

Brecht, Beatrix; Brecht, Leo

Working Paper

Analysis of remigrant behavior with the grouped Cox model

Diskussionsbeiträge - Serie II, No. 164

Provided in Cooperation with:

Department of Economics, University of Konstanz

Suggested Citation: Brecht, Beatrix; Brecht, Leo (1992) : Analysis of remigrant behavior with the grouped Cox model, Diskussionsbeiträge - Serie II, No. 164, Universität Konstanz, Sonderforschungsbereich 178 - Internationalisierung der Wirtschaft, Konstanz

This Version is available at:

<https://hdl.handle.net/10419/101550>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

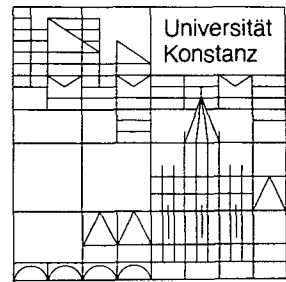
Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



Sonderforschungsbereich 178
„Internationalisierung der Wirtschaft“

Diskussionsbeiträge

Juristische
Fakultät

Fakultät für Wirtschafts-
wissenschaften und Statistik

Beatrix Brecht
Leo Brecht

**Analysis of Remigrant Behavior
with the Grouped Cox Model**

Postfach 5560
D-7750 Konstanz

Serie II — Nr. 164
November 1991

24. JAN. 1992

WS 113 - 164 AT

**ANALYSIS OF REMIGRANT BEHAVIOR WITH THE
GROUPED COX MODEL**

Beatrix Brecht
Leo Brecht

Serie II - Nr. 164

November 1991

Analysis of Remigrant Behavior with the Grouped Cox Model

Beatrix Brecht and Leo Brecht*

October 1991

Abstract:

The focus of this article is the application of the grouped Cox model to the investigation of influence factors of the inclination of guest workers in the Federal Republic of Germany to remigrate. The model on which this is based takes into consideration both the appearance of censored observations and the discrete-time raised length of stay, which the application of a continuous time sojourn model does not permit. The estimations were made with the program PRODISA (Program for Discrete Survival Analysis), which was developed in the context of this project. The Socio-Economic Panel of the DIW, Berlin, serves as the basis of the data.

Keywords: grouped Cox Model, time-discrete sojourn, covariate design, PRODISA, remigrant behavior, guest worker

*Our special thanks go to Dianne Mather who gave the translation the right touch.

Contents

1	Introduction	2
2	The Grouped Cox Model	3
2.1	Foundations	3
2.2	Estimation Equations	7
2.3	Asymptotic Statements	8
3	Empirical Analysis of Remigrant Behavior	10
3.1	The Socio-Economic Characteristics in the Panel Study	10
3.2	The Program PRODISA and the Choice of Design	12
3.3	Results	16
4	Conclusion	21
5	References	22

1 Introduction

Until now, empirical studies about the causes and reasons why guest workers living in Germany extend their stay or eventually remigrate have been conducted, especially from a sociological point of view, by means of surveys. The statistical methods used to analyse these surveys often only focus on the number of relative frequencies or average values, which carry weight depending on the size of the sample.

The survival analysis, however, permits the analysis of time intervals between successive events – in this case, the length of stay for guest workers in Germany – with individual data, where it is possible to model a dependency of the sojourns from the collected covariates. Therefore, the available information of the observed time periods is optimally used so that the effect of possible influence factors (such as age, sex, or a subjective feeling of well-being in Germany) on the length of stay can be analysed.

In continuation of the paper by *Brecht/Michels (1991)*, in which non-parametric estimation methods were used to analyse the sojourn-times, a discrete-time version of the well-known semiparametric Cox model, the *grouped Cox model*, will be taken as the basis for this contribution. The advantage of this method is that the influence of covariates on the length of stay of foreign workers can be explicitly estimated. In this case, censored observations are taken into consideration as well, i.e. an interesting change in state (in this context: the guest worker returns home to his native country) does not occur during the period under observation. In addition, not only the size and the direction of the influence of selected covariates are of interest, but also the possible differences that may exist among the analysed nationalities of the guest workers from Spain, Italy, Yugoslavia, Greece and Turkey.

The data that is used in this study comes from surveys conducted annually since 1984 by the German Socio-Economic Panel of the DIW, Berlin.

2 The Grouped Cox Model

2.1 Foundations

In many cases, no exact points of time can be given in which events or transitions occur, only time intervals. If one nevertheless uses a continuous-time model, this causes a great number of similar observation values - so called "ties" - to appear, and one generally obtains useless parameter estimations. Since the theoretical foundations, such as the derivation of asymptotic characteristics, proceed from the assumption that no ties are present, the theoretical foundation of this stastical method is invalid if a continuous-time model is still taken as a basis. For the modelling of discrete-time raised event times, the time axis is divided into T_0 intervals $(a_0, a_1], (a_1, a_2], \dots, (a_{T_0-1}, \infty)$. The intervals themselves are numbered from $t = 1, 2, \dots, T_0$. One can therefore observe a probability model with a positive random variable T , which can take on integer values from the set $\{1, \dots, T_0\}$. $\{T = t\}$ means that a transition has taken place in the interval $t = (a_{t-1}, a_t]$. The individuals that are at risk at the beginning of the interval t , remain at risk during the entire interval, and events that occur during the interval t , are interpreted as if they had occurred at point a_t . Furthermore, the covariates \mathbf{X} should remain constant in each interval t , and their values are established in each case at the beginning of an interval. Let T_1, \dots, T_n be independently and identically distributed random variables with a given discrete probability density.

The hazard rate for the discrete case can be defined as follows:

$$\lambda(t|\mathbf{X}) = P(T = t | T \geq t, \mathbf{X}) \quad \text{for } t = 1, \dots, T_0 \quad (2.1)$$

$T = t$ indicates that a transition has occurred in this interval. In (2.1), the conditional probability is rendered so that an observed individual undergoes a transition in the time interval t , given the covariates and given that the individual reached the beginning of the time interval.

One obtains for the conditional "survival function"

$$P(T > t | T \geq t, \mathbf{X}) = 1 - \lambda(t|\mathbf{X}). \quad (2.2)$$

It indicates the conditional probability that the time interval t has been "survived," or that no event has taken place in t , given the covariates and given that the time interval was reached.

The survivor function for the discrete case

$$S(t|\mathbf{X}) = P(T \geq t | \mathbf{X}) \quad (2.3)$$

is analogous to the continuous case the (unconditional) probability that an individual or a unit "experiences" the time interval t .

Nevertheless, the relation $S(t) = 1 - F(t)$ is no longer valid for the discrete case; however, the following relations can be indicated:

$$S(t|\mathbf{X}) = \prod_{k=1}^{t-1} (1 - \lambda(k|\mathbf{X})) \quad (2.4)$$

$$\begin{aligned} P(T = t|\mathbf{X}) &= P(T = t | T \geq t, \mathbf{X}) \cdot P(T \geq t|\mathbf{X}) \\ &= \lambda(t|\mathbf{X}) \cdot \prod_{k=1}^{t-1} (1 - \lambda(k|\mathbf{X})) \end{aligned} \quad (2.5)$$

$$\lambda(t|\mathbf{X}) = \frac{P(T = t|\mathbf{X})}{P(T \geq t|\mathbf{X})} = \frac{S(t|\mathbf{X}) - S(t+1|\mathbf{X})}{S(t|\mathbf{X})} \quad (2.6)$$

Included in (2.5) is the (unconditional) probability that an event occurred in time interval t , given the covariates.

The case of censored event times often occurs, i.e. no event occurs during the period of observation, or the individual is no longer available for the study. In that case, the tuple $(T_1, \delta_1), \dots, (T_n, \delta_n)$ is given with $T_i = \min(I_i, C_i)$ and with the censoring indicator δ_i , in which $\delta_i = 1$ or 0 , depending upon whether I_i is observed or not, and I_i and C_i are discrete positive random variables.

With the definition of the indicator function given in

$$\begin{aligned} Y_i(t) &= I\{T_i \geq t\}, \quad i = 1, \dots, n, \\ N_i(t) &= I\{T_i \leq t; \delta_i = 1\}, \quad i = 1, \dots, n., \end{aligned}$$

it is possible to indicate a closed term for the likelihood function of a discrete sojourn model. The indicator function $Y_i(t)$ is to be interpreted here in such a manner that it retains the value one as long as no event has occurred for the i -th individual. It holds true for the function $N_i(t)$ that the value one is only reached when an event has occurred. Furthermore, in order to formulate the likelihood function, one requires the concept of the risk set, which is given by

$$R(t) = \{i : Y_i(t) = 1\}. \quad (2.7)$$

Furthermore,

$$N(t) = \sum_{i=1}^n N_i(t), \quad Y(t) = \sum_{i=1}^n Y_i(t), \quad t = 1, \dots, T_0, \quad (2.8)$$

holds, and $\Delta N_i(t)$ is defined as

$$\Delta N_i(t) = N_i(t) - N_i(t-1), \quad N_i(0) = 0. \quad (2.9)$$

$\Delta N_i(t)$ should indicate whether a transition has occurred for the i -th individual in the t -th interval ($\Delta N_i(t) = 1$), or not ($\Delta N_i(t) = 0$).

Based on various assumptions, *Arjas/Haara (1987)* then define a likelihood function according to

$$L = \prod_{t \leq T_0} \prod_{i \in R(t)} P(\Delta N_i(t) = \Delta n_i(t) \mid \mathcal{G}_{t-1}). \quad (2.10)$$

There, the σ -algebra \mathcal{G}_{t-1} contains the previous history of the process up until the interval $t-1$, including the value of the covariates at the beginning of the interval t . In (2.10) a partial likelihood is defined with random censorship taken as a basis. Furthermore, the likelihood makes possible the treatment of ties which will inevitably appear often in the presence of discrete collected data. One should make certain that no ties are present during continuously raised event times and during continuous modelling of the hazard function. The estimation and testing theory based upon this always assumes that no ties are present. Furthermore, it is important that the covariates are raised at the beginning of each interval, but are then regarded as constant during the interval. The generalized situation that the covariates \mathbf{X} may be time-dependent is now considered. In addition, it is possible that the path of the covariates $\mathbf{X}(1), \dots, \mathbf{X}(t)$ as well as the time-independent covariate \mathbf{X}_0 can be included in the modelling. Altogether, one obtains a new covariate design with $\mathbf{Z}(t)$, so that the hazard function is given in

$$\lambda(t \mid \mathbf{Z}(t)).$$

The risk set is known at the beginning of every interval. The value of $\Delta N_i(t)$ can be understood as the result of a Bernoulli experiment, and

$$\begin{aligned} P(\Delta N_i(t) = 1 \mid \mathcal{G}_{t-1}) &= Y_i(t)P(T_i = t \mid T_i \geq t, \mathbf{Z}_i(t)) = Y_i(t)\lambda(t \mid \mathbf{Z}_i(t)), \\ P(\Delta N_i(t) = 0 \mid \mathcal{G}_{t-1}) &= 1 - Y_i(t)\lambda(t \mid \mathbf{Z}_i(t)). \end{aligned}$$

holds. The likelihood function can be rearranged with the application of the defined indicator function and with the covariate design $\mathbf{Z}(t)$ as follows:

$$\begin{aligned} L &= \prod_{t \leq T_0} \prod_{i \in R(t)} (Y_i(t)\lambda(t \mid \mathbf{Z}_i(t)))^{\Delta N_i(t)} (1 - Y_i(t)\lambda(t \mid \mathbf{Z}_i(t)))^{1 - \Delta N_i(t)} \\ &= \prod_{t \leq T_0} \prod_{i \in R(t)} \lambda(t \mid \mathbf{Z}_i(t))^{\Delta N_i(t)} (1 - \lambda(t \mid \mathbf{Z}_i(t)))^{1 - \Delta N_i(t)} \\ &= \prod_{t \leq T_0} \prod_{i=1}^n \lambda(t \mid \mathbf{Z}_i(t))^{Y_i(t)\Delta N_i(t)} (1 - \lambda(t \mid \mathbf{Z}_i(t)))^{Y_i(t)(1 - \Delta N_i(t))}. \end{aligned} \quad (2.11)$$

Since $\Delta N_i(t) = 0$, if $Y_i(t) = 0$, the right side of (2.11) is the same as

$$\prod_{t \leq T_0} \prod_{i=1}^n \lambda(t | \mathbf{Z}_i(t))^{\Delta N_i(t)} (1 - \lambda(t | \mathbf{Z}_i(t)))^{Y_i(t) - \Delta N_i(t)}. \quad (2.12)$$

In order to estimate a specific model, the specification of the implied hazard rate remains open. For this there are several possibilities:

A study by *Arjas/Haara (1988)* is, for example, based on the continuous version of the Cox model with the partial likelihood as the basis. This approach is used by *Arjas/Haara (1988)* to observe a discrete probability distribution

$$P_i(\beta, t) = \frac{Y_i(t) e^{\mathbf{Z}_i^t(t)\beta}}{\sum_{k=1}^n Y_k(t) e^{\mathbf{Z}_k^t(t)\beta}}, \quad t \in \{1, \dots, T_0\},$$

and to derive the inference about a partial likelihood approach with a log-likelihood function

$$l(\beta, t) = \sum_{\nu \leq t} \sum_{i=1}^n \log P_i(\beta, \nu) \Delta N_i(\nu)$$

from which structurally, the same log-likelihood function results, as in the continuous case (see *Anderson/Gill (1982)*). In this case, β provides the parameter vector belonging to the covariate design. The asymptotic statements carry over analogously as well.

In this study, we will now proceed from a further model approach. First, a discrete probability distribution will be derived from the continuous Cox model (cf. *Hamerle/Tutz (1989)*) for the model that is to be analysed here, and the ML approach will be applied to the discrete hazard rate. With this, the baseline hazard rate, which is taken as a basis in the continuous case, is included as a functional in the parameter estimation. Theoretical counting process arguments make it possible to derive conditions in which asymptotic statements are complied with; cf. *Brecht, L. (1991a)*.

If one considers a discrete model as is described in the preceding section, one sees that the hazard rate is derived from the continuous sojourn T_s (see *Hamerle/Tutz (1989)*) as follows, in the event that one starts from time-independent covariates \mathbf{X} :

$$\begin{aligned} P(T_s \geq a_t | \mathbf{X}) &= S(a_t | \mathbf{X}) = e^{-\exp(\mathbf{X}^t \theta_0) \int_0^{a_t} \lambda_0(u) du} \\ &= e^{-\exp(\eta_t + \mathbf{X}^t \theta_0)}, \quad \text{with } \eta_t = \ln \int_0^{a_t} \lambda_0(u) du. \end{aligned}$$

For the possibility that the discrete sojourn T takes on the value t , then

$$P(T = t | \mathbf{X}) = e^{-\exp(\eta_{t-1} + \mathbf{X}^t \theta_0)} - e^{-\exp(\eta_t + \mathbf{X}^t \theta_0)}$$

holds true, and with that,

$$\lambda(t | \mathbf{X}) = 1 - e^{-\exp(\gamma_t + \mathbf{X}^t \theta_0)}, \quad \text{with } \gamma_t = \ln(\exp(\eta_t) - \exp(\eta_{t-1})).$$

The choice of design is to be discussed later; the vector γ as well as the (time-dependent) covariates are summarized in the vector $\mathbf{Z}(t)$. One obtains the model

$$\lambda(t | \mathbf{Z}(t)) = 1 - e^{-e^{\mathbf{Z}(t)^t \beta}} \quad (2.13)$$

with the covariate design vector $\mathbf{Z}(t) := (e_t, \mathbf{X})$, in which $\mathbf{X} = \mathbf{X}(t)$ should be allowed. This model will be assumed in the following analysis and is referred to as the *grouped Cox model*.

A generalized approach to the grouped Cox model proceeds from the following specification (cf. *Hamerle/Tutz (1989)* or *Aranda-Ordaz (1983)*):

$$\ln(-\ln(1 - \lambda(t|\mathbf{X}))) = \gamma_t + \mathbf{X}^t \theta_0, \quad \alpha = 0,$$

$$\{[-\ln(1 - \lambda(t|\mathbf{X}))]^\alpha - 1\} / \alpha = \gamma_t + \mathbf{X}^t \theta_0, \quad \alpha \neq 0.$$

If one considers the special case of $\alpha = 1$, then the model is as follows:

$$\lambda(t|\mathbf{X}) = 1 - e^{-(1 + \gamma_t + \mathbf{X}^t \theta_0)}, \quad t \in \{1, \dots, T_0\}, \quad (2.14)$$

and can be regarded as a discrete-time version of the additive model

$$\lambda(t|\mathbf{X}) = \lambda_0(t) + \mathbf{X}^t \theta, \quad t \in [0, T_0].$$

For a model in which the covariates are time-independent, one obtains then with the covariate design vector $\mathbf{Z}(t)$ the term

$$\lambda(t|\mathbf{Z}(t)) = 1 - e^{-(1 + \mathbf{Z}(t)^t \beta)}. \quad (2.15)$$

2.2 Estimation Equations

The first and second derivations of the score function are needed in order to program the grouped Cox model. If one uses the previously derived general likelihood function (2.12) for discrete-time hazard functions, and if one implements the specification of the grouped Cox model from (2.13), then after taking the logarithm of $L(\beta, T_0)$, one obtains the term

$$\begin{aligned} \log L(\beta, T_0) &= \sum_{t=1}^{T_0} \sum_{i=1}^n \left\{ -(Y_i(t) - \Delta N_i(t)) e^{\mathbf{Z}_i(t)^t \beta} + \Delta N_i(t) \log(1 - e^{-e^{\mathbf{Z}_i(t)^t \beta}}) \right\} \\ &:= C(\beta, T_0). \end{aligned} \quad (2.16)$$

Differentiation with respect to the vector of the regression parameter β provides the score function given by

$$U(\beta, T_0) = \sum_{t=1}^{T_0} \sum_{i=1}^n \left\{ -(Y_i(t) - \Delta N_i(t)) \mathbf{Z}_i(t) e^{\mathbf{Z}_i(t)^t \beta} + \Delta N_i(t) \frac{\mathbf{Z}_i(t) e^{\mathbf{Z}_i(t)^t \beta} e^{-e^{\mathbf{Z}_i(t)^t \beta}}}{1 - e^{-e^{\mathbf{Z}_i(t)^t \beta}}} \right\}. \quad (2.17)$$

The maximum likelihood estimator of the vector of the regression parameter then follows by setting the equation of $U(\beta, T_0)$ to zero. If one calculates the second derivation of the log-likelihood function with respect to the vector of the regression parameter, then one obtains the expression

$$I(\beta, T_0) = \sum_{t=1}^{T_0} \sum_{i=1}^n \left\{ -(Y_i(t) - \Delta N_i(t)) e^{\mathbf{Z}_i(t)^t \beta} \right. \quad (2.18) \\ \left. + \Delta N_i(t) \frac{e^{\mathbf{Z}_i(t)^t \beta}}{(e^{e^{\mathbf{Z}_i(t)^t \beta}} - 1)^2} \left\{ e^{e^{\mathbf{Z}_i(t)^t \beta}} - 1 - e^{\mathbf{Z}_i(t)^t \beta} e^{e^{\mathbf{Z}_i(t)^t \beta}} \right\} \right\} \mathbf{Z}_i^{\otimes 2}(t),$$

in which $\mathbf{Z}_i^{\otimes 2}(t) = (\mathbf{Z}_{ij}(t) \mathbf{Z}_{ik}(t))_{j,k=1,\dots,q} = \mathbf{Z}_i(t) \mathbf{Z}_i^t(t)$.

$I(\beta, T_0)$ can be used as an estimation of the Fischer information matrix, and $-I(\beta, T_0)^{-1}$ provides the variance-covariance matrix of the parameter estimators.

The log-likelihood function $C(\beta, T_0)$ is concave in β , through which the existence of a global maximum is guaranteed. The proof results from the negative definiteness of the matrix of $I(\beta, T_0)$.

2.3 Asymptotic Statements

The proof of asymptotic statements of parameter estimators in time-discrete sojourn models has never before been completely given in literature. The asymptotic properties of a logistic approach originate from *Arjas/Haara (1987)*, and the statements for a grouped Cox model were proven in *Brecht, L. (1991a)*. In both papers, various conditions were derived that are necessary for the proof of the regular asymptotic statements about parameter estimators based on the maximum likelihood approach. Furthermore, sufficient conditions were derived in *Brecht, L. (1991a)* for the covariate process.

In conclusion, one obtains with the stipulations

$$S_{1,n}^{(2)}(\beta, t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) e^{\mathbf{Z}_i(t)^t \beta} e^{-e^{\mathbf{Z}_i(t)^t \beta_0}} \mathbf{Z}_i^{\otimes 2}(t), \\ S_{2,n}^{(2)}(\beta, t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) (1 - e^{-e^{\mathbf{Z}_i(t)^t \beta_0}}) \mathbf{Z}_i^{\otimes 2}(t)$$

$$\frac{e^{\mathbf{Z}_i(t)^t \beta}}{(e^{e^{\mathbf{Z}_i(t)^t \beta}} - 1)^2} \left\{ e^{e^{\mathbf{Z}_i(t)^t \beta}} - 1 - e^{\mathbf{Z}_i(t)^t \beta} e^{e^{\mathbf{Z}_i(t)^t \beta}} \right\},$$

the following assumption:

The matrix $\Sigma_0 := -\sum_{t=1}^{T_0} \{\Sigma_1(\beta_0, t) - \Sigma_2(\beta_0, t)\}$ is positive definite, in which it should hold true that

$$S_{j,n}^{(2)}(\beta, t) \xrightarrow{p} \Sigma_j(\beta, t) \quad \text{uniform in } \beta \in \mathbb{B}_0,$$

for limited, continuous, matrix-valued functions $\Sigma_j(\beta, t)$.

Under the sufficient hypotheses

(U1*) $(\mathbf{Z}_i, T_i, C_i), i = 1, \dots, n$ are independent identically distributed random variables.

(U2*) The covariate vectors $\mathbf{Z}(t)$ are almost certainly limited and measurable:

$$a \leq \mathbf{Z}(t) \leq b.$$

(U3*) The vectors of the covariates are linear independent.

one obtains the following statements:

(i) If $\hat{\beta}$ is the maximum likelihood estimator and thus the solution to

$$U(\hat{\beta}_n, T_0) = 0,$$

then β_0 is consistently estimable, i.e. it holds true that

$$\hat{\beta}_n \xrightarrow{p} \beta_0, \quad n \longrightarrow \infty. \quad (2.19)$$

(ii) The normalized log-likelihood function converges in distribution; it holds that

$$\frac{1}{\sqrt{n}} U(\beta_0, T_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_0), \quad n \longrightarrow \infty, \quad (2.20)$$

in which \mathcal{N} represents a multivariate normal distribution and Σ_0 as was previously shown.

(iii) The matrix $-\frac{1}{n} I(\beta, T_0)$ converges stochastically to the Fischer information; one obtains

$$-\frac{1}{n} I(\beta_n^*, T_0) \xrightarrow{p} \Sigma_0, \quad n \longrightarrow \infty, \quad (2.21)$$

$\forall \beta_n^*$ with $\beta_n^* \xrightarrow{p} \beta_0, \quad n \longrightarrow \infty.$

(iv) One obtains the convergence distribution in $\sqrt{n}(\hat{\beta}_n - \beta_0)$; it therefore holds true that

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_0^{-1}).$$

The proofs of each statement are taken from *Brecht, L. (1991a)* and are carried out with the application of the time-discrete martingal theory.

3 Empirical Analysis of Remigrant Behavior

3.1 The Socio-Economic Characteristics in the Panel Study

The German Socio-Economic Panel (SOEP) of the DIW (Deutsches Institut für Wirtschaftsforschung), Berlin, is a representative micro-longitudinal data collection for Germany. The random sample, first started 1984, furnishes about 1000 variables yearly from surveys in which about 12000 people in about 6000 households are asked every year. These variables include the composition of a household, occupation, mobility, income pattern, living conditions, as well as information about education, health or value judgements and satisfaction. A SOEP-East-Study was started in the former GDR in 1990, where about 2000 households and 4500 people were used as a basis.

In this study, six panel waves (1984-1989) from the SOEP-West were used, in which approximately 1400 households reported having a foreign head of household. Additional variables were considered for the people that form the subgroup of foreigners.

These variables include education (kind of school, professional training in Germany or abroad), transfer payments to home, year of migration to the destination country, planned length of stay, and official residence of the parents, children and married couples. Subjective questions were also asked about the feeling of belonging to a country, language proficiency, contact with Germans, and the nationality of a person's circle of friends.

In order to ascertain how these socio-economic characteristics influence the length of stay in Germany, one must determine when a remigration to the country of origin occurs, by examining the collection of data. The coding of the data into two categories – 1. foreigners who have remigrated to their homelands and 2. foreigners who are visiting their homelands for a longer period of time – and taking the year of immigration to the country allows the length of stay to be shown in annual intervals.

However, since the portion of remigrating guest workers is relatively small, a higher portion

of censored data is present. In order to keep the estimation from being biased, a subsample for each foreign group was formed so that the respective data set contains about 20 per cent censored data (see table 1).

This restriction should be taken into consideration when interpreting the results; as a control, an estimation with the entire sample appears to be meaningful.

Table 1: Subsample with 20% censored data

Homeland	Number of Guest workers	Number of Guest workers	Total
	Remigrated Back (uncensored)	Staying in Germany (censored, ca. 20%)	
Spain	75	19	94
Italy	105	26	131
Yugoslavia	41	10	51
Greece	61	15	76
Turkey	134	34	168

When estimating the influence of covariates, it must be taken into consideration that the number of parameters to be estimated depends on the choice of design. The increase in the number of parameters has the effect that the Newton scoring algorithm should no longer be carried out, since the matrix $I(\beta, T_0)$ can become singular. A large number of parameters is soon obtained, especially with a choice of design in which episode-specific parameters must be estimated for each episode as well. The flexibility which is kept open with this choice of design is obtained at the price of not being able to model as many covariate influences. Possible effects of the interaction between the covariates should also be taken into account, since they can become more complex as they increase in number. For this purpose, selective estimations were carried out, in which chiefly the influence of sex and the length of stay were analysed in order to make a comparison with the non-parametric methods (*Brecht/Michels, 1991*).

For the practical data processing of the SOEP with the relational data base system INGRES, cf. *Brecht, B. (1990)*. The program PRODISA, developed to carry out the estimations, is described in greater detail in the following section.

3.2 The Program PRODISA and the Choice of Design

The program PRODISA was developed for the analysis of time-discrete sojourns, among other things for the specified grouped Cox model. In contrast to traditional computer programs for discrete sojourn analysis, this one is characterized by the fact that time-dependent covariates can be included in the modelling. The program package GLAMOUR (developed at the University of Regensburg) assumes in the case of a discrete sojourn that time-independent variables are present, i.e. the covariates will not change their value between the observed intervals. This assumption is an important restriction of the modelling, especially in the case of socio-economic characteristics, for example the covariate of an individual's state of health or employment participation. Furthermore, the program PRODISA is characterized by the fact that it permits a flexible modelling of the covariate influence and therefore makes possible the selection of different designs which the number of regression parameters to be estimated implies. Moreover, one can distinguish between time lags of the first and higher orders, as well as the specification without time lags, i.e. with interval-dependent covariates. The program PRODISA is still in the development phase and will be extended to test procedures as well as to estimation procedures of other models, such as the logistic model and will be published elsewhere.

The estimation method is based on the likelihood function developed in section 2.1 and uses the Newton scoring algorithm to solve the score equations. The inversion of the matrix $I(\beta, T_0)$ from (2.18) is based on a Cholesky decomposition. In selecting a design, hence in determining the dependency of the covariates and their parameter, one can proceed as follows. One defines the path of covariates $\tilde{\mathbf{X}}(t) = (\mathbf{X}(1)^t, \dots, \mathbf{X}(t)^t)$, in which $\mathbf{X}(s)^t$ represents the time-dependent covariate vector in the s th interval, and one defines the parameter vectors θ_{ts} with the same dimensions as $\mathbf{X}(s)$

$$\theta_t = \begin{pmatrix} \theta_{t1} \\ \vdots \\ \theta_{tt} \end{pmatrix}. \quad (3.22)$$

The influence of the path of covariates $\tilde{\mathbf{X}}(t)$ of the time-dependent covariates on the hazard rate of the interval t can therefore be modelled by

$$h(\tilde{\mathbf{X}}(t)^t \theta_t).$$

Furthermore, $h : \mathbb{R} \rightarrow \mathbb{R}$ is positive and monotone increasing.

If X_0 is a time-independent covariate vector with the accompanying parameter vector θ_0 ,

then the following general design vectors can be specified:

$$\mathbf{Z}(t) := \begin{pmatrix} X_0 \\ 0 \\ \vdots \\ 0 \\ \bar{\mathbf{X}}(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{T_0} \end{pmatrix} \quad (3.23)$$

If one considers the grouped Cox model or the Aranda-Ordaz model, then one discovers that in this case, the parameters γ_t , $t \in \{1, \dots, T_0\}$ must be additionally estimated in each interval. For this, one defines the design vectors

$$\mathbf{Z}(t) := \begin{pmatrix} e_t \\ X_0 \\ 0 \\ \vdots \\ 0 \\ \bar{\mathbf{X}}(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \gamma \\ \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{T_0} \end{pmatrix}, \quad e_t, \gamma \in \mathbb{R}^{T_0}, \quad (3.24)$$

with e_t as the t th unit vector.

If one assumes the simpler case $\gamma_t \equiv \gamma_0, \forall t \in \{1, \dots, T_0\}$, the design vectors should then be selected to

$$\mathbf{Z}(t) := \begin{pmatrix} 1 \\ X_0 \\ 0 \\ \vdots \\ 0 \\ \bar{\mathbf{X}}(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \gamma_0 \\ \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{T_0} \end{pmatrix}, \quad \gamma_0 \in \mathbb{R}. \quad (3.25)$$

In this case, it is always assumed that the parameter vector γ_t , $t \in \{1, \dots, T_0\}$ is also estimated, and that time-independent covariates X_0 are raised as well.

The following special cases are therefore of interest for the modelling:

1. Without Time Lags (Interval-dependent Covariates)

In the case of a design without a time lag, only the value of the covariate vector that belongs to the observed duration t is taken into consideration. One obtains

$$\begin{aligned} \tilde{\mathbf{X}}(t) &= \mathbf{X}(t) \\ \mathbf{Z}(t) &:= \begin{pmatrix} e_t \\ X_0 \\ 0 \\ \vdots \\ 0 \\ \tilde{\mathbf{X}}(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \gamma \\ \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{T_0} \end{pmatrix} \text{ with } \theta_t \in \mathbb{R}, \gamma \in \mathbb{R}^{T_0}. \end{aligned} \quad (3.26)$$

2. Time Lags of the First Order

In the case of time lags of the first order, one obtains

$$\begin{aligned} \tilde{\mathbf{X}}(t) &= (\mathbf{X}(t-1)^t, \mathbf{X}(t)^t) \\ \mathbf{Z}(t) &:= \begin{pmatrix} e_t \\ X_0 \\ 0 \\ \vdots \\ 0 \\ \tilde{\mathbf{X}}(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \gamma \\ \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{T_0} \end{pmatrix}, \text{ with } \theta_t = \begin{pmatrix} \theta_{t,t-1} \\ \theta_{t,t} \end{pmatrix}. \end{aligned} \quad (3.27)$$

The design vector is determined by the time-dependent covariates which are raised in the interval t as well as in the pre-period $t - 1$.

3. Time Lags of Higher Orders

It holds for the order r :

$$\tilde{\mathbf{X}}(t) = (\mathbf{X}(t-r)^t, \dots, \mathbf{X}(t)^t)$$

and in this case the design vectors are

$$\mathbf{Z}(t) := \begin{pmatrix} e_t \\ X_0 \\ 0 \\ \vdots \\ 0 \\ \tilde{\mathbf{X}}(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \gamma \\ \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{T_0} \end{pmatrix}, \quad \text{with } \theta_t = \begin{pmatrix} \theta_{t,t-r} \\ \vdots \\ \theta_{t,t} \end{pmatrix}. \quad (3.28)$$

If one additionally considers categorical variables such as the sex of individuals as covariates, then a coding of the covariates must take place.

With coding possibilities, one distinguishes between a $(0, 1)$ coding, i.e.

$$X_{0q} = \begin{cases} 1, & \text{if category } q \text{ is present;} \\ 0, & \text{otherwise,} \end{cases}$$

with Q as categories, and $q = 1, \dots, Q - 1$, i.e. the Q th category serves as reference category, and as a second possibility, the so-called effect coding, given in

$$X_{0q} = \begin{cases} 1, & \text{if category } q \text{ is present;} \\ -1, & \text{if category } Q \text{ is present;} \\ 0, & \text{otherwise.} \end{cases}$$

The effect coding originates from the variance analysis, in which the sum of the effects of a variable is set equal to 0. Moreover, interaction effects, i.e. the combined influence of a given combination of two or more characteristics, come into question as an influence value. Here, however, a statement with an increasing covariate number becomes difficult.

The $(0, 1)$ coding is useful, especially with formulations containing metric as well as categorical covariates; at the same time, when choosing a reference category, one must make sure that the parameters θ_0 represent the "distances" of the j th distinction from the reference category.

For the special case of a dichotomous attribute, i.e., the covariate sex, one only obtains a dummy variable

$$X_{0pi} = \begin{cases} 0 & \text{"female"} \\ 1 & \text{"male"}. \end{cases},$$

In this case, X_{0pi} are the p th components of the time-independent covariate vector X_0 relating to the i th individual.

The design choice 1 was chosen for the empirical investigation of the remigration behavior, thus without a time lag. The sojourn has been divided into 5 intervals, which have proven to be relevant in the non-parametric study *Brecht/Michels (1991)* (cf. table 2).

Table 2: Interval Division of the Length of Stay of Foreigners in Germany

Interval Number	1	2	3	4	5
Length of Stay in Years	1 – 8	9 – 13	14 – 19	20 – 25	> 26

3.3 Results

The total sample, which is characterized by its large portion of censored data, is used as an underlying data basis for the empirical analysis. In addition to this, two new samples are selected which have a controlled portion of 50% and 20% censored data respectively. A significance test is carried out by the standardized coefficients of the covariates

$$\left| \frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \right|,$$

with $\hat{\beta}_k$ as a coefficient and $s(\hat{\beta}_k)$ as the estimated asymptotical standard deviation. According to the hypothesis $H_0 : \beta_k = 0$, these test statistics belong asymptotically to the standard normal distribution. Therefore, a significant influence of the covariates is present in the case of an amount higher than 1,96. The asymptotic behavior in the present discrete-time case follows from statement (iv) given in section 2.3.

I Estimations with the Total Sample

The covariate "sex" was chosen for the total sample, in which the coding

$$X_{01} = \begin{cases} 0 & \text{"female"} \\ 1 & \text{"male"} \end{cases}$$

was used. Table 3 provides the data record description and the estimation values which were obtained with the program PRODISA.

Table 3: Estimation of the Total Sample and the Covariate "Sex"

	Turkey	Italy	Yugoslavia	Greece	Spain
Sample	1217	612	574	444	399
censored (z=0)	1073	508	533	383	324
$\hat{\beta}(\text{Sex})$	0,347	0,331	0,096	0,190	0,334
variance	0,031	0,041	0,102	0,068	0,057
$\hat{\beta}/\text{S.E.}$	1,971	1,647	0,299	0,732	1,406

The larger influence of the male sex on the inclination to remigrate is recognizable. However, the estimation for Turkish individuals does show a significant value in this case. The largest variance occurs as expected with the Yugoslavians; the largest portion of censored data lies here. On the other hand, the smallest variance (Turkey) is connected with the largest sample. However, it should be taken into consideration with these statements that bias occurs as a result of the large portion of censored data. In the case of a reduction of the censored data, a significant change in the estimation values can occur in the amount as well as in the direction of the effect.

When one compares these results with the paper of *Brecht/Michels (1991)*, the shorter time of stay of the female sex established there cannot be verified. Nevertheless, in order to avoid a misinterpretation, one would have to model a dependency structure for the study of men and women to adequately take into consideration the paralellism of this remigration decision. The first theoretical approaches for this can be found in *Brecht, L. (1991b)*.

II Estimations with a Subsample (50 % Censored Data)

In order to demonstrate the effects of the large portion of censored data on the estimation, only the covariate "sex" is used at first, analogous to case I. This reduces the proportion of individuals without a change of state ($z=0$) to 50%.

Table 4: Subsample (50% Censored Data) and Covariate "Sex"

	Turkey	Italy	Yugoslavia	Greece	Spain
Sample	268	210	82	122	150
censored ($z=0$)	134	105	41	61	75
$\hat{\beta}(\text{Sex})$	0,113	0,077	-0,013	0,594	0,254
variance	0,031	0,041	0,106	0,069	0,057
$\hat{\beta}/\text{S.E.}$	0,638	0,380	-0,039	2,259	1,065

The estimations in table 4 indicate a significant influence of the Greek men on the inclination to remigrate; in the case of other countries, this influence is either weakened in relation to case I, or, in the case of the Yugoslavians, the direction of the influence has even changed. The variance, on the other hand, has stayed practically unchanged.

Another problem appears when an additional covariate "age" is included with the given data record. Table 5 shows that the women now tendentially have a greater inclination to remigrate, although none of the estimation values for the covariates "sex" are significant. On the other hand, "age" has a positive influence that is shown to be significant, except in the case of Italy.

Table 5: Subsample (50% Censored Data) and Covariates "Sex", "Age"

	Turkey	Italy	Yugoslavia	Greece	Spain
$\hat{\beta}(\text{sex})$	-0,196	-0,054	-0,137	0,335	-0,121
Var (sex)	0,032	0,046	0,105	0,073	0,060
$\hat{\beta}/\text{S.E. (sex)}$	-1,089	-0,251	-0,424	1,238	-0,494
$\hat{\beta}(\text{Age})$	0,028	0,008	0,024	0,019	0,026
Var (age)	$0,27 \cdot 10^{-4}$	$0,29 \cdot 10^{-4}$	$1,01 \cdot 10^{-4}$	$0,56 \cdot 10^{-4}$	$0,49 \cdot 10^{-4}$
$\hat{\beta}/\text{S.E. (age)}$	5,445	1,557	2,393	2,619	3,713

As a result of the interaction effect, the influence of the covariates "sex" is clearly weakened. The question arises, what happens when an additional covariate is included and the portion of censored data is further reduced.

III Estimations with a Subsample (20% Censored Data)

First a new subsample is taken in which the portion of individuals who have not remigrated is 20%. The covariates "sex" and "age" have been retained (table 6).

Table 6: Subsample (20% Censored Data) and Covariates "Sex", "Age"

	Turkey	Italy	Yugoslavia	Greece	Spain
Sample	168	131	51	76	94
censored (z=0)	34	26	10	15	19
$\hat{\beta}(\text{Sex})$	-0,078	0,005	-0,066	-0,108	-0,106
Var (Sex)	0,035	0,044	0,109	0,075	0,059
$\hat{\beta}/\text{S.E.}(\text{Sex})$	-0,415	0,023	-0,199	-0,393	-0,436
$\hat{\beta}(\text{Age})$	0,016	0,003	0,007	0,006	0,009
Var (Age)	$0,2 \cdot 10^{-4}$	$0,27 \cdot 10^{-4}$	$1,3 \cdot 10^{-4}$	$0,72 \cdot 10^{-4}$	$0,49 \cdot 10^{-4}$
$\hat{\beta}/\text{S.E.}(\text{Age})$	3,525	0,559	0,625	0,712	1,310

In comparison to table 5, only age in the case of Turkish individuals has a positive, significant influence; all other values, for age as well as for sex, must be rejected.

Finally, a third covariate "marital status" should be introduced. This is coded with

$$X_{03} = \begin{cases} 0 & \text{"not married"} \\ 1 & \text{"married"} \end{cases} ,$$

When considered as the only covariate in the model, it shows a positive, significant influence for Italy and Spain, i.e. married individuals have a greater inclination to remigrate than single individuals (table 7).

Table 7: Subsample (20% Censored Data) and Covariate "Marital Status"

	Turkey	Italy	Yugoslavia	Greece	Spain
$\hat{\beta}(\text{MARST})$	0,168	0,661	-0,038	-0,284	1,335
Var (MARST)	0,048	0,065	0,149	0,095	0,134
$\hat{\beta}/\text{S.E. (MARST)}$	0,763	2,589	-0,099	-0,922	3,645

When "marital status" is considered as an additional, third covariate along with "sex" and "age," the interaction effects are present again. As shown in table 8, the positive effect of the marital status "married" remains significant for the Italians; however, in the case of the Spaniards, it is lost. On the other hand, a negative influence is revealed to be significant for the Greeks; i.e. in this case, single people have a greater inclination to remigrate. At the same time, age becomes more important in the case of the Greeks; in comparison with table 6, it is now positively significant, in the case of the Spaniards as well. Sex does not have an influence on remigration for any nationality.

Table 8: Subsample (20% Censored Data) and Covariates "Sex", "Age", "Marital Status"

	Turkey	Italy	Yugoslavia	Greece	Spain
$\hat{\beta}(\text{Sex})$	-0,181	0,079	0,196	-0,183	-0,088
Var (Sex)	0,034	0,044	0,133	0,073	0,069
$\hat{\beta}/\text{S.E. (Sex)}$	-0,980	0,375	0,584	-0,678	-0,336
$\hat{\beta}(\text{Age})$	0,008	-0,008	-0,004	0,021	0,024
Var (Age)	$0,34 \cdot 10^{-4}$	$0,48 \cdot 10^{-4}$	$1,35 \cdot 10^{-4}$	$0,79 \cdot 10^{-4}$	$0,93 \cdot 10^{-4}$
$\hat{\beta}/\text{S.E. (Age)}$	1,431	-1,17	-0,384	2,401	2,486
$\hat{\beta}(\text{MARST})$	0,023	0,784	-0,005	-0,804	0,566
Var (MARST)	0,067	0,084	0,196	0,133	0,208
$\hat{\beta}/\text{S.E. (MARST)}$	0,088	2,712	-0,011	-2,21	1,241

The investigation of a further covariate "entitlement to a pension from the German pension insurance" unfortunately fails because of the too low sample size of those foreigners who replied to this in the Socio-economic Panel. Other variables that are time-dependent, such as income or sense of belonging to a certain nation, have been omitted so far.

4 Conclusion

The modelling of time-dependent covariates will be the subject of the future study. It should be considered, with what portion of censored data the analysis should be carried out, since the previous results show differences that are unacceptably large. In addition, the interaction effects that greatly influence the interpretation of the results should be taken into consideration. Furthermore, it would be desirable to be able to model a dependency structure which reflects the correlation, for example, of the decision to remigrate according to husband/wife. The first theoretical steps have been undertaken in this direction. These concern the analysis of general parallel stochastic processes that mutually influence each other and whose correlation structure should be investigated. With this, one comes to the topic multivariate sojourn models, which should be discrete-time or continuous-time estimated. For research results, see *Brecht, L. (1991b)*.

5 References

- ANDERSON, P.K., R.D. GILL : Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, Vol.10, No.4, 1982, pp.1100-1120.
- ARANDA-ORDAZ, F.J. : An extension of the proportional-hazards-model for grouped data. *Biometrics* 39, 110-118, 1983.
- ARJAS, E., A. HAARA : A logistic regression model for hazard: asymptotic results. *Scand. J. Stat.* 14, 1987, pp.1-18.
- ARJAS, E., A. HAARA : A note on the asymptotic normality in the Cox regression model. *Annals of Statistics*, Vol.16, No.3, 1988, pp.1133-1141.
- BRECHT, B. : Aufbau, Struktur und Anwendungen des Sozio-ökonomischen Panels in INGRES. Diskussionsbeitrag Nr.II-120, SFB 178, Universität Konstanz, 1990.
- BRECHT, B., P. MICHELS : Anwendungen nichtparametrischer Schätzverfahren für die Hazardfunktion bei zensierten Daten auf die Aufenthaltsdauer von Gastarbeitern in der Bundesrepublik. Diskussionsbeitrag Nr.II-137, SFB 178, Universität Konstanz, 1991.
- BRECHT, L. : Ein zeitdiskretes Modell zur Verweildaueranalyse: Regularitätsbedingungen und asymptotische Ergebnisse. Diskussionsbeitrag Nr. 127/s, Fakultät für Wirtschaftswissenschaften und Statistik, Universität Konstanz, 1991(a).
- BRECHT, L. : Ansätze zur semiparametrischen Regressionsanalyse multivariater korrelierter Verweildauermodelle. Diskussionsbeitrag Nr. 129/s, Fakultät für Wirtschaftswissenschaften und Statistik, Universität Konstanz, 1991(b).
- GLAMOUR , User's Guide, Version 2.0. Institut für Statistik und Wirtschaftsgeschichte, Universität Regensburg, 1990.
- HAMERLE, A., G. TUTZ : Diskrete Modelle zur Analyse von Verweildauer und Lebenszeiten. Campus Verlag, 1989.
- PRODISA (Program for Discrete Survival Analysis), Version 1.0. Fakultät für Wirtschaftswissenschaften und Statistik, Universität Konstanz, 1991.