_i\right) \tag{15.27}$$

$$= \left(\sum_{n=1}^{N}\sum_{t=1}^{T} \breve{\mathbf{x}}_{it}' \breve{\mathbf{x}}_{it}\right)^{-1} \left(\sum_{n=1}^{N}\sum_{t=1}^{T} \breve{\mathbf{x}}_{it}' \breve{y}_{it}\right), \tag{15.28}$$

where $\breve{y}_{it} = y_{it} - \theta \bar{y}_i$ and $\breve{\mathbf{x}}_{it} = \mathbf{x}_{it} - \theta \bar{\mathbf{x}}_i$ indicate that the data have been quasi-demeaned.

To derive θ, write the variance–covariance matrix in full matrix notation as $\boldsymbol{\Omega} \equiv \sigma_\epsilon^2 \mathbf{I}_T + \sigma_\alpha^2 \mathbf{j}_T \mathbf{j}_T'$. This can also be written as $\boldsymbol{\Omega} = \sigma_\epsilon^2[\mathbf{Q} + ((\sigma_\epsilon^2 + T\sigma_\alpha^2)/\sigma_\epsilon^2)(\mathbf{I}_T - \mathbf{Q})]$, where $\mathbf{Q} = \mathbf{I}_T - T^{-1}\mathbf{j}_T\mathbf{j}_T'$. This expression results because $\sigma_\epsilon^2(\mathbf{I}_T - T^{-1}\mathbf{j}_T\mathbf{j}_T') + (\sigma_\epsilon^2 + T\sigma_\alpha^2)\mathbf{I}_T - (\sigma_\epsilon^2 + T\sigma_\alpha^2)(\mathbf{I}_T - T^{-1}\mathbf{j}_T\mathbf{j}_T') = \sigma_\epsilon^2 \mathbf{I}_T + \sigma_\alpha^2 \mathbf{j}_T \mathbf{j}_T'$.

Note further that $\mathbf{Q} = \mathbf{Q}\mathbf{Q}'$ and $\mathbf{I}_T = \mathbf{I}_T\mathbf{I}_T'$. Therefore, $\boldsymbol{\Omega}^{-1/2} = 1/\sigma_\epsilon^2[\mathbf{Q} + (\sigma_\epsilon^2/(\sigma_\epsilon^2 + T\sigma_\alpha^2))^{1/2}(\mathbf{I}_T - \mathbf{Q})]$. Premultiplying the outcome by $\boldsymbol{\Omega}^{-1/2}$ gives the quasi-demeaned outcome $\breve{\mathbf{y}}_i = \mathbf{y}_i - [1 - (\sigma_\epsilon^2/(\sigma_\epsilon^2 + T\sigma_\alpha^2))^{1/2}]\bar{\mathbf{y}}_i$, which is just the compact form of $\breve{y}_{it} = y_{it} - \theta\bar{y}_i$ with $\theta = 1 - (\sigma_\epsilon^2/(\sigma_\epsilon^2 + T\sigma_\alpha^2))^{1/2}$.

The RE estimator therefore is the POLS estimator on the quasi-demeaned data. The model is usually estimated by feasible generalized least squares (FGLS). The procedure is called 'feasible' because consistent estimators for the variance of the two error components are available. A consistent estimator for σ_ϵ^2 is obtained by running a within (FE) regression and calculating the residual variance as $\hat{\sigma}_\epsilon^2 = \sum_{n=1}^{N}\sum_{t=1}^{T}[\hat{\ddot{\epsilon}}_{it}^2/(NT - N - K)]$. Using this result and residuals from a between regression of equation (15.3), one gets $\hat{\sigma}_\alpha^2 = \sum_{n=1}^{N}[\hat{\bar{\nu}}_{it}^2/(N - K) - (\hat{\ddot{\epsilon}}_{it}^2/T)]$. The variance of the unobserved effect results because the between regression leaves variation to the residuals that is due to α_i and ϵ_{it}, whereas the FE regression residuals contain only variation due to ϵ_{it}.

From the expression for θ, it follows that the RE estimates are in between POLS and FE estimates. The POLS estimator results when all the error variance is due to the idiosyncratic errors ($\theta = 0$). The FE estimator results when all the variance is due to the stable unobserved effects ($\theta = 1$). The advantage of RE over POLS is in terms of greater efficiency at the price of a stronger exogeneity assumption. The advantage of RE over FE is also greater efficiency given the orthogonality condition holds. If it does not hold, RE is inconsistent and FE should be preferred.

Panel regression as multi-level models

As mentioned in the introduction to this chapter, models for panel data can be formulated as multi-level models. In the case of individuals observed over time, there are two levels: person-years on the lower level (level 1) which are nested in persons (level 2). Note that multi-level models allow for individual-specific coefficients α_{ki}. Different specifications are possible. Each α_{ki} may be specified as a 'fixed coefficient' that does not vary over persons, $\alpha_{ki} = \gamma_k$, as a coefficient varying non-randomly across persons, $\alpha_{ki} = \mathbf{w}_{ki}\boldsymbol{\gamma}_k + \xi_{ki}$, or as a coefficient with variation that is purely random, $\alpha_{ki} = \gamma_k + \xi_{ki}$ (Cameron and Trivedi, 2005, pp. 846–847).[6]

Usually, models contain both fixed and varying coefficients. Such models are often classified as 'mixed-effects models', but 'mixed-coefficient models' would be a better term. In these models, there is a vector of variables $\mathbf{x}_{it}$ for which coefficients $\boldsymbol{\beta}$ are homogeneous across units, while for some variables $\mathbf{z}_{it}$ coefficients $\boldsymbol{\alpha}_i$ are individual-specific.[7] We specify a linear regression model on the lower level (level 1 equation):

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_{it}\boldsymbol{\alpha}_i + \epsilon_{it}. \tag{15.29}$$

For the individual-specific coefficients, we specify a general model on the upper level (level 2 equation):

$$\alpha_{ki} = \mathbf{w}_{ki}\boldsymbol{\gamma}_k + \xi_{ki}. \tag{15.30}$$

Conventional FE and RE models are special cases. The FE model assumes an individual-specific intercept that is non-random, $\alpha_{1i} = \gamma_{1i}$, while all other coefficients are 'fixed' (homogeneous across units). Plugging this into equation (15.29) gives $y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \gamma_{1i} + \epsilon_{it}$, which is just the model of equation (15.1) (albeit using different notation). The conventional RE model assumes a random individual intercept $\alpha_{1i} = \gamma_1 + \xi_{1i}$ (keeping all other coefficients fixed). Therefore, $y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \gamma_1 + \xi_i + \epsilon_{it}$, where γ_1 is a common constant that can be subsumed in $\mathbf{x}_{it}$. We see that this is the RE model.[8]

Using the multi-level framework, an obvious extension of the RE approach is to allow for individual variation not only in the individual intercepts, but also in coefficients. This idea leads to the random-slopes (RS) model, where the individual slope on time-varying regressors is modeled as a random coefficient, $\alpha_{2i} = \gamma_2 + \xi_{2i}$, and $\alpha_{1i} = \gamma_1 + \xi_{1i}$ is the random intercept (as in the RE model).

As an example, suppose we are interested in the effect of parental divorce (x_{it}) on children's cognitive ability (measured repeatedly by some IQ test, y_{it}). We envisage that children differ with respect to their level and growth in ability. For simplicity, we specify a linear trend over age (z_{it}). We end up with the model $y_{it} = \gamma_1 + \gamma_2 z_{it} + \beta x_{it} + \xi_{1i} + \xi_{2i}z_{it} + \epsilon_{it}$, where $\gamma_1 + \gamma_2 z_{it}$ reflect a common constant and trend and $\xi_{1i} + \xi_{2i}z_{it}$ are child-specific deviations from these. A further extension of the model would be to include a covariate such as parental education (w_i) in the level 2 equation to model group-specific differences in levels or growth. This way, we would specify a 'cross-level' interaction effect.

Raudenbush (2001) argues that the RS model should be used to estimate the effect of an event because it controls for unobserved heterogeneity regarding individual growth. However, this is not correct. The RS model imposes exogeneity assumptions on ξ_{1i} and ξ_{2i}. In fact, if the model is estimated by maximum likelihood, as it usually is, full distributional assumptions are needed. For consistency of RS estimates, ξ_{1i} and ξ_{2i} then have to be distributed according to a bivariate normal distribution which implies zero conditional mean. The model therefore is inconsistent if the parallel trends condition needed for consistency of FE is not met. Unlike FEIS, RS estimation cannot handle violations of this assumption.

Lagged dependent variable models

Panel models as considered so far induce correlation of the outcome over time. For instance, in the error components model (15.1) the stable person-specific unobservables (α_i) produce a correlation in y over time. The mechanism behind this correlation is simply that stable unobservables affect the outcome all the time and thus tend to produce similar outcomes over time. This outcome correlation over time produced by stable unobservables is sometimes called 'spurious state dependence'. However, often researchers argue that there might be a second source of correlation over time: 'true state dependence'. Here it is supposed that outcomes affect each other causally over time. The mechanism behind this is that outcomes tends to reproduce themselves.

Usually true state dependence is modeled by including a lagged dependent variable (LDV) on the right-hand side of a panel model:

$$y_{it} = \rho y_{it-1} + \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i + \epsilon_{it}, \tag{15.31}$$

where $|\rho| < 1$ (stationary outcome process) and everything else is as defined before. In this model ρ captures state dependence. A ρ close to one would indicate that the outcome process strongly tends to reproduce itself. In fact, it can be shown that such a model includes two sources of correlation over time: stable unobservables and true state dependence (Cameron and Trivedi, 2005, p. 763). It is important to understand that LDV estimators, unlike within estimators, are *not* designed to handle unobserved heterogeneity. Nor are they able to ameliorate it, let alone solve the problems associated with it, under general conditions. Instead, they aim to tackle the problem of true state dependence where the outcome in any period is determined by its past values.

Introducing an LDV in a panel regression seems intuitively appealing to most social scientists. Further, LDV models are termed often 'dynamic panel models', a term that signals superiority over the 'static panel models' that we have discussed so far. However, estimation of LDV models is inherently problematic. Most researchers naively estimate the LDV model (15.31) with POLS (POLS-LDV). Examples are given by Halaby (2004). However, POLS-LDV provides inconsistent estimators of both ρ *and* $\boldsymbol{\beta}$. Intuitively, this is because the regressor y_{it-1} is related to α_i by definition (α_i affects all outcomes of a person). So the estimate of ρ is biased. This is true even if α_i is purely random, that is, not related to any of the regressors in $\mathbf{x}_{it}$. The exogeneity condition required for POLS is violated by construction of the model. And even worse, this bias transfers to the estimate of $\boldsymbol{\beta}$. More exactly, whenever $\alpha_i \neq 0$ for some i and lagged outcomes y_{it-1} are correlated with $\mathbf{x}_{it}$, the POLS estimates of $\boldsymbol{\beta}$ are inconsistent as well (for more details see Angrist and Pischke, 2009, Section 5.4). In many applications, y_{it-1} and x_{it} will be correlated because covariates are trended. In fact, if we are interested in the effect of a binary treatment dummy (the effect of an event), this will necessarily be the case. In this setting, POLS-LDV is biased if $\boldsymbol{\beta} \neq 0$.

An idea that has intuitive appeal is to build on the within approach to eliminate the selection bias and construct an LDV estimator for the transformed data to deal with state dependence (FE-LDV). This idea has stimulated much methodological research, and several estimators have been proposed. However, an estimator that is consistent under general forms of selection into treatment is not available. Research has shown that FE estimation including a lagged dependent variable produces inconsistent estimators. This inconsistency has become known in the econometric literature as the 'Nickell bias' (Nickell, 1981). More recently, Phillips and Sul (2007) generalized Nickell's findings to cover the FE model with individual-specific time trends. Phillips and Sul give the formula for the bias of ρ and β and also show by simulations that detrending the data often increases the bias dramatically (p. 166). Clearly, within transformation of the data alone does not help to reduce the bias of the LDV model.

Therefore, several approaches that use instrumental variables have been suggested. Arellano and Bond (1991) suggested using further lags of the outcome variable as instruments in an FD model (AB-LDV). The FD model is

$$(y_{it} - y_{it-1}) = \rho(y_{it-1} - y_{it-2}) + (\mathbf{x}_{it} - \mathbf{x}_{it-1})'\boldsymbol{\beta} + (\epsilon_{it} - \epsilon_{it-1}). \tag{15.32}$$

AB-LDV uses y_{it-2} as an instrument for $y_{it-1} - y_{it-2}$. To increase efficiency AB-LDV uses additional lags as instruments (if available). The rationale for this is that lagged outcomes are unrelated to the error term in first differences ($\epsilon_{it} - \epsilon_{it-1}$) if sequential exogeneity holds. This assumption is weaker than the 'strict exogeneity' assumption from FE. Both strict and sequential exogeneity assume that past outcome levels are (conditionally) independent of the contemporaneous idiosyncratic error term. However, only strict exogeneity implies that this is also valid for future outcomes.

As N goes to infinity (keeping T fixed), the AB-LDV estimator becomes consistent if unobserved heterogeneity is purely random. Arellano and Bond (1991) show by simulation that the model works reasonably well even in small samples ($N = 100, T = 7$). However, if time-constant unobserved heterogeneity is related to covariates the estimator is inconsistent unless further conditions hold. Most importantly, it is required that the idiosyncratic errors are serially uncorrelated prior to first differencing (cf. Halaby, 2004). As we have noted earlier, this will often not be the case. Individual outcomes often exhibit strong persistence over time even after controlling for covariates. Then y_{it-2} will correlate with $\epsilon_{it} - \epsilon_{it-1}$ and thus will be an invalid instrument. Moreover, the estimator is known to suffer from bias due to weak instruments if $|\rho|$ gets large (close to one) or if $\sigma_\alpha^2/\sigma_\epsilon^2$ gets large (unobserved heterogeneity is very important) (Bun and Windmeijer, 2010; Blundell and Bond, 1998). Therefore, AB-LDV is also highly likely to fail in practice.

In conclusion, we see that LDV models pose inherent estimation problems that can hardly be solved with real-world data. Therefore, we would argue that LDV models are not useful at all. Fortunately, we also see theoretical reasons why researchers do not have to bother with LDV models. We would argue that the mechanism behind a correlation over time in many cases is due to stable unobservables, not to some kind of reproduction. That is, state dependence is ‘spurious’, not ‘true’. For instance, stable income levels over time are produced by (observed and unobserved) human capital levels, personality traits, and so on. We see no reason why one would want to invoke some mysterious ‘wage reproduction’. Therefore, we would argue that in many cases careful theoretical reasoning would lead to the conclusion that there is no need for an LDV specification.

Monte Carlo simulations comparing panel estimators

In this subsection, we will present simulation results in order to demonstrate the strengths and shortcomings of FE regression *vis-á-vis* alternative estimators. We compare FE to POLS and RE. Because we will also simulate scenarios with individual-specific slopes, we include FEIS and RS. Further, we will simulate situations with true state dependence and therefore also include POLS-LDV and AB-LDV.

The general setup of the simulations is as follows. We draw random samples with $N = 200$ and $T = 10$. We create artificial data based on the data generating process

$$y_{it} = \rho y_{it-1} + \beta x_{it} + \alpha_{1i} + \alpha_{2i} t + \epsilon_{it}. \tag{15.33}$$

Variable x_{it} is a binary treatment indicator. At $t = 0$, the treatment dummy is zero for all subjects. Half of the subjects are treated in later period $t = 5$ (all at the same time). From this point onward, the treatment indicator equals one for the treated.

The data generating process allows for unobserved heterogeneity and true state dependence. Heterogeneity will be modeled by including individual-specific intercepts α_{1i} and individual-specific slopes α_{2i}. We consider heterogeneity with respect to α_{1i} that is either purely random or systematically related to treatment (selection). We also model situations where there is selection into treatment with respect to not only α_{1i} but also α_{2i}. True state dependence will be modeled by setting $\rho \neq 0$, thereby allowing for individual outcomes that are ‘self-reproducing’ over time. The simulations cover cases with state dependence and either random heterogeneity or selection into treatment. We simulate data under several conditions (described more precisely

Table 15.2 Consistency of panel estimators under different scenarios (*N* large, *T* small)

Scenario	$\hat{\beta}_{POLS}$	$\hat{\beta}_{RE}$	$\hat{\beta}_{FE}$	$\hat{\beta}_{FEIS}$	$\hat{\beta}_{RS}$	$\hat{\beta}_{POLS\text{-}LDV}$	$\hat{\beta}_{AB\text{-}LDV}$
Heterogeneity w.r.t. α_{1i}	con.	con.	con.	con.	con.	inc.	con.
Selection w.r.t. α_{1i}	inc.	inc.	con.	con.	inc.	inc.	inc.
Selection w.r.t. α_{2i}	inc.	inc.	inc.	con.	inc.	inc.	inc.
TSD, het. w.r.t. α_{1i}	inc.	inc.	inc.	inc.	inc.	inc.	con.
TSD, sel. w.r.t. α_{1i} or α_{2i}	inc.	inc.	inc.	inc.	inc.	inc.	inc.

α_{1i}, individual-specific intercepts; α_{2i}, individual-specific slopes; TSD, true state dependence; con., consistent; inc., inconsistent.
Source: own compilation.

below) to investigate how unobserved heterogeneity and state dependence affect the estimates of our seven models. We focus on the estimate of the treatment effect β only.

Table 15.2 shows what we expect from the simulations. If unobserved heterogeneity is purely random, all of the estimators considered are consistent, except the POLS-LDV estimator. Of the consistent estimators, the RE estimator should be the most efficient. The RE estimator should, however, be inconsistent if unobserved heterogeneity is related to treatment (selection with respect to α_{1i}). This is the strength of FE regression. The conventional FE estimator should be consistent when the parallel trends condition holds. If there is self-selection with respect to α_{2i} the parallel trends assumption is violated and FE should be inconsistent. Of all estimators considered here, only FEIS should be able to deal with non-parallel trends. In the case with true state dependence, the AB-LDV estimator should be consistent provided unobserved heterogeneity is not systematically related to treatment. If there is both state dependence and selection with respect to α_{1i} or α_{2i}, all estimators considered should be inconsistent (even AB-LDV because it strongly depends on the condition of no serial correlation).

In the following, we describe the simulation scenarios more precisely and discuss the results of the simulations. For each of eight scenarios we draw 1000 samples and compute the mean of each of the seven estimates of β as well as the mean of their standard errors.[9] The treatment effect β is set to one, except in scenario (2), where it is set to zero (no treatment effect). In all simulations (except for the last scenario), ϵ_{it} is a Gaussian random variable. In scenario (8), however, we allow for serial correlation in the idiosyncratic errors. The results are shown in Table 15.3.

In scenarios (1)–(5) we model unobserved heterogeneity and assume that there is no true state dependence ($\rho = 0$). In scenarios (1) and (2) we assume that time-constant heterogeneity is purely random, that is, not related to treatment (α_{1i} is a standard normal random variable). Even in this arguably innocuous case without self-selection into treatment, there is one model that provides strongly biased estimates. The POLS-LDV estimator is consistent only if there is no treatment effect (column (2)), but inconsistent whenever the treatment effect is different from zero (column (1)). All other estimators provide correct answers most of the time. They all are consistent, but standard errors differ. In terms of efficiency, the random-effects models (RE or RS) would be the preferred choice as the standard errors are slightly smaller than those obtained by FE.

In scenario (3) we assume that there is selection into treatment with respect to time-constant unobservables. Here, we model α_{1i} as a normally distributed random variable with variance 1, but mean 1 for the treated and mean 0 for the untreated subjects. In this case of parallel trends, the FE and FEIS models provide the true value. POLS is heavily biased. The RE model provides an estimate that is in between POLS and FE, but is still substantially biased. The same holds for the RS model. POLS-LDV is biased as well. The AB-LDV estimator seems to perform very well,

Table 15.3 Simulation results for eight scenarios: mean of $\hat{\beta}$ for $N = 200$, $T = 10$ (1000 replications)

	(1) het. α_1 $\beta = 1$ $\rho = 0$	(2) het. α_1 $\beta = 0$ $\rho = 0$	(3) sel. α_1 $\beta = 1$ $\rho = 0$	(4) sel. α_2 $\beta = 1$ $\rho = 0$	(5) sel. α_2 $\beta = 1$ $\rho = 0$	(6) het. α_1 $\beta = 1$ $\rho = 0.5$	(7) sel. α_1 $\beta = 1$ $\rho = 0.5$	(8) sel. α_1 $\beta = 1$ $\rho = 0.5$, $\gamma = 0.2$
POLS	1.005	−0.006	1.896	2.495	1.257	1.557	3.314	3.297
	(0.139)	(0.139)	(0.140)	(0.890)	(0.892)	(0.269)	(0.272)	(0.280)
RE	1.004	−0.001	1.139	1.482	0.649	1.506	1.864	1.860
	(0.077)	(0.077)	(0.077)	(0.598)	(0.599)	(0.127)	(0.130)	(0.147)
RS	1.005	−0.001	1.131	0.917	0.904	1.303	1.239	1.202
	(0.078)	(0.078)	(0.078)	(0.126)	(0.126)	(0.103)	(0.104)	(0.111)
FE	1.004	−0.000	1.002	1.391	0.594	1.502	1.771	1.760
	(0.080)	(0.080)	(0.080)	(0.572)	(0.574)	(0.125)	(0.127)	(0.145)
FEIS	1.002	0.003	1.003	0.994	1.001	1.051	0.853	0.847
	(0.128)	(0.128)	(0.128)	(0.128)	(0.128)	(0.141)	(0.141)	(0.153)
POLS-LDV	0.644	−0.003	1.053	0.093	0.040	0.585	0.775	0.713
	(0.082)	(0.077)	(0.081)	(0.065)	(0.063)	(0.059)	(0.065)	(0.066)
AB-LDV	1.007	0.003	1.002	0.993	0.908	0.993	1.011	0.884
	(0.147)	(0.149)	(0.128)	(0.205)	(0.206)	(0.146)	(0.127)	(0.133)

Mean of panel-robust standard errors in parentheses. het. α_1, random unobserved heterogeneity; sel. α_1, selection into treatment with respect to α_{1i}; sel. α_2, selection into treatment with respect to α_{1i} and α_{2i}. For exact definition of simulation scenarios (1)–(8), see text and corresponding table in the online appendix on the volume's website.

Source: Simulated data.

but this is not true in general, for example, in settings with serially correlated errors (as shown below). Unlike AB-LDV, the FE and FEIS point estimates are consistent and even unbiased in the presence of serial correlation which affects only the standard errors. Hence, FE or FEIS would clearly be preferable. In fact, FE is the best choice because it is more efficient than FEIS.

In scenarios (4) and (5) there is selection into treatment due to both α_{1i} and α_{2i}. α_{1i} is modeled as in scenario (3). In scenario (4) we model α_{2i} as a normally distributed random variable with variance 1, but mean 0.1 for the treated and mean 0 for the untreated subjects. This results in diverging trends between treatment and control group. In scenario (5), α_{2i} is normally distributed with variance 1, but mean −0.1 for the treated and mean 0 for the untreated subjects. Hence, trends of the treatment and control group converge over time. Both scenarios simulate unobserved heterogeneity with respect to the outcome levels and growth curves, which both are related to treatment. In these cases, FE estimates are biased, as are all other estimators except FEIS. FEIS is the only model that provides estimates that are reasonably close to the true value.[10] Note that the RS model is substantially biased, though according to Raudenbush (2001) it is perfect for this situation. The reason is that RS assumes that the conditional means of the individual growth curves are equal across treatment groups. In our simulations, however, means differ systematically between treatment groups. In many real social science data it can be expected that the latter condition holds, and then the RS model provides biased estimates. Clearly, FEIS is the preferred choice if the parallel trends assumption does not hold.

Finally, in scenarios (6)–(8) we allow for state dependence. We set $\rho = 0.5$ so that there is substantial time-series dependence (but not strong enough to induce severe problems due to weak instruments). Scenario (6) assumes in addition that there is only random unobserved heterogeneity, defined as in scenarios (1) and (2). In scenario (7) there is selection into treatment due to α_{1i} (as in scenario (3)). Theoretically, if unobserved heterogeneity is purely random, the AB-LDV estimator is the only estimator that is consistent. This can be seen from the results of our simulations (column (6)). AB-LDV is the only estimator that is reasonably close to one. The estimator is even close to the true parameter if heterogeneity is related to treatment (column (7)). However, this result strongly depends on the assumption of no serial correlation in the idiosyncratic errors. To show this, we alter the setup of scenario (7) to allow for such serial correlation. In scenario (8) we assume a first-order autoregressive process $\epsilon_{it} = \gamma\epsilon_{it-1} + \nu_{it}$, where $\gamma = 0.2$ and ν_{it} is a Gaussian random variable. Serial correlation of this or even greater magnitude is often found with panel data. The results show that the AB-LDV model provides estimates that are heavily biased in such situations.

We have already emphasized that the POLS-LDV estimator is biased under all conditions. It is obvious that this estimator should not be used (nevertheless, it has been used very often in psychological and sociological research). However, that even the acclaimed AB-LDV estimator is biased under the very realistic scenario (8) is dramatic. This leads to the conclusion that hitherto no estimator has existed that allows a treatment effect to be estimated consistently if both true state dependence and self-selection are present in the data. Fortunately, as argued above, for theoretical reasons there is often no need to model true state dependence anyway.

AN APPLICATION: MARRIAGE AND HAPPINESS

In this section we want to apply the most important models discussed so far to real world panel data. We will use data from the German Socio-Economic Panel 1984–2009 (SOEP, 2010). The SOEP is a household panel with annual waves (for a description, see Wagner et al., 2007). Thus, for our analyses panels covering up to 26 waves are available. It is rare to have such long panels. However, it is obvious that long panels make the task of identifying the detailed nature of a causal effect easier. We will make ample use of this advantage.

Our research question is how marriage affects life satisfaction (happiness). The common-sense hypothesis would be that marriage makes people happy. However, there is the potential for self-selection: it is very plausible that it is the happier people who marry, because it is easier for them to find a partner (cf. Stutzer and Frey, 2005). So our working hypothesis is that POLS estimates of the marriage effect should be too high. An FE approach should provide lower estimates that perhaps are even close to zero.

This is a question on the effects of an event (marriage). It is in such questions that a within approach can unfold its full potential (cf. Allison, 1994). FE will compare happiness before and after a marriage within persons. Therefore, self-selection into marriage will not bias results.

Preparing the data

Every year since 1984 the SOEP questionnaire has concluded with the happiness question: ‘We would like to ask you about your satisfaction with your life in general.’ Respondents can answer on an 11-point scale from 0 (‘completely dissatisfied’) to 10 (‘completely satisfied’). This will be our outcome variable. It is an ordinal variable; however, we follow the standard approach in happiness research and treat it as a metric variable.[11]

Our treatment variable is marriage. We restrict our analysis to first marriages. An advantage of panel analysis is that it is possible to model and identify time-varying causal effects. Therefore, the analyst should think about how to model the time path of the causal effect. Simply adding an event dummy (as we did in the first section of this chapter) assumes that the causal effect is immediate and permanent. In most situations, however, one would want to use a more flexible modeling by adding an event time clock. One creates a new variable that measures time since the event has occurred and includes power terms thereof in the model (or, even more flexibly, splines or dummies). In our case we include 'years since marriage'.

Further, we include three controls. First, with panel data it is possible to separate age effects from cohort effects (see below). Therefore, including these variables in panel models should be standard. For didactic reasons, in a first step we include age only linearly. More detailed age/cohort analyses follow later. Second, we include household income as a time-varying variable (natural logarithm). Third, we include sex (time-constant).

Panel data preparation generally is much more complex than cross-sectional data preparation. Here we do not have the space to dwell on this in more detail.[12] However, we want to emphasize one important point, because it deviates from what one is used to from cross-sectional data preparation: *one should restrict the estimation sample to those persons who can potentially experience the treatment during the observation window*. In our case, treatment is first marriage. Thus, we restrict our estimation sample to those who are 'never married' when entering the SOEP. Those who are already married (or divorced, widowed, etc.) when entering the SOEP are dropped. Second, we exclude all person-years after the end of a marriage (separation, divorce, widowhood). Third, we exclude all persons who have only one person-year. In the SOEP (v26) there are 422,734 happiness person-years from 51,543 persons. In our estimation sample that is restricted as described above we are left with 122,919 person-years from 14,634 persons. Thus our estimation sample comprises only 29% of all available person-years.

This may seem strange because, coming from a between approach, social scientists are accustomed to 'hunt for each observation'. The standard approach is to base the between comparison on as many observations as available (sometimes missing data are even imputed). From a within approach, however, it seems reasonable to restrict the estimation sample to those person-years that potentially contribute to within estimation of the treatment effect. In the current context, these are person-years from persons who are 'never married' at the beginning and who have more than one person-year. Person-years of those never marrying are included because they provide the control group for estimating the common age effect. Furthermore, person-years after a marriage has dissolved contribute nothing to estimating the marriage effect and are therefore discarded.

Comparing different regression models

In this section we will compare the results of five regression models. POLS, RE and FE are canonical. We supplement the set by an extreme between approach: a regression that is run on the person-specific means is called 'between regression' (BE). The model is given in equation (15.3) and uses between variation only. Finally, we add another within approach: first differences regression (FD).[13]

For didactic reasons we simplify our modeling strategy and only include a marriage dummy. Before running a within panel regression, one should investigate whether there is enough within variation on the regressors. With 3793 marriages observed in our estimation sample, there is certainly enough within variation on the variable of main interest. It is no surprise that this is also the case for age and household income.

In Table 15.4 we present the results of our model comparison. Our estimates are based on 14,634 persons. BE regression is based on the person-specific means. The other regressions utilize the panel structure of the data and are based on the 121,919 person-years provided by these persons. However, FD is based on fewer cases, because we lose one person-year per person and another one for each gap in the panels (see above). There are 2614 gaps in the data. We lose 123 persons because they do not have any consecutive person-years.

To assess model fit we report overall R^2 for BE and POLS. For the other three models we report the within R^2. Our BE and POLS models are not very successful in explaining the overall happiness variation in the data. Too many other factors affect happiness. For RE, FE and FD it makes sense to report the within R^2, because it is the happiness variation within the persons that we want to explain with our time-varying regressors. The FE model, for instance, succeeds in explaining 1.6% of the within person happiness variation. More precisely, changes in marital status, age and household income can explain 1.6% of the happiness variation of a person over time. The remaining 98.4% is produced by other factors that change over time (health, labor market status, mood, weather, etc.). The strict exogeneity assumption states that these factors must not be correlated with marriage. Otherwise the within marriage effect estimate will be biased.

In all models (with the natural exception of BE) we estimate panel-robust standard errors. These are in most cases larger than conventional standard errors. This results in lower t-values. For instance, the conventional marriage effect t-value is 9.95, whereas in Table 15.4 we report a t-value of 7.37 (for FE). In any case, the marriage effect is highly significant.

Looking at the estimated marriage effect, we see that BE, in particular, provides a quite strong marriage effect: married people are happier by 0.34 scale points. However, the within estimates show that this is an overestimate: the happiness increase with marriage is much lower, namely only 0.14 scale points (FD). The most plausible explanation for this result is self-selection of the happy into marriage.

A subtle, but important point concerning the interpretation of regression coefficients from within models has to be made here. Whereas with standard between regression we compare different people, with within regression we investigate changes over time within the same people. Thus, standard regression coefficients tell us how people differ. For instance, the marriage coefficient in the BE model tells us that married people are on average happier by 0.34 points. However, regression coefficients from within models tell us how the outcome changes when treatment status changes. Thus, the FE estimate tells us that after a marriage, happiness increases by 0.17 points on average. It is obvious that such a change coefficient is closer to what we mean by a treatment effect than a coefficient that reports a difference between people only.

Concerning the age effect, we see that the within estimators are largest (in absolute terms): in 10 years happiness declines by 0.41 scale points (FE). BE, POLS and RE seem to underestimate the negative age effect. Below we will demonstrate that this underestimation is due to not controlling for cohort (older cohorts are happier) and to self-selection (the happy live longer).

Concerning the effect of household income, we see that between estimators strongly overestimate the effect. FE (and especially FD) provide much lower estimates. Again the most plausible reason for this are confounders: unobservables that increase both happiness and income. Nevertheless even the most conservative FD estimate predicts that with an income gain happiness increases significantly.

Finally, the effect of sex cannot be estimated by the within estimators because sex is time-constant. The other three models provide an estimate of the sex effect that is very similar across models: women are (marginally) happier by 0.05 scale points (RE). This effect is statistically significant at the 5% level.

The conclusion of our model comparison is that BE and POLS (and to a lower extent also RE) provide biased estimates. Obviously, using between regression models in happiness research has

Table 15.4 The effect of marriage on happiness: comparing five regression models

	(1) BE	(2) POLS	(3) RE	(4) FE	(5) FD
Marriage	0.343***	0.190***	0.098***	0.167***	0.143***
	(8.80)	(7.34)	(4.98)	(7.37)	(5.07)
Age	−0.009***	−0.014***	−0.027***	−0.041***	−0.062***
	(−9.25)	(−11.42)	(−24.78)	(−23.96)	(−23.09)
Household income (ln)	0.480***	0.325***	0.162***	0.125***	0.048**
	(25.84)	(23.18)	(15.28)	(10.16)	(3.22)
Woman	0.058**	0.060**	0.051*	–	–
	(2.84)	(2.65)	(2.49)		
(Within) R^2	0.024	0.027	0.014	0.016	0.015
Persons	14,634	14,634	14,634	14,634	14,511
Person-years	–	121,919	121,919	121,919	104,671

t-values in parentheses (based on panel-robust standard errors).
* $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Source: own computations

Data: SOEP 1984–2009 (v26)

the potential to yield grossly misleading results. Experience tells that this can be generalized: trying to infer causal effects from regression models that use variation between individuals leads to misleading conclusions in most cases. The reason is that people tend to self-select into (or out of) treatment. Therefore, it is generally not a good idea to compare different people to identify a causal effect. It is much better to try to identify the causal effect by looking at within person changes.

Presenting the results from a fixed-effects regression

In the previous subsection we modeled the causal effect very simply using an event dummy. We now want to use a more flexible modeling strategy. Instead of a single marriage dummy, we add a separate dummy for every year after marriage (cf. Allison, 1994). Altogether we add 25 'years since marriage dummies' to the FE model. Such a modeling strategy allows us to identify the time path of the causal effect in a very differentiated way. Further, to investigate whether there is an anticipation effect we also add a dummy for the year before marriage.

Presenting the results of such a regression model in a table would not be very helpful, because there are so many regression coefficients. In fact, some people would even argue that this is a waste of paper. It is much better to present the results in a graph. In principle there are three helpful graphs for interpreting regression models (see Chapter 10 in this volume): (i) one could plot the predicted values (profile plot); (ii) one could plot the regression coefficients (or more generally: the average marginal effects, AMEs) (effect plot); or (iii) one could plot the AMEs of X conditional on the values of Z (conditional effect plot). To show in a graph how the marriage effect evolves over time, we plot the regression coefficients of the marriage year dummies over time (with Stata's `marginsplot` command). We plot only over the first 10 years of marriage, because afterwards the estimates become very imprecise due to small numbers of cases.

As can be seen in Figure 15.3, in the year of marriage (year 0) happiness is higher by 0.38 scale points. The reference group is the average happiness in all person-years at least 2 years

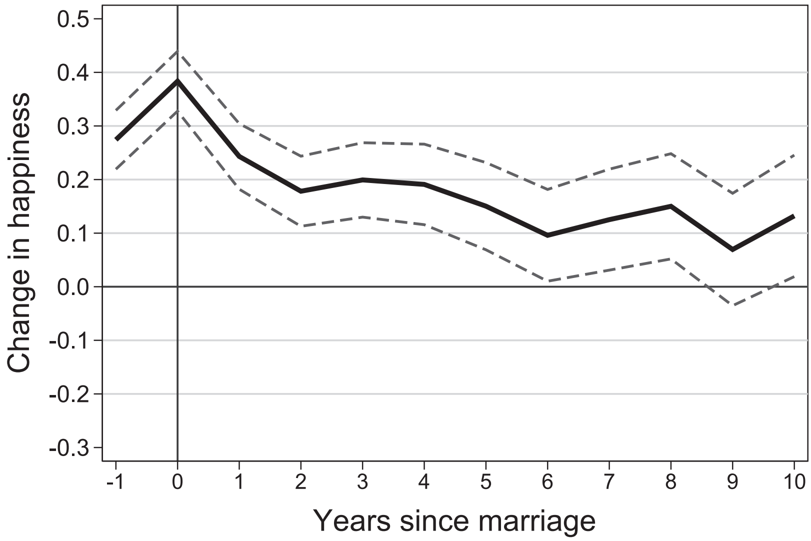


Figure 15.3 The time path of the marriage effect

before marriage. This marriage effect is much larger than that reported above (0.17). The reason is that the dummy marriage effect averages over all the (after) marriage effects we see in Figure 15.3, and many of these are much smaller. As we can see, the marriage effect quickly declines to 0.15 after 5 years. In the ninth year it becomes even insignificant, as indicated by the confidence interval (dashed line) that overlaps zero. In summary, this means that the marriage effect is immediate and strong. However, it is not persistent. It declines quickly and vanishes after about 10 years. Furthermore, we see a clear anticipation effect: in the year before marriage happiness is higher by 0.27 points.

We conclude that the modeling strategy suggested here allows us a very detailed investigation of the time path of a causal effect, including possible anticipation effects.

Modeling individual growth

So far we have emphasized that one advantage of panel data is that they can help in identifying causal effects. In the previous subsection, we saw that panel data even help in identifying the time path of a causal effect. Closely related is the second big advantage of panel data: panel data allow us to model individual dynamics, that is to say, with panel data we can model the development of the outcome over time. In most cases time will be 'age', but it could also be labor market experience, etc. Curves that describe the development of the outcome with age are called 'growth curves'.

A growth curve is easily modeled by including age terms in a regression model. Age dummies are most flexible, but age splines or age polynomials are also possible. Wunder et al. (2013) give a detailed discussion on the different modeling possibilities.

When modeling growth curves it is important to control for birth cohort (if one has panel data that were collected from several cohorts). As sociologists know, cohort effects are ubiquitous and therefore might spoil the age effects found. Cross-sectional data offer no solution here, because it is impossible to separate age effects from cohort effect. However, with panel data we can separate both effects due to repeated measurement of the outcome at different ages.

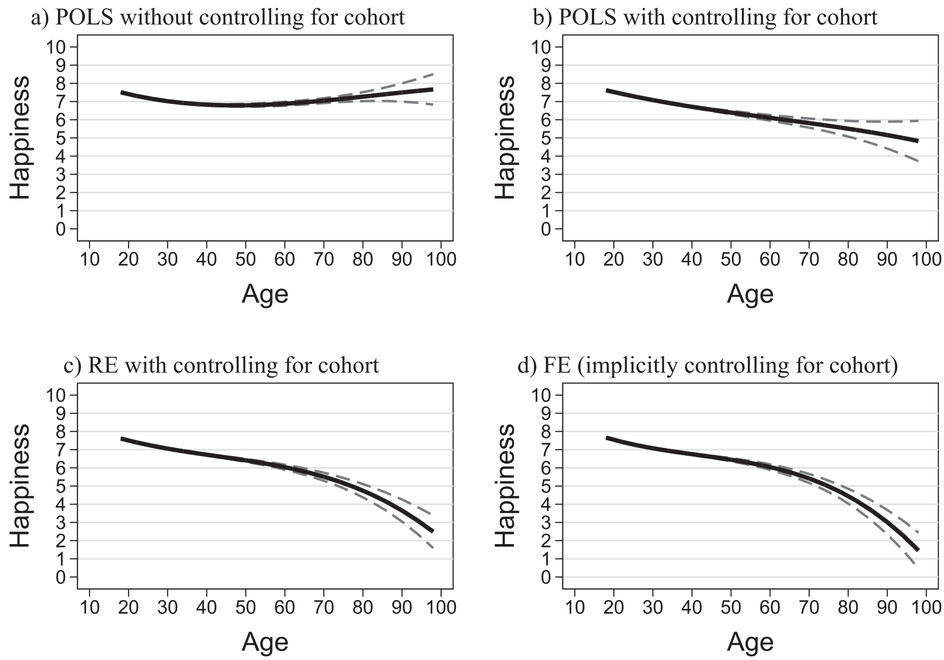


Figure 15.4 Comparing cubic growth curves obtained with different methods

When using FE to estimate a growth curve, cohort is automatically controlled for (because it is time-constant).

Growth curves are often estimated by RE modeling or structural equation modeling (latent growth curves). However, these modeling strategies potentially provide biased growth curves if there is self-selection: people are selected into different age groups according to their value on the outcome. This might happen because of differential mortality, or also because of differential panel attrition. Therefore, we recommend FE methodology for estimating growth curves. FE growth curves have the advantage that they provide a within age effect: the age effect is not estimated by comparing people of different ages, but by looking how the outcome changes when persons grow 1 year older. This avoids self-selection bias. Furthermore, Frijters and Beatton (2012) demonstrate that biased estimates of other regressors also might bias growth curves. In so far as FE helps in identifying unbiased effects of regressors, it will also help to estimate an unbiased growth curve.

To demonstrate growth curve estimation we estimate happiness regressions that include a linear, quadratic and cubic age term (cubic growth curve). As controls we include the marriage dummy and household income. After estimation we plot predicted values by age (profile plots using Stata's `marginsplot` command). Figure 15.4 presents the results of four different estimation strategies. In Figure 15.4a we present the happiness growth curve obtained with POLS. As can be seen, we get a U-shaped growth curve. Happiness declines slightly until age 50 and then increases again. Such a pattern has been reported in numerous studies, especially in the economics literature. However, common sense tells us this is a quite surprising result: it seems implausible that old people are happier than middle-aged people. Many studies show that happiness declines with deteriorating health. Therefore, because we do not control for health, happiness should decline with age. Meanwhile it has been shown by several studies (e.g. Frijters and Beatton, 2012; Wunder et al., 2013) that a U-shaped happiness growth curve is an artifact of not controlling for cohort and of not taking regard of potential self-selection.

Therefore, in Figure 15.4b we estimate the growth curve after controlling for cohort (by including 90 birth year dummies). We see that the growth curve now declines monotonically. This striking change in the growth curve pattern happens because there is a cohort effect in the data. As further inspection showed us, younger cohorts are on average less happy. In particular, the cohorts born after the Second World War are two full scale points less happy than the oldest cohorts born before the First World War.

However, we still do not see a pronounced old age decline in happiness. This might be due to self-selection at higher ages. Taking regard of self-selection, we further estimate RE (Figure 15.4c) and FE models (Figure 15.4d). In fact, we now see that there is a steep decline in happiness starting at about age 70. The decline estimated with FE is even steeper than that estimated with RE. Such a decline seems very plausible because, as Gerstorf et al. (2010) show, 3–5 years before death there is a 'terminal decline' in happiness.

The age–period–cohort problem

Until now we have ignored period effects. It is well known that models which do not include a period variable are potentially misspecified: age (and cohort) effects are potentially biased. Period effects (i.e. idiosyncratic events shortly before the panel interview took place and that affect the outcome) are ubiquitous. Therefore, it is generally recommended to include period effects in panel regression models. Including period effects in addition to age and cohort effects will not work, however. Age (in years), period (interview year), and cohort (birth year) are linearly dependent due to the relation $age = period - cohort$. Therefore, it is not possible to include all three terms in one regression model (the so-called age–period–cohort (APC) problem). In an FE model the problem already occurs if we include age and period, because cohort is implicitly controlled for.

As a solution one has to impose some kind of non-linear restrictions (APC restrictions). For instance, if one tries to estimate an FE model including age and interview year dummies, Stata automatically drops the last year dummy (in addition to the first year dummy that is dropped as a base). With this APC restriction perfect collinearity is broken and the model can be estimated. Unfortunately, this provides reasonable results only if the restriction is plausible (i.e. the mean outcomes are very similar in the first and last year). Otherwise, the growth curve will be misleading. Therefore, one should not use automatic APC restrictions. The researcher has to think very carefully about which APC restriction makes sense.

Therefore, we closely inspected the interview year dummies in models with a cohort restriction (cohort is missing): happiness in the years 1984, 1985, 1986, 1990, and 1991 turned out to be highest and at a very similar level. Hence, for our final model that includes period effects we define those five years as base category. Now we can estimate an FE growth curve that controls for cohort (implicitly) and period (with APC restrictions). In addition, we want to go beyond a parametric specification of the age growth curve. This is why we now include a full set of age dummies in the FE model instead of a cubic specification. The result can be seen in Figure 15.5.

Compared to the FE growth curve without controlling for period effects (Figure 15.4d), we see an important change: happiness stays essentially constant up to age 65. Only then does the decline in happiness begin. In addition, due to our dummy specification, we see two interesting details. There is a little 'happiness hump' around age 60. Wunder et al. (2013) argue that this is due to people anticipating and/or entering retirement. Further, there is a sharp happiness drop before age 20. This is due to a panel-conditioning effect, where respondents in their first three

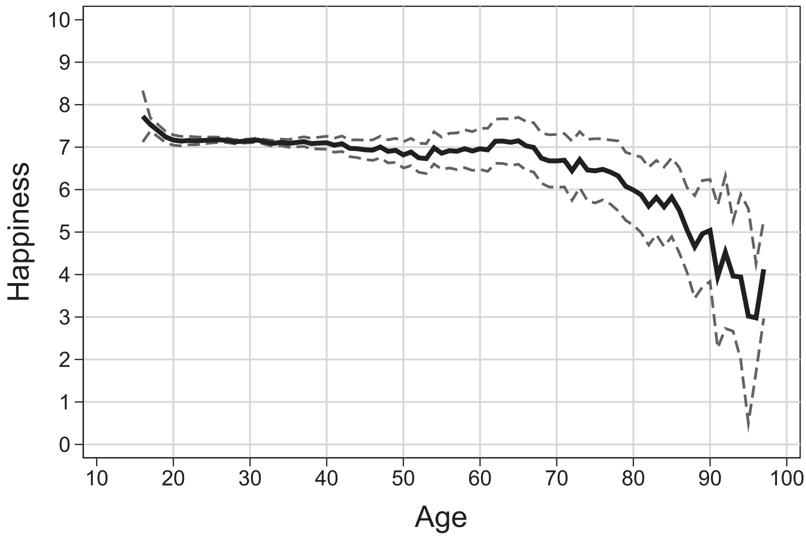


Figure 15.5 A non-parametric FE growth curve (base category years 1984, 1985, 1986, 1990 and 1991).

SOEP years report systematically higher happiness values. This can be confirmed by including first year dummies (not shown here).

One has to be cautious, however, in interpreting these results. We cannot be sure that our APC restriction ‘works’. Therefore, a better solution to the APC problem could be to impose substantive restrictions. Cohort and period dummies are only proxies for specific circumstances and events. For instance, people who have grown up during a war might have higher happiness levels throughout their lives after the end of the war. Another example is that Germans were very happy in 1990 and 1991 because of German unification and because their soccer team won the World Cup. Therefore, instead of the proxies one could introduce the substantive relevant variables directly. It seems difficult to get a handle on all variables that produce cohort effects. It might be easier to identify those (large-scale) events that produce period effects. Hence, we recommend including age and cohort terms, dropping the period terms, and instead including event dummies. Further important period variables are usually gross national product and the unemployment rate.

Finally, age is a proxy variable, too. It captures increasing human capital, maturation, declining health, and so on. Therefore, growth curves do not have a causal interpretation. Growth curves are a description of how the outcome develops over time. We hope that the reader of this section has been convinced that such a description is of value in itself. However, the mechanisms that produce the growth curve can also be uncovered by pushing the analysis further: including indicators for the mechanisms should flatten the growth curve. For instance, if the happiness growth curve is mainly due to the development of health, then the inclusion of health measures should make the growth curve flat (this is not the case with the SOEP data; cf. Wunder et al., 2013). Thus, the growth curve methodology as presented in this section can also be very helpful for causal analysis.

CAVEATS AND FREQUENT ERRORS

The main point with FE estimation is that it discards potentially ‘contaminated’ between variation. It uses only within variation to estimate the causal effect of an event. Thus, to identify the causal effect FE only requires the assumption that the within variation is exogenous. This is a weaker assumption, compared to POLS or RE, where both between and within variation have to be exogenous. Several implications follow:

1 Usually it is not a good idea to use also the between variation for causal inference (i.e. to use POLS or RE).
2 However, for descriptive purposes it makes sense to also use the between variation.
3 Because the FE estimator uses only the within variation, it is only generalizable to those units that show within variation.
4 If the within variation is not exogenous the FE estimator will yield biased estimates.

Concerning (1), in most non-experimental social science settings, between variation will be endogenous due to self-selection. Thus, for the purpose of causal analysis one should avoid models that rest on the assumption of exogenous between variation. However, some researchers are uncomfortable with discarding so much variation and therefore use RE or POLS. Two arguments are put forward in this context. (i) One should use RE estimation because it is more efficient. It is argued that estimators should be judged not only by bias, but also by their standard error (e.g. according to the mean square error). This argument certainly makes sense from a statistical perspective. Nevertheless, experience tells us that the bias introduced by using the between variation is so huge that the lower standard error is bought at an exorbitant price: estimates are so far off the true value that the whole analysis is misleading. (ii) The effects of time-constant variables are interesting, and therefore one should use RE estimation (or the hybrid model; cf. Firebaugh et al., 2013). Obviously, this argument is motivated by the common practice of reporting big regression tables with many coefficients. Again, this might spoil the whole analysis. If one is interested in the effects of time-constant variables, it makes much more sense to estimate group-specific growth curves (by including group–age interactions).

Nevertheless, and turning now to implication (2), in the social sciences not only causal questions are of value. Descriptive questions are also of prime interest. Some even argue that we should first have good descriptions of the social world, before we start analyzing causal effects. For descriptive purposes the between variation is important. Therefore, for descriptive purposes one should use POLS (not RE, because RE is 'biased' toward FE). For instance, if one wants to investigate the causal wage effect of migrating from West Germany to East Germany, then one should use only the within variation produced by those who actually migrate. However, if one simply wants to describe what the wage difference is between West and East Germans, one clearly should use all variation in the data (i.e. estimate the wage difference by POLS). In fact, panel data are suboptimal for such descriptive purposes, because between and within variation are mixed. Trend data from independent cross-sections are much better suited for descriptive analyses.

Concerning (3), by using only the variation of those who show within variation in treatment status, FE estimators cannot be generalized to the whole population. This point confuses many social researchers, because from cross-sectional research they are used to generalizing to the whole population (given that we have a random sample from the whole population). For instance, if we compare the wages of non-married and married people, we would generalize the wage differential found to the whole population. This makes sense for a descriptive research question.

However, a within estimator provides a treatment effect. And this can be generalized only to those who potentially experience such a treatment. For instance, the FE estimator on the marital wage premium uses only the information from those who have married during the observation period. Therefore, we can generalize the result found only to those in the population who marry. In the counterfactual literature this is called an 'average treatment effect on the treated' (ATT). For instance, those who do not marry might have quite different marriage effects (effect heterogeneity). Thus, the within marriage effect found cannot be generalized to them. However, this is not a shortcoming of FE estimation. It merely reflects the real-world fact that not everybody marries. And therefore we are satisfied with an ATT that only makes a statement on the causal effect of those who experience the treatment. Note that with an experimental design we would 'force' some people to marry, who in the real world would never marry. This would provide an 'average treatment effect' (ATE) that could be generalized to the whole population. From a substantive point of view, such an ATE does not seem to be preferable.

Policy recommendations, however, should be based on an ATE. For example, if we use an FE model and find a strong positive effect of a training program for some unemployed people on their subsequent wages, it is not clear whether we would find a similar effect for all other unemployed people who did not participate. Even after controlling for unobserved ability, participants might benefit more than other people would. In this case of 'differential treatment effects', generalizations to the whole population would be misleading (Morgan and Winship, 2007).

Finally, concerning (4), within estimation fails if strict exogeneity does not hold. Generally there are three sources of endogeneity (cf. Wooldridge, 2010, p. 321). (i) There may be time-varying confounders (unobservables that affect both the outcome and the treatment). For instance, in the context of our marital wage premium example, people might undergo cosmetic surgery and as a consequence have a higher probability of marriage and earn more money (beauty premium). Then what in fact is a beauty premium would erroneously be attributed to marriage. (ii) There may be simultaneity, that is, a change in the outcome may affect treatment (reverse causality). For instance, men might react to a wage increase with a higher probability of marriage. Again, FE would erroneously estimate a marital wage premium. (iii) There might be measurement errors in the treatment indicator. It is well known that this might also bias estimates.

If at least one of these mechanisms is at work, then the within variation used by FE to estimate the effect is no longer exogenous and FE fails to identify the true causal effect. All three of these sources of endogeneity are potentially present in most social science applications. Therefore, one should always be critical in the face of the FE estimates found and look for arguments why the strict exogeneity assumption might be violated. Only if no such arguments are found, one should provisionally accept the FE results.

Different methods have been suggested to deal with endogeneity in FE models (instrumental variables, structural equation modeling). Generally these methods rest on untestable assumptions and do not provide robust results. Research fields where these methods have been used abundantly are full of contradictory results. For example, Mouw (2006) gives an overview of longitudinal studies on the effects of social capital and drives this point home very nicely. Therefore, we do not recommend these methods. Instead of investing in complicated statistical modeling it is recommended to invest in 'shoe leather' – that is, to go out and collect better data that, for instance, include information on the supposed time-varying confounder or simply are measured with greater precision.

As we saw in the second section of this chapter, strict exogeneity can be weakened by allowing for individual-specific trends (FEIS). In many applications it seems likely that the parallel trends assumption does not hold, because there is self-selection not only on the outcome level but also on the outcome growth. For instance, Ludwig and Brüderl (2011) argue not only that there is self-selection of high earners into marriage, but also that those on a steeper wage profile self-select into marriage. If this is the case, FE overestimates the marital wage premium (see the simulations above). They actually find that the marital wage premium in Germany and the US vanishes if one uses FEIS. They interpret this as evidence that there is self-selection on wage growth and that no causal marital wage premium exists.

FURTHER READING

The main reference for FE regression models for panel data is Allison (2009). He discusses not only linear FE models, as we have in this chapter, but also non-linear FE regression models (fixed-effects logistic, count data, and Cox models). Halaby (2004) gives a short but excellent introduction to FE modeling, including a discussion of LDV models. In addition, he dissects the erroneous arguments that sociologists often put forward against FE regression (Rogosa, 1988, does the same with psychologists' arguments). Firebaugh et al. (2013) discuss panel models with a focus on the hybrid model. Econometric details on panel regression methods can be found, for instance, in Baltagi (2008), Cameron and Trivedi (2005) and Wooldridge (2010). Finally, introductions into the modern methods of causal inference – of which FE regression is one – can be found in Angrist and Pischke (2009) and Morgan and Winship (2007).

NOTES

* We thank Henning Best, Klaus Pforr, Patrick Riordan and an anonymous reviewer for careful reading and helpful comments on an earlier version of this chapter.

1 By 'unobserved heterogeneity' we mean unobserved heterogeneity that is correlated with the regressors (i.e. endogenous). In the following we will often use this somewhat loose wording which, however, is common in the social sciences.

2 The data (panelanalysis stylized.dta) and the Stata do-file (panelanalysis stylized.do) are available on the volume's website.

3 If one happens to be interested in estimates of inter-individual differences then one should apply LSDV. The estimates of the individual constants are unbiased under strict exogeneity, but inconsistent as $N \to \infty$ because, with T fixed, adding one parameter for each individual prevents convergence in probability. In general, one would like to have many observations per subject to get reliable estimates of the constants. Furthermore, for time-dependent processes, where process time might be age, tenure or marriage duration, each individual should be observed from the same point in time onward. (In event history parlance, this rules out left-censoring.) Otherwise the estimates of individual constants might be very misleading.

4 The procedure actually is an application of the Frisch–Waugh–Lovell theorem for partitioned regression. See Baltagi (2008) for a detailed exposition of the basic FE estimator using this framework.

5 The model actually is not as hungry as Morgan and Winship suggest. If we just want to allow for individual-specific constants and linear time trends, three observations are sufficient.

6 What we term 'fixed coefficients' here is usually called 'fixed effects'. This terminology is obviously confusing. What is meant is that the effect is assumed homogeneous across subjects. It is not a fixed effect in the sense used in this chapter.

7 We deviate here from the usual notation in the multi-level literature where the index of the lower level (t in this case) comes first.

8 When motivated by a multi-level framework, the conventional RE model is usually called the 'random-intercept' (RI) model and is estimated by maximum likelihood (instead of FGLS). In practice, maximum likelihood and FGLS often yield very similar coefficients in large samples.

9 The Stata do-file (panelsim.do) used to run the simulations is available for download on the volume's website.

10 The AB-LDV provides reasonable estimates in scenario (4) merely by coincidence, as further simulations with different values for the conditional means of α_{2i} showed us.

11 An anonymized version of the data file that we extracted from the SOEP (Happiness Anonym.dta) and the Stata do-file (Happiness Regressions.do) are available on the volume's website.

12 A tutorial on panel data preparation can be found on our 'Teaching Materials' page under the heading 'Working with panel data': `http://www.ls3.soziologie.uni-muenchen.de/teach-materials`.

13 Another potential candidate would be the 'hybrid model'. This model provides the FE estimates for time-varying regressors, and RE-type estimates for time-constant regressors (Allison, 2009; Firebaugh et al., 2013). However, the effect of time-constant regressors such as sex should not be interpreted as causal effects in this model. Generally, there is much confusion over how to interpret the estimates of this model. For causal analysis the model is therefore not useful.

14 The marriage effects estimated by BE, POLS, and RE are even biased downwards, because we do not control for cohort in these models (FE and FD do implicitly). This biases the age effect upwards (see below), which in turn biases the marriage effect downwards.

REFERENCES

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

Allison, P. D. (2009). *Fixed Effects Regression Models*. Thousand Oaks, CA: Sage.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58, 277–297.

Arránz Becker, O., Salzburger, V., Lois, N. and Nauck, B. (2013). What narrows the stepgap? Closeness between parents and adult (step)children in Germany. *Journal of Marriage and Family*, 75(5), 1130–1148.

Baltagi, B. (2008). *Econometric Analysis of Panel Data*, 4th edn. Chichester: Wiley.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87, 115–143.

Bun, M. J. G. and Windmeijer, F. (2010). The weak instrument problem of the system GMM estimator in dynamic panel data models. *Econometrics Journal*, 13, 95–126.

Cameron, A. and Trivedi, P. (2005). *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Firebaugh, G., Warner, C. and Massoglia, M. (2013). Fixed effects, random effects, and hybrid models for causal analysis. In S. L. Morgan (ed.), *Handbook of Causal Inference for Social Research* (pp. 113–132). New York: Springer.

Frijters, P. and Beatton, T. (2012). The mystery of the U-shaped relationship between happiness and age. *Journal of Economic Behavior & Organization*, 82, 525–542.

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology*, 36, 21–47.

Gerstorf, D., Ram, N., Mayraz, G., Hidajat, M., Lindenberger, U. and Wagner, G. (2010). Late-life decline in well-being across adulthood. *Psychology and Aging*, 25, 477–485.

Halaby, C. (2004). Panel models in sociological research: theory into practice. *Annual Review of Sociology*, 30, 507–544.

Hinz, T. and Gartner, H. (2005). Geschlechtsspezifische Lohnunterschiede in Branchen, Berufen und Betrieben. *Zeitschrift für Soziologie*, 34, 22–39.

Kézdi, G. (2004). Robust standard error estimation in fixed-effects panel models. *Hungarian Statistical Review*, 9, 96–116.

Legewie, J. (2012). Die Schätzung von kausalen Effekten: Überlegung zu Methoden der Kausalanalyse anhand von Kontexteffekten in der Schule. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 64, 123–153.

Ludwig, V. and Brüderl, J. (2011). *Is there a Male Marital Wage Premium? Resolving an Enduring Puzzle with Panel Data from Germany and the U.S*. Paper presented at RC 28 meeting in Essex.

Morgan, S. and Winship, C. (2007). *Counterfactuals and Causal Inference*. New York: Cambridge University Press.

Mouw, T. (2006). Estimating the causal effect of social capital: A review of recent research. *Annual Review of Sociology*, 32, 79–102.

Nickell, S. J. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49, 1417–1426.

Phillips, P. C. B. and Sul, D. (2007). Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *Journal of Econometrics*, 137, 162–188.

Polachek, S. and Kim, M.-K. (1994). Panel estimates of the gender earnings gap: Individual-specific intercept and individual-specific slope models. *Journal of Econometrics*, 61, 23–42.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, 52, 501–525.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie (ed.), *Methodological Issues in Aging Research* (pp. 171–209). New York: Springer.

SOEP (2010). *Data for years 1984-2009, version 26*. doi:10.5684/soep.v26.

Stock, J. H. and Watson, M. W. (2008). Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica*, 76, 155–174.

Stutzer, A. and Frey, B. S. (2005). Does marriage make people happy or do happy people get married? *Journal of Socio-Economics*, 35, 326–347.

Wagner, G. G., Frick, J. R. and Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP) – scope, evolution and enhancements. *Schmollers Jahrbuch*, 127, 139–169.
Wooldridge, J. (2010). *The Econometrics of Cross-Section and Panel Data*, 2nd edn. Cambridge, MA: MIT Press.
Wunder, C., Wiencierz, A., Schwarze, J. and Küchenhoff, H. (2013). Well-being over the life span. *Review of Economics and Statistics*, 95, 154–167.