Chapter 6
Dataset Design and Estimation Methods

All models are wrong, but some models are useful.
(George E.P. Box 1979)

Abstract The phenomenon under investigation guides the data collection process, the structural design of the dataset as well as the choice of empirical methods (Blossfeld et al. 2007, p. 4). The data collection process for the German laser industry database was described above. Now, we will turn our attention to the two latter points. Chapter 6 is divided into two sections: Section 6.1 presents the two compiled datasets. On the one hand, an event history dataset was constructed to analyze the propensity and timing of laser source manufacturers to cooperate and enter the German laser industry innovation network. On the other hand, a panel dataset was employed to analyze the determinants of firms in the German laser industry from various angles. Section 6.2 provides an overview and general discussion of estimation methods which were applied in Part IV of this book. We start with a brief discussion on non-parametric event history analysis models using continuous time, followed by an introduction of econometric models for panel count data.

6.1 Design and Scope of the Compiled Datasets

In essence, there are three types of dataset designs: cross-sectional datasets, panel datasets and event history datasets (Blossfeld et al. 2007, pp. 5–21). Due to the aim of this study focus is placed on the two latter dataset designs.

6.1.1 Dataset I: Event History Data Structure

The use of event history analysis methods requires a relatively demanding data design. Firstly, the compilation of an event-oriented longitudinal dataset requires the appropriate choice of time intervals and information on the origin state and destination state (Blossfeld et al. 2007, p. 42). Secondly, data has to be organized in

M. Kudic, *Innovation Networks in the German Laser Industry*,
Economic Complexity and Evolution, DOI 10.1007/978-3-319-07935-6_6

an event-oriented design in which each record is related to a particular duration in a predefined state (Blossfeld et al. 2007, p. 42). Finally, precise time-tracking of start and end dates is needed for the event under investigation to analyze the transitions from one state to another.

Some issues, however, require particular attention. The most notable is the setup of the analytical framework. Both dimensions of the analytical framework – "state space" and "time space" – have to be defined carefully to avoid misspecification (Blossfeld et al. 2007, p. 38). The choice of these two dimensions is driven by theoretical considerations and determines the choice of the empirical estimation models. In general, there are four types of models: the "single episode model", "multi state model", "multi episode model", and "multi state multi episode model" (Blossfeld et al. 2007, p. 39).[1] For the purpose of this study we employed a design that is based on the following considerations. On the one hand, we seek to understand what factors determine a firm's propensity to cooperate for the first time and enter the laser industry innovation network. On the other hand, we are interested in the factors that affect the length of time until the first cooperation occurs. Consequently, we constructed a single-episode event history dataset that provides the basis for conducting a non-parametric spell-duration analysis.

The single-episode event history dataset for the German laser industry is constructed and organized as follows: The time axis is defined on the basis of century months. All firm foundation dates as well as all start and end dates of cooperation events are given in century months. The unit of analysis is the firm. In cases where the number of censored observation units is small, it is acceptable to simply exclude them (Allison 1984, p. 11). Thus, firms founded before 1990 were excluded from the dataset to avoid left truncation and left censoring problems (Blossfeld and Rohwer 2002, pp. 39–41). Starting from the full population of 233 LSMs in our sample we identified 39 firms which were founded before 1990. Thus, a total of 194 firms were potentially at risk for conducting the first cooperation event. Out of this population we ended up with a total of 112 firms with at least one cooperation event during the observation period. The event of interest is the first cooperation for all laser manufacturing firms which are at risk in the time period between 1990 and 2010. The dataset allows us to analyze the transition from the origin state ("no cooperation") to the destination state ("first cooperation"). These two states allow us to define the risk set. At the same time the initial cooperation event marks the firm's entry into the network. Repeated events were

[1] Single episode models allow for event transitions from the origin state to the destination state whereas multi-state models allow us to analyze event transitions from the origin state to multiple destination states. In contrast, multi-episode models allow for repeated events or event transitions over time. Finally, multi-state multi-episode models can be applied to analyze both repeated episodes and repeated events. For further details see (Blossfeld et al. 2007).

not considered. Firms were basically considered to have two ways of entering the German laser industry innovation network. The first cooperation event can be either participating in a *Foerderkatalog* project or in a *CORDIS* project. Both types of cooperation event were coded separately by using a dummy variable. All event occurrence dates and durations were recorded in century months. Variables were grouped in the following categories: organizational, relational and contextual.[2]

Organizational variables[3] were included in the dataset to account for heterogeneity across firms. In particular these variables are firm origin *[origin_ev]*[4] and firm size *[firmsize_cat_ev]*.[5] Additionally, a simple dummy variable was created to differentiate between "young" and "mature" firms *[firmage_ev]*.[6] Relational variables were included in the dataset. Both types of publicly funded R&D cooperation projects, *Foerderkatalog* projects as well as *CORDIS* projects, were coded separately *[coop_type_ev]*.[7] Occurrence dates and duration were recorded in century months. Moreover, a set of contextual variables was included in the dataset. For the first set of geographical variables, we split the sample into four geographical regions. We included a set of geographical location variables in our dataset *[region_ev]*[8] indicating whether a firm is located in the northern, southern, eastern or western part of Germany. Finally, we included a set of cluster variables *[clu_ev]* in our dataset indicating whether a firm was located inside or outside of a densely

[2] Note that the following variables in the event history dataset can differ from those that were coded and used in the panel dataset. The suffix "_ev" indicates an event oriented variable.

[3] Organizational level variables were coded at the date of a firm's population entry and considered to be time-invariant for the purpose of the non-parametric spell-duration analysis.

[4] Origin dummies are coded on the basis of the following categories: origin_ev1 = new foundation; origin_ev2 = PRO spin-off; origin_ev3 = LSM spin-off; origin_ev4 = other background, such as: spin-offs from other types of organizations, name change and post-merger firm formations.

[5] The following five size categories were used: firmsize_cat_ev1 = "micro firm" = 1–9 employees; firmsize_cat_ev2 = "small firm" = 10–49 employees; firmsize_cat_ev3 = "medium firm" = 50–249 employees; firmsize_cat_ev4 = "larger firms" = more than 250 employees. This categorization is drawn upon the definition proposed by the European Commission (2005). Missing data for the number of employees were extrapolated based on employee data for the same firms but other observation windows.

[6] We used the mean age of the firms (97 months) in the observation period to split the sample. Definition of "young firms": firmage_ev = 0 if firmage_ev < = 97 months (8.1 years); "mature firms": firmage_ev = 1 if firmage_ev > 97months (8.1 years).

[7] The variable coop_type = 1 in the case of a *CORDIS* project; coop_type = 2 in the case of a Foerderkatalog project.

[8] Definition of the four geographical regions: region_ev1 = Baden-Württemberg (BW) Bavaria (BY); region_ev2 = Bremen (HB) Hamburg (HH) Schleswig-Holstein (SH); region_ev3 = Berlin (B) Brandenburg (BB) Mecklenburg-Western Pomerania (MV) Saxony (S) Thuringia (TH) Saxony-Anhalt (SA); region_ev4 = North Rhine-Westphalia (NW) Lower Saxony (NS) Rhineland-Palatinate (RP) Saarland (SR) Hessen (H). The variable for Saarland had to be omitted due to the non-existence of LSMs in this federal state throughout the entire time period.

crowded region. The four geographical clusters were identified based on the descriptive analysis in Sects. 7.1.2 and 7.1.3.[9]

6.1.2 Dataset II: Panel Data Structure

Panel data methods require the data to be in long form, meaning that each individual time pair in the dataset is a separate observation (Cameron and Trivedi 2009, p. 274). The panel dataset for the German laser industry is constructed and organized as follows: The unit of analysis is the firm year meaning that each firm in the sample is observed for each year. Thus, we decided in favor of annual time intervals for the purpose of this study. The panel is unbalanced due to a considerable proportion of firms entering the sample after 1990 (i.e. new foundations, spin-offs etc.) or leaving the sample before 2010 (i.e. mergers, bankruptcies etc.). Unbalanced data usually causes no significant complications as most empirical methods are designed to handle both balanced and unbalanced panel data (Cameron and Trivedi 2009, p. 230).

Over the course of 21 years we have a total of 233 laser source manufacturers (LSMs) and 2,645 firm years. Thus, we have an average of 11.35 observations per firm. The dataset contains time-variant as well as time-invariant variables organized in content-specific groups of explanatory variables.[10]

The first group of variables encompasses all firm-specific variables. A linear firm age variable *[firmage]* as well as a squared firm age variable *[firmage_sq]* were generated on the basis of firm entry and firm exit dates. Both age variables in our panel dataset are recorded on an annual basis. Data on yearly turnover for each firm provides the basis for the calculation of a time-invariant average turnover variable for each firm in the sample *[avgturnover]*. A set of dummy variables was constructed to account for the origin of firms in the sample *[origin]*[11] and a second set of dummies was generated to account for differences in legal status across firms *[leg_stat]*.[12] In addition, the yearly number of employees was used to code a set of firm size dummies *[firmsize_cat]*.[13]

[9] These clusters are defined as follows: planning regions: 72, 73, 74, 76 & 77 = clu_ev_bw, located in Baden-Württemberg; planning regions: 86, 90 & 93 = clu_ev_bay, located in Bavaria: planning regions: 54 & 56 = clu_ev_thu, located in Thuringia; planning region 30 = clu_ev_B, located in Berlin.

[10] Note that we had to choose a more detailed categorization for most of the panel data variables due to the theoretical considerations (cf. Chaps. 10, 11 and 12).

[11] Origin dummies are coded on the basis of the following categories: origin1 = new foundation; origin2 = name change; origin3 = post merger firm; origin4 = PRO spin-off; origin5 = LSM spin-off; origin6 = spin-offs from other types of organizations.

[12] Legal status dummies are coded on the basis of the following categories: leg_stat1 = GmbH; leg_stat2 = GmbH & Co; leg_stat3 = GmbH & Co KG; leg_stat4 = OHG; leg_stat5 = AG; leg_stat6 = other.

[13] The following five size categories were used: firmsize_cat1 = "micro firm" = 1–9 employees; firmsize_cat2 = "small firm" = 10–49 employees; firmsize_cat3 = "medium firm" = 50–249 employees; firmsize_cat4 = "large firms" = 250–749 employees; firmsize_cat5 = "very large

The second group of variables encompasses all geographical measures on a firm level, regional level and industry level. All geographical variables are coded on the basis of annually updated address data for three types of laser-related organizations: LSMs, PROs and LSPs. At the firm level a set of geographical dummy variables for each LSM was generated indicating the federal state in which the firm is located *[fed_state]*.[14] Two types of geographical co-location measures *[coloclsm, colocpro]* were included in the dataset which were calculated on the basis of the localized density measures outlined above (cf. Sect. 5.3.1). Additionally we split the PRO sample into two sub-samples.[15] Localized geographical density measures were generated by calculating all distances between LSM and PROs in both sub-samples separately *[colocpro_appl, colocpro_basic]*. In order to get a picture of geographical concentration tendencies at the industry level, HHI indices were calculated for LSMs, LSPs and PROs (cf. Sect. 5.3.2) and included in the dataset *[hhi_lsm; hhi_lsp; hhi_pro]*.

The third group of variables encompasses all cooperation-related variables. Data on publicly funded cooperation projects from both "*Foerderkatalog*" as well as "*CORDIS*" databases were used to generate cooperation counts *[coopcnt_fk; coopcnt_c]*, cumulative cooperation counts *[coopcum_fk; coopcum_c]* and cooperation funding *[coopfund_fk; coopfund_c]* on an annual basis. Based on these measures several combined variables were generated which include both project types *[coopcnt_fkc; coopcum_fkc; coopfund_fkc]*. All cooperation funding variables are measured in thousand euros.

The fourth group of variables encompasses network variables calculated at three levels of analysis[16] (cf. Sect. 5.2). The following network level variables were included in the dataset: overall network density *[nw_density]*, network size *[nw_size]*, clustering coefficients *[nw_clust]* a weighted clustering coefficient *[nw_wclust]*, a network fragmentation measure *[nw_compcnt]* and an average reachability measure *[nw_areach]*. The next set of network variables allows us to quantify the structural configuration of ego network characteristics for each firm in the sample. In particular these ego network variables measure the ego network size *[ego_size]*, the ego network density *[ego_density]* and two ego network-based brokerage indicators *[ego_broker; ego_nbroke]*. The last set of network variables

firms" = more than 750 employees. Missing data for the number of employees was extrapolated based on employee data for the same firm but for other firm years.

[14] Definition of federal state dummies: fed_state1 = Baden-Württemberg (BW); fed-state2 = Bavaria (BY); fed_state3 = Berlin (B); fed_state4 = Brandenburg (BB); fed_state5 = Bremen (HB); fed_state6 = Hamburg (HH); fed_state7 = Hessen (H); fed_state8 = Mecklenburg-Western Pomerania (MV); fed_state9 = Lower Saxony (NS); fed_state10 = North Rhine-Westphalia (NW); fed_state11 = Rhineland-Palatinate (RP); fed_state12 = Saarland (SR); fed_state13 = Saxony (S); fed_state14 = Saxony-Anhalt (SA); fed_state15 = Schleswig-Holstein (SH); fed_state16 = Thuringia (TH). The variable for Saarland had to be omitted due the non-existence of laser source manufacturers in this federal state throughout the entire time period.

[15] The full population of 149 PROs is a relatively heterogonous group of organizations. Some predominantly focus on applied research whereas others mainly conduct basic research.

[16] We used UCI-Net 6.2 if not otherwise stated (Borgatti et al. 2002).

measures the strategic network positioning of each firm within the network. The following network centrality measures were included in the dataset: degree centrality *[ctr_degree],* betweenness centrality *[ctr_between],* two reach-based measures *[ctr_2step; ctr_ard]* and two power-related measures *[ctr_ev; ctr_bon]*. The last group of variables encompasses innovation indicators measured by patent count variables. Patent application counts *[pacnt]* and patent grant counts *[pgcnt]* were recorded on an annual basis and included in the dataset. Lag variables were generated for both variables with a lag of 1 year *[pacnt1; pgcnt1]*, 2 years *[pacnt2; pgcnt2]* and 3 years *[pacnt3; pgcnt3]*.

6.2 Introducing Selected Econometric Methods

In this section we turn our attention to longitudinal econometric methods. The discussion addresses general issues connected to event-history and panel data estimation methods. Specific issues are addressed in the context of their application later this book.

6.2.1 Event History Analysis Methods

There are basically three classes of event history methods: non-parametric methods, parametric methods and semi-parametric methods (cf. Allison 1984, p. 14). Non-parametric models do not make distributional assumptions. Parametric models assume that the time until an event of interest occurs follows a specific distributional form. Semi-parametric models make no assumption with regard to the distribution of event time. But these models require a specification of a regression model with a specific functional form (cf. Allison 1984, p. 14).

As stated above, we focus on non-parametric event history methods to analyze cooperation events of German laser source manufacturers. Non-parametric estimation methods do not make any assumptions about the distribution of the process under investigation and are well suited for an initial analysis of a specific phenomenon (Blossfeld et al. 2007, p. 58). The most commonly used non-parametric approach is the Life-Table method. However, this approach has some notable limitations (Blossfeld and Rohwer 2002, p. 56). First, the method requires the pre-specification of fixed and discrete time intervals. Second, to ensure the reliability of estimates conditional for each interval, the Life-Table method is usually applied in the case of a relatively large number of episodes. To overcome these restraints an alternative non-parametric approach has been proposed, i.e. the Product-Limit estimator, also known as the Kaplan-Meier method (Kaplan and Meier 1958).

The Kaplan-Meier method is a non-parametric empirical method that estimates the survivor function based upon longitudinal event data (Cleves et al. 2008, p. 93). In general, the survival function gives the probability of surviving past time t, or to put it in another way, the probability of failing after time t (ibid). The Kaplan-Meier method provides some notable advantages. The method is straightforward to use, requires only weak assumptions and allows us to analyze non-repeated events in single-episode event history data. In this study we are interested in the German laser source manufacturers' propensity to cooperate for the first time and enter the innovation network as well as the length of time until the first cooperation occurs on average. Consequently, based on the risk set specified above we define and interpret the survival function as follows: the survival function estimates the firms' probability of having the first cooperation event after time t. In our case the survival function reflects the probability of moving from the origin state ("no cooperation") to the destination state ("first cooperation") at a given point in time. In addition, both the variance and the confidence interval can be calculated by using Greenwood's variance formula of the survival function and the asymptotic variance of the logarithm of the survival function respectively.[17]

In some cases the hazard rate[18] or the cumulative hazard rate function is of interest rather than the survival function itself. We focus on the latter concept as it allows us to measures the overall risk that has been accumulated up to time t (Cleves et al. 2008, p. 8). There is a simple relationship between the survival function and the cumulative hazard rate.[19] A simple interpretation of the cumulative hazard rate is that it records the number of times we would theoretically expect to observe the occurrence of an event (Cleves et al. 2008, pp. 13–15). It is important to note that cumulative hazards must be interpreted in the context of repeated events regardless of whether the event of interest is, due to its very nature, repeatedly observable or not (ibid).[20] The commonly used method to calculate the cumulative hazard rate is the Nelson-Aalen estimator. The reason is that the Nelson-Aalen estimator exhibits better small-sample properties than the Kaplan-Meier estimator (Cleves et al. 2008, p. 108).

Non-parametric estimation methods provide the opportunity to compare survivor functions. The overall population can be divided into two or more subgroups by using an indicator variable to analyze whether the probability of failing after time t significantly differs among these subgroups. The indicator variable defines the membership in a particular subgroup based on firm-specific characteristics

[17] For an in-depth description of the calculation methods see Cleves et al. (2008, p. 96).

[18] The hazard rate function h(t) can vary from zero to infinity and is also known as the conditional failure rate. It is defined as "[. . .] the (limiting) probability that the failure event occurs in a given interval, conditional upon the subject having survived to the beginning of the interval, divided by the width of the interval" (Cleves et al. 2008, p. 7).

[19] The cumulative hazard function H(t) is defined as: $H(t) = -\ln\{S(t)\} = \int_0^t h(u)\ du$, where S (t) represents the survival function and h() gives the hazard function (Cleves et al. 2008, p. 107).

[20] To illustrate this point, cooperation events can occur several times in a firm's lifespan whereas other events such as firm exits occur only once.

(Blossfeld et al. 2007, p. 76). Using these preparatory steps, separate survival functions are calculated for the members of each subgroup. We apply this approach to analyze the extent to which firm-specific characteristics affect the cooperation behavior over time.

The simplest way to check for statistically significant differences in survivor functions is to calculate and compare the confidence intervals for the estimated survivor functions. The survivor functions are said to be significantly different as long as the confidence intervals are not overlapping (Blossfeld et al. 2007, p. 76).

The more comprehensive approach is to calculate a test statistic.[21] For the purpose of this study we make use of the most commonly applied test statistics, i.e. the Log-Rank test, Cox test, Wilcoxon-Breslow test and Tarone-Ware test. These tests are designed to compare globally defined overall survival functions (Cleves et al. 2008, p. 123). Even though these tests provide, in most cases, relatively similar results, it can be useful to calculate and compare alternative test statistics. One reason for this is that some tests (e.g. Wilcoxon-Breslow) emphasize differences in survivor functions at the onset of the observation period whereas other test statistics (e.g. Log-Rank) stress the differences at the end of the observation period (Blossfeld et al. 2007, p. 81). Several alternative test statistics have since been proposed. For instance the Cox test is very similar to the Log-Rank test whereas the Tarone-Ware test, like the Wilcoxon-Breslow test, puts more weight on earlier time slots.[22] Common to all these test statistics is that they are χ^2-distributed with m-1 degrees of freedom.[23] The tests are based on the null hypothesis that the survivor functions do not differ significantly from each other (Blossfeld et al. 2007, p. 81). A significant test result indicates that the null hypothesis must be rejected (ibid), or to put it another way, the rejection of the null hypothesis based on a significant test result supports the alternative hypothesis that the compared survivor functions differ significantly from one another.

6.2.2 Econometric Methods for Panel Count Data

The econometric analysis of firm innovativeness based on patent data requires the use of a particular category of estimation methods, so-called count data methods. Patent data, which is the same as other types of count data,[24] takes discrete non-negative integer values (Wooldridge 2002, p. 645) and is typically highly

[21] For a description of the general approach to construct test statistics to compare non-parametric survival functions, see Blossfeld and Rohwer (2002, pp. 79–81).

[22] For an in-depth discussion on other further tests (e.g. Peto-Peto-Prentice test or the Fleming-Harrington two-parameter test), see Cleves et al. (2008, pp. 122–128).

[23] The degree of freedom is determined by the number of pre-specified subgroups. Thus, the variable m takes the value 2 in the case of two subgroups (Blossfeld et al. 2007, p. 81).

[24] For a brief overview of other count variables analyzed in economics and social science see (Wooldridge 2002, p. 645).

skewed making the use of conventional linear models inappropriate (Cincera 1997, p. 266). Moreover, a considerable fraction of patent data observations takes on the value zero so that a natural log transformation of the dependent variable in linear models is not possible (Wooldridge 2002, p. 645). Consequently, for count data it is advisable to model the population regression directly and to choose a functional form that ensures positivity for the vector of explanatory variables and any parameter values possible (ibid). As long as the dependent variable has no upper limit, an exponential function is an appropriate choice to meet these requirements (ibid). In general, count data models can be used to analyze cross-sectional as well as longitudinal data. Due to the aim of the present study, focus is placed on the longitudinal models. In their seminal work on the analysis of firm-level R&D activities, Hausman et al. (1984) propose econometric models which are in many cases still the model of choice for the analysis of longitudinal patent count data.

We will start with a brief discussion on panel data characteristics. Even though panel data methods are more complicated than cross-sectional methods, they provide considerable additional value as they allow data to be analyzed that encompasses both variation across individual units as well as variation over time (Cameron and Trivedi 2009, p. 229). The fundamental advantage of panel data is that it allows great flexibility in modeling differences across individual units (Greene 2003, p. 284). For instance, a set of firms is by no means homogeneous as all firms involved differ from one another in several dimensions. Panel data allows us to cope with the problem of unobserved heterogeneity (Kennedy 2003, p. 302). Both fixed and random effects models are associated with some notable advantages and disadvantages that are discussed later. Panel data alleviates multicollinearity problems by creating more variability through combining variation across individual units and across time (Kennedy 2003, p. 302). In addition, panel data allows the analysis of dynamic adjustments which can be crucial in understanding economic phenomena (ibid). Panel data consists of repeated measurements at different points in time, usually observed in regular time intervals, on a well-defined set of individual units such as a spatially and sectorally delimited set of firms. In other words, panel data are repeated observations of the same cross section of firms (Wooldridge 2002, p. 7).

In general, one can distinguish between short panels (i.e. many individual observations across a few time periods) and long panels (few individual observations across many time periods) or panels where the cross-sectional and the time series dimension are roughly of the same magnitude (Cameron and Trivedi 2009, p. 230; Wooldridge 2002, p. 7). Most panel data methods can handle both balanced as well as unbalanced panels (Cameron and Trivedi 2009, p. 230). In the first case, the full set of individual units is observed over all time periods whereas in the second case a considerable fraction of individual units are observed for fewer time periods. Due to the aim of this study and the structural features of the laser industry panel dataset (cf. Sect. 6.1.2) we focus below on single equation models for analyzing short, unbalanced panels.

Now we turn our attention to the choice of model. In many cases the explanatory variables of our count data model reveal empirical evidence for overdispersion.[25] There are several ways to deal with overdispersion in count data models. Commonly, overdispersion that is induced by unobserved heterogeneity is accounted for by estimating Negative Binomial models (NB model) instead of the intuitive standard Poisson model. The NB model is more general than the Poisson model because it allows for increased dispersion by incorporating an additional parameter α. The variance is a linear function of the mean that can be transformed into a Poisson model (Cameron and Trivedi 1986). The NB model reduces to the Poisson Model as $\alpha \rightarrow 0$ (Winkelmann 2003). It enables us to deal with a predominance of zero and small integer values (Cameron and Trivedi, 1986). Finally, we have to take a brief look at the differences between models that allow for a correlation between the explanatory variables and the time-invariant component of the error term, and models which require the unobserved effects and the explanatory variables to be completely uncorrelated (Wooldridge 2002, p. 668). In the former case we have a fixed effects model and in the latter case a random effects model. Pioneering work on the analysis of unobserved effects in panel count data has been conducted by Hausman et al. (1984) who developed a fixed effects and random effects model under full distributional assumption (Wooldridge 2002, p. 668). Panel data shows two types of variation. Variation from one observation to another observation for an individual unit is called "within-variation" whereas the variation from one individual unit to another individual unit is called "between-variation". The fixed effects model ignores variation across individuals and uses only within-variation for all individual units over all observation windows (Kennedy 2003, p. 307). It is important to note that the coefficient of the regressor in fixed effects models will be incorrectly estimated or not identified with little or no within-variation (Cameron and Trivedi 2009, p. 238). Fixed effects models have some important advantages. The fixed effects estimator is unbiased because it includes dummy variables for the different intercepts and is more robust against selection bias problems compared to the random effects estimator (Kennedy 2003, p. 304). However, fixed effects models also have two considerable drawbacks. Firstly, all time-invariant explanatory variables are thrown out because the estimation procedure fails to estimate a slope coefficient for variables that do not vary within an individual unit (Kennedy 2003, p. 304). Secondly, using only within-variation leads to less efficient estimates and the model loses explanatory power (Cameron and Trivedi 2009, p. 259). The random effects model compensates for this disadvantage.

The random effects estimator takes advantage of within-variation as well as between-variation in panel data (Cameron and Trivedi 2009, p. 256) by using cross-sectional variation in panel data and by running OLS estimation on the average values for each individual unit in order to calculate a (matrix) weighted average of both between-estimators and within-estimators (Kennedy 2003, p. 307). The random effects model has several advantages compared to the fixed effects

[25] The procedure to check for overdispersion was proposed by Cameron and Trivedi (1990).

model. On the one hand random effects estimators make better use of the information values of patent data and generate efficient estimates with higher explanatory power. In addition, a random effects estimator can generate coefficient estimates of both time-variant and time-invariant explanatory variables (Kennedy 2003, p. 307). However, these advantages are not without a downside. The major drawback of the random effects model is that correlations between the error term and explanatory variables generate biased estimates (Kennedy 2003, p. 306). In other words, the random effects estimator generates inconsistent results when the model assumptions are violated.

In summary, the main difference between the estimation techniques is that fixed effects models allow for correlation between the unobserved individual effect and the included explanatory variables whereas random effects models require the unobserved individual effect and the explanatory variables to be uncorrelated (Greene 2003, p. 293).

The question remains whether fixed effects models or random effects models should be applied. Hausman (1978) has proposed a specification test to select the appropriate model. The basic idea of the Standard Hausman specification test is to test the null hypothesis that the unobserved effect is uncorrelated with the explanatory variables (Greene 2003, p. 301). In the case that the null hypothesis cannot be rejected, both the fixed effects estimates and the random effects estimates are consistent and the model of choice is the random effects model due to its higher explanatory power. Under the alternative, random effects and fixed effects estimators differ and it can be argued that the latter model is the appropriate choice (Cameron and Trivedi 2009, p. 260).

Nonetheless, choosing the model based on the Standard Hausman specification test is controversial for two reasons. First, there is an interdisciplinary controversy of whether consistency should be preferred over efficiency. Microeconometric literature advocates the use of fixed effects models whereas most other branches of applied statistics tend to give preference to random effects models due to their higher explanatory power (Cameron and Trivedi 2009, p. 230). Second, the Standard Hausman specification test itself has serious shortcomings because it requires the random effects estimator to be efficient (Cameron and Trivedi 2009, p. 261). In the case of unbalanced panels, a robust version of the specification test can alleviate the latter point (Cameron and Trivedi 2009, pp. 261–262).

References

Allison PD (1984) Event history analysis – regression for longitudinal event data. Sage, London

Blossfeld H-P, Rohwer G (2002) Techniques of event history analysis – new approaches to causal analysis. Lawrence Erlbaum, London

Blossfeld H-P, Golsch K, Rohwer G (2007) Event history analysis with Stata. Lawrence Erlbaum, London

Borgatti SP, Everett MG, Freeman LC (2002) Ucinet for windows: software for social network analysis. Analytic Technologies, Harvard

Box GEP (1979) Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN (eds) Robustness in statistics. Academic, New York
Cameron CA, Trivedi PK (1986) Econometric models based on count data: comparisons and applications of some estimators and tests. J Appl Econ 1:29–53
Cameron CA, Trivedi PK (1990) Regression based tests for overdispersion in the Poisson model. J Econ 46(3):347–364
Cameron CA, Trivedi PK (2009) Microeconometrics using Stata. Stata Press, College Station
Cincera M (1997) Patents, R&D, and technological spillovers at the firm level: some evidence from econometric count models for panel data. J Appl Econ 12(3):265–280
Cleves MA, Gould WW, Gutierrez RG, Marchenko YU (2008) An introduction to survival analysis using Stata, 2nd edn. Stata Press, College Station
European Commission (2005) The new SME definition – user guide and model declaration. Enterprise and Industry Publications, Brussels
Greene WH (2003) Econometric analysis, 5th edn. Prentice Hall, Upper Saddle River
Hausman JA (1978) Specification tests in econometrics. Econometrica 46(6):1251–1271
Hausman JA, Hall BH, Griliches Z (1984) Econometric models for count data with an application to the patents – R&D relationship. Econometrica 52(4):909–938
Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
Kennedy P (2003) A guide to econometrics. Blackwell, Oxford
Winkelmann R (2003) Econometric analysis of count data, 4th edn. Springer, Heidelberg/New York
Wooldridge JM (2002) Econometric analysis of cross sectional and panel data. MIT Press, Cambridge, MA