Statistical and Empirical Issues in the Analysis of Duration Data

Doctoral Thesis

Author: Anna Espinal Berenguer
Supervisor: Albert Satorra Brucart

Departament d'Economia i Empresa
Universitat Pompeu Fabra
October, 2001

Dipòsit legal: B.8967-2003
ISBN: 84-688-2274-4

Aquesta tesi és el fruit d'un llarg procés en el qual no he participat en solitari. És per això que vull fer un agraïment molt especial a algunes persones :

De la UPF, a l'Albert Satorra com a director que m'ha plantejat problemes interessants, al Jaume García com a revisor de la segona part i perquè ha confiat sempre en mi, al Michael Greenacre com a corrector de l'anglès, al Frederic Udina per l'ajut tècnic, al Robert Díez per facilitar-me els temes administratius, a la Lydia García pel seu ajut en tràmits i terminis, i al Joan Trench pel seu suport en els temes informàtics.

De la família, al Pitu que s'ha deixat atabalar i ha fet rutllar la casa en tot moment, a l'Àurea i la Sira que han "entès" que la mare hagués d'anar a treballar alguns dies que no tocava, i a les àvies i avis que han ajudat en aquells dies que es necessitaven més hores del compte.

Als col.legues del GRASS començant per la Lupe, que m'ha revisat la feina més d'un cop aportant-t'hi idees molt interessants, i a tots els altres pel caràcter encoratjador dels seus comentaris.

Als amics, que han viscut aquest projecte amb gran interès i ànims perquè arribés a ser realitat.

A tots ells els dono les gràcies i els dedico aquest treball.

Contents

Introduction

The background of this thesis is the analysis of time-to-event data, that is data related with the individual time elapsed in a certain situation or state. Examples of these kind of data comes from diverse fields such as medicine, biology, public health, epidemiology, engineering, economics and demography. In economics, two examples are the time on unemployment or the duration of an individual in a certain job. The main feature of these data is the issue of censoring, which occurs when the periods of time for some individuals cannot be completely observed. The presence of censored observations requires the use of specific techniques and analyses, usually named Survival Analysis (e.g. Klein & Moeschberger, 1997).

Survival analysis comprises a set of specialized statistical methods used to study response time data. In analyzing such data the main goals are to determine the length of time intervals spent in a state, and the transition probabilities from the current situation to the next entered state.

Even though survival analysis arises from the analysis of life tables in demography (see, e.g. Berkson & Gage, 1952, Cutler & Ederer 1958, Geham 1969) and studies of mortality in biostatistics sciences (see, e.g. Irwin 1942, Armitage 1959, Pike, 1966, Peto & Lee 1973), it is amenable to a wide range of questions in fields such as epidemiology, social sciences and economics. Indeed, data related to employment and occupational careers are sequences of duration times in several states that may be considered as survival data. The main advantages of this kind of analysis compared to others commonly used in econometrics (linear regression, limited dependent variable models or time series) are the ease of computing the probabilities of transition between two different states and the absence of assumptions of implicit hypotheses required by the other approaches. The development of survival analysis adapted to econometric data is named the analysis of Duration Models with most contributions starting in the seventies. Some of the important references are Lancaster (1979), Nickel (1979), Heckman & Borjas (1980), Heckman & Singer (1984), Kiefer (1988), Lancaster (1990), Han & Hausman (1990), and Narendranathan & Stewart (1993).

Our work is focused on two aspects of the survival analysis. In Part I our emphasis is on linear regression model when the response variable is censored (Breiman, Tsur & Zemel, 1993). Here we propose a method for estimating the coefficients of such linear models when covariates contain measurement error. In Part II we propose the use of survival techniques in the analysis of Spanish labor histories. From these multivariate data we make several analyses to study the duration of the feasible states in the labor market.

0.1 Brief introduction to survival analysis

The interest of these statistical tools is mainly based on two distinguishing features of time data. Firstly, duration times are non-negative values, usually describing a highly skewed distribution, and therefore the assumption of a normally distributed variable may not be valid. Secondly, the true duration is not always observed. Indeed, if data are collected in a certain period of time, at the end of the study some subjects remain in the same state before any change has occurred. Their episodes of time, or spells, are then partly recorded. This characteristic is known as censoring and it is the most important reason for using the special methods developed in survival analysis.

Censored observations may appear in many situations and due to different mechanisms (see e.g. David & Moeschberger, 1978), the most usual types of censoring being as follows:

- An observation is said to be right-censored if it is recorded from its beginning until a well defined time before its end point. For instance, if we follow for several months a set of individuals who became unemployed at a certain known date, some of them will become employed and the others will still remain in the same situation. For the latter group, we only know that the whole period of unemployment runs past the end point of the follow-up. See Klein & Moeschberger (1997) for a discussion about types of right censoring.

- An observation is said to be left-censored if the starting point of its spell is unknown, being recorded only from a well defined point until its end. For instance, using the same example about unemployment data, left censoring occurs when the entry date is before the start of the follow-up period, that is when an unemployment period is recorded when it is already started.

- An observation is said to be interval-censored if it is only known that the event occurs within a time interval but the exact point is unknown. This situation is less usual with econometric data and it is mainly found in medical studies. For instance in a study designed to know when people infected with the AIDS virus develop AIDS, the patients are periodically examined with negative results until they have the first positive examination. In such a case, the exact time of developing AIDS is only known to be between the two last dates.

In economic data two are the usual cases. On the one hand, the cohort studies like the Cohort Study of the Unemployment in Britain (e.g. Nickel et al. 1989) where the individuals are followed for a period of time and at the end some of them are still at the same situation (censored observation) and others have been changed and, therefore have a complete observation. On the other hand, data have been collected on a certain moment on time so that all the observations are censored. In this case it is required specialized method (see Salant, 1977 and Flinn, 1986).

We shall now introduce some notation and the basic concepts used in survival analysis. Even though there is not sure a standard notation because survival analysis is used in several fields we introduce what we consider the most usual in econometrics.

First of all we are going to introduce in a formal way the time variable that defines the individual duration in a certain situation. That is, let T be a continuous and non-negative random variable with density function $f(t)$ and distribution function $F(t)$. In the analysis of duration data two specifications of the distribution of T are very useful, the survival function, which is the probability of an individual remaining in a certain state beyond time t, and the hazard rate (function) which is the chance an individual ends the current state in the next instant.

The survival function is defined as

$$S(t) = P(T \geq t) = 1 - F(t),$$

and it is the probability of having an episode larger than t or surviving in a certain situation beyond t. In economics the knowledge of the survival function may be used to approximate the proportion of long-term unemployed people, the proportion of employees with a large tenure job or the probability of retirement of older people.

Another fundamental concept in survival analysis is the hazard function, also known as the inverse of the Mill's ratio in economics, which is defined as

$$\lambda\,(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

and it represents the instantaneous probability of changing the current state. Here we note that $\lambda\,(t)\Delta t$ approximates the conditional probability of leaving in the next instant the state that has been occupied during the past t time units. This function is very useful for describing the way in which the chance of ending a current state is changing with time.

Note that if we know any of the four functions, $f(t)$, $F(t)$, $S(t)$ and $\lambda(t)$ the other three can be uniquely determined. Some interesting relationship among them are

$$\begin{aligned} \lambda\,(t) &= \frac{f\,(t)}{S\,(t)} \\ S\,(t) &= \exp\left(-\int_0^t \lambda\,(u)du\right) = \exp\left(\Lambda\,(t)\right) \end{aligned}$$

where $\Lambda(t)$ is the cumulative hazard function, and

$$f\,(t) \quad = \quad \lambda\,(t)\,\exp[-\Lambda\,(t)],$$

and hence, $\lambda\,(t) = \lambda$, $\forall t \geq 0$ if and only if $S(t) = e^{-\lambda t}$ and $f(t) = \lambda e^{-\lambda t}$, i.e. T has the exponential distribution with parameter λ.[1]

The estimation of the distribution of T is one of the main goals in the analysis of survival data. However, we note that the standard approximation coming from the empirical distribution function cannot be used in survival analysis because of censoring. In the next we introduce a non-parametric estimation of the survival function for right-censored data.

We assume a type I censoring, a right-censored mechanism where the whole spell is observable only if it ends prior to some specific time. That is, for each individual

[1] For general models with non-constant hazard functions see Cox & Oakes (1984) or Klein & Moeschberger (1997).

we have a fixed censoring time denoted by C. Thus, the observed duration data can be conveniently represented by pairs of random variables (Z, δ), where δ indicates whether Z corresponds to a complete observation $(\delta = 1)$ or is censored $(\delta = 0)$, and Z equals T if it is observed, and $Z = C$ for the censored observations, i.e. $Z = \min\{T, C\}$. Thus a sample of n individual durations may be expressed as $\{(z_i, \delta_i),\ i = 1, \cdots, n\}$.

The standard estimator of the survival function, taking censoring into account, was proposed by Kaplan & Meier (1958). Two quantities are required in order to obtain such an estimator. Let $t_{(1)} < t_{(2)} < \cdots < t_{(D)}$ be the ordered times corresponding to the complete observations. Let d_j be the number of individuals with $t_i = t_{(j)}$ and let N_j be the number of individuals at risk at time $t_{(j)}$ (i.e. the number of individuals with $z_i \geq t_{(j)}$). Then, the estimator proposed is defined as

$$\hat{S}\,(t) = \begin{cases} 1 & \text{if} \quad t < t_{(j)} \\ \\ \prod\limits_{t_{(j)} \leq t} \left[1 - \frac{d_j}{N_j}\right] & \text{if} \quad t_{(j)} \leq t \end{cases}$$

where for t beyond the largest observation this estimator is not well defined (see Efron, 1967 and Gill, 1980). Estimators of $F(t)$ or $\Lambda(t)$ are straightforward once $\hat{S}\,(t)$ has been obtained. There is another non-parametric estimator of $\Lambda(t)$ proposed in Nelson (1972) and Aalen (1978) which may be an alternate estimator of $S(t)$ and it may also provide crude estimates of the hazard rate $\lambda(t)$.

Several examples of the Kaplan-Meier estimator will be given in the Part II of the thesis, where we approximate the survival probabilities of remaining in three states of the labor market (self-employment, wage-earner and non-working).

We shall now make some comments about modeling when additional characteristics are also known about the individuals. In this case it is possible to analyze the effect of these explanatory variables (or covariates) on the durations. Two approaches are commonly used in survival analysis. The first can be viewed as an extension of the classical linear regression approach. Thus, the logarithm of the durations $Y = \ln T$ is modeled as

$$y = \mathbf{x}'\boldsymbol{\beta} + w, \tag{1}$$

where $\mathbf{x}$ is the vector of covariates, $\boldsymbol{\beta}$ is the vector of regression coefficients and w is the residual term. This model is called the accelerated failure time model. A second class of models is based on the hazard function. Thus, given the covariates, the conditional hazard rate is modeled as the product of a base-line hazard rate (a function of t) and a non-negative function of the covariates, that is

$$\lambda\,(t/\mathbf{x}) = \lambda_0\,(t)\; g\,(\mathbf{x}'\boldsymbol{\beta}).$$

Most applications use $g\,(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$ proposed by Cox (1972). This model is also called the proportional hazards model. Both models are introduced in Section 1.1 and more details about advantages and disadvantages of both families of models are in Klein & Moeschberger, (1997).

0.2 Survival analysis in econometrics

Periods of unemployment, time to re-enrol in school, time developing a profession or the age of retirement are economic examples of time response variables. Econometric methods for analyzing these type of data are known as econometric duration analysis, and these are based on survival analysis techniques. Even though the main results have been obtained in the last two decades Silcock (1954) was already using the hazard function in the study of employment durations. We now review some of the literature about duration analysis.

As key references we single out Lancaster (1979), Heckman & Singer (1984) and Kiefer (1988). The first one is a pioneer work about unemployment analysis linking search theory and duration analysis. It starts assuming an exponential distribution for the duration of unemployment which is generalized to a Weibull distribution. Later on Lancaster (1979) allows for unobserved heterogeneity due to error of specification possibly due to the omission of relevant regressors. The second reference by Heckman & Singer (1984) contains an introduction to the main ideas and concepts of duration models, emphasizing the distinctive features of econometric data with respect to data analyzed in biostatistics or reliability. They also report three examples of duration models that generalize the more standard discrete choice theory (see also Sueyoshi, 1995). The paper says that the discrete choice models such as logit and probit, when is defined for one time interval, are of a different functional form when applied to another time unit, if they are defined at all. They also emphasize that continuous time models are invariant to the time unit used to record the available data. Therefore a common set of parameters can be used to generate probabilities of events occurring in intervals of different length. For this reason the use of continuous time duration models has become widespread in economics. The paper of Kiefer (1988) is a good starting reference to the issue of the analysis of economic duration data. It defines the specific concepts of the topic and reviews the most important contributions in model specification, estimation, and hypothesis checking.

The main development of duration models has been in the analysis labor data. In this context interest is focused on the duration of individuals in a certain state (e.g. employed, unemployed or out of the labor force) and which transition to another state has taken place. Flinn & Heckman (1982a) present a duration model that accommodates as special cases the model of Jovanovic (1979) and the index function model widely used in labor economics. General surveys are provided by Lancaster (1990) who deals with model building and inference for the econometric analysis of transition data. Devine & Kiefer (1991) is a more specific reference about labor economics analysis comparing duration models with job search theory. Two more references partly related with the issue of duration analysis are Heckman & Singer (1988) and Florens, Ivaldi, Laffont & Laisney (1990). The first one is about the longitudinal analysis of labor market data, with a general review of duration analysis and some specific chapters about heterogeneity (Chamberlain, 1988), counting processes (Andersen, 1988) and the analysis of data about transition to work (Mare & Winship, 1988). Florens, Gérard-Varet & Werquin (1990) contains a survey of micro-econometrics. We also draw attention to the chapter about general concepts in duration analysis (Florens, 1990) and an empirical study of unemployment (Florens et al. 1990).

The topics of interest when dealing with duration analysis are mainly the transition to a new state, the effects of unobserved heterogeneity and the identifiability problems of the more usual models. When unemployed individuals may experience some competing events such as employment, out of labor force or enrollment to school the theory of competing risks models (David & Moeschberger, 1978) may be applied, allowing differences among the entering states. Flinn & Heckman (1982b) showed that the competing risks model can be applied to the three-state (employment, non-employment and non-market activity) model of labor force dynamics. Narendranathan & Stewart (1993) using a competing risks model for analyzing unemployment durations, showed that the effect of income is biased if there is no distinction among transitions. Han & Hausmans (1990) specified a flexible parametric proportional competing risks model which allows unrestricted correlation among the risks. Recent references establishing a competing risks model are McCall (1997) which analyzes the determinants of full-time and part-time reemployment; Dolton & O'Neil (1996) which distinguish three transitions from unemployment: to a job, a training placement or to signing-off unemployment benefits; and Mealli, Pudney & Thomas (1996) which specify a competing risks model when there is a natural limit on the duration of some state.

Unobserved heterogeneity is found when some relevant individual information is not available. Heckman & Singer (1984) showed the effects of ignoring heterogeneity among individuals using simple examples. More about this issue are in Flinn & Heckman (1982) and Elbers & Rider (1982). Two specific papers are Van den Berg & Van Ours (1996) which is about unobserved heterogeneity different from duration dependence and McCall (1994) who proposes a proportional hazards testing under unobserved heterogeneity.

The study of identification problems are centered on the two main classes of models used in survival analysis. Ridder (1990) deals with the generalized accelerated failure time models while Heckman & Honoré (1989) is focused on the identifiability of competing risks models or Heckman & Singer (1984) for the proportional hazards model.

0.3 Main results

The common issue of the thesis is the survival analysis which is used in two different ways. In this section we summarize the main results obtained in each part.

The first part of the thesis is focused on linear regression models with two features: a dependent variable possibly censored, and the explanatory variable contaminated with measurement error. As expected, the well known techniques to estimate the regression coefficients of linear models with explanatory variables measured with error (Fuller, 1987), give biased estimates due to censoring. We propose a methodology which produces consistent estimates of the regression coefficients based on errors-in-variables methods but taking into account the censoring of the dependent variable. Since the estimation is performed in two stages, we have called it the *two-step estimator.*

The two-step estimator is a procedure easy to implement in practical applications. It combines two methods of estimation already existing but used separately: on the one hand, the procedures of estimation for linear models with censoring and, on the other hand, the estimation methods for linear models with measurement error. Standard errors of the estimator are computed using the Bootstrap method. The performance of the the proposed estimator is studied carrying out Monte Carlo studies varying the sample size, the magnitude of the measurement error and the proportion of censoring. From the results of these simulations we conclude that the two-step estimator remains unbiased in any scenario, and looking at the expected sampling variability, the empirical values match the theoretical ones.

The second part of the thesis is about duration analysis, the term usually used in econometrics to name the analysis of time-to-event data. We analyze data about Spanish labor market histories during the period 1980-1993. The main goal is the analysis of the duration of three types of labor spells (self-employment, wage-earner and non-working). We start with non-parametric approaches of the survival functions for several sets of spells. In the statistical analysis, we carry out several studies: First we analyze the duration of the first spells of the labor history; second, we carried out separate analysis for each state (self-employment, wage-earner and non-working) using also competing risks models to allowing differences among the possible transitions; third, we have considered the five early spells of the labor histories taking into account the dependencies between observations of the same individual.

PART I: Censored Linear Models with Measurement Errors on Covariates

Typical analysis of survival data assesses the impact of several explanatory variables on the time duration response variable. The standard methodology for such analysis assumes that the explanatory variables, or covariates, are measured without error. This assumption is often violated in practice, since we frequently encounter covariates that clearly suffer from measurement error (for example, income, dietary fat consumption, learning skills, exposure levels, etc.).

The effect of measurement error on covariates has been mainly studied in the case of linear and non-linear models. See Fuller (1987) and Carrol, Ruppert & Stefanski (1995) for a general overview. In survival analysis the topic of measurement error in covariates has not been too much studied. Some important exceptions are Cheng & Wang (2001), Jiang, Turnbull & Clark (1999), Kulich & Lin (2000), Nakamura (1992) and Prentice (1982). The recent paper by Cheng & Wang (2001) generalize the linear transformation models (see Dabrowska & Doksum, 1988) to accommodate measurement error on covariates. That is they assume a proportional odds model for the survival time and a linear relationship between the observed and the true covariates. There is also a sensitivity analysis using various measurement error reliability ratios. In Jiang et al. (1999) consider a regression analysis for repeated event where is taking account the presence of measurement error on covariates and modelling the possible unobserved heterogeneity as random effects. The paper of Kulich & Lin (2000) assumes an additive hazards model (see Breslow & Day, 1987) and proposes consistent estimates for the regression coefficients that are asymptotically normal distributed. The last two papers deal with the proportional hazards model proposed by Cox (1972). Nakamura (1992) shows the effects of the measurement error on the relative risk estimates and, the paper by Prentice (1982) develops a modified partial likelihood for consistent estimation of the parameters of interest. However, in both procedures the estimate are obtained after minimizing a function and this could sometimes be cumbersome in practical applications. Other work has paid attention also to measurement error on the duration variable (see Holt, McDonald & Skinner, 1991).

In this first part of the thesis we deal with the analysis of data where a response variable is right censored and some covariates are contaminated with measurement error. Assuming a linear model we focus on obtaining consistent estimates of the regression parameters. In Espinal & Satorra (1996) we showed that for the regression coefficients, the bias of the estimates obtained from the observed covariates increases when the amount of measurement error increases. Moreover, even when only one covariate has measurement error, all the estimates for the regression coefficients may be biased. Motivated by these results we develop a procedure for estimating accelerated failure time models when the covariates are subject to measurement error. We assume a log-linear model with a right-censored response, and a set of covariates some of them measured with error. Thus, our model allows data that have two sources of non-observability. On the one hand, the response variable may be censored and, on the other, covariates may be contaminated with error. For this, we propose a sequential procedure for consistent estimation of the regression parameters that takes measurement error into account. It is a sequential method which uses techniques from survival analysis followed by methods based on measurement error models. A practical advantage of this two-step estimation procedure is that the estimates may be obtained using standard software packages.

The first step of the method deals the issue of censoring. We take a method of estimation valid for linear models with a censored response based on the paper of Buckley & James (1979). The method used is proposed in Schneider & Weissfeld (1986). We apply this procedure to estimate the log-linear model using the observed covariates. From the estimates of the regression parameters we compute the linear predictions for the response. Then, a consistent estimate of the moments matrix between the covariates and the true response may be obtained.

The second step takes into account the measurement error of covariates. We assume a linear relationship between the observed and the true covariates. Then we advocate the linear regression models for the method of estimation, taking into account measurement error on covariates. In order to avoid the usual identifiability problems of these models, we assume that the covariance matrix of measurement errors is known. However, this assumption could be relaxed (see Fuller, 1987). The estimates of the regression parameters of the log-linear model are computed using the observed covariates and the estimated moments matrix obtained in the previous step.

The performance of the two-step estimator is studied using simulated data. We carried out some Monte Carlo simulations varying sample sizes, proportion of censoring and amount of measurement error.

Finally, standard errors are also obtained. Even though for the case of uncensored data the standard errors from the normal theory are still valid, they are not so in the presence of censoring, and we require bootstrap methods.

This part is structured as follows. In Chapter 1 we review the methods of estimation for survival models. We start with fully parametric models based on maximum likelihood estimation when a family of distributions is specified up to a vector of parameters. We note the impact of the presence of censored observations on the form of the likelihood function. In particular, we discuss the likelihood function for the families of proportional hazards and accelerated failure time models. Next, we focus on linear models with a censored dependent variable. We introduce methods for estimating unknown parameters based on the ordinary least squares which are mainly iterative procedures due to censoring.

In Chapter 2 we discuss the topic of measurement error. We define the measurement error on covariates as a random variable which is the difference between the true and the observed value of the covariate. In Section 2.1 we describe the estimation procedures for the regression parameters that takes measurement error into account. In Section 2.2 our attention is on work dealing with survival analysis and measurement error on covariates. Here there is a brief summary of the main findings of previous papers dealing with this issue.

In Chapter 3 we introduce the proposed two-step estimator. In Section 3.1 we introduce the linear model assumed between the complete response and the true covariates and, we also define the mechanism of censoring for the response and the measurement error equation of covariates. In order to motivate our procedure, in Section 3.2 we do a Monte Carlo study in order to show the effects of ignoring measurement error on covariates for survival data. In Section 3.3 we describe the two steps of the method for obtaining consistent estimations of the regression parameters. We indicate how to take into account both the presence of censoring and measurement error. The two-step estimator emerges as a result of iterating the previous two steps. Finally we indicate how to implement this methodology and discuss the standard errors of the estimators. The performance of the two-step estimator is analyzed in Section 3.4 using Monte Carlo methods.

Chapter 1

Estimation in Survival Models

The standard methods for estimating parameters in regression type of models have to be modified in order to accommodate the usual censoring of the failure time data. For this reason in this chapter we review the main characteristics of methods for estimating parametric or semi-parametric survival regression models.

Section 1.1 contains the fully parametric methods based on maximum likelihood when a family of survival distributions is specified up to a vector of parameters. We note the impact of censoring on the form of the likelihood function. That is, it has two blocks due to the contributions of the censored and the complete failure times. In particular, we discuss the likelihood function for the families of the proportional hazards (PH) and the accelerated failure time (AFT) models. We give an introduction to these procedures and describe some basic tools of survival analysis. A more detailed discussion about these methods of estimation and general approaches to survival analysis can be found in Kalbfleisch & Prentice (1980), Lawless (1982), Cox & Oakes (1984), Blossfeld, Hamerle & Mayer (1989), Collett (1995), Andersen, Bogard, Gill & Keiding (1993) and Klein & Moeschberger (1997).

In Section 1.2 we focus on linear models with a censored response. We present least squares (LS) methods for estimating regression parameters. The presence of censoring is not innocuous and induces bias in the LS estimates obtained from the observed data. To correct this bias several modifications of the LS have been developed. Miller (1976) followed by Buckley & James (1979) are the starting references. More papers related to this issue are by Koul, Susarla & Van Ryzin (1981) and Schneider & Weissfeld (1986) among others. A review of these methods as well as the hypotheses assumed for each of them are discussed. Even though these procedures have been seldom used in survival analysis, their ease of implementation and the popularity of LS in fields such as econometrics, motivates our interest in them. Moreover, these related methods to LS estimation are very appropriate when the effects of covariates on the failure time can be formulated as a log-linear model.

1.1 Standard methods

In the context of survival models, the theory of maximum likelihood (ML) can be applied to parametric models for time data with arbitrary censoring mechanism.

Let us consider failure times as observations of a non-negative continuous random variable T with Type I censoring. We denote by C a random variable independent of T such that $T > C$ corresponds to censoring. Here we also assume a single observation for each individual.

Let $F(t;\boldsymbol{\theta})$ be the distribution function of T known up to the vector of parameters, $\boldsymbol{\theta}$; that is, $F(t;\boldsymbol{\theta}) = P(T \leq t;\boldsymbol{\theta})$. For survival data, the contribution of subjects to the likelihood depends on whether or not the individuals have a complete or a censored observed time. That is, a subject with a complete failure time t_i, contributes to the likelihood with the density function, $f(t_i;\boldsymbol{\theta})$. However, the contribution of a subject with a censored time c_i is the probability of surviving beyond c_i, that is $P(T > c_i;\boldsymbol{\theta})$. Let $S(t;\boldsymbol{\theta}) = P(T > t;\boldsymbol{\theta}) = 1 - F(t;\boldsymbol{\theta})$ be the survival function. Then the full likelihood function for a sample of n independent individuals is

$$\mathcal{L}(\boldsymbol{\theta};t,c) = \prod_{i\in U} f(t_i;\boldsymbol{\theta}) \prod_{i\in C} S(c_i;\boldsymbol{\theta}), \tag{1.1}$$

where $i \in U$ and $i \in C$ denote, respectively, the product over the subsamples of uncensored and censored observations.

If we denote the observed time as $z_i = \min\{t_i, c_i\}$ and the indicator of censoring as $\delta_i = \mathbf{1}_{\{t_i \leq c_i\}}$, we obtain an alternative expression for (1.1)

$$\mathcal{L}(\boldsymbol{\theta};z) = \prod_{i=1}^{n} \left[f(z_i;\boldsymbol{\theta})\right]^{\delta_i} \left[S(z_i;\boldsymbol{\theta})\right]^{1-\delta_i}$$

Using the hazard function defined as $\lambda(z_i;\boldsymbol{\theta}) = f(z_i;\boldsymbol{\theta})/S(z_i;\boldsymbol{\theta})$, the likelihood function may be expressed as

$$\mathcal{L}(\boldsymbol{\theta};z) = \prod_{i=1}^{n} \left[\lambda(z_i;\boldsymbol{\theta})\right]^{\delta_i} \; S(z_i;\boldsymbol{\theta})$$

where taking logarithms and using the result[1]

$$\log S(t; \boldsymbol{\theta}) = -\int_0^t \lambda(u; \boldsymbol{\theta})\, du = -\Lambda(t; \boldsymbol{\theta}),$$

the log-likelihood for right censored survival data is

$$L(\boldsymbol{\theta}; z) = \sum_{i=1}^{n} (\delta_i \, \log \lambda(z_i; \boldsymbol{\theta}) - \Lambda(z_i; \boldsymbol{\theta})) \tag{1.2}$$

which is the likelihood function expressed in terms of the hazard function. The ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained maximizing (1.2) and it has well known properties of consistency and efficiency (see, e.g. Cox, 1979).

Results presented until now involve estimates for the unknown parameters of a univariate distribution for T. However, it is also interesting to assess the effects of some explanatory variables on failure time. This is possible when for each individual i of a sample, there is available a p–vector of covariates, $\mathbf{x}_i = (x_{1i}, \cdots, x_{pi})'$. We assume no missing values in the data.

The ML methodology proposed through equation (1.2) is again applicable. In this case, the likelihood function involves a parameter vector $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\beta})$, where $\boldsymbol{\psi}$ denotes the parameters of the specified survival distribution, and $\boldsymbol{\beta}$ denotes the regression coefficients associated with the covariates. Thus, the hazard function will depends not only on $\boldsymbol{\psi}$ but also on $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\beta})$, so that the log-likelihood will have the same expression (1.2) with unknown parameters $(\boldsymbol{\psi}, \boldsymbol{\beta})$.

Usually, estimates of $\boldsymbol{\beta}$ are of principal importance because the goal of the analysis is to describe the time elapsed before an event occurs among different sets of subjects. For this setting there are two families of models. A brief discussion of the ML procedure for each of them is given in the next two sections.

1.1.1 The proportional hazards models

This model was proposed by Cox (1975) and it has become a very popular model in survival analysis. The model assumes that the hazard function of the random variable T, for a fixed vector of covariates $\mathbf{x}$, is

$$\lambda(t; \mathbf{x}, \boldsymbol{\psi}, \boldsymbol{\beta}) = \lambda_0(t; \boldsymbol{\psi}) \, \phi(\mathbf{x}; \boldsymbol{\beta}) \tag{1.3}$$

[1] In the absolutely continuous case, $\lambda(t; \boldsymbol{\theta}) = \frac{-S'(t;\boldsymbol{\theta})}{S(t;\boldsymbol{\theta})} = -\frac{d}{dz} \log S(t; \boldsymbol{\theta})$.

for some functions $\lambda_0(\cdot)$ and $\phi(\cdot)$ such that $\phi(0) = 1$. Since $\lambda_0(\cdot)$ is equal to the hazard function when $\mathbf{x} = 0$ is a non-negative function, it is named the baseline hazard function. Note that no particular form of the probability distribution for T is assumed.

Property (1.3) implies that the effect of a certain value of a covariate over the probability of leaving a state, is proportional to the hazard function of a reference value. That is, given a sample of individuals, (1.3) establishes for each moment t, a time-constant factor of proportionality between the hazard functions of any pair of subjects say i and j. The constant factor of proportionality is given by

$$\frac{\lambda(t; \mathbf{x}_i, \boldsymbol{\psi}, \boldsymbol{\beta})}{\lambda(t; \mathbf{x}_j, \boldsymbol{\psi}, \boldsymbol{\beta})} = \frac{\phi(\mathbf{x}_i; \boldsymbol{\beta})}{\phi(\mathbf{x}_j; \boldsymbol{\beta})}, \tag{1.4}$$

hence the name “proportional hazards”. Thus, for a discrete variable defining a finite number of subsamples, it allows us to compute the relative risks of leaving a state of one subsample with respect to the others.

In order to estimate the parameters of this model by ML, we could substitute (1.3) in the log-likelihood function (1.2). In any case, we note that this function cannot be maximized without assuming a specific form for $\phi(\cdot)$ and for $\lambda_0(\cdot)$. Cox (1972) proposed a procedure in order to estimate parameters β when $\lambda_0(\cdot)$ remains arbitrary and $\phi(\mathbf{x}; \boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$. This method is based on a modified likelihood function called the partial likelihood developed in Cox (1975). Hence, the proportional hazards model is also known as the Cox regression model.

Now, we summarize the partial likelihood approach (see Kalbfleisch & Prentice, 1980 for more details). Let $t_{(1)}, \cdots, t_{(k)}$ be the oredered distinct failure times for a sample of n individuals (i.e. there are $n - k$ censored observations). For each $t_{(j)}$, let $\mathcal{R}(t_{(j)}) = \{i : z_i \geq t_{(j)}\}$ be the risk set, that is, the set of subjects still alive just before $t_{(j)}$. Then the probability that the failure $t_{(j)}$ is on individual (j) as observed, is

$$P_{t_{(j)}} = \frac{\exp(\mathbf{x}'_{(j)}\boldsymbol{\beta})}{\sum\limits_{i \in \mathcal{R}(t_{(j)})} \exp(\mathbf{x}'_{(i)}\boldsymbol{\beta})}.$$

Then contribution of each uncensored $t_{(j)}$ to the likelihood the partial likelihood function is

$$L^\star(\boldsymbol{\beta}) = \prod_j P_{t_{(j)}} = \prod_j \frac{\exp(\mathbf{x}'_{(j)}\boldsymbol{\beta})}{\sum\limits_{i \in \mathcal{R}(t_{(j)})} \exp(\mathbf{x}'_{(i)}\boldsymbol{\beta})}$$

which depends only on $\boldsymbol{\beta}$, so that its maximization does not involve $\lambda_0(\cdot)$. In this sense, the proportional hazards model is a semi-parametric model.

Even though it is not exactly a likelihood function, the value $\hat{\boldsymbol{\beta}}$ that maximizes $L^\star(\boldsymbol{\beta})$ defined in (1.5) is the ML estimate. In particular, $\hat{\boldsymbol{\beta}}$ is a consistent estimate of $\boldsymbol{\beta}$ and it is asymptotically normally distributed. Cox (1975) presented the proof of these properties.

1.1.2 The accelerated failure time models

A second broad family of models for analyzing survival data establishes a relationship between the failure time and the covariates. Indeed it is assumed that the effects of covariates are on the time scale, that is, different values of covariates lead to shortening the duration of a time interval or hastening the occurrence of the event of interest. For instance, looking at the survival function, the effects of two values of a covariate x, say $x^{(1)}$ and $x^{(2)}$, may be written as $S(t; x^{(1)}) = S(\tilde{t}; x^{(2)})$. Thus, the probability of survival at time t if the covariates take a reference value (usually it is zero and corresponds to some standard set of conditions) is equal to the probability of survival in $\tilde{t}$ when covariates take a different value. Thus the effects of covariates "accelerate" the time at which events occur. This class of models is widely used in engineering (see, e.g. Nelson, 1989).

In terms of random variables, these models assume that

$$T = \frac{T_0}{\phi(\mathbf{x}; \boldsymbol{\beta})} \tag{1.5}$$

where T_0 denotes the failure time under the standard conditions $x = 0$, and $\phi(\cdot)$ is a positive function such that $\phi(0) = 1$, so that its natural form is $\phi(\mathbf{x}; \boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$.

Using (1.5) and letting $\lambda_0(\cdot)$ denote the hazard function of T_0, then the hazard function of T is[2]

$$\lambda(t; \mathbf{x}, \boldsymbol{\psi}, \boldsymbol{\beta}) = \lambda_0(t\, \phi(\mathbf{x}; \boldsymbol{\beta});\ \boldsymbol{\psi})\ \phi(\mathbf{x}; \boldsymbol{\beta}) \tag{1.6}$$

which in general does not have the proportional property defined in (1.3). Relationship with the PH model and properties of these assumption are discussed in Cox & Oakes (1984).

[2] By simple algebra of transforming random variables, we see that the hazard rates $\lambda(\cdot)$ and $\lambda_0(\cdot)$ are related in the absolutely continuous case as $\lambda(t; \mathbf{x}, \boldsymbol{\psi}, \boldsymbol{\beta}) = f(t; \mathbf{x}, \boldsymbol{\psi}, \boldsymbol{\beta})/S(t; \mathbf{x}, \boldsymbol{\psi}, \boldsymbol{\beta}) = \lambda_0(t\, \phi(\mathbf{x}; \boldsymbol{\beta}))\phi(\mathbf{x}; \boldsymbol{\beta})$, where $f(\cdot)$, denotes the probability density function.

Taking logarithms in (1.5) and using $\phi(\mathbf{x};\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$ it can be proved that the assumption of accelerated failure time is equivalent to assuming the log-linear model defined as

$$\log T = \mu_0 - \mathbf{x}'\boldsymbol{\beta} + \epsilon \tag{1.7}$$

where $\mu_0 = \mathrm{E}(\log T_0)$ and ϵ is a random variable. Therefore, specifying a distributional family for ϵ, unknown parameters of (1.7) may be estimated by ML substituting (1.6) in (1.2).

Since equation (1.7) is a standard linear model, other estimation methods for these specific models could also be considered. However, the usual presence of censoring in survival data requires some modification of the standard methods. The next section contains a review of procedures based on least squares when censored observations of the response variable are included.

1.2 Censored linear models

Often, to analyze the effects of a set of explanatory variables on a dependent variable, a linear relationship is assumed, that is,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \cdots, n \tag{1.8}$$

where $y_1, \cdots, y_n$ are independent realizations of Y, $\mathbf{x}_i$ is the $p \times 1$ vector of covariates for individual i and $\epsilon_1, \cdots, \epsilon_n$ are independent and identically distributed (i.i.d.) random variables with unspecified distribution function F with finite mean not necessarily equal to zero and finite variance σ_ϵ^2.

The usual estimates of the unknown parameters $\boldsymbol{\beta}$ in (1.8) are computed using the least squares theory (LS). The interest of these procedures is mainly for two reasons: the ease of implementation and the absence of parametric assumptions about the distribution function for the residuals.

In survival analysis, however, there are at least two features that we need to take into account before applying LS. First, the responses usually corresponding to times are non-negative so that, the usual assumption of normality for the random variables $\epsilon_1, \cdots, \epsilon_n$ is not the most appropriate. Second, some of the values are not observable due to the presence of censoring. Even though the first point is usually solved by taking an increasing transformation of the time variable, for instance $Y = \log T$, the problem of censoring is not so easily solved.

This section reviews the methods based on LS for estimating the regression parameters of linear model (1.8) when Y is a right censored random variable. From now on we will refer to it as the Censored Linear Model (CLM).

We will assume right censored data Type I, so that we do not observe y_i but the pairs (z_i, δ_i), where

$$z_i = \min\{y_i, c_i\}$$

and

$$\delta_i = \begin{cases} 1 & \text{if } y_i \leq c_i \\ 0 & \text{otherwise} \end{cases}$$

with $c_1, \cdots, c_n$ independent realizations of a random variable C, assumed to be independent of Y. Note that $c_1, \cdots, c_n$ are not the censored times but a transformation. It depends on the transformation Y taken for the failure times, i.e. if $Y = \log T$ then $c_1, \cdots, c_n$ are the log-transformed censored times.

The estimation methods for CLM are mainly based on the paper due to Miller (1976), where a simple model with one-dimensional covariate defined as

$$y_i = \alpha + x_i\beta + \epsilon_i, \quad i = 1, \cdots, n \tag{1.9}$$

is considered. This work is the starting point of LS methods for estimating regression coefficients of linear models that take into account a censoring mechanism. However, there are some previous papers (see, e.g. Zippin & Armitage, 1966; Glasser, 1967; Mantel & Myers, 1971) where the mean survival time is related to an independent variable by means of a linear relationship.

1.2.1 A first approximation: Miller (1976)

A first modification of the LS estimator on order to accommodate censored data was suggested by Miller (1976). The proposed method is based on the definition of the LS estimates for (1.9) as those values a and b that minimize $\sum_i (y_i - a - bx_i)^2$. That is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n} (y_i - a - bx_i)^2 = \int \epsilon^2 \, d\hat{F}_{ab}(\epsilon) \tag{1.10}$$

where $\hat{F}_{ab}(\epsilon)$ is the empirical distribution of ϵ obtained from the data (in this case equal to $1/n$, for each subject).

Using the product-limit estimator (Kaplan & Meier, 1958) of a distribution function with right censored data, Miller (1976) defined the Kaplan-Meier Least Squares estimators (KMLS), say $\hat{\alpha}^{KM}$ and $\hat{\beta}^{KM}$, as the values a and b that minimize

$$\int e^2 \, d\hat{F}_{ab}^{KM}(e) \tag{1.11}$$

where $\hat{F}_{ab}^{KM}(e)$ is the product-limit estimate of F (the distribution function of ϵ) computed with the uncensored and the censored residuals defined as $e_i = z_i - a - bx_i$, $i = 1, \cdots, n$. That is,

$$\hat{F}_{ab}^{KM}(e) = 1 - \prod_{i;e(i) \leq e} \left(\frac{n-i}{n-i+1} \right)^{\delta_i}$$

where $e_{(1)}, \cdots, e_{(n)}$ are the residuals in increasing order.

For fixed a and b the integral (1.11) is the weighted sum of squares given by

$$\frac{1}{n} \sum_{i \in U} w_i(a,b) \, (y_i - a - bx_i)^2 \tag{1.12}$$

where $i \in U$ denotes the subsample of uncensored observations and $w_i(a,b)$ is the weight assigned to $e_i = y_i - a - bx_i$ by the product-limit estimate applied to $\{e_i, \; i = 1, \cdots, n\}$. The estimators $\hat{\alpha}^{KM}$ and $\hat{\beta}^{KM}$ are obtained by minimizing (1.12).

Even though this method of estimation for the regression coefficients seems a good suggestion it has at least two difficulties. On one hand, it is very hard to study its asymptotic properties analytically. Moreover, it is possible to find examples where for some censoring patterns, $\hat{\alpha}^{KM}$ and $\hat{\beta}^{KM}$ do not converge to the true α and β as $n \to \infty$. A sufficient condition to guarantee the asymptotic consistency of $\hat{\alpha}^{KM}$ and $\hat{\beta}^{KM}$ is that

$$G_x(c + \beta x) = G_0(c) \tag{1.13}$$

where G_x is the distribution function for the censoring variables C on the value x of the independent variable. On the other hand, although the method can be generalized to multiple regression, it is computationally difficult to obtain the Kaplan-Meier Least Squares with more than one regressor. These two reasons motivate a modification of the KMLS that is also introduced by Miller (1976), briefly described as follows. From the LS estimates computed for the uncensored observations only, an iterative procedure may be applied in order to find $\hat{\beta}$ which satisfies $\hat{\beta} = \Phi(\hat{\beta})$, where

$$\Phi(\hat{\beta}) = \frac{\sum\limits_{i \in U} w_i^\star(0,\hat{\beta})\, y_i\, (x_i - \overline{x}^\star)}{\sum\limits_{i \in U} w_i^\star(0,\hat{\beta})\, (x_i - \overline{x}^\star)^2}$$

and $w_i^\star(0,\hat{\beta}) = w_i(0,\hat{\beta}) / \sum\limits_{i \in U} w_i(0,\hat{\beta})$ are the normalized weights that sum to one and, $\overline{x}^\star$ and $\overline{y}^\star$ denote the weighted averages computed with $w_i(0,\hat{\beta})$ for y_i and x_i, respectively. However, Miller (1976) points out that this procedure does not always converge.

1.2.2 Main result: Buckley & James (1979)

One of the most studied modifications of LS for a censored dependent variable has been proposed by Buckley & James (1979).

Using simulated data, Buckley & James (1979) viewed that under the condition (1.13), the iterative estimate of Miller (1976) and the LS estimate for the slope based only on the uncensored observations perform similarly. In that way they proposed an estimate that in some sense tries to relax this assumption. The method is also restricted to the simple linear model, $y_i = \alpha + x_i\beta + \epsilon_i$, and it is based on the normal equations instead of the sum of squared residuals.

Assuming right censored data with $c_1, \cdots, c_n$ known constants, consider a new dependent variable $Y^\star = \delta Y + (1-\delta)\mathrm{E}(Y/Y > C;\, x)$ which satisfies $\mathrm{E}(Y^\star) = \alpha + x\beta$. Then, the usual estimate of β is computed from the normal equation

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i^\star - x_i\beta) = 0. \tag{1.14}$$

However, due to censoring, $\mathrm{E}(Y/Y > C;\, x)$ is unknown and therefore $\{y_i^\star,\ i = 1, \cdots, n\}$ is also a set of unknown values. At this point a self-consistency approach is used (Efron, 1967), which suggests estimating $\mathrm{E}(Y/Y > C;\, x)$ from $\hat{F}_{ab}^{KM}$, the Kaplan-Meier estimator of F based on the residuals $e_i = z_i - a - bx_i$, where a and b are the LS estimates computed using uncensored data only. That is, censored observations are replaced by,

$$\hat{y}_i^\star(b) = bx_i + \sum_{j \in U} w_{ij}(0,b)\, (y_j - x_j b) \tag{1.15}$$

where

$$w_{ij}(0,b) = \begin{cases} \frac{v_j(b)}{(1-\hat{F}^{KM}_{0,b}(c_i - x_i b))} & \text{if } e_i(0,b) < e_j(0,b) \\ \\ 0 & \text{otherwise} \end{cases}$$

and $v_j(b)$ is the probability mass assigned by $\hat{F}^{KM}_{ab}$ to the uncensored residual $e_j(a,b)$.

Then, the proposed estimator of $\hat{\beta}$ is the limit of the following iterative procedure:

$$\hat{\beta}^{(k+1)} = \frac{\left(\sum_{i \in U} y_i \, (x_i - \overline{x}) + \sum_{i \in C} \hat{y}^{\star}_i(\hat{\beta}^{(k)}) \, (x_i - \overline{x})\right)}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \quad k = 1, 2, \cdots \qquad (1.16)$$

where $\sum_{i \in C}$ denotes summation over censored observations only. After convergence has been reached and $\hat{\beta}$ has been obtained, the intercept can be computed, as

$$\hat{\alpha} = n^{-1} \left(\sum_{i \in U} y_i + \sum_{i \in C} \hat{y}^{\star}_i(\hat{\beta}) \right) - \overline{x}\hat{\beta}.$$

The method just described will provide approximate solutions because it does not ensure to achieve convergence and it may finish oscillating between two values. However we note that the final values are closer approximations to the solution than using the method proposed by Miller (1976). The properties of the estimator are proved by Buckley & James (1979) in a heuristic way, but the displayed simulations tend to support their arguments. The above approach was the motivation for the following extensions of this extensively studied method.

Miller & Halpern (1982) point out the satisfactory performance of the Buckley & James estimator in some simulations and empirical studies. In particular, Miller & Halpern (1982) compared the performance of this estimate with two other proposed modifications of LS (Miller, 1979 and Koul, Susarla & Van Ryzin 1981) in order to accommodate censoring data.[3] From these results they concluded that the estimator defined in Buckley & James (1979) is the best method because they found that the other procedures had methodological weaknesses. Also Heller & Simonoff (1990) showed using simulations that the Buckley & James estimator was preferred in

[3] Koul, Susarla & Van Ryzin (1981) proposed another procedure for estimating the unknown coefficients of a CLM. The method is based on the normal equations where responses are replaced by estimates related to the distribution function of the censoring variable.

terms of consistency compared to Miller (1976) and Koul, Susarla & Van Ryzin (1980).

James & Smith (1984) showed that under some regularity conditions which avoid restrictions on the censoring patterns, the estimated slope proposed by Buckley & James (1979) is weakly consistent. They also point out that in practice the estimator of the intercept tends to be biased downwards.

The paper due to Ritov (1990) gives asymptotic properties of an estimator close to the Buckley & James (1979). This is a theoretical paper using the Counting Processes approach (see, e.g. Andersen et al. 1993). It also proves the asymptotic equivalence between the proposed estimate and the one suggested almost simultaneously by Tsiatis (1990).

Lai & Ying (1991) proved large sample properties for the slight modification of the Buckley & James estimator. They get around the problems caused by the instability of the upper tail of the Kaplan-Meier estimate of the distribution F (see Efron (1967) and Gill (1980) for properties of the Kaplan-Meier estimate) using a weighted function. Introducing their modified Kaplan-Meier in (1.15), they obtain new estimates for the censored responses that can be used in (1.16) to compute the updated estimator. The consistency and the asymptotic normality of the limiting estimator is proved under certain regularity conditions. To this end Lai & Ying (1991) used the methodology of stochastic integrals (see Andersen, Borgan, Gill & Keiding, 1993). Moreover they also extend the estimator as well as its properties to the multiple regression model with random right censoring.

Another work related with the Buckley & James estimator is due to Schneider & Weissfeld (1986). A new estimator for the variance of the error term based on the uncensored and censored observations is proposed. Schneider & Weissfeld (1986) introduce an alternative way of replacing censored observations that differs from (1.15).

1.2.3 Suggested procedure: Schneider & Weissfeld (1986)

The procedure described in this section is a modification of Buckley & James (1979) in the way the censored observations are used. Moreover, it also applies to multiple regression. The estimator proposed in Schneider & Weissfeld (1986) is based on the random variable

$$Y^{\star} = \delta Y + (1-\delta)\mathrm{E}\,(Y \mid Y > C \;;\; \mathbf{x}'\boldsymbol{\beta}) \tag{1.17}$$

for which $\mathrm{E}(Y^{\star}) = \mathrm{E}(Y)$. From here, the estimate of regression parameters based on the LS is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{\star}$ where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)'$ and $\mathbf{y}^{\star} = (y_1^{\star}, \cdots, y_n^{\star})'$. However, the conditional expectation $\mathrm{E}\,(Y \mid Y > C \;;\; \mathbf{X}'\boldsymbol{\beta})$ that appears in (1.17) will be unknown if no distribution function is specified for the residual term ϵ. Motivated by this point, Schneider & Weissfeld (1986) developed an estimate of this conditional expectation without imposing any assumptions about the parametric form of the residual's probability density function. This method is, as they also said, in the spirit of the EM algorithm (see Dempster, Laird & Rubin, 1977). It is an iterative procedure, initialized with the LS estimate computed from the observed data $\{(z_i, \mathbf{x}_i),\ i = 1, \cdots, n\}$, and then proceeding with the following two steps:

1. Given $\hat{\boldsymbol{\beta}}^{(k)}$, the conditional expectation $\mathrm{E}_{(y_i,\boldsymbol{\beta})} = \mathrm{E}\,(Y_i \mid Y_i > C_i \;;\; \mathbf{x}_i'\boldsymbol{\beta}) = \mathrm{E}(\mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \mid \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i > C_i \;;\; \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta} + \mathrm{E}(\epsilon_i \mid \epsilon_i > C_i - \mathbf{x}_i'\boldsymbol{\beta} \;;\; \mathbf{x}_i'\boldsymbol{\beta})$ is estimated as

$$\hat{\mathrm{E}}_{(y_i,\hat{\boldsymbol{\beta}}^{(k)})} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(k)} + \frac{\sum\limits_{j \in E_i^{(k)}} e_j}{M_i^{(k)}} \tag{1.18}$$

   where the second term on the right is the mean of the residuals $e_j = z_j - \mathbf{x}_j'\hat{\boldsymbol{\beta}}^{(k)}$ larger than residual $e_i = c_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(k)}$ corresponding to the censored observation i. By $M_i^{(k)}$ we denote the cardinal of the set by $E_i^{(k)} = \{j : e_j > e_i\}$.

2. From (1.18), the updated value of $\hat{\boldsymbol{\beta}}^{(k)}$ is computed using standard LS with dependent variable $y_i^{\star}(\hat{\boldsymbol{\beta}}^{(k)}) = \delta_i y_i + (1-\delta_i)\hat{\mathrm{E}}_{(y_i,\hat{\boldsymbol{\beta}}^{(k)})}$ instead of the observed z_i, that is

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\;\mathbf{y}^{\star}(\hat{\boldsymbol{\beta}}^{(k)}) \tag{1.19}$$

   with $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)'$ and $\mathbf{y}^{\star}(\hat{\boldsymbol{\beta}}^{(k)}) = \Big(y_1^{\star}(\hat{\boldsymbol{\beta}}^{(k)}), \cdots, y_n^{\star}(\hat{\boldsymbol{\beta}}^{(k)})\Big)'$.

Steps 1 and 2 are iterated until convergence is achieved. The existence of the limiting value $\hat{\boldsymbol{\beta}}$ as well as its consistency are only shown using numerical simulations by Schneider & Weissfeld (1986).

In our opinion for practical applications, there are three important points for suggesting the estimator due to Schneider & Weissfeld (1986) rather than the estimator proposed by Buckley & James (1979):

- The iterative procedure proposed by Schneider & Weissfeld (1986) incorporates both the censored and the uncensored residuals in contrast with the Buckley & James method which uses only uncensored residuals.
- The ease of implementation
- The Schneider & Weissfeld estimator still remains valid for the multiple regression model.

1.2.4 Other types of censoring

The results described until now have assumed right-censored data. However there are some results for estimating CLM for other types of censoring. Here we emphasize two recent papers suggesting methodologies based on LS taking into account interval censoring and doubly censoring data (see, e.g. Lawless, 1985 or Kalbfleisch & Prentice, 1980 for dealing with the issue of censoring).

Rabinowitz, Tsiatis & Aragon (1995) use regression theory on estimating the coefficients of a linear model with an interval-censored response variable. Indeed, they assumed that the log-transformed survival times are equal to a linear combination of the covariates plus independent and identically distributed residuals. Then a procedure for estimating regression coefficients based on score statistics is described. Moreover, the approach is also close to the one due to Buckley & James (1979).

On the other hand Zhang & Li (1996) considered a linear regression model with a doubly censored response. They proposed a methodology for estimating regression parameters analogous to the one suggested by Ritov (1990) for the case of right-censoring. Some sufficient conditions for the asymptotic consistency and normality of the estimators are also given.

Chapter 2

Measurement Errors in Linear Models

The issue of measurement error in variables is based on the unobservability of the true values of the variables. It may be studied from several points of view. We focus on the context of linear regression models where some of the explanatory variables contain measurement error.

Let's assume a standard linear model between a response variable Y and a vector of explanatory variables $\mathbf{X}^\star$, such that

$$Y = \mathbf{X}^{\star\prime}\boldsymbol{\beta} + W \tag{2.1}$$

where W is a random variable usually with zero expectation and finite variance represented by σ_w^2.

In order to introduce measurement errors on covariates we also assume that instead of $\mathbf{X}^\star$, we observe the variables $\mathbf{X}$ which are linearly related, that is

$$\mathbf{X} = \mathbf{X}^\star + \mathbf{U}$$

where $\mathbf{U}$ is a vector of random variables usually normal distributed with zero mean and covariance matrix Σ_{uu}. Hence, the relationship between the response variable Y and the observed variables $\mathbf{X}$ is given by

$$Y = \mathbf{X}'\boldsymbol{\gamma} + \epsilon \tag{2.2}$$

where $\boldsymbol{\gamma}$ are the regression coefficients we are able to estimate by LS methods. We note there is a relationship between parameters $\boldsymbol{\gamma}$ and the parameters $\boldsymbol{\beta}$ of

(2.1). In order to see how these parameters are related we follow the arguments due to Cochran (1968). We restrict our attention to the case of one non-constant covariate and we will compute the covariance between Y and X, $cov(Y, X)$, from both models. When there is only one covariate the models (2.1) and (2.2) may be written respectively as,

$$Y = \beta_0 + X^\star \beta + W \tag{2.3}$$

and

$$Y = \gamma_0 + X\gamma + \epsilon. \tag{2.4}$$

From model (2.4), we have

$$cov(Y, X) = cov(\gamma_0 + X\gamma + \epsilon, X) = \gamma \ var(X) + cov(\epsilon, X) \tag{2.5}$$

where $var(X)$ denotes the variance of variable X.

On the other hand, from model (2.3) and using $X = X^\star + U$, we have

$$\begin{aligned} cov(Y, X) &= cov(\beta_0 + X^\star\beta + W, X) = cov(\beta_0 + X^\star\beta + W, X^\star + U) \\ &= \beta \ var(X^\star) + \beta \ cov(X^\star, U) + cov(W, X^\star) + cov(W, U). \end{aligned} \tag{2.6}$$

Now if we set (2.5) equal to (2.6), we can isolate γ as follows

$$\gamma = \frac{\beta \ var(X^\star) + \beta \ cov(X^\star, U) + cov(W, X^\star) + cov(W, U) - cov(\epsilon, X)}{var(X)} \tag{2.7}$$

where we have made the usual assumption for linear regression models, namely that W is independent of $X^\star$ and ϵ is independent of X, and where we have computed $var(X^\star)$ and $cov(U, X^\star)$ in terms of X and U. This leads to the well known expression

$$\gamma = \frac{\beta \ [var(X) - cov(U, X)] + cov(W, U)}{var(X)}. \tag{2.8}$$

From these results it is clear that estimation methods taking into account the presence of measurement errors are required in order to obtain unbiased estimators of the regression parameters of the true model (2.1). Such methods have been studied extensively and there are two extensive reviews of the literature due to Fuller (1987) and Carrol, Ruppert & Stefanski (1995) dealing with linear and non-linear models, respectively.

In this chapter we introduce in Section 2.1 a brief review of the estimation methods in linear models with measurement errors on covariates. Section 2.2 presents previous work about survival analysis and measurement errors.

2.1 Estimation in linear measurement error model

Here we introduce the methods for estimating the coefficients of linear regression models when some explanatory variables are measured with error.

The goal is to obtain unbiased estimates of the regression coefficients β in model (2.1) using the observed linear model (2.2) and equation (2.8). We restrict attention to the simple regression model. Assuming that random variables W and U are independent, then

$$\gamma = \beta \, \frac{var(X) - cov(U, X)}{var(X)}. \tag{2.9}$$

where, assuming that $X^\star$ and U are independent random variables, we have $cov(U, X) = \sigma_u^2$, so that

$$\gamma = \beta \, \frac{var(X) - \sigma_u^2}{var(X)} = \beta \, \frac{var(X^\star)}{var(X)} = \beta \; k \tag{2.10}$$

where

$$k = \frac{var(X^\star)}{var(X)}$$

is known as the *reliability ratio*. Note that k ranges from 0 to 1 and it represents the amount of measurement error in the following sense: $k = 1$ means $var(X^\star) = var(X)$ and indicates no measurement error on X; while small values of k mean large $var(X)$ compare with $var(X^\star)$ and therefore presence of measurement error on X. Hence, the LS estimator $\hat{\gamma}$ of γ in (2.2) is a biased estimator of β. However, if k is known, an unbiased estimator of β may be obtained as

$$\hat{\beta} = \hat{\gamma} \; k^{-1}. \tag{2.11}$$

We note that this estimate also applies in the case of the multiple regression model (see Fuller 1987).

The presence of measurement error on covariates involves identifiability problems in the model (2.1). Indeed, as Fuller (1987) points out, in order to obtain consistent estimators for parameter β some specification of additional information with respect to the distribution of (Y, X) is needed. The assumption of a known reliability ratio allows us to identify the model, but other conditions may be used. For instance, to use the method of moments to estimate β a frequent assumption is to consider that the variance of measurement error, that is σ_u^2, is known. Thus if we consider that the model is given by

$$\begin{aligned} Y &= \beta_0 + X^\star\beta + W \\ X &= X^\star + U \end{aligned} \tag{2.12}$$

then the method of moments gives

$$\begin{aligned} \hat{\beta} &= (m_{xx} - \sigma_u^2)^{-1}\, m_{xy} \\ (\hat{\sigma}^2_{x^\star}, \hat{\sigma}^2_w) &= (m_{xx} - \sigma_u^2, m_{yy} - \hat{\beta} m_{xy}) \\ (\hat{\mu}_x, \hat{\beta}_0) &= (\overline{x}, \overline{y} - \hat{\beta}\overline{x}) \end{aligned} \tag{2.13}$$

where knowledge of σ_u^2 allows us to construct a one-to-one mapping from the minimal sufficient statistic to the vector of unknown parameters $(\hat{\mu}_x, \hat{\sigma}^2_{x^\star}, \hat{\beta}_0, \hat{\beta}, \hat{\sigma}^2_w)$. Note that under normality the method of maximum likelihood is equivalent to the method of moments.

The estimates already presented may also be generalized to multiple regression and similar expression are still valid without the hypothesis of normality. For the case of non-linear models see Carroll et al. (1995).

Even though we have presented a restrictive case of the estimation in linear models with measurement errors on covariates, we are not going to introduce more general cases because we do not want to concentrate too much on this issue.

2.2 Survival analysis and measurement errors

Even though the issue of measurement errors has been largely developed in regression analysis, as far as we know, in survival analysis there is little work on this topic. Moreover, the methodologies already proposed are focused on the proportional hazards model (Cox, 1972) and on the additive hazards model (e.g. see Breslow & Day, 1987).

The initial work is due to Prentice (1982). Assuming a proportional hazards model defined in (1.3) for the true data $(t, X^\star)$:

$$\lambda(t; X^\star) = \lambda_0(t)\ \exp(X^\star\beta) \tag{2.14}$$

where $\lambda_0(\cdot) \geq 0$ is the baseline hazard function, it examines the possible effect of the measurement error on estimated relative risks. It also develops an improved risk estimator under certain assumptions about the error distribution.

If X denotes the observed covariate, it is proposed to estimate the parameters β and $\lambda_0(\cdot)$ in (2.14) coming from inference on the hazard function $\lambda(t; X)$, which is amenable to direct estimation. The basic assumption that allows $\lambda(t; X^\star)$ and $\lambda(t; X)$ to be related asserts a conditional independence, given $X^\star$, of failure rate at t and X; that is,

$$\lambda(t; x^\star, x) = \lambda(t; x^\star) \tag{2.15}$$

From this assumption,

$$\lambda(t; X) = \mathrm{E}[\lambda(t; X^\star, X) \mid T \geq t,\ X] = \mathrm{E}[\lambda(t; X^\star) \mid T \geq t,\ X] \tag{2.16}$$

where using (2.14) we obtain:

$$\lambda(t; X) = \lambda_0(t)\ \mathrm{E}[\exp(X^{\star\prime}\beta) \mid T \geq t,\ X] \tag{2.17}$$

which still assumes proportionality of the hazards. However, note that the presence of $\{T \geq t\}$ in the conditioning event will usually imply some dependence of the relative risk function $\mathrm{E}[\exp(x^{\star\prime}\beta) \mid T \geq t,\ X]$ on the baseline hazard function $\lambda_0(t)$.

In this paper a partial likelihood for (2.17) is derived using the argument of Cox (1975) after some assumptions on the censoring mechanism of the data are made. In that way the partial likelihood function can be used in a standard manner for estimating β just specifying the error distribution $f(x^\star \mid T \geq t,\ x)$. Two cases are considered: first when the error distribution does not depend on β and $\lambda_0(t)$, and second, when errors are normally distributed.

Nakamura (1992) also proposes a method for estimating a proportional hazards model under the presence of measurement error on covariates. He establishes the relationship $X = X^\star + U$ where $X^\star$ is the true covariate and U are random variables with zero mean and covariance matrix Σ_{uu}. The proposed procedure of estimation for parameters β is based on the score function obtained from the partial likelihood. The log of the partial likelihood defined in (2.14) is given by

$$l(\beta, t, X^{\star}) = \sum_{i} X^{\star\prime}_{i}\beta - \ln S_i(t;\ \beta, X^{\star}) \tag{2.18}$$

where $S_i(t;\ \beta, X^{\star})$ is the survival function of T. Hence, the score function is

$$V(\beta, t, X^{\star}) = \sum_{i} X^{\star\prime}_{i} - \frac{S_i'(t;\ \beta, X^{\star})}{S_i(t;\ \beta, X^{\star})}. \tag{2.19}$$

Even though it is well known that estimates obtained from (2.19) are unbiased, when X is used instead of $X^{\star}$ the resulting estimate β_x is asymptotically biased. From here, Nakamura (1992) proposes a correction of this bias using a function $V^{\star}(\beta, t, X)$ whose expectation with respect to U given t and $X^{\star}$ coincides with $V(\beta, t, X^{\star})$. The function $V^{\star}(\beta, t, X)$ is called the corrected score function and β such that $V^{\star}(\beta, t, X) = 0$ is a corrected estimate. The procedure is based on the hypothesis that the covariance matrix Σ_{uu} is known. The properties of the proposed estimate are shown numerically.

A more recent paper about measurement errors on covariates in survival analysis is due to Kulick & Lin (1998). They establish the additive hazard function given by

$$\lambda(t; X^{\star}) = \lambda_0(t) + X^{\star\prime}\beta \tag{2.20}$$

where $\lambda_0(\cdot)$ also remains unspecified. Moreover, they also model the observed covariate X as a linear function of the true covariate $X^{\star}$ plus a random error and only impose moment conditions on the measurement error distribution. The error variance may depend on the true covariate through an arbitrary linear or quadratic variance function.

In that context, Kulick & Lin (1998) develop a class of asymptotically unbiased estimating functions for the regression coefficients β. They obtain these estimating funcions from an existing pseudo-score function without measurement error by incorporating a bias-correction term. The resulting estimator is proved to be consistent and asymptotically normal.

Finally another paper related with the issue of measurement errors in survival analysis is by Holt, McDonals & Skinner (1991). In this paper measurement error is contained in the response variable of a linear regression model. However, the covariates are considered well observed.

Chapter 3

Censored Linear Model with Measurement Errors on Covariates

In this chapter we are concerned with the estimation of the regression parameters in a censored linear model when the covariates are measured with error.

A frequent problem in statistics is to obtain the estimates of the regression parameters, that is, to assess the effects of a set of covariates on a response variable. In survival analysis, the presence of censoring requires specialized methods for estimating unknown parameters. For linear models, we emphasize the procedures which are modifications of Least Squares (LS) procedures in order to accommodate censored values of the response (see, e.g. Miller 1976, Buckley & James 1979, Koul, Susarla & Van Ryzin 1981 and Schneider & Weissfeld 1986). A common assumption underlying these methods is that covariates are measured in a precise way. However, many characteristics observed in practical applications are difficult to be measured exactly and the true value is contaminated with measurement error.

The study of linear models with explanatory variables containing measurement error is a topic of interest since the past century (see, e.g. Adcock, 1877, 1878 and Kummell, 1879). Even though there is a wide range of methodologies for estimating the regression parameters taking into account measurement errors (see Fuller, 1987 or Carrol, Ruppert and Stefanski, 1995), all of them are based on the values for the dependent variable when no censoring is present.

We propose a method for estimating censored linear models with measurement errors on covariates based on a combined procedure that merges known results from measurement error theory together with methods for censored data. We describe a two-step approach for obtaining consistent estimates of the regression parameters. In the first step we compute linear predictions for the censored values based on methods for estimating linear models with censored response. In the second step we compute the estimators of the regression coefficients based on estimators obtained in the previous step.

In Section 3.1 we describe the model. Specifically, we consider a log-linear model with a set of covariates and a linear relationship that defines the presence of measurement error on covariates. In fact we note that this is an accelerated failure time model (see Section 1.1 in Chapter 1) with errors in variables, which from now on we will refer to as AFTME.

The motivation of a procedure for estimating regression parameters that takes measurement error into account is given in Section 3.2. Indeed, a Monte Carlo simulation shows the effects of ignoring measurement error on covariates when standard procedures of estimation for survival models are used.

Section 3.3 develops the two-step estimator for the regression coefficients of the AFTME already described. We remark that the proposed estimator is easy to implement with real data because it can be obtained using standard statistical software. We also describe the Bootstrap method for computing standard errors.

Numerical studies are reported in Section 3.4. We use simulated data in order to show the performance of the proposed estimator. For this we have carried out several Monte Carlo studies where we varied sample sizes, levels of measurement error and proportion of censored observation.

Finally in Section 3.5 we propose some extensions of the two-step estimator.

3.1 The model

We consider a non-negative and continuous random variable T (this is time elapsed in a certain state) and a set of explanatory variables $\{X_1^\star, \cdots, X_p^\star\}$, also called covariates.

Let $Y = \log T$ be the log-transformation of the true duration T. Consider $y_1, \cdots, y_n$, independent realizations of Y such that y_i is related to the vector of covariates $\mathbf{x}_\mathbf{i}^\star$ as

$$y_i = \mathbf{x}_i^{\star\prime}\boldsymbol{\beta} + w_i, \quad i = 1, \cdots, n \tag{3.1}$$

where $\boldsymbol{\beta}$ is the vector of unknown parameters and $w_1, \cdots, w_n$ are i.i.d. realizations of a disturbance term W of variance σ_w^2 and mean not necessarily zero. We assume that W and $X_j^\star, j = 1, \cdots, p$ are independent random variables.

As usual in survival analysis, we allow the presence of censoring. In particular we assume a right censorship model (see Kalbfleisch & Prentice, 1980). That is, our observable duration for the ith individual consists of the values

$$z_i = \min\{y_i, c_i\}, \quad i = 1, \cdots, n \tag{3.2}$$

together with the indicator of censoring

$$\delta_i = \mathbf{1}_{\{y_i \leq c_i\}}, \quad i = 1, \cdots, n$$

where $c_1, \cdots, c_n$ are independent realizations of a random variable C (in this case c_i represents the log-transformed censored time for individual i). Here we assume that the censoring mechanism is not informative (see Tsiatis, 1975). The indicator of censoring δ_i equals 0 for the censored observations and 1 when the true duration is observed.

The model defined by (3.1) and (3.2) stated for analyzing data of the form $\{(z_i, \delta_i, \mathbf{x}_i^{\star\prime}), \quad i = 1, 2, \cdots, n\}$ is usually known as the censored linear model (see, e.g. Breiman, Tsur & Zemel, 1993). From now we will refer to it as CLM.

Here we consider a CLM including one more assumption in order to accommodate the possible presence of measurement errors in the covariates. Thus we assume that variables $X_j^\star$ may be unobservable, with only the observed covariates X_j, $j = 1, \cdots, p$, being available. The relationship between the observed covariates $\mathbf{x}_i$ for the ith individual and the true value of the covariates $\mathbf{x}_i^\star$ is defined by the measurement error model:

$$\mathbf{x}_i = \mathbf{x}_i^\star + \mathbf{u}_i, \quad i = 1, \cdots, n \tag{3.3}$$

where $\mathbf{u}_1, \cdots, \mathbf{u}_n$ are i.i.d. realizations of the random vector $\mathbf{U} = (U_1, \cdots, U_p)'$ with zero mean and known covariance matrix Σ_{uu}.[1] We also assume that $\mathbf{U}$ is independent of $\mathbf{X}$ and W.

The main goal now is to estimate the regression coefficients in the model defined by (3.1), (3.2) and (3.3). The next sections introduce a sequential procedure with two

[1] This assumption could be relaxed, however some additional hypotheses on the error terms u_i are needed (see Fuller, 1987).

steps for obtaining consistent estimates of $\boldsymbol{\beta}$ when recorded data are $\{(z_i, \delta_i, \mathbf{x}_i'),\ i = 1, \cdots, n\}$. However, we shall first motivate the method we propose, by showing the effects of ignoring the presence of measurement error.

3.2 The effects of measurement error: Monte Carlo illustration

In this section we are going to show what happens when we use standard methods of estimation in survival analysis in the presence of measurement error in covariates.

Let $(z_i, \delta_i, \mathbf{x}_i^\star), i = 1, \cdots, n$ be the survival data we want to analyze. However, due to the difficulties of measuring covariates $X^\star$ there are only available variables X. Here we show the effects on the standard estimators of the regression coefficients, when X is used instead of $X^\star$. To this end, we generate data with covariates subject to measurement error. Then a model that ignores the presence of measurement error will be estimated using maximum likelihood, and the bias of the estimators will be assessed.

We consider a 2-dimensional vector of covariates $\mathbf{X}_i^\star = (X_{1i}^\star, X_{2i}^\star)'$ which are i.i.d. realizations of $\mathcal{N}(\mathbf{0}, \mathrm{diag}(\sigma_{x_1}^2, \sigma_{x_2}^2))$. The observed covariates are taken to be $X_{1i} = X_{1i}^\star + \epsilon_{1i}$ and $X_{2i} = X_{2i}^\star$, where the measurement errors $\{\epsilon_1\}$ are i.i.d. normally distributed of zero mean and variance $\sigma_{\epsilon_1}^2$. The vector of observed covariates for the ith individual is augmented with a constant of 1, that is $\mathbf{x}_i = (1, x_{1i}, x_{2i})'$.

We first consider uncensored duration times t_i simulated as independent observations from T_i, a random variable with a Weibull distribution[2] with shape parameter $\alpha = 2$ and scale parameter $\gamma_i = \exp(\mathbf{x}_i^{\star\prime}\boldsymbol{\beta}^\star)$, where $\mathbf{x}_i^\star = (1, x_{1i}^\star, x_{2i}^\star)'$ and $\boldsymbol{\beta}^\star = (3, 1, 1)'$. The values t_i are censored according to a Type II censoring mechanism;[3] that is, the observed duration for individual i is $z_i = \delta_i t_i + (1 - \delta_i) t_{(m)}$. We record also the censoring indicator δ_i ($\delta_i = 1$ when t_i is uncensored and 0 otherwise).

The Monte Carlo study considers variation on the sample size n and the variance $\sigma_{\epsilon_1}^2$ of the measurement error variable ϵ_1. The sample size n takes the values 100,

[2] Assuming T_i to have a Weibull distribution with parameters α and $\gamma_i = \exp(\mathbf{x}_i^{\star\prime}\boldsymbol{\beta}^\star)$, a linear relationship like (3.1) may be established between $\log t_i$ and $\mathbf{x}_i^\star$ where the regression parameters $\boldsymbol{\beta}$ are related through the Weibull parameters $\boldsymbol{\beta}^\star$ and α by the expression: $\boldsymbol{\beta} = -\boldsymbol{\beta}^\star/\alpha$.

[3] After sorting the survival times in increasing order, $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ and for a given value $m \leq n$, all $t_{(r)}$ with $r > m$ are censored to be equal to $t_{(m)}$.

500 and 1000, while $\sigma^2_{\epsilon_1}$ is varied so that the reliability ratio $k = \sigma^2_{x_1^\star} / \left(\sigma^2_{x_1^\star} + \sigma^2_{\epsilon_1}\right)$ ranges from $k = 1$ (no measurement error) to $k = 0.2$ (80% of the variance of X_1 is due to measurement error). The percentage of censoring c is fixed at $c = 20\%$. Each Monte Carlo run was based on 500 replications.

Table 3.1 shows the empirical bias of the estimator of $\boldsymbol{\beta}^\star$ obtained for the different values of k and n. From this table we see that the bias of the estimators of the components of $\boldsymbol{\beta}^\star$ increases with the decrease of the reliability ratio. That is, as the amount of measurement error increases, the bias of the usual estimators of $\boldsymbol{\beta}^\star$ also increases. This behavior is observed for the three sample sizes. Note that even though only x_1 is affected by measurement error, the estimator of $\beta_2^\star$ is also affected by bias.

Table 3.1: Monte Carlo results: Bias of the estimators when ignoring the presence of measurement error

k	$\hat{\beta}_0^\star$			$\hat{\beta}_1^\star$			$\hat{\beta}_2^\star$		
	100	500	1000	100	500	1000	100	500	1000
1	.06	.03	.01	.01	.01	.00	.01	.00	.01
.8	−.24	−.33	−.33	−.25	−.28	−.28	−.06	−.10	−.10
.6	−.54	−.57	−.57	−.48	−.50	−.50	−.14	−.17	−.17
.4	−.69	−.78	−.78	−.67	−.69	−.69	−.20	−.23	−.22
.2	−.87	−.90	−.96	−.85	−.85	−.86	−. 25	−.27	−.27

NOTE: Percentage of censoring $c = 20\%$. Population value of parameters $\beta_0^\star = 3, \beta_1^\star = 1, \beta_2^\star = 1$.

These results show that under the presence of measurement error on covariates, regression parameters have to be estimated using procedures that accommodate the possible errors. Thus, in the next section we describe a procedure that gives consistent estimates of regression coefficients for AFT models with covariates possibly contaminated with error.

3.3 The two-step estimator

In this section we develop a two-step estimator that gives unbiased estimates of the regression coefficients of the model defined by (3.1), (3.2) and (3.3). The method modifies the standard procedures of estimation for linear measurement error models in order to account for censoring.

The first step of the method takes into account the presence of censoring in the data. We take the model given by equations (3.1) and (3.2) with $\mathbf{x}_i$ instead of $\mathbf{x}_i^\star$, that is, ignoring the measurement error. Thus we have a CLM and the methodologies described in Chapter 1 may be applied. As a result of this step we obtain a consistent estimator of the matrix of mean squares and products of Y and the observed covariates $X_j, j = 1, \cdots, p$, say $\hat{\kappa}_{xy}$.

Second step consists in estimating a linear measurement error model defined by (3.1) and (3.3). Here, a consistent estimator of the regression coefficients is obtained using the methods for estimating error-in-variables models. However we point out that the method is slightly modified because it involves the covariance matrix of Y and $\mathbf{X}_j$ and we propose the use of the matrix $\hat{\kappa}_{xy}$ computed in the first step.

Once both steps have been performed, unbiased estimators of the regression coefficient in model (3.1) are obtained.

3.3.1 Estimated censored values: Step 1

In this step we ignore the presence of measurement error in the sense that we state the survival model defined by

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i \\ z_i &= \min\{y_i, c_i\} \\ \delta_i &= \mathbf{1}_{\{y_i \le c_i\}} \end{aligned} \tag{3.4}$$

where $\mathbf{x}_i$ is the vector of explanatory variables for individual i (here we are using the observed values of them) and $\boldsymbol{\gamma}$ are the regression coeficients. We note that the change of notation for the parameters is because, as pointed out in Chapter 2, in the observed model (3.4) the parameters are not the same as those in the true model defined in (3.1) and (3.2).

We note that (3.4) is a linear model with a censored response, therefore it is a CLM like the model defined in Chapter 1. Thus, consistent estimates of parameter $\boldsymbol{\gamma}$ may be obtained using the methodologies described in the same chapter. Based on the properties of the procedure, we propose the method due to Buckley & James (1979). However, as it noted by Lai & Ying (1991), this estimator presents unstability problems for large values of the response variable since it is based on the Kaplan-Meier estimator (see Breslow & Crowley, 1974). In order to avoid this problem, for the ease of implementation in practical situations including multiple regression, we suggest using the modification proposed by Schneider & Weissfeld (1986) (see page 17 for details about this algorithm). Thus, the estimator of $\boldsymbol{\gamma}$, say $\hat{\boldsymbol{\gamma}}$, is obtained applying this method to the model defined in (3.4). However, we point out the use of observed covariates instead of the true values, therefore we have to check the properties of this modified estimator of Schneider & Weissfeld (1986). In Section 3.4 we analyze the performance of this estimator and the empirical results show that $\hat{\boldsymbol{\gamma}} \xrightarrow{P} \boldsymbol{\gamma}$ even though $\hat{\boldsymbol{\gamma}}$ is a biased estimator of $\boldsymbol{\beta}$ due to the use of $\mathbf{X}$ instead of $\mathbf{X}^{\star}$.

As we describe above, in this step we want to deal with the censoring of the response variable. For this reason, we are not interested in the estimator $\hat{\boldsymbol{\gamma}}$ but in computing the linear predictors, conditional on $\mathbf{x}_i$, for z_i coming from model (3.4). Indeed, from $\hat{\boldsymbol{\gamma}}$ may be obtained $\hat{z}_i = \mathbf{x}_i'\hat{\boldsymbol{\gamma}}$, $i = 1, \cdots, n$. This leads to the following result for the values $(\hat{z}_1, \cdots, \hat{z}_n)$:

Result 1:

The $(\hat{z}_1, \cdots, \hat{z}_n)$ are "good" estimators of the censored response variable in the sense that

$$\hat{\boldsymbol{\kappa}}_{xy} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \hat{z}_i \tag{3.5}$$

is a consistent estimator of $\mathbf{k}_{xy} = \mathrm{E}(XY)$, where Y is the true response variable. That is,

$$\hat{\boldsymbol{\kappa}}_{xy} \xrightarrow{P} \mathbf{k}_{xy}.$$

□

This result arises because $\hat{\boldsymbol{\kappa}}_{xy} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \hat{z}_i = (n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i') \hat{\boldsymbol{\gamma}}$. From here, using the consistency of $\hat{\boldsymbol{\gamma}}$, we have $\hat{\boldsymbol{\kappa}}_{xy} \xrightarrow{P} \mathbf{k}_{xx}\boldsymbol{\gamma}$ where $\boldsymbol{\gamma} = \mathbf{k}_{xx}^{-1}\, \mathbf{k}_{xy}$. Thus, the consistency of $\hat{\boldsymbol{\kappa}}_{xy}$ is proved.

The usefulness of the estimator $\hat{\boldsymbol{\kappa}}_{xy}$ just defined is based on the following argument. If the response variable of a linear model is censored, for the observed z_i the cross products matrix $\mathbf{K}_{xz} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i z_i$ is not a consistent estimator of $\mathbf{k}_{xy}$. Indeed, $z = (z_1, \cdots, z_n)'$ contains a known proportion of censored values different from the true durations (i.e. $z_i = c_i$ for those subjects with a censored time). Here we note that the estimator is useful for any set of p covariates. Thus we have defined a consistent way of estimating the cross product matrix between a variable Y, possibly censored, and a set of explanatory variables $X_j, j = 1, \cdots, p$ when there is a linear relationship among them.

3.3.2 Errors-in-variables model: Step 2

In this step we compute the estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, using methodologies for estimating linear measurement error models described in Chapter 2. The proposed procedure is based on the estimator $\hat{\boldsymbol{\kappa}}_{xy}$ defined in step 1.

We consider the errors-in-variables model

$$\begin{aligned} y_i &= \mathbf{x}_i^{\star\prime}\boldsymbol{\beta} + w_i \\ \mathbf{x}_i &= \mathbf{x}_i^{\star} + \mathbf{u}_i. \end{aligned} \tag{3.6}$$

where the covariance matrix of $\mathbf{U} = (U_1, \cdots, U_p)$, denoted by Σ_{uu}, is known. Then for the standard case where y_i are observed for all $i = 1, \cdots, n$, a consistent estimator of $\boldsymbol{\beta}$ is defined as (see Fuller, 1987)

$$\hat{\boldsymbol{\beta}} = (\mathbf{K}_{xx} - \Sigma_{uu})^{-1}\,\mathbf{K}_{xy} \tag{3.7}$$

where $\mathbf{K}_{xx} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'$ and $\mathbf{K}_{xy} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i$. Here we note that $\mathbf{K}_{xx}$ and $\mathbf{K}_{xy}$ denote the matrix of the raw mean squares and products.

The consistency of (3.7) is proved with next results. Let $\mathbf{k}_{xx} = \mathrm{E}(\mathbf{XX})$ be the cross products matrix of $\mathbf{X}$ and let $\mathbf{k}_{xy} = \mathrm{E}(\mathbf{X}Y)$ be the cross products matrix of $\mathbf{X}$ and Y. Then $\mathbf{K}_{xx} \xrightarrow{P} \mathbf{k}_{xx}$ and $\mathbf{K}_{xy} \xrightarrow{P} \mathbf{k}_{xy}$. Moreover, for a real-valued function g, continuous at $(\mathbf{k}_{xx}, \mathbf{k}_{xy})$, we have (e.g. Rao, 1973)

$$g(\mathbf{K}_{xx},\ \mathbf{K}_{xy}) \xrightarrow{P} g(\mathbf{k}_{xx},\ \mathbf{k}_{xy}). \tag{3.8}$$

Now, assuming that $(\mathbf{K}_{xx} - \Sigma_{uu})$ is a non-singular matrix and taking $g(\mathbf{K}_{xx}, \mathbf{K}_{xy}) = (\mathbf{K}_{xx} - \Sigma_{uu})^{-1}\,\mathbf{K}_{xy}$ we have that

$$\hat{\boldsymbol{\beta}} = g(\mathbf{K}_{xx}, \mathbf{K}_{xy}). \tag{3.9}$$

Therefore using result (3.8) it can be proved that $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$. Moreover, note that property (3.8) allows us to define consistent estimators of $\boldsymbol{\beta}$ by simply replacing the $\mathbf{K}_{xy}$ used in (3.7) for any consistent estimator of $\mathbf{k}_{xy}$. Hence next result emerges.

Result 2:

The proposed estimator of $\boldsymbol{\beta}$ defined as

$$\hat{\boldsymbol{\beta}} = (\mathbf{K}_{xx} - \Sigma_{uu})^{-1} \, \hat{\boldsymbol{\kappa}}_{xy} \tag{3.10}$$

is a consistent estimator of $\boldsymbol{\beta}$, where $\hat{\boldsymbol{\kappa}}_{xy}$ is the estimator computed in step 1.

□

The consistency of (3.10) comes from property (3.8) taking $g(\mathbf{K}_{xx}, \hat{\boldsymbol{\kappa}}_{xy}) = (\mathbf{K}_{xx} - \Sigma_{uu})^{-1} \, \hat{\boldsymbol{\kappa}}_{xy}$.

3.3.3 The proposed procedure

The estimator emerging from steps 1 and 2 is a consistent estimator of the regression parameters of model (3.1), (3.2) and (3.3). It is called the two-step estimator.

The two-step estimator is a procedure easy to implement in practical applications. Even though the first stage is carried out without using standard methodology, our second step may be computed using standard software. In fact, the two-step estimator combines two methods of estimation already existing but used separately. On the one hand, the procedures of estimation for linear models with censoring and, on the other hand, the estimation methods for linear models with measurement error on covariates.

The available software for the second step and the ease of implementing the procedure of the first step, imply that the issue of measurement errors on covariates may be handled without a huge effort for some survival models. Indeed, the two-step estimator may be applied to our data as follows:

1. Let $(z_i, \mathbf{x}_i, \delta_i), i = 1, \cdots, n$ be the observed data where z_i is the observed duration time, $\mathbf{x}_i$ represents the vector of covariates and δ_i the indicator of censoring.

2. **Step 1**.

   i) Compute the LS estimates, say $\boldsymbol{\gamma}_0$, of model

   $$z_i = \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i.$$

   ii) Apply the iterative procedure of Schneider & Weissfeld (1986) described in Section 2.3 of Chapter 1. Start with $\boldsymbol{\gamma}_0$ as the initial values and let $\hat{\boldsymbol{\gamma}}_{(\mathrm{sw})}$ be the obtained estimator.

   iii) Compute $\hat{z}_i = \mathbf{x}_i'\hat{\boldsymbol{\gamma}}_{(\mathrm{sw})}$.

   iv) Compute $\hat{\boldsymbol{\kappa}}_{xy} = n^{-1}\sum_{i=1}^n \mathbf{x}_i\hat{z}_i$.

3. **Step 2**.

   Compute the measurement error estimator using $\hat{\boldsymbol{\kappa}}_{xy}$ instead of $\mathbf{K}_{xy} = n^{-1}\sum_{i=1}^n \mathbf{x}_i y_i$. That is, $\hat{\boldsymbol{\beta}} = (M_{xx} - \Sigma_{uu})^{-1}\,\hat{\boldsymbol{\kappa}}_{xy}$, where Σ_{uu} is the known covariance matrix of the measurement error term.

The performance of the estimator is shown using simulations in Section 3.4.

3.3.4 Standard errors

In order to obtain the standard errors of the two-step estimator we will start assuming uncensored observations only. In such a case, asymptotic robust standard errors may be computed using the normal theory estimates. Indeed, Satorra (1992) proved that even though $\mathbf{X}^{\star}$ and the disturbance term in model (3.1) and (3.3) are not normally distributed, the standard errors using normal theory are asymptotically correct.

However, we remark that, in the presence of censoring, the usual formulae for standard errors in linear measurement error models do not apply. This is due to step 1 of the estimation procedure, where the observed durations are replaced by the estimated values of the true duration. Even though the asymptotic standard errors are not straightforward to obtain, the asymptotic normality could be proved by making a slight modification in step 1. Indeed, Lai & Ying (1991) give the asymptotic covariance matrix and also prove the asymptotic normality for a modification of the Buckley & James (1979) estimator. However, based on the ease of computation in practical application, we define our two-step estimator using the

Scheider & Weissfeld (1986) procedure for accomplishing step 1. Thus in this case, we advocate computing standard errors using bootstrap methods (see, e.g. Efron & Tibshirani, 1993). This methodology has been implemented for the case of simple linear regression model as follows:

1. From the observed data

   $$\mathcal{D} = \{(z_i, \delta_i, x_{1i}), \quad i = 1, \cdots, n\}$$

   we select $B = 50$ independent bootstrap samples

   $$\{\mathcal{D}_b, \quad b = 1, \cdots, 50\}$$

   each of size $n = 1000$.

2. For each bootstrap sample $\mathcal{D}_b$, we compute the two-step estimates

   $$\hat{\beta}(b) = \begin{pmatrix} \hat{\beta}_0(b) \\ \hat{\beta}_1(b) \end{pmatrix}$$

3. We compute the bootstrap standard errors s_b as:

   $$s_b = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}(b) - \hat{\beta}(\cdot) \right)^2 \right\}^{1/2}$$

   where

   $$\hat{\beta}(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}(b)$$

In Table 3.5 of Section 3.4 we show 5% and 10% tails of the empirical distribution of the z-statistic of the two-step estimator defined in (3.10). Those results indicate that these empirical values remain close to the theoretical ones when there is censoring in the response and measurement error on covariates.

3.4 Monte Carlo studies

In this section we are going to describe some numerical studies as well as the common data generating process.

We consider a non-constant covariate such that $\mathbf{x}_i^\star = (1, x_{1i}^\star)'$ where the $\{X_{1i}^\star,\ i = 1, \cdots, n\}$ are i.i.d. $\mathcal{N}(0, \sigma_{x_1}^2)$. The uncensored response variables $y_i = \mathbf{x}_i^{\star\prime}\boldsymbol{\beta} + w_i,\ i = 1, \ldots, n$ where the $\{W_i,\ i = 1, \ldots, n\}$ are i.i.d. random variables with an extreme value distribution, with shape parameter set to be 2 and scale parameter equal to 1. We set $\boldsymbol{\beta} = (3, 1)'$.

The measurement error is introduced by taking the observed covariates to be $X_{1i} = X_{1i}^\star + u_{1i},\ i = 1, \cdots, n$, where $\{U_{1i},\ i = 1, \cdots, n\}$ are i.i.d. distributed normally random variables with zero mean and variance $\sigma_{u_1}^2$. Thus, the observed model is defined as $y_i = \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i,\ i = 1, \cdots, n$.

On the other hand, the values $\{y_i,\ i = 1, \cdots, n\}$ are right censored according to a Type I censoring mechanism. For this, we generate the censored values from the i.i.d. random variables $C_1, \cdots, C_n$ independent of $Y_1, \cdots, Y_n$. The probability distribution for $C_i,\ i = 1, \cdots, n$ is Uniform $[0, b]$ with parameter b fixed in terms of c, the desirable percentage of censoring in the observed sample. Hence we generate the observed response values as $z_i = \min\{y_i, c_i\},\ i = 1, \cdots, n$.

In order to implement the two-step estimation proposed in Sections 3.3.1 and 3.3.2 we follow the procedure described in 3.3.3. That is, step 1 consists in obtaining $\hat{\boldsymbol{\gamma}}$ from the model $z_i = \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i$. Step 2 is accomplished by computing the measurement error estimator of model $y_i = \mathbf{x}_i^{\star\prime}\boldsymbol{\beta} + w_i$ with $x_{1i} = x_{1i}^\star + u_{1i}$. From here the expected estimator $\hat{\boldsymbol{\beta}}$ is obtained.

The Monte Carlo study considers variation on the sample size n and the variance $\sigma_{u_1}^2$ of the measurement error variable U_1. The sample size n takes the values 100, 500 and 1000, while $\sigma_{u_1}^2$ is varied so that the reliability ratio[4] $k = \sigma_{x_1^\star}^2 / \left(\sigma_{x_1^\star}^2 + \sigma_{u_1}^2\right)$ ranges from $k = 1$ (no measurement error) to $k = 0.2$ (80% of the variance of x_1 is due to measurement error). The percentage of censoring, c is approximately equal to 20%. Each Monte Carlo run was based on 500 replications.

All the simulation studies have been implemented using MATLAB (1997) and all the files are available.

The results of the simulations are in the next subsections. Firstly we present the performance of the estimates used in step 1 of the procedure. That is, we study the method of Schneider & Weissfeld (1986) when covariates may be contaminated with measurement error. Afterwards we display the results for the estimator $\hat{\mu\boldsymbol{\kappa}}_{xy}$ defined in (3.5). Finally, the performance of the two-step estimator is displayed.

[4] see Section 3.2 for details about the reliability ratio

Here we present results for the case of uncensored data only and for a 20% of censored observations. The bias as well as the empirical probabilities for the tails of the z-statistic are given.

3.4.1 The effects of the measurement error on the SW estimator

The method proposed by Schneider & Weissfeld (1986) gives consistent estimates of the regression parameters for censored linear models. However the procedure is described assuming that covariates are fixed and free of measurement error. Here, using simulated data, we analyze the performance of the SW estimator when covariates may be measured with error.

For the case of uncensored data, the LS estimates of the regression parameters coming from the observed data, that is, ignoring measurement error, are biased (see Chapter 2). Indeed, if the true model is

$$y_i = \mathbf{x}_i^{\star\prime}\boldsymbol{\beta} + w_i, \quad i = 1, \cdots, n$$

but we estimate

$$y_i = \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i, \quad i = 1, \cdots, n$$

where $\mathbf{x}_i = \mathbf{x}_i^\star + \mathbf{u}_i, i = 1, \cdots, n$, the LS estimator, say $\hat{\boldsymbol{\gamma}}_{(\mathrm{LS})} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'y)$ with $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)'$ and $\mathbf{y} = (y_1, \cdots, y_n)'$ is such that $\hat{\boldsymbol{\gamma}} \xrightarrow{P} \boldsymbol{\gamma} = \mathbf{k}_{xx}^{-1}\mathbf{k}_{xy}$.

The simulations carried out in this section show that when censored observations are also included, the SW estimator of the regression parameters behaves similarly to the LS estimator for the uncensored case. That is, even though $\hat{\boldsymbol{\gamma}}_{(\mathrm{SW})}$ is a biased estimator of $\boldsymbol{\beta}$ it satisfies $\hat{\boldsymbol{\gamma}}_{(\mathrm{SW})} \xrightarrow{P} \boldsymbol{\gamma}$.

In order to compare the behavior of $\boldsymbol{\gamma}$ and the LS estimator, we restrict to a simple regression model where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$. In Table 3.2 we compare the empirical bias of the estimator $\hat{\gamma}_{1(\mathrm{SW})}$ and the bias for the LS estimator $\hat{\gamma}_{1(\mathrm{SW})}$ obtained in the case of uncensored data only. With respect to the parameter γ_1 of the observed model, we emphasize that both estimators remain unbiased.

Table 3.2: Bias of the LS estimator without censoring and the SW estimator for censoring under measurement error

	100		500		1000	
k	$\hat{\gamma}_{1(\mathrm{LS})}$	$\hat{\gamma}_{1(\mathrm{SW})}$	$\hat{\gamma}_{1(\mathrm{LS})}$	$\hat{\gamma}_{1(\mathrm{SW})}$	$\hat{\gamma}_{1(\mathrm{LS})}$	$\hat{\gamma}_{1(\mathrm{SW})}$
1	–.0014	–.0109	.0008	–.0056	–.0005	–.0073
.8	.0017	-.0056	.0027	–.0035	–.0012	–.0050
.6	.0003	–.0039	–.0001	–.0053	–.0008	–.0054
.4	–.0009	–.0060	–.0020	–.0060	.0006	–.0042

NOTE: Sample size n = 100, 500, 1000. Reliability ratio k = 1, .8, .6, .4. The percentage of censoring c is approximately 20%. Population values of $\gamma_1 = k$.

3.4.2 Consistency of the estimator $\hat{\boldsymbol{\kappa}}_{xy}$

In step 1 of the two-step estimator an estimator of $\mathbf{k}_{xy}$ is defined as

$$\hat{\boldsymbol{\kappa}}_{xy} = n^{-1} \sum_{i=1}^{n} x_i \hat{z}_i,$$

where $\hat{z}_i = x_i' \hat{\boldsymbol{\gamma}}_{(\mathrm{SW})}$. The consistency of $\hat{\boldsymbol{\kappa}}_{xy}$ is required in order to get a consistent estimator of $\boldsymbol{\beta}$ obtained after the two steps of the proposed methodology are accomplished.

Results in Table 3.3 show the empirical bias of $\hat{\boldsymbol{\kappa}}_{xy}$. We present the results assuming a 20% censoring and varying the sample size n = 100, 500 and 1000. We point out that $\hat{\boldsymbol{\kappa}}_{xy}$, computed from the observed values of the response variable, defines an unbiased estimator of $\mathbf{k}_{xy} = \mathrm{E}(XY)$, the cross products matrix of $\mathbf{X}$ and the true response variable, Y.

3.4.3 Performance of the two-step estimator

Before applying the foregoing procedure to real-life examples some Monte Carlo studies have been carried out. The results indicate that the estimator remains unbiased for several sample sizes, amounts of measurement error and proportions of censored observations.

Table 3.3: Bias of $\hat{\boldsymbol{\kappa}}_{xy}$

k	100	500	1000
1	–.0038	–.0066	–.0086
.8	–.0056	–.0081	–.0060
.6	–.0155	–.0038	–.0057
.4	–.0108	–.0083	–.0075

NOTE: Sample size $n = 100, 500, 1000$. Reliability ratio $k = 1, .8, .6, .4$. The percentage of censoring c is approximately 20%. Population values of $k_{xy} = 1$.

Table 3.4 gives the results about bias and standard errors for the case of $c = 0$, that is non-censoring. The second and sixth columns contain the bias of the estimations for parameter β_0 and β_1, respectively. We note that the estimates are unbiased regardless of the value of the reliability ratio k, that is, the size of measurement error of X_1.

With regard to the expected sampling variability of the two-step estimator, asymptotic robust standard errors have been computed using the normal theory estimates. Results about sampling variability are displayed in the columns of 5% and 10% in Table 3.4. They contain respectively, the empirical probability of the 5% and 10% tails for the $z-$statistic. These results show that for large sample sizes the empirical percentiles match the theoretical ones.

The results obtained for data including censored observations (i.e. $c \neq 0$) are summarized in Table 3.5. We consider a Type I of right censoring and the censored observations are $c = 20\%$ approximately. The results for such a case were computed using the two-step estimator developed in Section 3.3.

As we can see from the second and sixth columns of Table 3.5, the two-step estimator is also unbiased regardless the values of the reliability ratio, k. Thus the results suggest that the proposed procedure of estimation give consistent estimates of the regression parameters of model defined by (3.1), (3.2) and (3.3).

The next issue we are interested are the standard error of the two-step estimator. As pointed out in Section 3.3.4, under the presence of censoring, usual formulae for standard errors in linear models with measurement error, do not apply. Thus we

Table 3.4: Monte Carlo results for uncensored data. Sample size $n = 1000$

	$\hat{\beta}_0$				$\hat{\beta}_1$			
k	B($\hat{\beta}_0$)	V(z)	5%−tail	10%−tail	B($\hat{\beta}_1$)	V(z)	5%−tail	10%−tail
1	.001	1.04	5.80	10.20	.001	.87	4.80	8.00
.8	.000	1.06	4.60	11.60	.003	.88	3.40	7.40
.6	.001	.98	4.20	8.60	.000	1.05	5.40	12.20
.4	.001	.97	5.40	10.00	.002	.97	5.40	11.80

NOTE: B($\cdot$) is the bias of the estimator, V($\cdot$) denote the estimated variance of the z-statistic, and 5%−tail, 10%−tail are the empirical P($|z| > 1.96$) and P($|z| > 1.65$), respectively. Population values of parameters are $\beta_0 = 3, \beta_1 = 1$.

propose to use the bootstrap method.

Once the bootstrap standard errors have been obtained, we compute the z-statistics for β_0 and β_1, i.e. $z = \text{B}(\hat{\beta}_0)/s_b(\hat{\beta}_0)$ and $z = \text{B}(\hat{\beta}_1)/s_b(\hat{\beta}_1)$, respectively. The 5% and 10% columns in Table 3.5 show the empirical probability of $|z| > 1.96$ and $|z| > 1.65$, respectively. From these results we note that empirical values agree with the theoretical ones. From here we conclude that the bootstrap methodology allows us to obtain consistent standard error of the regression parameters for censored linear models with measurement error on covariates.

3.5 A more general two-step estimator

An assumption behind the two-step estimator just described is that the covariance matrix of the measurement error, Σ_{uu}, is known. However for real data this matrix is not available, so other methods have to be applied. We consider two cases: methods using consistent estimates of the Σ_{uu} and instrumental variables estimation. Thus the two-step estimator has to be modified just in the second step.

If S_{uu} denotes an unbiased estimator of Σ_{uu}, then the estimator analogous to (3.7) is defined as

$$\hat{\boldsymbol{\beta}} = (\mathbf{K}_{xx} - S_{uu})^{-1}\mathbf{K}_{xy}.$$

Table 3.5: Monte Carlo results with 20% of Type I of censoring.

	$\hat{\beta}_0$				$\hat{\beta}_1$			
k	B($\hat{\beta}_0$)	V(z)	5%−tail	10%−tail	B($\hat{\beta}_1$)	V(z)	5%−tail	10%−tail
	$n = 100$							
1	−.011	1.04	6.20	11.20	−.011	1.09	6.60	10.60
.8	−.009	.95	4.00	9.20	.000	.92	3.60	8.80
.6	−.004	.79	3.20	5.80	.027	.85	4.20	8.00
.4	−.017	.52	1.40	3.80	.068	.48	1.80	4.20
	$n = 500$							
1	−.003	1.04	5.20	10.40	−.006	1.07	6.20	10.80
.8	−.008	1.09	6.80	11.80	−.006	1.04	6.00	10.60
.6	−.004	.98	5.00	10.00	.002	1.03	5.00	10.00
.4	−.010	.87	5.00	10.00	.005	1.01	4.60	9.60
	$n = 1000$							
1	−.005	1.14	7.40	12.20	−.006	1.02	5.80	12.40
.8	−.007	1.03	6.40	11.80	−.007	1.17	6.40	11.20
.6	−.008	1.06	6.20	11.60	−.009	1.05	7.20	12.20
.4	−.011	1.02	6.20	11.40	−.008	1.01	6.20	11.80

NOTE: B($\cdot$) is the bias of the estimator, V($\cdot$) denotes the estimated variance of the z-statistic and 5%−tail, 10%−tail are the empirical P($|z| > 1.96$) and P($|z| > 1.65$), respectively. Population value of parameters $\beta_0 = 3, \beta_1 = 1$.

which is a consistent estimator of the regression parameter $\boldsymbol{\beta}$ and is asymptotically normal under the hypothesis of normality (these properties are in Fuller, 1987). In that case the two-step estimator should be obtained replacing $\mathbf{K}_{xy}$ by $\hat{\boldsymbol{\kappa}}_{xy}$ in the previous equation.

The other case that is where it is not possible to have an estimator of the measurement error covariance matrix. Here, if in addition to the observed data $(y_i, \mathbf{x}_i)$ we also observe a third set of variables denoted by $\mathbf{w}_i$ that is known to be correlated with $\mathbf{x}_i^\star$, we can use the method of the instrumental variables. In such a case it follows that we can estimate parameters $\boldsymbol{\beta}$ in model (3.1) and (3.3) using

$$\hat{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{X})^{-1}\,\mathbf{W}'\,\mathbf{y} \tag{3.11}$$

where $\mathbf{W} = (w_1, \cdots, w_n)'$, $\mathbf{X} = (x_1, \cdots, x_n)'$ are the observed covariates and $\mathbf{y} = (y_1, \cdots, y_n)$ are the response variables. An overview of this topic as well as the properties of this estimator are in Fuller (1987) and in Carrol et al. (1995). In such a case the two-step estimator has to be modified in both steps. That is, in step 1 we compute $\hat{\boldsymbol{\kappa}}_{wy} = n^{-1}\sum_{i=1}^{n}\ \mathbf{w}_i\hat{z}_i$ as a consistent estimator of $E(\mathbf{W}Y)$. The second step uses (3.11) with $\hat{\boldsymbol{\kappa}}_{wy}$ instead of $\mathbf{W}'\,\mathbf{y}$.

Another assumption made in our model is that the measurement error model is given as

$$\mathbf{x}_i = \mathbf{x}_i^\star + \mathbf{u}_i, \quad i = 1, \cdots, n$$

assuming that $\mathbf{U}_i$ are independent of $\mathbf{X}_i^\star$. However in certain situations the measurement error can be correlated with the true value. In such a case Fuller (1987) gives the modifications of (3.7) required to obtain consistent estimates.

PART II: Statistical Analysis of Spanish Labor Histories

In this part we develop an empirical study about labor histories of workers of the Spanish labor market in the period 1980-1993. The interest of this analysis comes from the new flexible kind of contracts (temporary versus fixed) introduced in 1984 (see Segura, Durán, Toharia & Betolila, 1991), as well as the high rates of unemployment in the beginning of 80's together with the recovery period at the end of the decade.

We focus on the dynamics and the transitions among labor states done by subjects entering into the labor market in 1980. To this end we consider a sample of labor histories, that is, individual sequences of episodes elapsed in several states since the individual enters the labor market at any date between 1980 and 1993. For these purposes we use duration models (see e.g. Kiefer, 1988 for a complete review about this topic).

The analysis of Spanish duration data mainly has been related to unemployment periods or working episodes, taking for each individual the duration of a single period as the response variable. Some rellevant references are due to Andrés (1993), Gil, Martín & Serrat (1994), Ahn & Ugidos-Olazabal (1995), Antolín (1995), Blanco (1995), García-Fontes & Hopenhayn (1995), Bover, Arellano & Bentolila (1997), Carrasco (1997), García (1997) and Gonzalo (1998). The goals of these papers may be summarized in three points: first, the analysis of the time variable (the duration in a certain situation) just in the distributional sense (the usual hypothesis of normality for the response variable is not appropriate); second, assessing the effects of a set of variables on this duration and, third, the study of the possible transitions once the episode has finished. Specifically, Gil, Martín & Serrat (1994) analyze the duration of unemployment establishing a competing risks model (that is several states are taking account once the event has finished) with a Weibull distribution for the response variable. Bover, Arellano & Bentolila (1997) focus on the duration of unemployment and they used the hazard function (that is the instantaneously rate of ending an episode) in order to estimate the effects of unemployment benefit duration and the business cycle on the duration of unemployment periods. On the other hand, Carrasco (1997) analyze the characteristics that determine transitions to self-employment and the duration in this labor state (this work is restricted to the case of discrete time). Finally, Gonzalo (1998) studies how the probability of being unemployed varies along a period of time.

However, as far as we know the analysis of the whole sequence of spells that a subject has experienced during his/her labor life seems to be much less studied. Here we would like to draw attention to a recent paper due to Arranz & Muro (1999) about recurrent unemployment where an analysis is made of the correlation between a past paid unemployment period with current and future paid unemployment spells. We note that they do not have a single period for the individuals but a vector of the durations in each of the three periods of paid unemployment.

From our point of view, one of the reasons because the analysis of duration have not been applied to sequences of episodes in the Spanish labor market, is due to the few availability of datasets about the labor market. Indeed, the main sources of information about labor are collected by the Spanish Institute of Statistics (INE). The datasets related to labor market are coming from the Labor Force Survey (EPA) and the Continuous Family Expenditure Survey (ECPF). Briefly we are going to emphasize the main features of both data bases. The EPA is defined as a quarterly continuous investigation about families. It collects data about the labor force market as well as the state when individual is out of the force market. The sample consists of 60,000 families per quarter, which belong to a rotating panel up to six quarters. The labor situation of all members in a family and a few set of personal characteristics are recorded. The ECPF is focused on family expenditures as well as demographic and wealth characteristics of the families. In that sense it has more rich information than EPA but not related with the labor status. It is also a rotating panel where families are involved in up to eight quarters. Moreover it also contains discrete (once per quarter) information about the labor market situation and income.

The data just described contain at least two problems when we are interested in the time elapsed in a labor state. On the one hand, shorter durations than a quarter are not observed and previous history is not always available. On the other hand, the observed period of subjects is too short in order to study duration in the states of the labor market and transitions among them.

In the studies we present in this part the data comes from a big data set of the Spanish Social Security named “Fichero Técnico de Afiliados a la Seguridad Social”. This file contains detailed information about the complete sequence of contribution periods to the Social Security by nearly 1,000,000 workers in Spain. In this way the data allows the analysis of several aspects related to the dynamics of the duration of the spells occupied for the individuals inside the labor market. Indeed the starting and the ending point of all episodes are available as well as three kind of states: self-employment, wage-earner or both simultaneously, in addition to the gaps between them. However, in spite of the richness of time episodes and states, we note that the file contains only a small set of individual characteristics namely gender, age, province and a variable related with the job category. This is in contrast with other sources of information related with the labor market in Spain as the EPA or ECPF briefly described above. Our final dataset used in the analyses contains the labor histories of 8,000 individuals affiliated to the Social Security system from January 1980 until July 1993.

This second part of the thesis is structured in two chapters. Chapter 4 displays some features of the data. In Section 4.1 we show the results of the descriptive analysis for all the variables. We have considered two separate sets of variables: personal and job characteristics and the time variables which contain the duration of all spells in the sequence. In Section 4.2 we have applied non-parametric techniques coming from survival analysis (see e.g. Klein & Moeschberger, 1997) in order to estimate distributional characteristics of the spell's duration. We emphasize that as in the standard analysis for survival data, our time variables have two remarkable characteristics: they may contain censored values and, at the end of each episode alternative transitions may be reached. We distinguish a first analysis for the univariate case, that is taking a single kind of episodes for each individual. Here we display separate analyses according to certain episodes in the whole sequence, the starting date or the state occupied. A second analysis takes into account differences due to the different transitions at the end of certain episodes. Here we compare the estimated survival probabilities of certain kinds of episodes according to the transition done at the end. A complete table of variables used is in the Appendix at the end of this part.

In Chapter 5, we present a comparison of several techniques used when data related to the labor market are analyzed. Here we focus on statistical inference and we discuss the most important contributions coming from our data. Because our data contain information related to a large period of time, we have analyzed some calendar effects, the state dependence of the previous history on the present spell and the differences between the states that can be occupied along the labor history. Thus the most important point about our results is related to dynamic aspects of the labor histories in Spain.

Chapter 4

Data Analysis

In this chapter we introduce the data we will use afterwards and we point out their most important features using standard descriptive analysis and non-parametric techniques.

Firstly, we would like to emphasize that the main goal of using these data is the analysis of durations elapsed in certain states of the labor market. Indeed, for each subject of the sample we do not consider only a single period but a sequence of episodes corresponding to their labor history from 1980 to 1993. Among the variables, we distinguish between two sets: personal characteristics and time durations.

The available data file contains all the contributions that each individual has been made to the Social Security system while he was in the labor market. Hence, it is possible to rebuild the individual labor histories elapsed during the analyzed period in terms of the duration of episodes, types of job, some causes of finishing a certain spell or the starting and ending dates of episodes. Therefore we emphasize that we may define a multivariate vector of durations for each individual.

Even though the dataset have information about working spells, we also consider the periods of time defined between two consecutive working episodes. We named these gaps *non-working spells* and include unemployment and non-contributing episodes. There are lots of causes behind a person has a non-contributing spell, so it is difficult to modeling those intervals under a single criteria. To that respect we only consider two categories: Volunteer causes like maternity or study and non-volunteer causes like to be dismissed from a work. In spite of everything we include the non-working episodes in our study because we consider that they are also components of the whole labor history.

Besides the time durations, a few set of personal characteristics about the individuals are also available. Moreover, we also introduce two macro-economic indicators: the quarterly values of the Spanish unemployment rate and the gross domestic product. The descriptive analysis of these variables is displayed in Section 4.1.

Section 4.2 is focused on the non-parametric analysis of duration time variables. That is, we draw some features of the distribution of duration variables. This analysis is developed using survival techniques. The goal here is to use graphical methods in order to show differences in the durations among the several episodes. Thus, for instance in the analysis of non-working spells, we would like to know the probability of changing job in the first three months or how high is the probability of remaining at the same state for more than two years. To this end, we draw the product limit estimators (see Kaplan & Meier, 1958) for several sets of durations.

Finally, we summarize three outstanding conclusions of these data. First, the behavior of the first working spell is different compared with the subsequent ones. Second, there is an effect of the calendar period in which an episode started. Third, there are differences of remaining in a job if it corresponds to a self-employed or a wage-earner spell.

4.1 Descriptive analysis

The data used in the analyses consist of individual periods of time contributing to the Spanish Social Security between 1980 and 1993, a service falling under the Ministry of Work and Social Security. The complete file contains the whole labor histories of about 38 million of people since they entered the labor market. That is, the available information is the sequence of time episodes as well as a set of characteristics of all different labor states carried out by each individual. There are also some personal characteristics of individuals.

Our dataset is a sample of the complete file. It contains information about 50399 spells which describes the labor histories of 8986 individuals who started to work between January 1980 and July 1993. From now on, we will denote by $i = 1, \cdots, n$ the number of individuals (i.e. $n = 8986$), and $j_i = 1, \cdots, J_i$ the number of episodes per subject i. Thus, $\sum_i J_i = 50399$ is the total number of episodes or spells.

The episodes of our data are defined by three variables,

$$(beg_{j_i}, end_{j_i}, dur_{j_i}) \qquad j_i = 1, \cdots, J_i.$$

The first two variables are calendar dates and they represent, respectively, the dates of starting and finishing the episode. After a transformation both are measured in days elapsed from a common origin.[1] Hence, we define the total duration of each episode by subtracting beg_{j_i} from end_{j_i}. That is we obtain the variables dur_{j_i} as the days elapsed in state $j_i = 1, \cdots, J_i$ for individual $i = 1, \cdots, n$. The analysis of variables $\{dur_{j_i}, j_i = 1, \cdots, J_i\}$ is the main goal of the study.

We emphasize two important features about the time variables. First, there is no a single response variable for each subject in the sample but a sequence. Second, the variables dur_{j_i} may be censored if they correspond to the last spells of the individual's sequence. That is, the last episode could finish after July 1993, the end point of the study period. Thus, we observe some values that do not correspond to the true duration of this time interval.

The descriptive statistics of the main variables are in Table 4.3.

4.1.1 Personal and job characteristics

Even though our data mainly contain information about duration of spells, it has been also possible to define some characteristics about the episodes, as well as a few characteristics about the individuals.

Within an individual sequence, we distinguish between *working episodes* and *non-working episodes*. The first ones correspond to periods where subjects are working in a certain kind of job and therefore are contributing to the Social Security system. The non-working episodes include unemployment periods as well as gap intervals elapsed between two consecutive jobs where the subject does not contribute to the Social Security system. Here we remark that the category of non-working episodes is imprecise and includes a large number of situations (e.g. illness, study periods, dismissing or maternity).

For the working spells, we distinguish two kind of different episodes according to the type of job, that is self-employed and wage-earner.

[1] Due to the use of the SAS program, all the dates in the data are measured in days from January 1960.

Hence, we define three dummy variables *self, w-e, bothc* which equal 1 for spells of self-employed, wage-earner and both contributions, respectively. Their frequency distribution is given in the next table:

Table 4.1: Working spells

Variable	Frequency	Percent
self	811	1.6
w-e	33502	66.5
bothc	809	1.6

For the non-working episodes we distinguish four categories according to the type of transition to a non-working spell: after a period of job the individual chooses do not work for a while (volunteer), because a dismissal of the previous job, due to the TLI (Temporary Labor Incapacity) has finished and other reasons. The next table summarizes this variables.

Table 4.2: non-working spells

Variable	Frequency	Percent
volunteer	3938	25.78
dismissal	7770	50.86
end of TLI	3046	19.94
others	523	3.42

Because our data correspond to the period 1980-1993 which includes important changes in the Spanish labor market, we also take into account the initial date where the spell starts. Figure 4.1 displays the distribution of the spells with respect to the year of starting within our study period. We remark the increasing shape of the distribution. There are mainly two reasons to explain this shape: First, new types of temporary short contracts were introduced in Spain from 1984, and second the analyzed individuals started to work in 1980, so probably they are still looking for a definitive work.

Another variable defined is related with the position of a certain episode inside the entire sequence of intervals defining the labor history. This variable is named

Figure 4.1: Bar chart according to starting dates

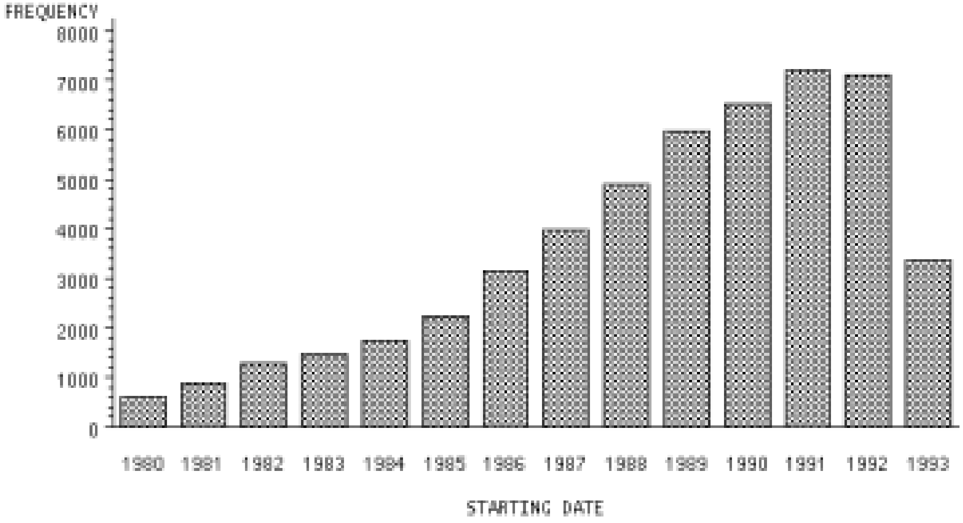


Figure 4.2: Bar chart for variable *nepis*

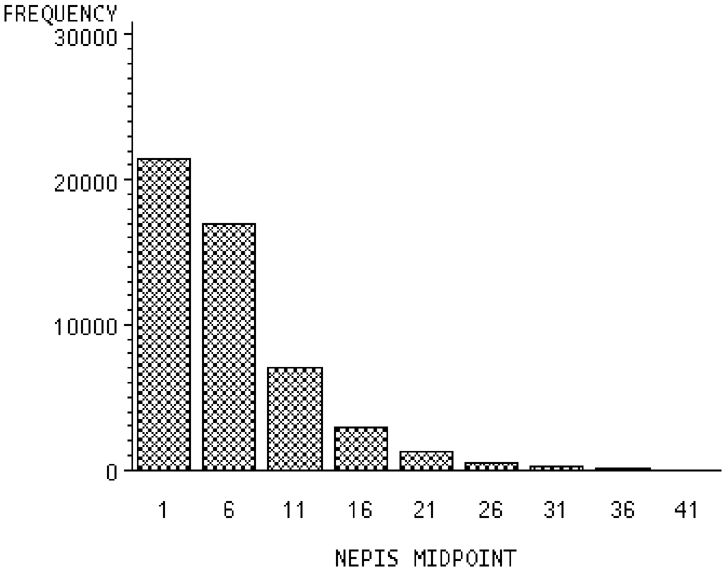


Table 4.3: Descriptive statistics of data

variable	Mean	StDev	Minimum	Maximum
sex	1.46	0.49	1.00	2.00
age	26.19	9.88	14.00	78.00
unemp	0.18	0.02	0.10	0.23
gdp	2.80	2.03	–1.67	6.11
self	0.02	0.12	0.00	1.00
w-e	0.66	0.47	0.00	1.00
non-work	0.30	0.46	0.00	1.00
nepis	6.02	5.48	1.00	43.00
nselfe	0.05	0.33	0.00	7.00
nw-ee	3.96	3.63	1.00	43.00
nnonce	1.92	2.08	0.00	19.00

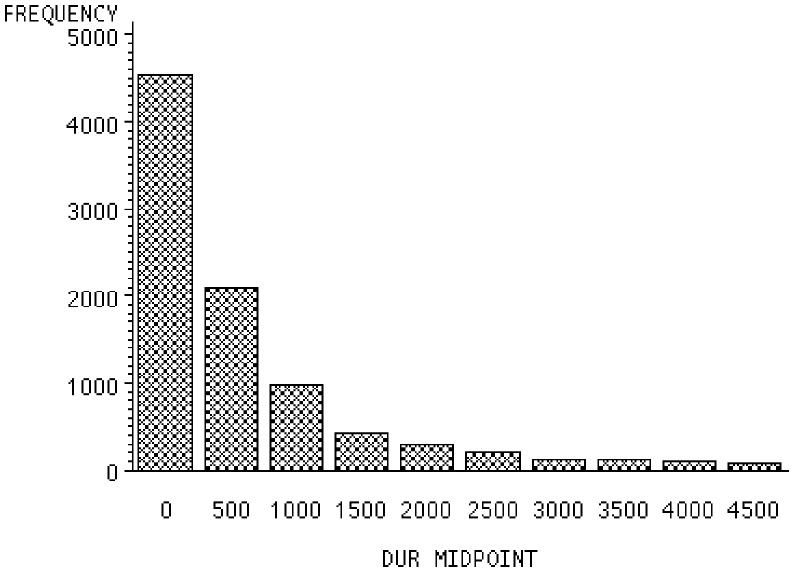
nepis and its frequency distribution is displayed in Figure 4.2. Even though its average is around 6 spells, it has a maximum of 42 spells. Therefore there are some individuals with a large amount of spells in their labor history. This is likely because our sample contains people starting to work in 1980, so that they are young individuals which still do not have a definitive job.

Also related to the number of episodes we have *nselfe, nw-ee, nbothe* and *nnonce*. They are count variables that for a given spell are, respectively, the number of previous episodes of self-employed, wage-earner, both contributions and non-working.

With respect to personal characteristics we have the gender of the subject (*sex*) and the age at the beginning of the episode (*age*).

Even though the set of variables describing characteristics of individuals is limited, the interest of the dataset is on the sequence of durations describing the labor histories. Thus, it is possible to analyze temporal aspects, as well as the impact of changes in the labor regulations (for instance, a new contractual policy) introduced in a calendar date of this period of time. We also can control the effect of macroeconomic factors. In particular we are using two variables in our analy-

Figure 4.3: Bar chart for variable dur_1



sis: Unemployment rate (*unemp*) and the Gross Domestic Product (*gdp*) defined by quarterly values.

4.1.2 Time variables

In order to point out the main features of duration variables given by $\{dur_{j_i}, j_i = 1, \cdots, J_i, i = 1, \cdots, n\}$, we carried out several analyses. From the 50399 spells, we have 8986 first spells, 6873 second spells, 5590 third spells, 4705 fourth spells and 3950 fifth spells.

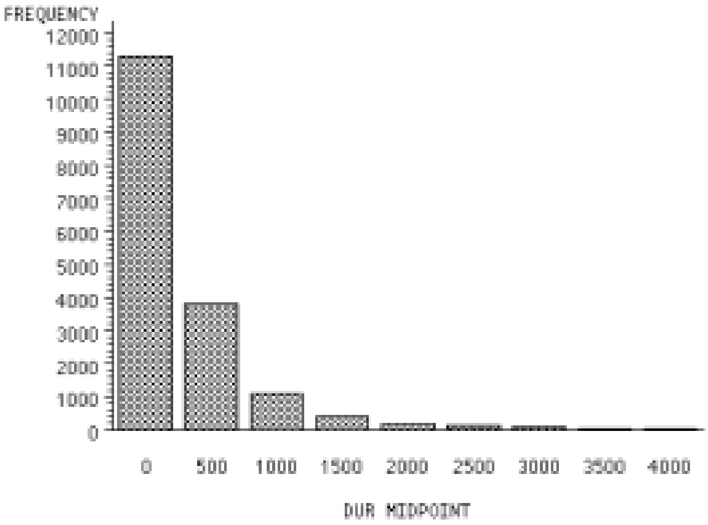
We start with variable dur_1, that is, the subsample of durations of the individual's first episode. We point out four important characteristics:

- All of the first episodes are working spells. As it is usual with variables related to the time, they have distribution functions that are highly asymmetric to the right. Here we emphasize the histogram displayed in Figure 4.3.

- The mean of the complete periods[2] is very small and equals to 398.55 days, with a standard deviation of 582.63 days; the quartiles measured in days are

[2] The sample size in this case is 6942.

Figure 4.4: Bar chart for the duration of the first non-working spell



$Q3 = 487$, $median = 181$ and $Q1 = 63$. We note that for the following episodes the mean becomes smaller.

- It contains around 23% of right censoring. However only the last spells of the sequences may be censored.

The analysis of variables dur_{j_i} which correspond to the states of self-employment, wage-earner and both contributions are in the next table:

Table 4.4: Mean duration of working spells

Variable	mean	stdev	Q3	median	Q1
self	197.54	470.09	119.25	29.00	15.50
w-e	636.24	897.70	807.00	244.00	88.50

For the first non-working episodes the histogram of variable dur_{j_i} is in Figure 4.4. The mean of complete periods is 320.86 days and the standard deviation equals 486.67 days.

4.2 Empirical survival functions

Here we introduce a non-parametric analysis for the duration time variables. Using survival techniques we obtain estimators of the empirical distribution function (equivalently the survival function, see Chapter 1) of the duration variables. We note that the presence of censored values do not allow the use of the standard empirical distribution function but a modification introduced by Kaplan & Meier (1958). The next section briefly describes this estimator.

4.2.1 Univariate analyses: The product limit estimator

In this section we do not consider the whole sequence of spells but separate univariate analyses for several types of episodes.

Let T be the random variable representing the duration of a certain type of spell. We denote by $F(t) = P(T \leq t)$ the distribution function and $S(t) = 1 - P(T > t)$ the survival function of T.

The main goal of this section is to estimate $S(t)$ without making any parametric assumption about the distribution of T. Because $S(t) = 1 - F(t)$, a standard approximation could be to compute the empirical distribution function using only the given sample. However, as it occurs in survival analysis, variable T may include censored values, so that for some subjects the true value of the time variable is unknown. For this kind of data, Kaplan & Meier (1958) suggested a procedure for obtaining an estimator of $S(t)$, named the Product-Limit (PL) or Kaplan-Meier estimator, which allows for censored and uncensored values.

In what follows we describe the Kaplan-Meier estimator for the simplest case of an univariate sample of times with Type I censoring. That is, the observed times denoted by $z_1, \cdots, z_n$ are defined as $z_i = \min\{t_i, c_i\}$ where t_i is the true time of individual i, also named failure time, and c_i is the corresponding censored value. Let $t_{(1)}, \cdots, t_{(L)}$ be the ordered subsample of the different observed failure times (i.e. if there are no ties, L will be the sample size of the uncensored observations). Then the Kaplan-Meier estimator of the survival function $S(t)$ is

$$\hat{S}(t) = \prod_{l:t \geq t_{(l)}} \left(\frac{n_l - d_l}{n_l} \right) \tag{4.1}$$

where n_l denotes the number of individuals in the sample such that $z_i \geq t_{(l)}$ and d_l

is the number of subjects with a failure time equal to $t_{(l)}$.

We note that if there were no censored observations, i.e. $z_i = t_i, \ i = 1, \cdots, n$, the PL estimator defined in (4.1) reduces to the ratio between the last numerator at each t_i and the first denominator. Thus, it is a decreasing step-function with discontinuities of magnitude $1/n$ at each t_i, that is, one minus the empirical distribution function.

We also point out a problem of the PL method that arises if there exist censored times larger than the largest true time. In that case, the PL estimator does not go to zero and it is not possible to estimate the survival function for values of the time beyond the largest failure time.

A nice development of the PL estimator as well as its properties can be found in Kaplan & Meier (1958), the original paper where it was introduced. However, it is also introduced in all the common literature about Survival Analysis (see, e.g. Lawless, 1982; Cox & Oakes, 1984; Collet, 1994 or Klein & Moeschberger, 1997).

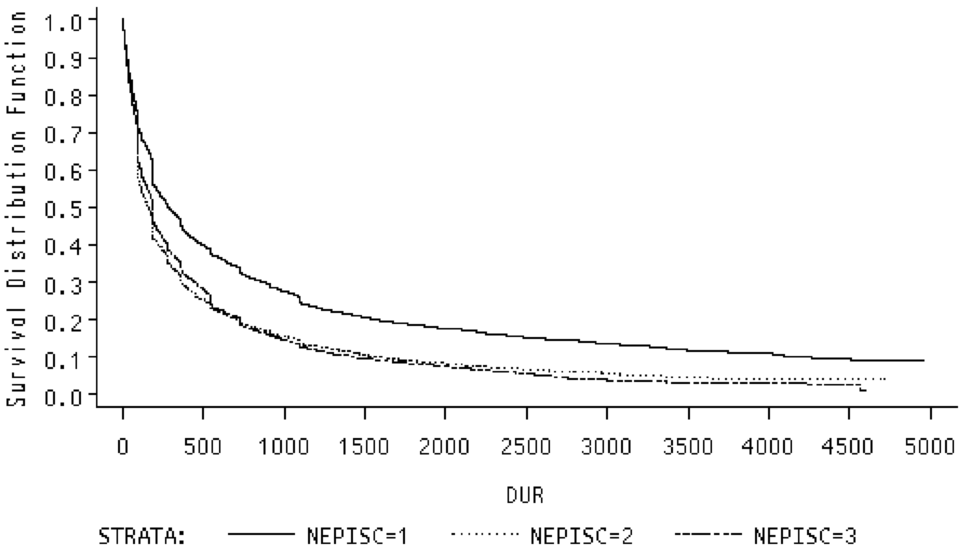
In our dataset, instead of computing the total survival function using the whole vector of duration times, we focus on separate analyses after the data have been stratified according to several criteria. In fact, this is the usual use of the PL estimator in order to be able to compare the survival curves among several subsamples. In addition to the visual display of the estimated survival functions from the graphical plots, some statistical tests may also be used in order to decide if can be accepted differences among the survival curves obtained from the data. For instance, if we have divided the sample in two subsamples, say 1 and 2, with survival functions $S_1(t)$ and $S_2(t)$, respectively, one wants to test the hypothesis $H_0 : S_1(t) = S_2(t)$ for all t. The most relevant statistics to test H_0, are the log-rank test and the Wilcoxon test. Details about these tests are given for instance by Lawless (1982).

4.2.2 Empirical results of univariate analysis

We start these analyses focusing in the first episodes of the labor histories. Figure 4.5 shows the estimated survival functions of the first three working episodes. This plot suggest that the probability of remaining in the first episode is higher compared with the probabilities of the second and third working episodes which remain very similar.

However, we emphasize that if we take into account the type of spells, the be-

Figure 4.5: PL estimators for the first three contributed spells



havior of the first three episodes is sometimes different. If we only consider self-employment episodes the estimated survival curve for the second spell is above the one for the first episode (see Figure 4.6). Thus for the self-employment status the probability of remaining longer in that state is higher in the second episode than for the first one. Looking at the first three unemployment periods (we restrict our attention to spells shorter than 3 years because a longer period usually means a non-contributing spell) we observe in Figure 4.7 differences among the episodes for durations shorter than a year. However, the estimated survival curves are very similar for longer durations, that is, there are no differences in the probability of remaining in one of the first three unemployment spells after a year of duration.

In contrast, the results for the first three non-working spells are much more different and the survival curves for the three episodes are almost parallel. Thus, the probability of remaining in this kind of episodes goes down each time the subject has a non-working spell. We display this last result in Figure 4.8.

The next analysis involves non-working spells only. García-Fontes & Hopenhayn (1995) wonder if there are differences in the survival of the non-working spells according to the causes of this situation. Here we study the effect of having a non-working spell depending on whether the individual wanted or not to have a non-working spell.

Figure 4.6: PL estimators for the first two self-employment spells

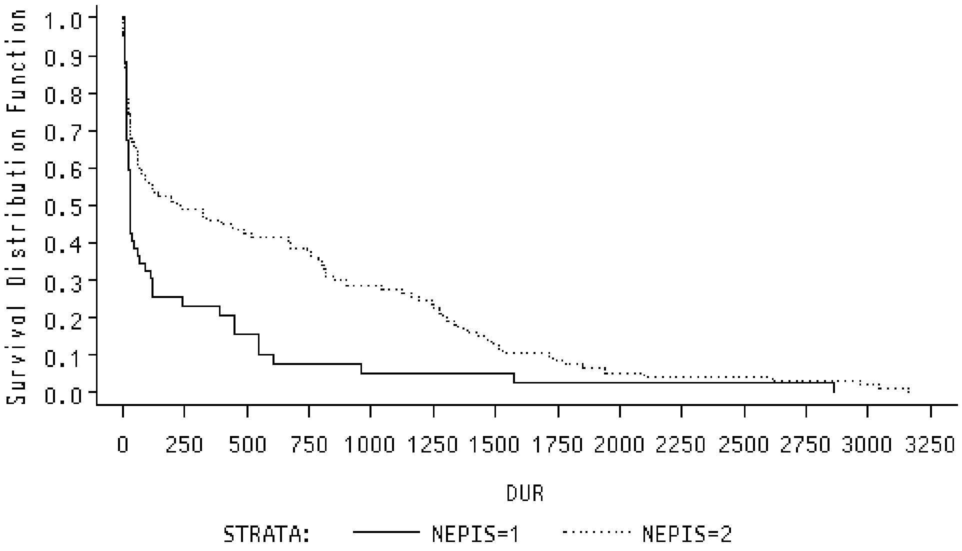


Figure 4.7: PL estimators for the first three unemployment spells

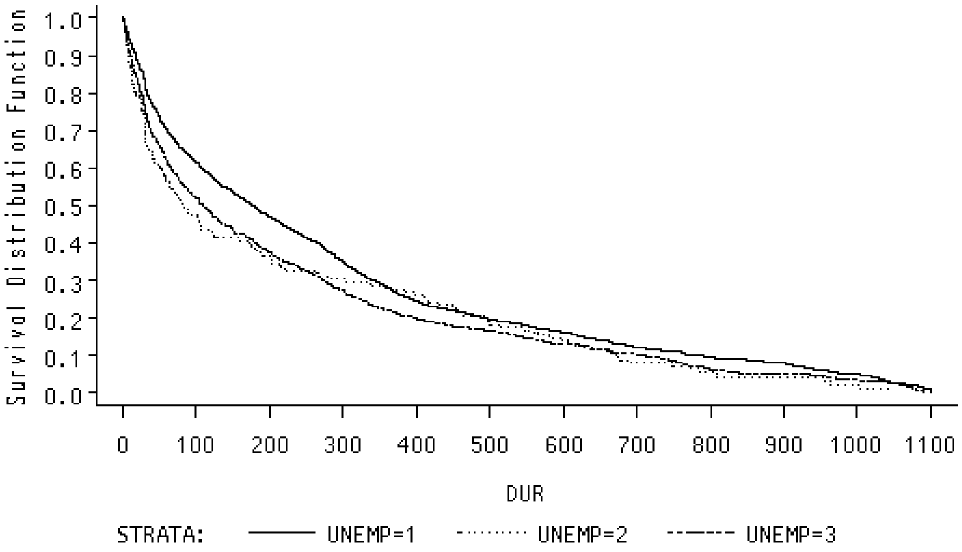
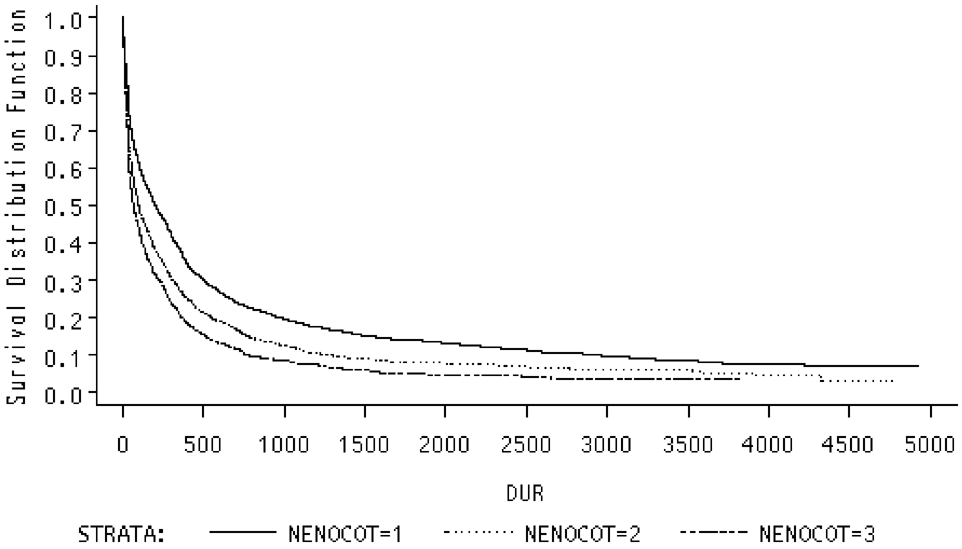


Figure 4.8: PL estimators for the first three non-working spells

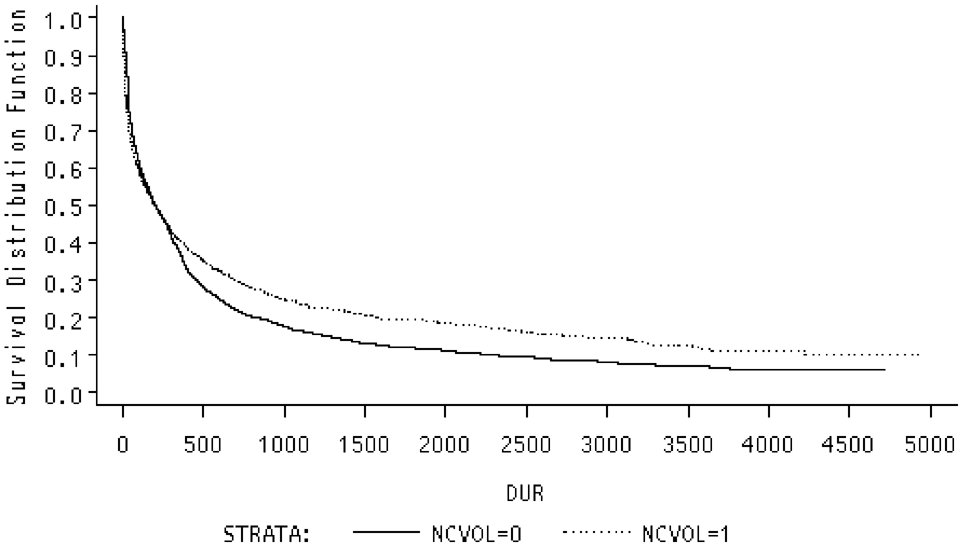


That is, for each non-working episode we have kept the cause of ending the previous working episode and have classified it into two categories: non-volunteer causes (unemployment and end of TLI) and volunteer causes (see the Appendix for details about these categories). For the first episode in the non-working status, Figure 4.9 shows that subjects who decide to be out of the labor force remains longer in this situation that people which have a non-volunteer spell of non-working. However, this changes for the subsequent episodes where the differences appear to vanish (see Figure 4.10).

As well as analyzing the different types of spells in our sample, our goal is also to look at the effects of the calendar date. That is, taking into account the moment where a certain episode started in the period 1980-1993. We start with the analysis of the behavior of the first spell through the period 1980-1993. Here we obtain a similar behavior for the working and the non-working episodes. In that sense, Figure 4.11 shows that survivor function of remaining on the first episode is higher if it starts on the intervals 1980-1984 than after 1984. Here we point out that some changes in the labor market were introduced in Spain from 1984.

More generally, we stratify our sample in two subsamples of spells according to their beginning date. The time periods we have analyzed are: 1980-88, 1985-93.

Figure 4.9: PL estimators for the first non-working spell



Because of the nature of the working and non-working spell is really different, we analyze them separately. Thus, for the working episodes, the estimated survival function of episodes starting between 1985 and 1993 is lower than the survivor curve of episodes starting from 1980 until 1984. Thus it looks like the probability of remaining in a job is higher if it started at early 80's. These results are shown in Figure 4.12.

For the non-working spells the analysis including all of them is quite similar to the one obtained for the working episodes. However, if we consider only short spells (i.e. less than 3 years) Figure 4.13 shows that both survival curves join from 2 years onwards.

The last analysis take into account the working episodes. Here if we compare the survivor curves of self-employment and wage-earner, we do not obtain significant differences. However, if we analyze only the first episodes, then the probability of remaining in a job is higher if the individual is a wage-earner than if he or she is a self-employed. Figure 4.14 shows the survival curves in this case. For the subsequent spells no differences are again obtained.

Figure 4.10: PL estimators for the second non-working spell

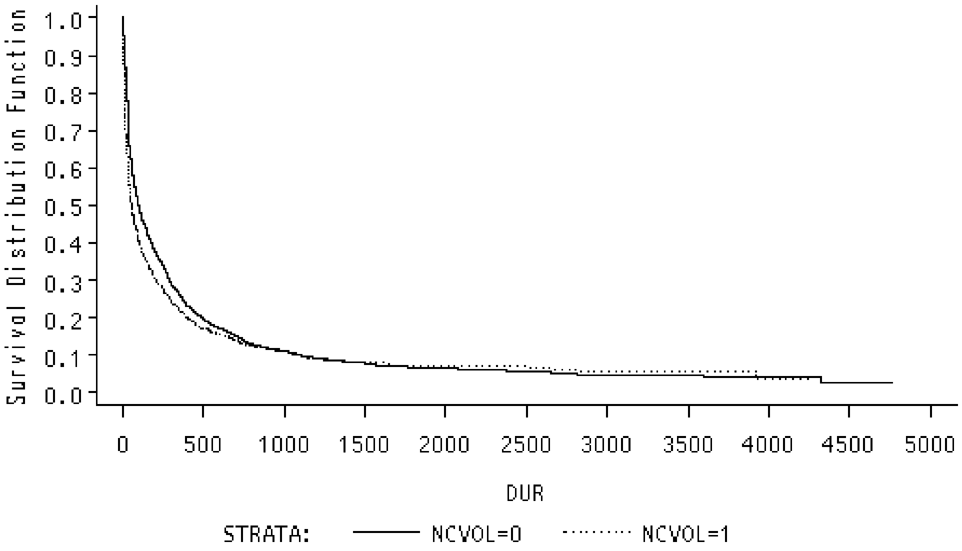


Figure 4.11: PL estimators for the first job by starting dates

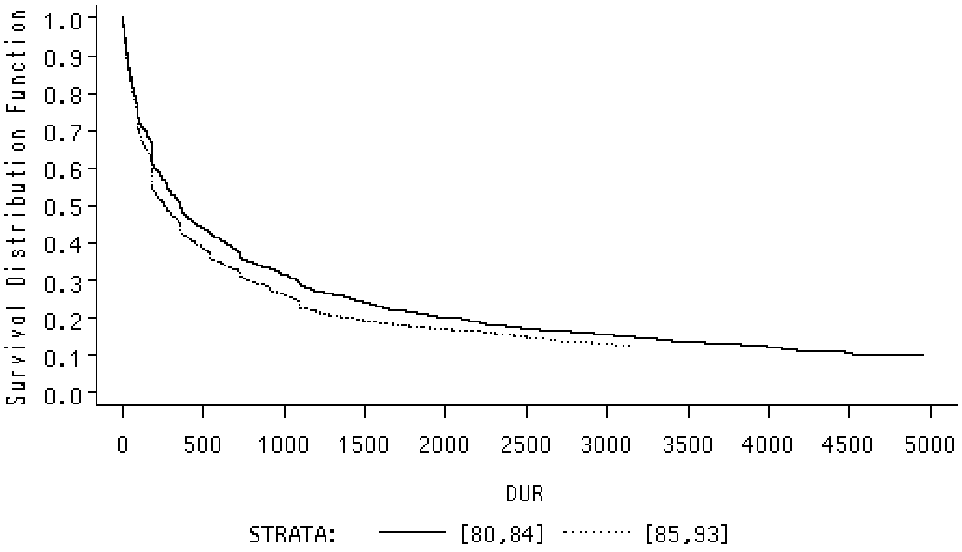


Figure 4.12: PL estimators by starting dates for the working spells

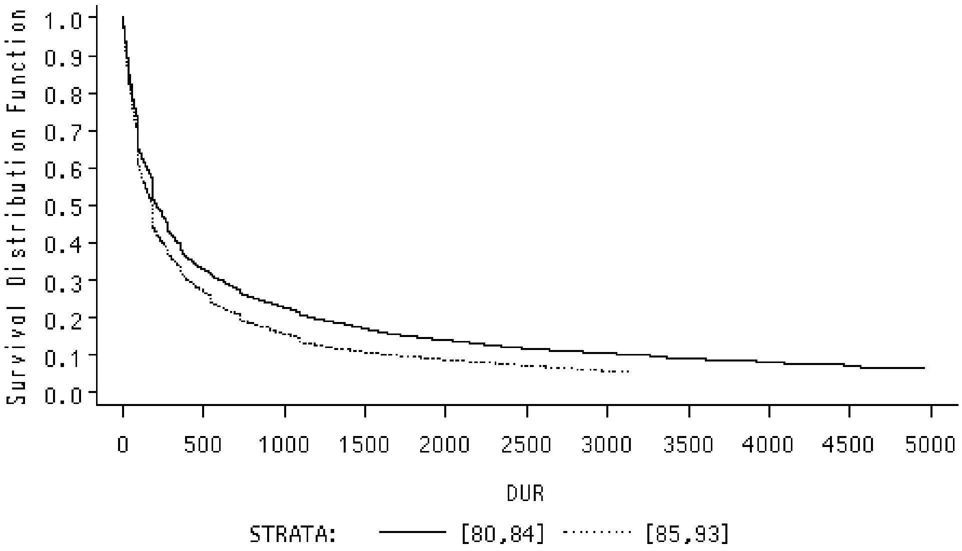


Figure 4.13: PL estimators by starting dates for the short non-working spells

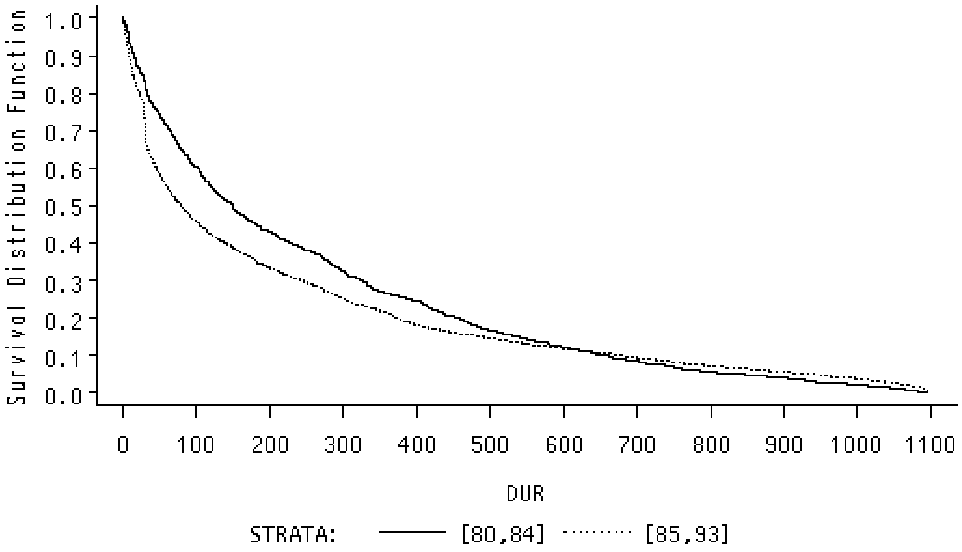
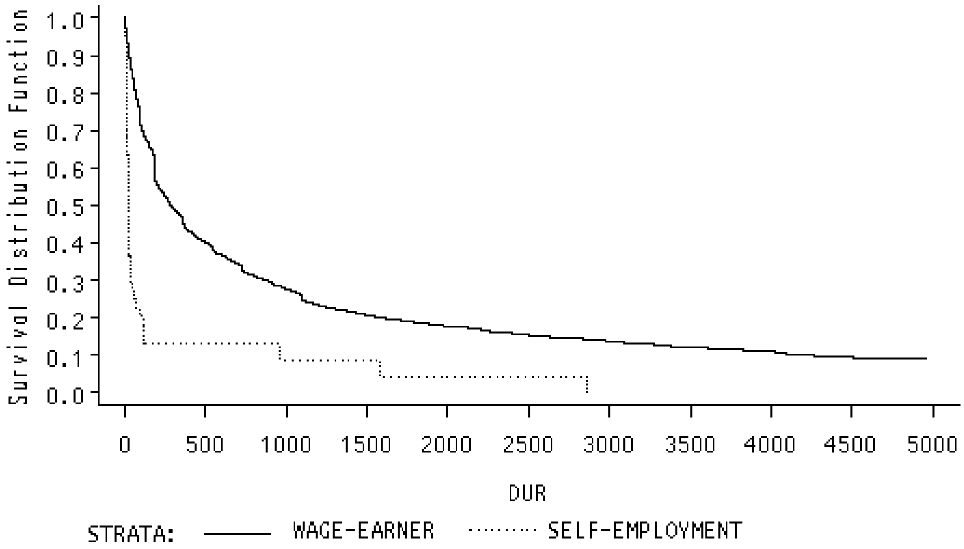


Figure 4.14: PL estimators for the first working episode



4.2.3 Competing risks analyses

In the non-parametric analyses just displayed, we have not distinguished the survival curves among the several destination states that a subject can reach after an elapsed spell. As we described before we have individual sequences of spells where all possible states are available at each transition (i.e. we have four feasible states: self-employment, wage-earner, both-contributions and non-working, that we will denote by $j = 1, 2, 3, 4$). In that way it is possible to analyze the different probabilities of having a specific transition according to the current state. Note that this kind of analysis does not compare the survival curves of types of episodes but the survivors, for a given class of spells, according to their transitions at the end of the episode.

The study of survival data taking account the specific transition at the end of a spell is known as competing risks analysis (e.g. Klein & Moeschberger, 1997). Here we are going to introduce a generalization of the Product Limit estimator taking account of the destination state (see Kalbfleisch & Prentice, 1980)

Let $t_{j(1)}, \cdots, t_{j(L_j)}$ be the subsample of the observed failure times of type $j =$

$1, \cdots, J$ in increasing order. Then, the Kaplan-Meier estimator of $S_j(t)$ is

$$\hat{S}_j(t) = \prod_{l:t \geq t_{j(l)}} \left(\frac{n_{jl} - d_{jl}}{n_{jl}} \right). \tag{4.2}$$

where n_{jl} denotes the number of sample individuals such that $z_i \geq t_{j(l)}$ and d_{jl} is the number of subjects with a failure time of type j equal to $t_{j(l)}$. The estimator defined in (4.2) is the Kaplan-Meier estimator regarding times that have transitions different than j as censored. Even though (4.2) do not correspond to estimators of the survival functions and are thus usually named pseudo survival functions (see Kalbfleisch & Prentice (1980) for a discussion, and Allison (1995) for practical applications about them) it is possible to estimate the overall survival function as[3]

$$\hat{S}(t) = \prod_{j=1}^{J} \hat{S}_j(t). \tag{4.3}$$

4.2.4 Empirical results of competing risks analysis

We are going to apply the competing risks analysis to several cases in our data. In this case we have $J = 4$ because the possible states are wage-earner, self-employment, both contributions and non-working and all the transitions are allowed except non-working to non-working.

With respect to the first episode we display in Figure 4.15 the Product Limit estimators according to the three possible destinations (self-employment, wage-earner and non-working). Here we can see that a non-working spell has the highest probability to be reached after the first episodes of the labor history. On the other hand if the second episode is of self-employment the individual will have the highest probability of remaining in the first state. These results are in agreement with the fact that the first episodes are mainly of wage-earner.

Another analysis taking into account the destination state is about the starting dates. That is, Figures 4.16 and 4.17 show the survival of the states reached after a spell starting before 1984 or after 1984, respectively.

We emphasize that the difference between both cases is about the behavior of the wage-earner and non-working states. While for the periods starting before 1984 the destination to a non-working spell has the highest survival and different from the

[3] Here we are assuming that there are no ties with the times of different types of transitions.

Figure 4.15: PL estimators for the first episodes according to the destination state

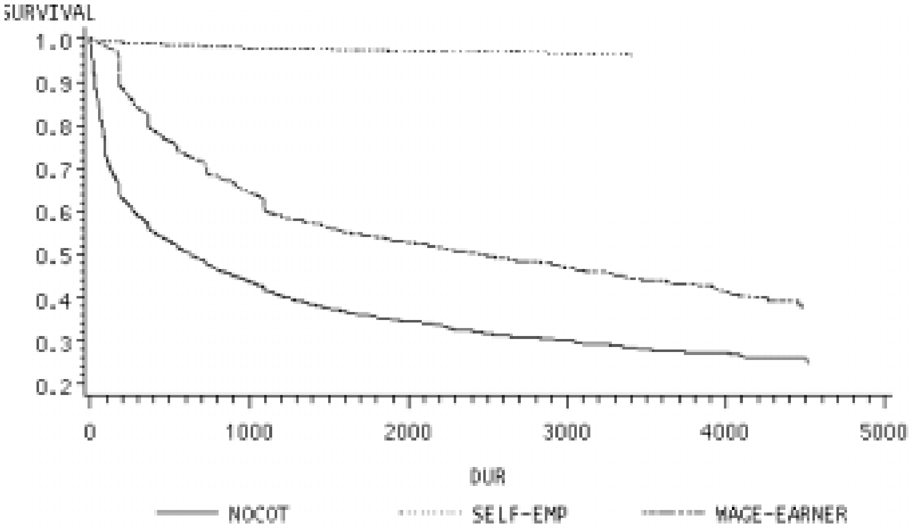


Figure 4.16: PL estimators for the episodes starting before 1984 according to the destination state

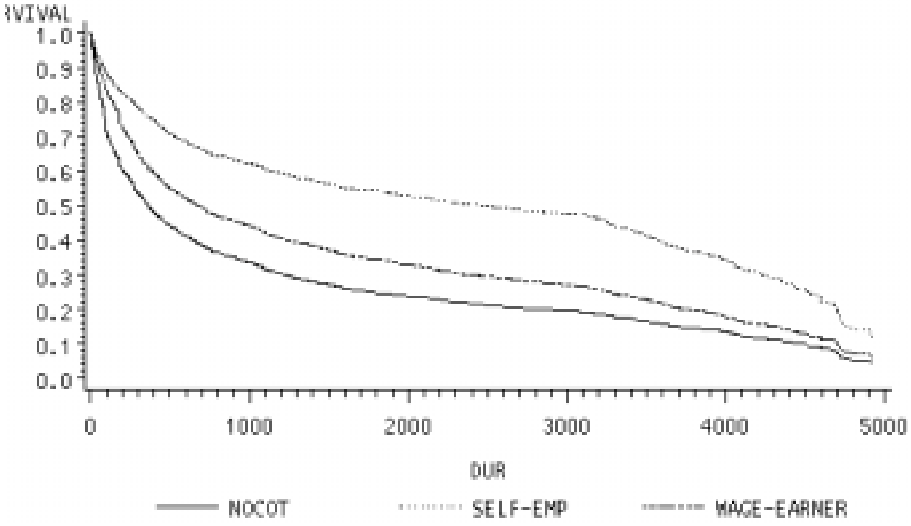
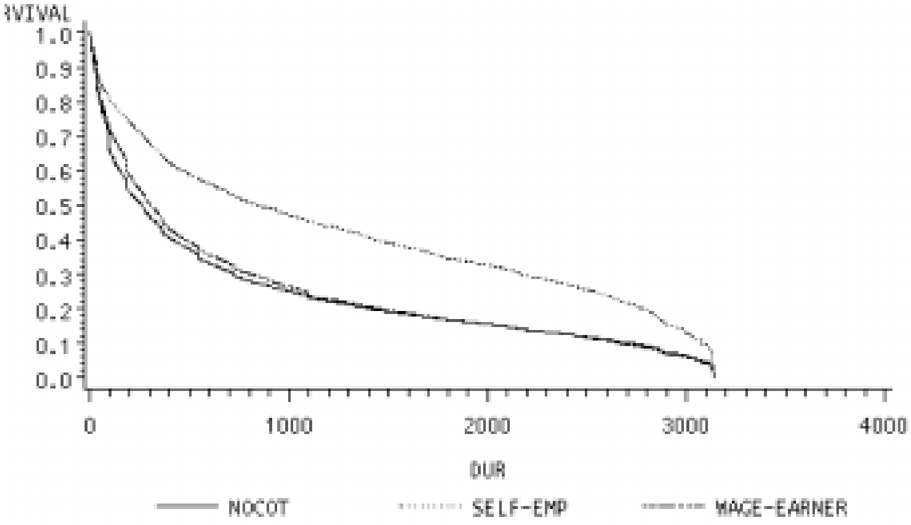


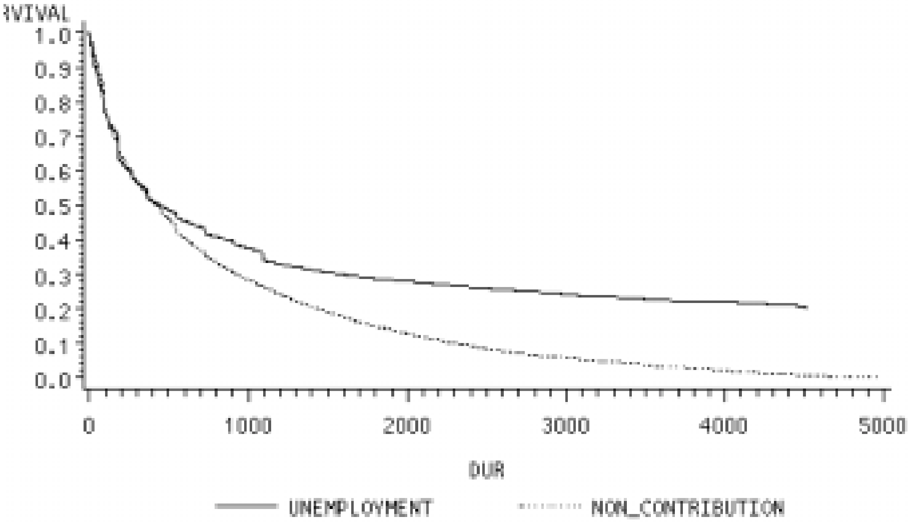
Figure 4.17: PL estimators for the episodes starting after 1984 according to the destination state



destination to a wage-earner, for the periods starting after 1984 both states behave similarly.

A final analysis is about working episodes followed by non-working episodes. Here we distinguish between unemployment and non-contributed episodes. The results are displayed in Figure 4.18 where the main feature is that unemployment has lower probability to be reached than other kinds of non-working spells. That is, a working individual has higher probability of surviving in this situation if the next state is unemployment than if the final state is a non-contributing episode.

Figure 4.18: PL estimators for working episodes according the following non-working spell



Chapter 5

Statistical Analysis of Labor Histories

In this chapter we analyze the duration of the episodes within an individual's labor history. From a set of characteristics classified into three categories (individual characteristics, previous labor history and economic indicators) we study which are the main factors determining the duration of the episodes.

We consider entire labor histories of a sample of individuals entering the labor market in 1980. As we described in the previous chapter, the data we are using come from the Social Security information files and their main attraction is their richness with respect to longitudinal information. Indeed, the file contains the entire sequence of individuals' periods of contribution to the Spanish social security. That is, for each individual it is possible to build the labor history elapsed on a time interval. The disadvantage of these data compared with other databases such as EPA or ECPF (see the introduction for more information about these data) is the small set of available explanatory variables.

Most of the previous studies dealing with duration time data about Spanish labor are focused on a single episode elapsed in a well defined state (usually unemployment, but also working episodes or periods out of the labor force). Thus, the main goals have been: the analysis of the time variable only in the distributional sense, to assess the effects of a set of variables on that duration, and the study of possible transitions at the end of the considered episode. As far as we know, the analysis of the entire sequence of spells that a subject has experienced along his/her labor life seems to be less studies. However, we emphasize the paper by Arranz & Muro (1999) where there is an investigation about recurrent unemployment. They analyzed a sequence of the first three episodes of unemployment using data from INEM (Spanish Institute of unemployment).

This chapter is based on the analysis of duration os spells, and it is structured as follows: in Section 5.1 we describe the theoretical methodologies used in the analysis of our datasets. In Section 5.2 we report an analysis for the first episode of the labor sequences. From the descriptive analysis shown in the previous chapter, the first contribution to the Social Security has its own behavior. In Section 5.3 the analysis is about types of episodes. Here we carry out separate duration analysis for the longitudinal data according to wage-earner, self-employment or non-contribution spells. We also compare our results with previous studies about the labor market in Spain. In Section 5.4 we analyze the five early spells of the labor histories. We start with separate analysis for each of the five episodes, first ones, second ones, and so on. We compare these results with the analysis using a pooled sample of all the observations. At the end we compute corrected estimations taking into account possible dependencies among the durations of the spells belonging to the same labor history. Finally, in Section 5.5 the conclusions from these results are summarized.

5.1 Statistical models

In this section we introduce the main models for the study of multivariate survival data, that is, when data involve more than one failure time on each subject. As usual we distinguish two cases: first, the competing risks problem where there is more than one cause of failure; second, when a sequence of spells one following the other is available for each individual.

5.1.1 Competing risks model

A first extension of the univariate survival analysis is the case where an episode may be ending due to several causes. Thus, in that case there is not only a non-negative random variable T but a pair of random variables (T, J) where T represents the failure time and J represents the destination state reached when the failure time takes place. Some classical references about this topic are David & Moeschberger (1978) and Prentice, Kalbfleisch, Peterson, Flournoy, Farewell & Breslow (1978). We also emphasize a new reference by Crowder (2001) which gives a very complete overview about this issue for continuous and discrete time, as well as several approaches such as the Bayesian and the counting process.

The analyses of competing risks data are based on the cause-specific hazard functions given by

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t,\ J = j \,/\, T \geq t)}{\Delta t} \tag{5.1}$$

for $j = 1, \cdots, m$ where m is the finite number of feasible states. Then, assuming that only one destination state may be reached at each failure time, the total hazard function is defined as

$$\lambda(t) = \sum_{j=1}^{m} \lambda_j(t) \tag{5.2}$$

which, as usual, is the instantaneous rate of observing a failure time t conditional on the failure time not having occurred before.

Even though from (5.2) we can define the survival function for T as

$$S(t) = \exp\left[-\int_0^t \lambda(u)\, du\right], \tag{5.3}$$

we note that from (5.1) the functions defined as

$$G_j(t) = \exp\left[-\int_0^t \lambda_j(u) du\right], \quad j = 1, \cdots, m \tag{5.4}$$

do not correspond to any survival function[1] for $j > 1$. This is based on the fact that $G_j(t) \neq \int_t^\infty \lambda_j(u)\, S(u)\, du$. Some authors refer to those functions as "pseudo" survival functions. In spite of the point we just mentioned, from (5.4) it is possible to define the overall survival function

$$S(t) = \prod_{j=1}^{m} G_j(t). \tag{5.5}$$

[1] However, Allison (1995) suggests an interpretation as a survival function, after defining the random variable T_{ij} as the time at which the jth event type either occurred to the ith individual or would have occurred if other event types had not preceded it. Then, (5.4) are the cause-specific survival functions and they give the probability that transition to j occurs later than time t.

The theory related to estimation procedures, as well as the use of covariates presented in Chapter 1 for the univariate case, may be directly extended to the case of competing risks models. In that way the likelihood function is completely defined by the cause-specific hazard function. Indeed if $\lambda_j(t;\theta)$ denotes the cause-specific hazard function depending on the unknown parameter θ, the total likelihood function for a sample of n individuals is

$$L(\theta;t) = \prod_{i=1}^{n} (\lambda(t_i;\theta))^{\delta_i} \prod_{j=1}^{m} \exp\left[-\int_0^{t_{ij}} \lambda_j(u;\theta)\,du\right] \tag{5.6}$$

where δ_i denotes the indicator of censoring ($\delta_i = 1$ if observation is complete and 0 otherwise).

Note that upon rearrangement the likelihood (5.6) factors into a component for each j. Moreover, each of these factors is precisely the same as those that would be obtained considering as censored observations all the failures with destination state other than j .

5.1.2 Multi-state and multi-episode model

This section introduces the method of maximum likelihood for the analysis of sequences of durations. We briefly emphasize the most relevant results we will use in the analysis of our data presented later on. Some general references about repeated events and multi-state processes are Blossfeld, Hamerle & Mayer (1989), Hamerle(1989) and Petersen (1995), Hougaard (1999). More in the context of economics we emphasize the papers by Elbers & Rider (1982), Heckman & Singer (1984) and Honoré (1993).

Here we assume that the multivariate survival data are represented by a sequence of pairs (T_j, Z_j), where T_j represents the failure time and Z_j the state for the jth spell, $j = 1, 2, \cdots$. The failure times are considered non-negative and continuous random variables that define an increasing sequence, $T_1 \leq T_2 \leq \cdots$, and the state variables $\{Z_j : j = 1, 2, \cdots\}$ are defined as a sequence of discrete stochastic variables in a finite state space with m possible values.[2] It is also assumed that the m possible states are mutually exclusive.

The usual analysis for this multivariate duration time data is carried out by modeling the hazard function. From here, the probability law of the transition

[2] For a more general case, the number of destination states could change among episodes.

process as well as the duration distributions can be calculated. For a complete specification of the hazard function, we have to define three points. As in the univariate case, we have to decide the time dependence of the hazard function as well as the dependence of the covariates. In addition we have to decide how much of the previous history is included. In this respect Heckman & Borjas (1980) defined four types of dependences. The first type is Markovian, so that the transitions depend solely on the current state in which the individual is located. The second type is termed "occurrence dependence". It assumes that the probability that an individual changes to a specific state depends of the number of previous spells he has been in that state. The third type is named "duration dependence" and considers that the probability of remaining in a certain state depends on the length of the time interval that the individual has already been in that state, that is it depends on the current duration. Finally, they defined the "lagged duration dependence" when the probability of remaining in a state depends on the previous failure times in that state.

The hazard function for multi-episode and multi-state data is defined by modeling the specific transition probabilities for each destination state and each episode. The cause-specific hazard functions for the jth ($j = 1, 2, \cdots$) failure on a study subject is defined for $t \geq t_{j-1}$ as

$$\lambda^j_{z_j}(t; \mathbf{x}_j, H_{j-1}) = \lim_{\Delta t \to 0} \frac{P(t \leq T_j < t + \Delta t,\ Z_j = z_j \mid T_j \geq t;\ H_{j-1}, \mathbf{x}_j)}{\Delta t} \tag{5.7}$$

where $z_j = 1, 2, \cdots, m$, are the possible states, $\mathbf{x}_j$ is the vector of covariates for the jth failure and $H_{j-1} = \{(t_l, z_l),\ l = 1, 2, \cdots, j-1\}$ is the previous history of the time process until t_{j-1}. Note that $\lambda^j_{z_j}(t; \mathbf{x}_j, H_{j-1})$ will be identically zero for $t < t_{j-1}$. Equation (5.7) gives the rate for the jth episode at which a transition to state z_j occurs at duration t, given no transition prior to t and given that state z_{j-1} was occupied immediately prior to t.

The total hazard function for each failure $j = 1, 2, \cdots$, that is, the risk of any transition from state z_{j-1} in the jth episode is defined by

$$\lambda^j(t; \mathbf{x}_j, H_{j-1}) = \sum_{z_j=1}^{m} \lambda^j_{z_j}(t; \mathbf{x}_j, H_{j-1}). \tag{5.8}$$

From here the survival function $S^j(t; \mathbf{x}_j, H_{j-1}) = P(T_j > t \mid \mathbf{x}_j, H_{j-1})$ and the joint density function of the random vector (T_j, Z_j) may be respectively obtained as

$$S^j(t; \mathbf{x}_j, H_{j-1}) \;=\; \exp\left[-\int_{t_{j-1}}^{t} \lambda^j(u; \mathbf{x}_j, H_{j-1})\, du\right], \quad t \geq t_{j-1} \tag{5.9}$$

$$f^j_{z_j}(t; \mathbf{x}_j, H_{j-1}) \;=\; \lambda^j_{z_j}(t; \mathbf{x}_j, H_{j-1})\; S^j(t; \mathbf{x}_j, H_{j-1}), \quad t \geq t_{j-1} \tag{5.10}$$

From a parametric point of view, after assuming a statistical model for (T_j, Z_j), the next step is to estimate the unknown parameters of the distribution family. In this case the method of the Maximum Likelihood is the most commonly used. As in the univariate case introduced in Chapter 1, the likelihood function is completely characterized by the hazard function, that is, by the cause-specific hazard functions related to each destination state and each episode.

Here we derive the likelihood function for multivariate survival data. We assume that each individual $i = 1, 2, \cdots, n$ has a sequence of J_i failure times given by $t_{1_i} \leq t_{2_i} \leq, \cdots, t_{J_i}$. Let δ_{j_i} be the indicator of censoring for the jth failure of individual i such that $\delta_{j_i} = 0$ if t_{j_i} is a censored observation and $\delta_{j_i} = 1$ otherwise.[3] Thus, the likelihood contribution of individual i is

$$\mathcal{L}_i = f(H_{J_i} \mid z_0)\;\left[S^{J_i}(t_{J_i}; \mathbf{x}_{J_i}, H_{J_i-1})\right]^{\delta_{J_i}} \tag{5.11}$$

where $f(H_{J_i} \mid z_0) = f(t_{J_i}, z_{J_i}, \mathbf{x}_{J_i}, \cdots, t_{1_i}, z_{1_i}, \mathbf{x}_{1_i} \mid z_0)$ is the joint density of $\{(t_{j_i}, z_{j_i}), j = 1, 2, \cdots, J_i\}$ given that individual i is in state z_0 at time t_0, and $S^{J_i}(t_{J_i}; x_{J_i}, H_{J_i-1})$ is the survival function for the last episode.

Using properties of the conditional probabilities, (5.11) can be written as[4]

$$\mathcal{L}_i = \prod_{j=1}^{J_i} f(t_{j_i}, z_{j_i} \mid \mathbf{x}_{j_i}, H_{j_i-1})\, g(\mathbf{x}_{j_i} \mid H_{j_i-1})\;\left[S^{J_i}(t_{J_i}; \mathbf{x}_{J_i}, H_{J_i-1})\right]^{\delta_{J_i}}$$

where $g(\mathbf{x}_{j_i} \mid H_{j_i-1})$ is the marginal distribution of the covariates and it does not depend on the parameters we are interested in.[5] Thus, we need to maximize

$$\mathcal{L}_i = \prod_{j=1}^{J_i} f(t_{j_i}, z_{j_i} \mid \mathbf{x}_{j_i}, H_{j_i-1})\;\left[S^{j_i}(t_{j_i}; \mathbf{x}_{j_i}, H_{j_i-1})\right]^{\delta_{j_i}}$$

[3] In fact, $\delta_{j_i} = 1, j = 1, \cdots, J_i - 1$ because only the last episode may be censored.

[4] See Hamerle (1989).

[5] If $g(\mathbf{x}_j \mid H_{j-1})$ contains some required parameters, one has to specify a parametric form for it.

and using the relationships (5.9) and (5.10) one obtains

$$\mathcal{L}_i = \prod_{j=1}^{J_i} \left[\lambda^{j_i}_{z_{j_i}}(t_{j_i}; \mathbf{x}_{j_i}, H_{j_i-1})\right]^{\delta_{j_i}} \exp\left[-\int_{t_{j_{i-1}}}^{t_{j_i}} \lambda^{j_i}(u; x_{j_i}, H_{j_i-1})\, du\right] \qquad (5.12)$$

Then the complete likelihood based on data $\{(t_{j_i}, z_{j_i}), j = 1, 2, \cdots, J_i;\ \delta_{j_i};\ \mathbf{x}_{j_i}\}$ for the sample of $i = 1, 2, \cdots, n$ individuals is the product of the terms defined in (5.12), that is,

$$\mathcal{L} = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J_i} \left[\lambda^{j_i}_{z_{j_i}}(t_{j_i}; \mathbf{x}_{j_i}, H_{j_i-1})\right]^{\delta_{j_i}} \prod_{z_l=1}^{m} \exp\left[-\int_{t_{j_{i-1}}}^{t_{li}} \lambda^{j_i}_{z_l}(u; \mathbf{x}_{j_i}, H_{j_i-1})\, du\right]\right\}$$

From this likelihood we can estimate fully parameterized models, when the cause-specific hazard function are written as $\lambda^{j}_{z_j}(t; \mathbf{x}_j, H_{j-1}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of unknown parameters.

For the purpose of estimation we note that, for each individual, the total likelihood defined in (5.12) can be factorized according to two criteria. On the one hand, in the case of data about repeated events irrespective of the state, the total likelihood of the entire sequence is obtained as a product of the density of the first spell, density of the second spell, and so forth until the last spell[6]. On the other hand, if we consider multi-state data, each term of the total likelihood refers to a single cause-specific irrespective of the episode. Therefore if there are no unobserved variables common or correlated across the spells, estimates of the parameters involved in a certain cause-specific hazard can be obtained from separate cause-specific hazards for each spell or destination state.

5.2 Analysis of the first spell

Looking at the descriptive analysis presented in Chapter 4, we see that the first episodes among the entire sequence of labor periods seems to be outstanding. Here we analyze the duration of the first contributing spell of the labor histories.

We establish parametric models for the response variable *dur* (this is dur_1 for the first episode of the labor histories) and we estimate the effect of a set of covariates. For the analysis of the first episode we use as covariates individual characteristics,

[6] If the last spell is censored the contribution to the likelihoof function is not the density but the survival function

economic indicators and variables defined from the episode itself. Furthermore in the study of the first time on a certain state we also introduce explanatory variables related with the previous history and the dynamics of the process.

5.2.1 A duration model for dur_1

In order to clarify the data we are using in the later duration analysis, first of all we are going to introduce the main features of the sample we have used.

The sample contains $n = 8983$ observations[7] corresponding to individuals starting to contribute to the Social Security system in the period 1980-1993.

The dependent variable is denoted by T such that $t_i = dur_{1_i},\ i = 1, \cdots, n$ defined by the days elapsed in the first contributed episode. The average value of this variable is $\bar{t} = 633.70$ days and the standard deviation is $\sigma_t = 896.42$ days. Moreover we note that the 22.8% of the observations are censored.

The set of explanatory variables used contains:

1. Individual's characteristics: *sex* (1 correspond to male and 2 is female) and *age* categorized as 6 dummy variables *age1-age6* corresponding to the age groups $[14, 22], [23, 27], [28, 32], [33, 42], [43, 54]$ and older than 55.

2. Spanish economic indicators: *unemp* the quarterly rate of unemployment, and *gdp* the quarterly value of the Gross Domestic Product. For each episode these two variables take the values corresponding to the quarter where the episode began.

3. Spell's characteristics. This is a set of 0/1 dummy variables defined as: *w-e* with value 1 for the wage-earner spells,[8] *pre84* with value 1 if the spell started before 1984, *volunt* with value 1 if the spell ends because of volunteer causes, *equarter1-equarter4* indicators of the quarter where the episode ends and *squarter1-squarter4* indicators of the quarter where the episode starts. In addition there is the variable *trans* with value 0 for the transition to a self-employment, 1 to wage-earner and 2 to a non-contribution spell.

[7] We note that there is one observation per individual since we only consider the first spell of the labor history.

[8] Note that such as we have defined the non-contributed spells, the first episode of the labor history can not be of this kind.

Table 5.1: Descriptive statistics of the explanatory variables

variable	Mean	StDev	Minimum	Maximum
sex	1.54	0.50	1.00	2.00
age	28.95	15.32	14.00	78.00
unemp	0.18	0.02	0.10	0.23
gdp	2.86	1.94	–1.67	6.11
w-e	0.99	0.07	0.00	1.00
pre84	0.24	0.43	0.00	1.00
volunt	0.19	0.39	0.00	1.00
equarter1	0.17	0.38	0.00	1.00
equarter2	0.17	0.38	0.00	1.00
equarter3	0.45	0.50	0.00	1.00
equarter4	0.20	0.40	0.00	1.00
squarter1	0.23	0.42	0.00	1.00
squarter2	0.26	0.44	0.00	1.00
squarter3	0.26	0.44	0.00	1.00
squarter4	0.25	0.43	0.00	1.00
trans[a]	1.67	0.50	0.00	2.00

[a] The values are computed over the uncensored values

The summary statistics of these variables are in Table 5.1. From here we emphasize: the large number of people in the younger category due to the fact that the sample contains individuals who started to contribute to the Social Security in 1980; the large percentage of wage-earner jobs; the 23.7% of the observations belong to individuals with a single contribution, so that, the first episode is also the last of the labor history; approximately 75% of spells started after 1984; around 45% of episodes finished in the third quarter of the year, while the starting dates are uniformly distributed among all the quarters.

The analysis of the duration of the first episode was carried out using the SAS

System (see Appendix B). We assume a Weibull model (see Appendix A)for the random variable $dur_1 = T$ conditioned on the covariates. Table 5.2 displays the estimates of parameters for the explanatory variables as well as for the shape and scale parameters of the distribution.

From these results we emphasize the following conclusions[9] for the duration of the first spell of the labor history:

- older individuals have larger first episodes,
- an increment of 0.1 in the rate of unemployment means durations 0.9% higher,
- when the start of the episode is prior to 1984 the duration is 78% higher than for episodes starting after 1984,
- if the transition from the first episode is to a wage-earner job then the duration is 46% longer than if the episode goes to a non-contributing spell,
- volunteer causes of ending shorten the duration by 20%,
- starting to contribute in the third quarter reduces the duration of the episode by 30%

The goodness-of-fit of the model we have just presented is based on a graphical method of the residuals. We note that in survival models several kinds of residuals have been proposed (see Collet, 1994) but the most suitable for this purpose are the Cox-Snell residuals defined as in Section 5.1. The residual plot for the Weibull model fitting the data related to the first episodes is given in Figure 5.4 of Appendix D.

5.3 Analysis according to type of spell

Based on the richness of our data about characteristics of the episodes, here we show different analyses for the three sets of spells according to their type (wage-earner, self-employment and non-working). That is we try to know the main characteristics which better explain the duration of episodes depending on the type of contribution to the social security system. Due to the factorization of the total likelihood function

[9] The interpretations have to be made controlling for the other covariates.

Table 5.2: Estimates of the model for the first episode

variable	Estimate	Std Error	$p-$value
intercept	6.102	0.30	0.0001
sex	–0.035	0.03	0.3056
age1	–2.644	0.09	0.0001
age2	–2.429	0.10	0.0001
age3	–2.251	0.11	0.0011
age4	–1.875	0.11	0.0001
age5	–1.463	0.12	0.0001
unemp	2.243	0.70	0.0001
gdp	0.007	0.01	0.5687
pre84	0.596	0.05	0.0001
volunt=0	0.195	0.04	0.0001
equarter1	0.144	0.04	0.0027
equarter2	0.005	0.04	0.9148
equarter3	1.383	0.04	0.0001
squarter1	0.062	0.04	0.1984
squarter2	–0.074	0.05	0.1163
squarter3	–0.367	0.04	0.0001
trans=0	0.138	0.16	0.3818
trans=1	0.383	0.04	0.0001
Scale	1.38	0.01	

NOTE: The log-likelihood for the Weibull model is -14689.34

defined in (5.12), we note that under some conditions, estimates of each cause-specific rate can be obtained by separate analyses. Even though this procedure is valid provided there are no restrictions on the parameters across cause-specific hazards and there are no unobserved variables common to, or correlated across, the hazards, it allows us to obtain naive estimates which approximate the behavior of the duration of these episodes. We remark that these naive estimates may suffer the same biases as in the case of ignoring unobserved heterogeneity.

In the three sets of spells, defined for each type of contribution, we have not included the episodes which are the first of the labor sequences already analyzed in the previous section.

The statistical analyses undertaken here have two goals. On the one hand, we have fitted a parametric model for the duration of spells where there is also a set of covariates (variables related with the spells, personal characteristics and economic indicators). Thus we may compare the effect of these variables on the duration depending on each kind of spell. On the other hand, we are also interested in the differences according to the possible transitions at the end of the spell. That is, we carried out competing risks analyses for the three feasible states available in our data.

5.3.1 Self-employment versus wage-earner spells

Here we consider data coming from the two kind of contributed episodes: self-employment and wage-earner spells. The analysis of working episodes in Spain seems to be largely ignored. Thus, as far as we know very few papers deal with this issue in the framework of the duration analysis. We emphasize the papers of Carrasco (1997) for episodes of self-employment and García-Fontes & Hopenhayn (1996) and García-Pérez (1997) for wage-earner jobs.

The work of Carrasco (1997) is concerned with the factors influencing the decision of entry into self-employment and, the analysis of such episodes distinguishing exit into employment from exit into unemployment. The dataset used in this paper comes from the ECPF (Continuous Family Expenditure Survey) for the period 1985-1991. Hence, the data do not contain information about time interval but discrete points in the whole period. Moreover this survey is only related to the heads of household. The set of covariates used contains three groups of variables: previous labor market situation[10], the quarterly unemployment rate and three variables related to the number of quarters that the subject has been self-employed.[11] In the analysis they focus on the estimation of the hazard function for a discrete time model. The main conclusions about their results are: first, the hazard rate is decreasing with the duration of the self-employment spell; second, the effects of being employed in the previous episode reduces the hazard of leaving the self-employment state; and third, higher unemployment leads to higher risk of quitting the self-employment spell. Our analysis extends to the period 1980-1993, defines the exact duration of the self-employment spells using a continuous time variable, and carries out a competing risks analysis for the feasible transitions after the current spell.

We fitted a generalized gamma distribution (see Appendix C) for the dependent variable defined as the duration of self-employment episodes. In addition to the variables introduced in Section 5.2, the set of covariates contains variables related to the previous history. Thus we use *empl* for the situation of the previous episode (1 is self-employment, 2 is wage-earner and 3 is non-contribution), *nselfe, nw-ee, nnonce* which are the the number of previous episodes of self-employed, wage-earner, and non-working respectively, and *lotimese, lotimewe, lotimenc* the logarithms of the previous time elapsed in self-employment, wage-earner and non-working respectively. The rate of unemployment is introduced in this case using 5 dummy variables *atur1, atur2, atur3, atur4, atur5* according to the levels $[0.10, 0.16), [0.16, 0.17), [0.17, 0.185), [0.185, 0.20), [0.20, 0.21), [0.21, 0.22]$.

The estimated coefficients of covariates are in Table 5.3. From the results we emphasize the significant values for the variables referred to previous history, the transition at the end of the spell, seasonal effects coming from the quarter of starting and ending the episode and the rate of unemployment. For the previous history we see that the longer the prior duration on self-employment, the longer the duration of the current spell. Furthermore coming from self-employment reduces the hazard of leaving the self-employment spell. Transition to a self employment is the most likely and going to wage-earner increases by 35% the duration of the current self-employment spell. For the rate of unemployment, low values mean shorter durations.

[10] She uses a dummy variable equals 1 in case of being employed before and 0 when the subject had an unemployment episode.

[11] Here she uses three dummies for the first three quarters because of the small number of durations longer than 3 quarters.

Table 5.3: Estimates for the duration of self-employment

variable	Estimate	Std Error	p−value
intercept	6.218	0.36	0.0001
sex	–0.019	0.09	0.8396
age1	–0.375	0.17	0.0278
age2	–0.123	0.16	0.4538
age3	–0.183	0.17	0.2998
age4	–0.226	0.20	0.2524
age6	–0.114	0.31	0.7168
unemp1	–0.358	0.19	0.0664
unemp2	–0.318	0.14	0.0229
unemp3	–0.641	0.14	0.0001
unemp4	0.065	0.19	0.7311
unemp5	–0.366	0.16	0.0224
pre84	0.191	0.16	0.2418
empl=1	0.392	0.13	0.0019
empl=2	0.093	0.14	0.5131
nselfe	–0.408	0.07	0.0001
nw-ee	–0.056	0.02	0.0086
nnonce	0.033	0.03	0.3456
lotimese	0.143	0.04	0.0001
lotimewe	0.015	0.02	0.5448
equarter1	0.839	0.12	0.0001
equarter2	–0.468	0.15	0.0019
equarter3	0.571	0.14	0.0001
squarter1	0.362	0.12	0.0036
squarter2	0.369	0.13	0.0058
squarter3	0.126	0.13	0.3415
trans=0	–0.692	0.13	0.0001
trans=1	–0.513	0.15	0.0006
Scale	1.152	0.03	
Shape	0.335	0.11	

NOTE: The log-likelihood for the generalized gamma model is −1329.06

The goodness of fit for the generalized gamma model has been analyzed using the residual of Cox-Snell. In Figure 5.5 of Appendix D we display the log-survivor plot versus the residuals. The graph is approximately linear so it is valid to assume this model.

The significant values for the estimated coefficients of variable *trans* motivate the analysis of competing risks according to the three feasible states at the end of a self-employment spell. The possible transitions are to self-employment, wage-earning and non-working. Using the theoretical framework introduced in Section 5.1, we have assumed a model for each transition considering as censored observations all the failures with destination state other than the analyzed. Appendix E contains tables displaying all the estimates and statistics for each destination state. Here we summarize the main differences among the three models.

- Transition to self-employment: the number of previous non-working episodes is significant with a positive coefficient. The duration increases if the previous spell was one of working.
- Transition to wage-earner: it is characterized by the null effect of the rate of unemployment, and longer duration are associated with the episodes coming from a non-working state.
- Transition to non-working: the variables *sex* and *age* become significant and the variable *empl* now has no effect. In this case women have shorter duration on self-employment, almost 50% less than men; and younger individuals have shorter durations.

Let's now to focus on the analysis of wage-earning spells. Two papers have already dealing with this type of episodes. The analysis by García-Fontes & Hopenhayn (1996) used a duration model in order to emphasize the effects of the changes introduced in the Spanish labor market from 1994. We emphasize that they define the dependent variable as the duration of the "match" between a worker and a certain firm. The paper due to García-Pérez (1997) analyzed rates of leaving employment using a sample of wage-earner episodes and a set of covariates related to personal characteristics and economic indicators. Here we notice that these two studies also used data coming from the the social security contributions, the same file that we are using. Our analysis extend these papers in the sense that we use the previous history of individuals and we also analyze competing risks models for the different transitions.

The analysis of the wage-earner spells assumes a log-normal model[12] with scale parameter equal to 1.1 (≥ 1) and therefore the hazard function highly increases until a maximum and then decreases asymptotically to 0. The main differences with respect to the duration of the self-employment spells are with respect to variable *pre84*, which is now highly significant with a positive coefficient (the expected duration is 15 percent greater for those episodes starting prior 1984), and with respect to the variables related with the previous non-working spells which in this case are significant while the previous durations on self-employment now have no effect.

From Table 5.4 we emphasize the negative coefficient for *empl*=2 (wage-earner) with respect to the other previous states, thus having a previous spell of wage-earner shortens the duration of the current one. The total time elapsed in this state has a positive effect on duration while the total time spent in non-working has a negative effect. With respect to seasonal effects we note that a longer duration is expected for episodes ending in the third quarter and starting in the first or fourth.

The goodness of fit for the log-normal model has been analyzed using the residual of Cox-Snell. In Figure 5.6 of Appendix D we display the log-survivor plot versus the residuals. The graph is approximately linear so it is valid to assume this model.

The estimates of the competing risks analyses for the wage-earner state are also given in Appendix E. The estimated coefficients of variable *trans* displayed in Table 5.4 show evidence of significant differences among transitions. Indeed the negative values of the estimates mean a shorter expected duration for the transitions to a working spells. In the next we points out the most relevant differences among transitions.

- Transition to self-employment: the variable *sex* has a highly significant coefficient which reveals the low probability of women becoming self-employed in the period 1980-1993. The estimated coefficients for variable *empl* give longer durations for those spells coming from non-working. There is no effect of variable *pre84*.

- Transition to wage-earner: the log-normal distribution gives a worse fit than the generalized gamma model. All the variables related with the total du-

[12] It is assumed that the $\log T$ is a normal random variable.

Table 5.4: Estimates for the duration of wage-earner

variable	Estimate	Std Error	$p-$value
intercept	4.768	0.08	0.0001
sex	–0.004	0.02	0.7978
age1	–0.329	0.04	0.0001
age2	–0.193	0.04	0.0001
age3	–0.031	0.04	0.4942
age4	–0.031	0.04	0.1456
age6	0.188	0.07	0.0001
unemp1	–0.091	0.03	0.0099
unemp2	–0.089	0.03	0.0013
unemp3	–0.099	0.03	0.0003
unemp4	–0.014	0.04	0.7341
unemp5	–0.102	0.03	0.0014
gdp	0.015	0.01	0.0015
pre84	0.138	0.03	0.0001
empl=1	–0.005	0.10	0.9625
empl=2	–0.401	0.02	0.0001
nselfe	–0.056	0.07	0.4514
nw-ee	–0.016	0.00	0.0001
nnonce	–0.082	0.01	0.0001
lotimese	0.043	0.03	0.2113
lotimewe	0.222	0.01	0.0001
lotimenc	–0.012	0.00	0.0066
equarter1	–0.108	0.02	0.0001
equarter2	–0.081	0.02	0.0009
equarter3	0.724	0.02	0.0001
squarter1	–0.010	0.02	0.6614
squarter2	–0.153	0.02	0.0001
squarter3	–0.275	0.02	0.0001
trans=0	–0.436	0.09	0.0001
trans=1	–0.288	0.01	0.0001
Scale	1.099	0.01	

NOTE: The log-likelihood for the log-normal model is –28876.59

rations on each of the feasible states are significant (positive coefficients for self-employment and wage-earner and negative value for non-working). Moreover, the wage-earner state of the prior episode give higher risks of leaving the current wage-earner episode.

- Transition to non-working: there is no effect from the prior state (*empl* is not significant). For the total time elapsed in each state there is only a positive coefficient for the wage-earner jobs, but the number of previous episodes of non-working has no effect.

5.3.2 Non-working spells

As we already explained, the non-working episodes have been defined as the gap periods between two jobs. Therefore, included in this kind of spells are many situations such as unemployment, illness, temporal incapacity or other volunteer reasons. In that sense the analysis we introduce in this section is restricted to the non-working episodes which correspond with unemployment periods coming after a wage-earning job.

The issue of unemployment in Spain has been largely studied. In the most part of the papers have been used duration analysis and the covariates have mainly been socio-economic variables. We emphasize the papers by Andrés, García, Jiménez (1989) Cebrián, García-Serrano, Muro, Toharia & Villagómez (1995) and Bover, Arellano & Bentolila (1996).

The main goal here is to state the main factors which explain the duration of an unemployment episode which belongs to an individual labor history. We want to extend the results of García-Perez (1997) in two directions: to include previous labor history, and a competing risks analysis according to whether the following episode is on self-employment or wage-earner. To do this, we define some new variables already used in García-Perez (1997): *bempl3* dummy variable equals 1 if the previous spell was longer than 3 years, *lodur, lodur2* which are the logarithms of the previous durations and its square value, and *qual12, qual34* defining the kind of previous job according the category: engineers and bachelor degrees, and medium degrees.

A Weibull model is assumed for the duration of unemployment. The first result we would like to note is that the estimated parameter α is 0.11 (≤ 1) and therefore defines an increasing hazard function. Thus we conclude that longer unemployment durations means lower probability to leave it. The estimates and the statistics for covariates are given in Table 5.5. We would like to emphasize the significance of variables related with the previous episodes (*lodur, lodur2, lotime*) and variable *sex*, with shorter duration for women, but we will see an important changing in the competing risks model.

The goodness of fit for the Weibull model has been analyzed using the residual of Cox-Snell. In Figure 5.7 of Appendix D we display the log-survivor plot versus the residuals. The graph is approximately linear so it is valid to assume this model.

The competing risks analysis establishes differences between the possible transitions. The complete tables of the estimates and statistics are in Appendix E, however here we summarize the main features:

- Transition to self-employment:[13] for the significant variables we note that the duration of unemployment is higher for women than for men. However, variables like *qual12, qual34, unemp* or *lodur* are not significant in this case.

- Transition to wage-earner: here the Weibull model does not fit the data, so we use the generalized gamma. The main difference is in the shape of the hazard function. Related to the variables, we find that the type of previous job and the qualification become significant as well as the previous history, the seasonal factors, the rates of unemployment and the gross domestic product.

5.4 Analysis of the five early spells

In the previous sections we have dealt with duration analyses of several subsets of spells treating them separately. However, some observations may belong to the same labor history of an individual, and there may be possible correlations among them. In such a case, as we point out in Section 5.1 the naive estimates as well as their standard errors obtained from separate analysis according to a certain criteria (kind of spells, job number or whatever) may be biased. In some sense it has the same effect as ignoring the unobserved heterogeneity (e.g. Heckman & Singer, 1985).

The main goal here is to correct for possible dependence among observations of our data. To achieve this we restrict our sample to labor histories with no more than five episodes. In our opinion more than five labor spells in the study period

[13] We points out that the sample is really small (92 uncensored durations).

Table 5.5: Estimates for the duration of unemployment

variable	Estimate	Std Error	$p-$value
intercept	10.636	0.45	0.0001
sex	–0.069	0.03	0.0360
age1	–0.014	0.10	0.8983
age2	–0.192	0.11	0.0698
age3	–0.138	0.11	0.2080
age4	–0.207	0.11	0.0696
qual12	0.061	0.03	0.0854
qual34	0.013	0.04	0.7632
unemp	2.592	0.72	0.0003
gdp	0.111	0.01	0.0001
pre84	0.138	0.03	0.0001
bempl3	–0.114	0.15	0.4635
lodur	–0.597	0.18	0.0013
lodur2	0.058	0.01	0.0016
lotimewe	–0.150	0.02	0.0001
lotimenc	–0.005	0.01	0.4972
equarter1	0.026	0.04	0.5628
equarter2	0.215	0.04	0.0001
equarter3	0.231	0.05	0.0001
squarter1	–0.105	0.04	0.0209
squarter2	–0.023	0.05	0.6363
squarter3	–0.050	0.04	0.2180
trans	–3.584	0.10	0.0001
Scale	0.901	0.01	

NOTE: The log-likelihood for the Weibull model is -4931.78

(1980-1993) means very short durations with very heterogeneous data. Moreover, using labor histories with up to five spells means 5729 individuals (almost 64% of the sample) with a total of 30104 spells.

The duration analysis of the labor histories is carried out in three stages:

1. Separate analysis for each successive event. The first point we would like to note is that this approach makes no assumptions about independence between the spells on the same individual. However, it is valid if the hazard function of each spell does not depend on unobserved variables common or correlated across spells. On the other hand, there are a lot of parameters to estimate and the results are difficult to interpret. Indeed the effect of a given covariate may vary greatly from one spell to another. An additional problem is the length bias: the duration of the fifth episodes will probably be shorter because these belong to individuals who already had four episodes in the period 1980-1993.

2. Treating each spell as a distinct observation, pooling all the episodes and estimating a single model. In this analysis one has to take into account that spells coming from the same individual tend to be more alike than two randomly chosen observations. Therefore, not taking into account this information means that some unobserved heterogeneity in the sample is not included in the analysis.

3. Correcting the possible dependence among the spells because they belong to the same labor history. A first approach to detect the dependence among spells is suggested by Allison (1998). It is a simple *ad-hoc* method consisting in estimating a model for a certain duration where the lengths of the previous spells are included as covariates. The estimated coefficients of these variables give an idea about their significance.

   Once the three stages just described are accomplished, a method for estimating multivariate survival data is needed. In our analysis we use the procedure proposed by Wei, Lin & Weissfeld (1989) taking into account the dependence among the observations. It is based on the Cox proportional hazards model (see Chapter 1) and these authors have shown that the resulting estimates are asymptotically normal with a covariance matrix that can be consistently estimated.

5.4.1 Separate analysis

From the five spells histories we carried out five duration analyses according to the number of the episode. Thus we had five subsambles defined from the values of the variable *nepis* with dependent variables *dur1, dur2, dur3, dur4* and *dur5*, respectively. The main descriptive statistics of these variables are in Table 5.6.

The duration analyses were carried out assuming parametric models with a set of covariates related to individual characteristics (*sex* and *age* categorized into the categories already defined in Section 5.2), economic indicators (*unemp* and *gdp*) and the features of the current spell (*w-e, pre84, equarter1-equarter4, squarter1-squarter4* and *trans*) all of them defined in Section 5.2.

The fitted model in each case assumes a Weibull distribution. Tables 5.7-5.11 give the coefficient estimates and the associated statistics for the log-linear models with dependent variables $ldur1 = \log(dur1), ldur2 = \log(dur2), ldur3 = \log(dur3), ldur4 = \log(dur4)$ and $ldur5 = \log(dur5)$, respectively. However, note that is possible to convert these Weibull estimates to the estimates of the log-hazard function by dividing by the scale estimate and changing the sign. Therefore we have proportional hazards models for *dur1* to *dur5*.

From these separate analysis some of the variables have a similar behavior for the five spells (*age, unemp, pre84*), while others are quite different (*w-e, trans*). The remaining variables are significant for some spells and non-significant for others (*sex, gdp, equarter squarter*). One of the relevant coefficients is for the variable *pre84* which gives an expected duration 75% longer for the spells started prior to 1984. The transition at the end of episode is also a very significant variable: going

Table 5.6: Descriptive statistics of the variables *dur1-dur5*

variable	n	Mean	StDev	Q_1	Median	Q_3	Maximum
dur1	8986	433.47	655.56	69	182	521	4958
dur2	6873	341.93	566.99	52	130	364	4919
dur3	5590	326.28	514.76	41	122	366	4717
dur4	4705	307.73	490.29	45	125	365	4762
dur5	3950	313.16	488.60	43	124	361	4323

to wage-earner shortens the duration of the first job but increases the duration of the successive episodes. There is also a seasonal effect, given by variables *equarter, squarter* which gives longer duration to the episodes starting on the third quarter of the year. From the macro-economic indicators we note the positive effect of the rate of unemployment as well as the gross domestic product. Finally, for the individual characteristics gender is significant only for the third and fourth spells while the effect of *age* decreases with the number of the episode.

The shape of the hazard function, that is the rate of leaving an episode, is decreasing for all the spells, with $\alpha = -0.276, -0.288, -0.306, -0.275$ and -0.291, respectively. This coefficient is computed using the relationship $\alpha = (1/\sigma) - 1$ and can be interpreted as follows for the first spell: a 1% increase in the duration of the episode produces a 0.28% decrease in the hazard of leaving the first spell.

5.4.2 Pooled analysis

Here we consider the 30104 observations as a sample of randomly chosen observations. Therefore we fitted a single model with dependent variable *dur* irrespective of the number of spell. We used the same set of covariates as in the previous section and again a Weibull model was assumed. The estimated coefficients and the statistics are given in Table 5.12.

For this large sample all variables except the gender of individuals are highly significant. The duration of a certain episode of labor histories is increased by starting prior 1984, ending in the first or the third quarter with a transition to wage-earner and with the increase of unemployment and gross domestic product. On the other hand, *age*, having a wage-earning job and starting in the third quarter tends to short the duration of the episodes.

5.4.3 Correcting the dependence

In this section we consider two analyses. First, the main goal is to detect if there is dependence or not among the length of the spells belonging to the same individual. To do this we establish several models for each spell where the duration of the previous episodes are introduced as covariates. In the second analysis we computed corrected pooled estimates taking into account for the dependencies among durations coming from the same labor history. We also tested if there were significant

Table 5.7: Estimates for *dur1*

variable	Estimate	Std Error	$p-$value
intercept	6.552	0.27	0.0001
sex	–0.041	0.03	0.2333
age1	–2.662	0.10	0.0001
age2	–2.453	0.10	0.0001
age3	–2.270	0.12	0.0001
age4	–1.899	0.11	0.0001
age5	–1.489	0.12	0.0001
unemp	2.085	0.70	0.0031
gdp	0.004	0.01	0.7325
w-e	1.847	0.21	0.0001
pre84	0.559	0.05	0.0001
equarter1	0.138	0.05	0.0041
equarter2	0.004	0.05	0.9256
equarter3	1.419	0.05	0.0001
squarter1	0.054	0.05	0.2625
squarter2	–0.064	0.05	0.1716
squarter3	–0.367	0.05	0.0001
trans	–0.377	0.04	0.0001
Scale	1.381	.01	

NOTE: The log-likelihood for the Weibull model is -14708.62

Table 5.8: Estimates for *dur2*

variable	Estimate	Std Error	$p-$value
intercept	4.301	0.23	0.0001
sex	0.057	0.04	0.1390
age1	–1.066	0.13	0.0001
age2	–0.967	0.14	0.0001
age3	–0.653	0.15	0.0001
age4	–0.659	0.15	0.0001
age5	–0.483	0.16	0.0022
unemp	4.857	0.84	0.0001
gdp	0.023	0.01	0.0796
w-e	–1.037	0.05	0.0001
pre84	0.796	0.06	0.0001
equarter1	0.119	0.05	0.0266
equarter2	0.140	0.05	0.0086
equarter3	1.210	0.06	0.0001
squarter1	–0.023	0.05	0.6610
squarter2	–0.013	0.05	0.8202
squarter3	–0.041	0.05	0.4227
trans	1.041	0.04	0.0001
Scale	1.404	0.01	

NOTE: The log-likelihood for the Weibull model is −11578.67

Table 5.9: Estimates for *dur3*

variable	Estimate	Std Error	$p-$value
intercept	4.711	0.27	0.0001
sex	–0.149	0.04	0.0006
age1	–0.968	0.17	0.0001
age2	–0.822	0.17	0.0001
age3	–0.568	0.18	0.0019
age4	–0.640	0.19	0.0007
age5	–0.434	0.19	0.0257
unemp	5.303	0.97	0.0001
gdp	0.073	0.01	0.0001
w-e	0.251	0.05	0.0001
pre84	0.654	0.07	0.0001
equarter1	0.165	0.06	0.0063
equarter2	–0.023	0.06	0.6980
equarter3	1.430	0.06	0.0001
squarter1	–0.001	0.05	0.9823
squarter2	–0.014	0.06	0.8148
squarter3	–0.275	0.06	0.0001
trans	0.207	0.04	0.0001
Scale	1.442	.02	

NOTE: The log-likelihood for the Weibull model is -9936.65

Table 5.10: Estimates for *dur4*

variable	Estimate	Std Error	$p-$value
intercept	3.985	0.29	0.0001
sex	0.107	0.04	0.0169
age1	–0.694	0.19	0.0004
age2	–0.632	0.20	0.0013
age3	–0.428	0.20	0.0380
age4	–0.346	0.21	0.1070
age5	–0.055	0.22	0.8024
unemp	5.020	1.05	0.0001
gdp	0.042	0.01	0.0034
w-e	–0.182	0.05	0.0003
pre84	0.703	0.07	0.0001
equarter1	0.280	0.06	0.0001
equarter2	0.035	0.06	0.5687
equarter3	1.252	0.07	0.0001
squarter1	–0.004	0.06	0.9442
squarter2	–0.020	0.06	0.7516
squarter3	–0.111	0.06	0.0677
trans	0.576	0.04	0.0001
Scale	1.380	0.02	

NOTE: The log-likelihood for the Weibull model is −8078.21

Table 5.11: Estimates for *dur5*

variable	Estimate	Std Error	$p-$value
intercept	4.349	0.36	0.0001
sex	–0.014	0.05	0.7712
age1	–0.605	0.25	0.0142
age2	–0.441	0.25	0.0743
age3	–0.333	0.25	0.1922
age4	–0.238	0.26	0.3677
age5	–0.205	0.28	0.4576
unemp	2.606	1.22	0.0335
gdp	0.054	0.01	0.0006
w-e	0.107	0.05	0.735
pre84	0.359	0.08	0.0001
equarter1	0.026	0.07	0.7149
equarter2	–0.037	0.07	0.5957
equarter3	1.376	0.07	0.0001
squarter1	0.250	0.07	0.0003
squarter2	0.121	0.07	0.0841
squarter3	–0.224	0.07	0.0014
trans	0.421	0.05	0.0001
Scale	1.411	0.02	

NOTE: The log-likelihood for the Weibull model is -6801.49

Table 5.12: Estimates for the Weibull model of the pooled observations

variable	Estimate	Std Error	$p-$value
intercept	5.899	0.11	0.0001
sex	0.014	0.02	0.4492
age1	–1.892	0.07	0.0001
age2	–1.791	0.07	0.0001
age3	–1.583	0.07	0.0001
age4	–1.430	0.07	0.0001
age5	–1.148	0.08	0.0001
unemp	3.702	0.41	0.0001
gdp	0.043	0.01	0.0001
w-e	–0.086	0.02	0.0001
pre84	0.752	0.03	0.0001
equarter1	0.101	0.03	0.0001
equarter2	0.024	0.03	0.3445
equarter3	1.423	0.03	0.0001
squarter1	0.025	0.03	0.3233
squarter2	–0.011	0.03	0.6846
squarter3	–0.192	0.03	0.0001
trans	0.328	0.02	0.0001
Scale	1.431	0.01	

NOTE: The log-likelihood for the Weibull model is -51940.61

differences of the parameters among the episodes of a labor history.

In order to detect if the duration of a given spell depends on the length of the previous ones we can formulate a model where the set of covariates includes the durations of the episodes prior to the analyzed one. Thus we estimate a Weibull model for each of the durations *dur2*-*dur5* and, for instance, in the analysis of the third episode the dependent variable of the log-linear model is *ldur3* while variables *ldur1* and *ldur2* are covariates.

The results for these analyses are given in Tables 5.13, 5.14, 5.15 5.16. The first point we note is the smaller log-likelihoods for the fitted models including the logarithms of the previous durations. Therefore if we use the Akaike information criterion (AIC) for choosing a simple model which fits the data, we find that the more appropriate parametric models are those with variables *ldur1*-*ldur4* as covariates. However, we emphasize that the effect of these covariates on the duration varies according to the episode number. Indeed, for the fifth episodes, the duration of the fourth episode is not significant. On the other hand, the estimated coefficients of the previous durations have not the same sign for all the episodes.

From these results we see that there is some kind of dependence of the previous durations when a given spell is analyzed. However it is difficult to understand this dependence. The method introduced by Wei, Lin & Weissfeld (1989) gives a solution when sequences of spells have to be considered. The procedure is based on modelling each marginal distribution by a Cox proportional hazards model. Thus without imposing any structure on the dependence among spells the method allows the computation of estimates of the regression coefficients which are asymptotically normal with robust variance estimates. Therefore the obtained estimates are computed using the pooled sample but taking into account for the possible dependence among observations.

To implement this methodology we used a SAS code described by Allison (1998) where the output contains the corrected estimates of the coefficients, the standard errors and the statistics for testing the null hypothesis that for each covariate all coefficients, one for each spell of a labor history, are equal. The results for our data are in Table 5.17. Note that the estimated values of the coefficients correspond to the log-hazard model. Therefore, in order to see differences with the pooled estimations we have to divide by σ and change the sign of the coefficients in Table 5.12. The main point is that coefficients tend to be greater than in the naive pooled estimates for

Table 5.13: Estimates for the Weibull model with covariate *ldur1*

variable	Estimate	Std Error	$p-$value
intercept	4.887	0.24	0.0001
sex	0.039	0.04	0.3008
age1	–1.179	0.13	0.0001
age2	–1.068	0.14	0.0001
age3	–0.727	0.15	0.0001
age4	–0.743	0.15	0.0001
age5	–0.537	0.16	0.0006
unemp	5.062	0.84	0.0001
gdp	0.020	0.01	0.1189
w-e	–0.865	0.05	0.0001
pre84	0.786	0.06	0.0001
equarter1	0.123	0.05	0.0218
equarter2	0.140	0.05	0.0082
equarter3	1.210	0.06	0.0001
squarter1	–0.008	0.05	0.8718
squarter2	–0.015	0.05	0.7910
squarter3	–0.051	0.05	0.3171
trans	1.047	0.04	0.0001
ldur1	–0.110	0.01	0.0001
Scale	1.400	0.01	

NOTE: The log-likelihood for the Weibull model is −11547.10

Table 5.14: Estimates for the Weibull model with covariates *ldur1, ldur2*

variable	Estimate	Std Error	$p-$value
intercept	3.967	0.29	0.0001
sex	–0.123	0.04	0.0043
age1	–0.787	0.17	0.0001
age2	–0.671	0.17	0.0001
age3	–0.497	0.18	0.0063
age4	–0.509	0.18	0.0069
age5	–0.372	0.19	0.0534
unemp	4.489	0.97	0.0001
gdp	0.075	0.01	0.0001
w-e	0.342	0.05	0.0001
pre84	0.667	0.06	0.0001
equarter1	0.129	0.06	0.0304
equarter2	–0.013	0.06	0.8195
equarter3	1.408	0.06	0.0001
squarter1	–0.007	0.05	0.9065
squarter2	–0.016	0.06	0.7801
squarter3	–0.281	0.06	0.0001
trans	0.194	0.04	0.0001
ldur1	–0.110	0.01	0.0001
ldur2	–0.036	0.01	0.0077
Scale	1.430	0.01	

NOTE: The log-likelihood for the Weibull model is −9862.71

Table 5.15: Estimates for the Weibull model with covariate *ldur1-ldur3*

variable	Estimate	Std Error	$p-$value
intercept	3.902	0.32	0.0001
sex	0.091	0.04	0.0405
age1	–0.628	0.19	0.0012
age2	–0.556	0.19	0.0045
age3	–0.354	0.20	0.0848
age4	–0.277	0.21	0.1946
age5	–0.016	0.22	0.9406
unemp	5.034	1.05	0.0001
gdp	0.036	0.01	0.0100
w-e	–0.155	0.05	0.0001
pre84	0.711	0.07	0.0001
equarter1	0.235	0.06	0.0002
equarter2	0.029	0.06	0.6371
equarter3	1.245	0.07	0.0001
squarter1	0.000	0.06	0.9965
squarter2	–0.032	0.06	0.6098
squarter3	–0.131	0.06	0.0302
trans	0.596	0.04	0.0001
ldur1	0.050	0.02	0.0019
ldur2	0.062	0.01	0.0001
ldur3	–0.109	0.01	0.0001
Scale	1.371	0.02	

NOTE: The log-likelihood for the Weibull model is -8037.09

Table 5.16: Estimates for the Weibull model with covariate *ldur1-ldur4*

variable	Estimate	Std Error	$p-$value
intercept	3.074	0.40	0.0001
sex	0.038	0.05	0.4558
age1	–0.368	0.24	0.1310
age2	–0.253	0.24	0.2997
age3	–0.192	0.25	0.4452
age4	–0.093	0.26	0.7210
age5	–0.063	0.27	0.8149
unemp	1.693	1.22	0.1645
gdp	0.062	0.01	0.0001
w-e	0.117	0.06	0.0484
pre84	0.378	0.08	0.0001
equarter1	0.028	0.07	0.6851
equarter2	–0.052	0.07	0.4460
equarter3	1.355	0.07	0.0001
squarter1	0.248	0.07	0.0003
squarter2	0.133	0.07	0.0559
squarter3	–0.211	0.06	0.0023
trans	0.421	0.05	0.0001
ldur1	0.115	0.01	0.0001
ldur2	0.001	0.01	0.9713
ldur3	0.145	0.01	0.0001
ldur4	–0.025	0.02	0.1223
Scale	1.389	0.02	

NOTE: The log-likelihood for the Weibull model is -6738.94

Table 5.17: Corrected estimates for the five early spells

variable	Estimate	Std Error	2-sided $p-$value
sex	0.011	0.01	0.3911
age1	1.153	0.04	0.0000
age2	1.041	0.04	0.0000
age3	0.843	0.05	0.0000
age4	0.779	0.05	0.0000
age5	0.608	0.05	0.0000
unemp	–2.159	0.29	0.0000
gdp	-0.013	0.00	0.0020
w-e	0.136	0.02	0.0000
pre84	–0.349	0.02	0.0000
equarter1	–0.092	0.02	0.0000
equarter2	–0.017	0.02	0.3054
equarter3	–0.889	0.02	0.0000
squarter1	–0.023	0.02	0.1693
squarter2	0.013	0.02	0.4469
squarter3	0.140	0.02	0.0000
trans	-0.168	0.01	0.0000

the variables related to the kind of spell while for variables related to the individuals the coefficients tend to be smaller. This is emphasized for the variables *age*, *w-e* and *trans*.

5.5 Conclusions

In this section we are going to summarize the results we have obtained from the analysis of the Spanish labor histories in the period 1980-1993.

The sample we used comes from the social security dataset and it contains the labor spells of individuals starting to work in the analyzed period. We emphasize the richness on longitudinal information of these database compare with the standard data coming from the EPA and the ECPF described in the introduction of these second part. These data has been used very few times before our analysis, we emphasize the papers by García-Fontes & Hopenhayn (1996) and García-Pérez (1997). The data consist of contributed spells to the social security system, being available the type of contribution, the starting and ending date and the transition at the end of the spell. Therefore it has been possible to analyze the duration of three different types of episodes: self-employment, wage-earning and non-working taking into account for the destination state at the end of the spell (competing risks models). Moreover we also have considered the longitudinal information available in the dataset in order to analyze not only a single duration but a set of duration per individual. As a first analysis for a given duration we have introduced the duration of the previous spells as covariates. Afterwards a model taking into account that several durations belong to the same individual history has been estimated. As far as we know there are very few papers about analysis of multiple duration and competing risks for Spanish labor data. We just emphasize two papers both of them dealing with unemployment episodes: Gil, Martin & Serrat (1994) establishing a competing risks model and Arranz & Muro (1999) analyzing three episodes of unemployment. Thus we emphasize the contribution of the analysis already presented in this chapter because: first, it is analyzed the duration of episodes which are the first spells of labor histories; second, we made the analysis of the duration of three types of spells of the labor market using previous history and calendar variables as covariates and distinguishing according to their transitions at the end; finally, in the last analysis we have considered a methodology for analyzing multiple durations allowing for possible dependencies among them. Standard analysis previous to this one pooled all the observations and established duration models with all the observations.

The statistical analysis of the data is focused on the duration of the spells. We stated parametric models for the duration conditioned to a set of covariates which are related to individual characteristics, the spell and the previous labor history, and economic indicators. We carried out three different analysis:

1. Duration of the first episodes of the labor history. From the descriptive analysis of Chapter 4 we saw that theses episodes deserve special attention.

2. Duration of the episodes according to the type of them. In the data there are available three types of episodes: self-employment, wage-earner and non-working.

3. Duration of the five early episodes as longitudinal data, that is taking into account the possible correlations among observations which belong to the same individual.

In the first analysis we have a sample of durations, one for each individual. Thus we carried out a standard duration analysis, that is, we stated a log-linear model where the log-transformed duration is linearly related with a set of covariates. We assumed a Weibull distribution for the duration of the spell. The fitted model gave a decreasing hazard function, therefore larger durations have smaller probability of leaving.

In the second analysis, we stated separate models for each of the subsamples according the type of episodes. Here we have considered all the spells which are not a first one and we introduced as covariates the labor history previous to the current spell. Moreover we also used competing risks models in order to see differences among destination states at the end of a certain spell.

The third analysis compare three situations for the labor histories of five episodes: Analyzing each of the episodes separately, that is, the first ones, the second ones, and so on; the pooled analysis where all the observations have been analyzed together as independent ones; and the analysis correcting possible dependencies due to several observations come from the same individual.

The most relevant conclusions of each of the analyses are described as follows:

- **First spell**. We fitted a Weibull model with a decreasing hazard function for the duration. The statistical significant covariates are age, rate of unemployment at the beginning of the spell, starting date, the cause of finishing the spell and the transition at the end. Larger spells correspond to those starting prior 1984, having a wage-earning destination state and ending due to a non-volunteer cause. Related to other characteristics, older people have longer durations, and the rate of unemployment at the beginning of the episode has a positive sign in the duration.

- **Type of spell**. We fitted several models according to the type of spell. We assumed a generalized gamma model for the duration of self-employment spells, while for the wage-earner we assumed a log-normal model with a non-monotone hazard function. Finally for the non-working spells, we assumed a Weibull model with an increasing hazard function. The significant variables for the three models are related with the destination state and the previous labor history (type of previous episodes, the total time elapsed in other states than the analyzed and the number of times has been on each type of episode). The competing risks analyses allowed differences among the destination states. For the self-employment spells we found differences according to the transition is working or non-working. For the wage-earning spells variable sex become significant with a positive coefficient when the transition is to self-employment, and age is significant for the transition to non-working. Finally for non-working spells, sex is significant with a positive value for the transition to self-employment and the previous duration is significant for the wage-earning destination.

- **Five early spells**. In this analysis we proceed in three stages. First we carried out separate analyses for episodes according the place on the labor sequence. Here we obtained a high variation among the estimated coefficients for the variables depending on the episode correspond to the first, the second or so on. In the second step we performed estimation on the pooled observations, that is, without taking into account that some of them belonged to the same labor history. For this analysis we found positive coefficients for age, starting date prior 1984 and transition, and negative coefficient for the wage-earner spells. Third, we analyzed the data introducing the possible dependence between durations using the previous durations as a covariates and we obtained significant coefficients for the previous durations. Therefore we fitted a model taking into account the dependencies among observations and we obtained corrected estimates of the coefficients and standard errors. The main factors which lead to larger durations are: not to be a wage-earning spell, starting date prior 1984, to have a transition to a wage-earning or non-working, ending in the first or third quarter, low rates of unemployment at the beginning of the spell. With respect to individual characteristics we does not find significant differences between males and females and older people has larger durations than younger subjects.

Even though duration analysis has been largely used for Spanish data, overall for unemployment data, as far as we know the analysis of several durations for each individual has been only used in Arranz & Muro (1999). The study we have carried out contributes in two directions to the analysis of the Spanish labor data. On the one hand, we have considered competing risks models to see differences among transitions at the end of each episode of the labor history. On the other hand, we have fitted a model for more than one duration per individual taking into account the possible correlations existing among them.

5.6 Appendix A: The Weibull regression

Consider a duration variable T which has a Weibull distribution with parameters γ and α. That is the density function of T is given by

$$f(t) = \gamma\, \alpha\, t^{\alpha-1}\, \exp(-\gamma t^{\alpha}).$$

When in addition to the duration variable is also collected a set of explanatory variables or covariates, we may establish a regression model to analyzed the effects of these factors on the duration. A usual model is the log-linear model where the response variables is $Y = \log T$ which is linearly related with $\mathbf{X}$ a vector of the covariates, that is

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \sigma w_i, \quad i = 1, \cdots, n \tag{5.13}$$

where w_i is a random disturbance term and $\boldsymbol{\beta}$ and σ are parameters to be estimated. Note that when the disturbance term has a standard extreme value distribution (see Blossfeld, Hamerle, Mayer, 1989) the duration variable T follows a Weibull distribution with parameters $\gamma = \exp(\mathbf{x}_i' \boldsymbol{\beta}^{\star})$ with $\beta_k{}^{\star} = -\beta_k/\sigma$, $k = 1, \cdots, p$, and $\alpha = 1/\sigma$.

5.7 Appendix B: Brief introduction to the SAS® System

The results obtained in this part are carried out with the SAS® system (`www.sas.com`) which is an integrated system of software providing complete control over data access, management, analysis and presentation.

The SAS/STAT software contains three procedures concerned with the analysis of continuous duration data:

- **LIFETEST** is designed for the analysis of univariate time data and produces life tables and graphs of the estimated survival functions. It does not produce estimates of the parameters.
- **LIFEREG** estimates regression models with censored continuous time data under several alternative distributional assumptions.
- **PHREG** uses Cox's partial likelihood method to estimate regression models with censored data.

5.7.1 The LIFETEST procedure

We mainly used this procedure in Chapter 4. It allowed to obtain the product limit estimator using only the time values and the censoring indicator. The following SAS code shows how to get the product limit estimator:

```
proc lifetest;
time dur*cens(0);
run;
```

where `dur` is the time variable (the value may be complete or censored) and `cens(0)` is the indicator of censoring with the value that corresponds to a censored observation in parentheses.

Besides the estimated survival function $\hat{S}(t)$, with this procedure you can also get the plots of the log survival, $-\log \hat{S}(t)$, the log-log survival, $\log(-\log \hat{S}(t))$, and the empirical hazard. Moreover we can test the differences in survival functions between groups as well as test whether quantitative covariates are associated with survival time.

5.7.2 The LIFEREG and PHREG procedure

Both of these procedures are used for estimating regression models with a censored dependent variables related to a set of covariates.

Because we have mainly used the LIFEREG procedure in the previous chapters, we are going to focus on it. It is used for computing the estimates of the accelerated failure time models using the method of maximum likelihood. Indeed, LIFEREG estimates a log-linear model where the dependent variable is $\log T$, that is

$$\log T_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\epsilon_i, \quad i = 1, \cdots, n \tag{5.14}$$

where ϵ_i is a random disturbance term and $(\boldsymbol{\beta}, \sigma)$ are the parameters to be estimated. The SAS code to estimate this model is for the general case:

```
proc lifereg;
model dur*cens(0)=var1 var2 var3 ... /dist= ;
run;
```

where `var1 var2 var3...` are the covariates contained in the vector $\mathbf{x}_i$ and `dist=` expects the assumed distribution for T, for instance `dist=` Weibull.

Here we note that the output obtained from LIFEREG is related to $\log T$, however in survival analysis it is also of interest to control the effect of the covariates on the hazard function, $\lambda(t)$. To be able to do this we have to know the relationship between the coefficients in equation (5.14) and

$$\log \lambda(t) = g(t; \psi) + \mathbf{x}_i'\boldsymbol{\beta}^\star; \tag{5.15}$$

The two special cases of exponential and Weibull distributions come from this equation. For the exponential distribution $g(t; \psi) = 1$ and $\boldsymbol{\beta}^\star = -\boldsymbol{\beta}$, and for the Weibull model $g(t; \psi) = \alpha \log t$ and $\boldsymbol{\beta}^\star = -\boldsymbol{\beta}/\sigma$

Moreover, for the estimated values of the coefficients in the output of LIFEREG there are also the Wald tests for testing the hypothesis that each coefficient is 0. These are obtained by dividing each coefficient by its standard error and squaring the result.

More extended details about these procedures are given by Allison (1998).

5.8 Appendix C: Generalized gamma model

This distribution was firstly proposed by Stacey (1962) but was defined in a way which presented several problems in statistical inference procedures. Later a work due to Prentice (1974) proposed a re-parameterization of the density function and this eliminated some of the difficulties.

The generalized gamma distribution has a probability density function (p.d.f.) given by

$$f(t) = \frac{\alpha\, \lambda^{\gamma}\, t^{\alpha\gamma-1} \exp\left(-\lambda\, t^{\alpha}\right)}{\Gamma(\gamma)},\ t \geq 0 \tag{5.16}$$

with parameters $\alpha > 0$, $\gamma > 0$ and $\lambda > 0$. This family of distributions includes as special cases the exponential ($\alpha = \gamma = 1$), the Weibull ($\gamma = 1$), the Gamma ($\alpha = 1$) and tends to the log-normal as $\gamma \longrightarrow \infty$.

The re-parameterized function comes from considering $Y = \log T$ and setting $u = -\alpha^{-1} \log \lambda$ and $b = \alpha^{-1}$. It follows that

$$W = \frac{Y-u}{b}$$

has a log-gamma distribution with p.d.f.

$$\frac{\exp\left(\gamma w - e^{w}\right)}{\Gamma(\gamma)},\quad -\infty < w < \infty \tag{5.17}$$

From here, if we consider

$$W_1 = \gamma^{1/2}\left(W - \log\gamma\right) = \frac{Y-\mu}{\sigma}$$

where

$$\sigma = \frac{b}{\gamma^{1/2}} \quad \text{and} \quad \mu = u + b\log\gamma,$$

then the p.d.f. for W_1 only depends on parameter $\gamma > 0$ and it is given by,

$$f(w_1) = \frac{\gamma^{\gamma-1/2} \exp\left(\gamma^{1/2} w_1 - \gamma \exp(w_1 \gamma^{-1/2})\right)}{\Gamma(\gamma)},\quad -\infty < w_1 < \infty \tag{5.18}$$

From here the p.d.f. for $Y = \log T$ may easily be computed.

A further re-parameterization defines μ and σ as above but takes $g = \gamma^{-1/2}$. The generalization of this case to include $g < 0$ due to Prentice (1974) considers that W_1 has p.d.f. defined by

$$f(w_1) = \frac{|g|(g^{-2})^{g^{-2}} \exp\left(g^{-2}(g w_1 - e^{g w_1})\right)}{\Gamma(g^{-2})},\quad -\infty < w_1 < \infty \tag{5.19}$$

where $-\infty < g < \infty,\ g \neq 0$.[14] Using the transformations already described, we obtain the p.d.f. for our variable T of interest

$$f(t) = \frac{|g| \left(g^{-2}\, t^{g/\sigma}\, e^{-g\mu/\sigma}\right)^{g^{-2}} \exp\left(-g^{-2}\, t^{g/\sigma}\, e^{-g\mu/\sigma}\right)}{t\, \sigma\, \Gamma(g^{-2})}, \quad t \geq 0 \tag{5.20}$$

where g, μ and σ are parameters. Hence, the survival function is given by,

$$S(t) = \begin{cases} I(t^{g/\sigma}e^{-g\mu/\sigma}, g^{-2}) & \text{if} \quad g < 0 \\ 1 - I(t^{g/\sigma}e^{-g\mu/\sigma}, g^{-2}) & \text{if} \quad g > 0 \end{cases} \tag{5.21}$$

where $I(\cdot,\cdot)$ is the incomplete gamma function.[15] From the quotient of functions (5.20) and (5.21) the hazard function is obtained.

When some covariates are also available and we are interested in studying the effects of those variables on T, then parameter μ may be re-parameterized as $\mu = X'\beta$, where X denotes the vector of covariates and β the unknown parameters. In fact, this corresponds to assuming a linear relationship between $Y = \log T$ and X given by

$$Y = X'\beta + \epsilon \tag{5.22}$$

where ϵ is a random variable independent of X and with a p.d.f. defined in (5.18).

The effects of covariates on the hazard function are shown in Figure 5.1 and Figure 5.3. Indeed, for a given value of parameters g, σ and μ, the hazard function increases faster until the maximum is reached and then it decreases slowly to zero. Once we have fixed parameters g and σ, the effects of varying μ are mainly on the maximum. That is, for large values of μ (Figure 5.1) the hazard function takes very small values and the maximum is reached at value that increases with μ. However, for small values of μ (Figure 5.3), including negative values, the maximum of the hazard function is taken for t close to zero so that the hazard is almost decreasing for all t.

[14] More details about this are given by Lawless (1982).

[15] The incomplete gamma function for a general case of parameters a and b is defined by

$$I(a,b) = \frac{\int_0^a\ u^{b-1}\, e^{-u}\, du}{\Gamma(b)}.$$

Figure 5.1: Hazard function for $\mu = 5$

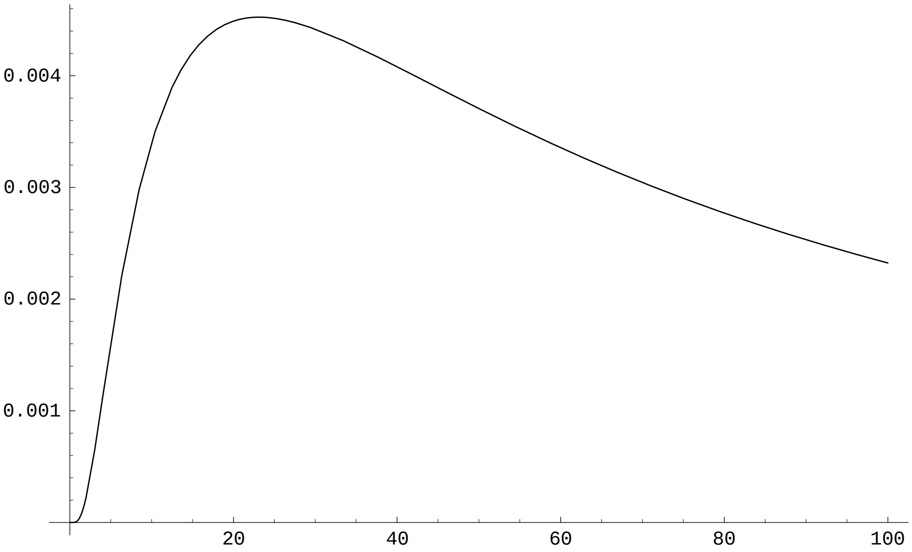


Figure 5.2: Hazard function for $\mu = 0$

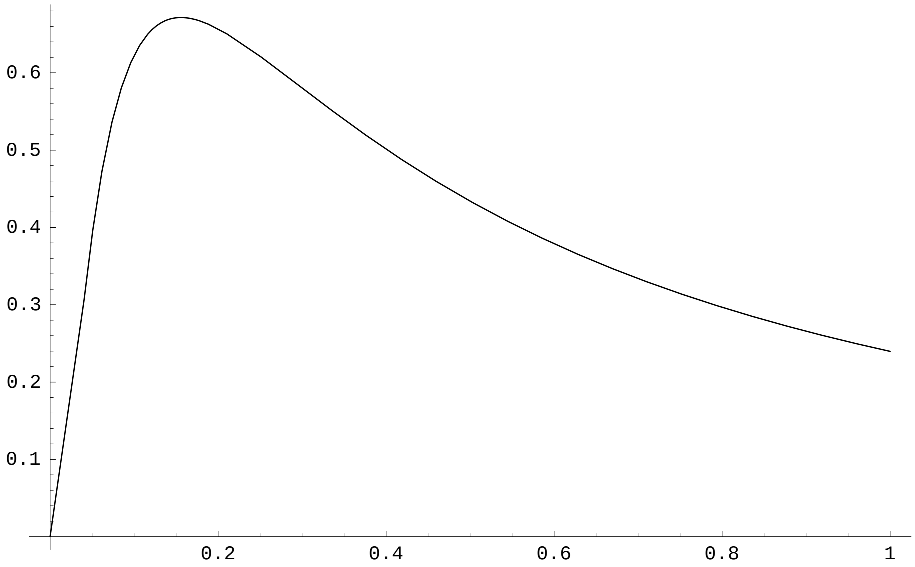
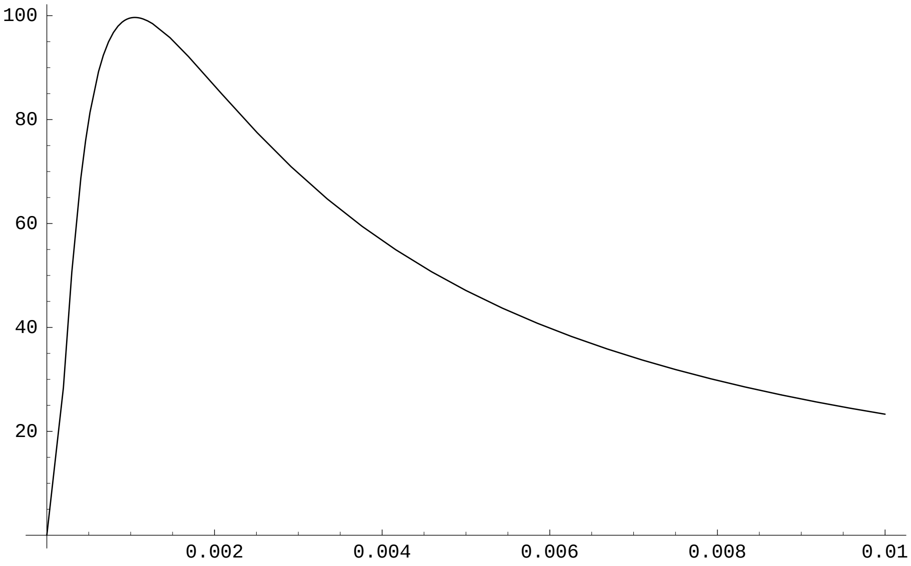


Figure 5.3: Hazard function for $\mu = -5$



5.9 Appendix D: Residual plots

Here we display the plot of the residuals for the fitted models in the chapter. In the graphics we have plotted the Cox-Snell residuals defined as $e_i = -\log \hat{S}(t_i/\mathbf{x}_i)$ where $\hat{S}(t)$ is the estimated probability of surviving to time t, based on the fitted model. If the fitted model is correct the e_i have approximately an exponential distribution with the parameter equals 1. Therefore the plot of residuals against $-\log \hat{S}(t)$, the Kaplan-Meier estimator of the survival function, should be approximately a straight line.

Figure 5.4: Residual plot for the Weibull model of *dur1*

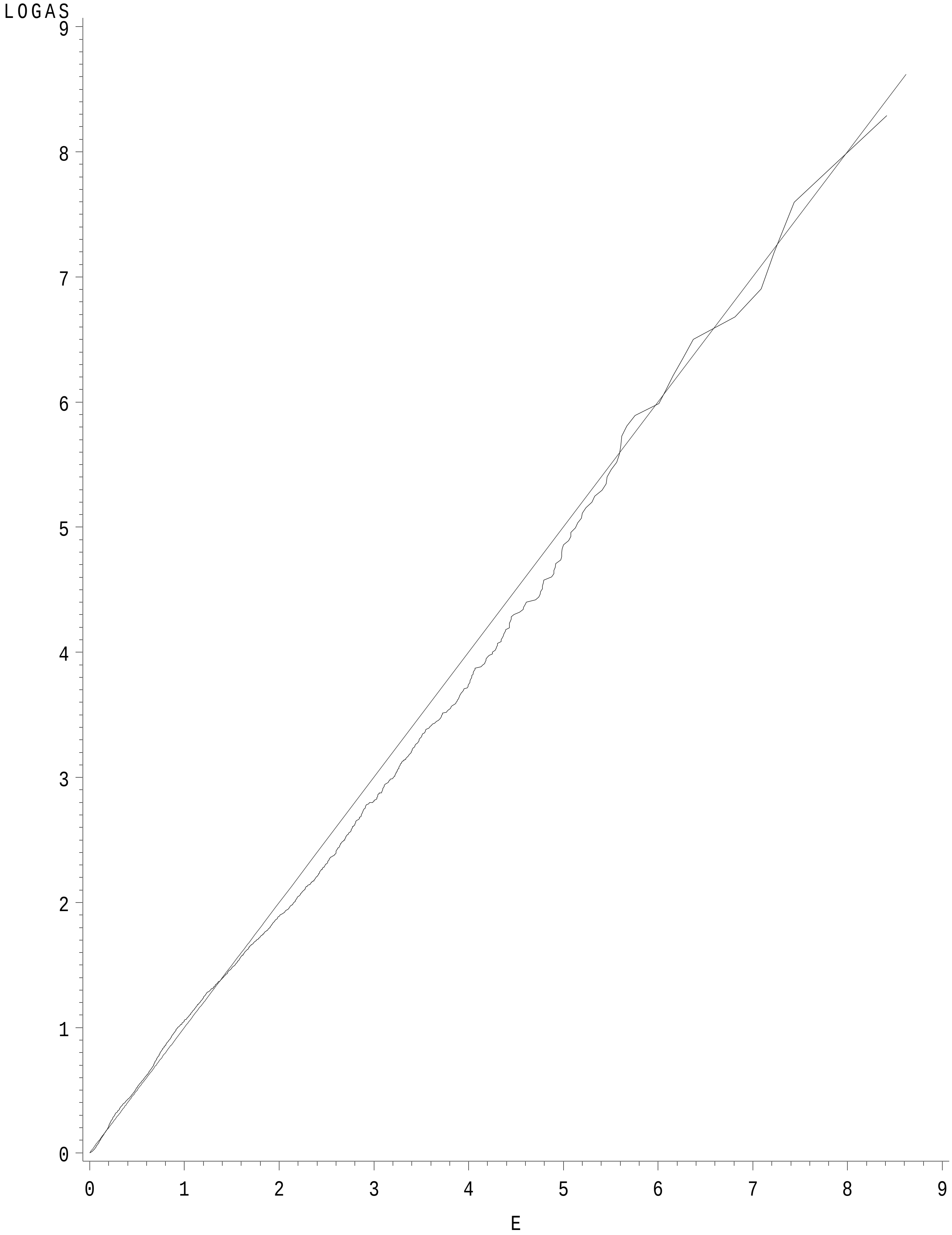


Figure 5.5: Residual plot for the generalized gamma model for the self-employment duration

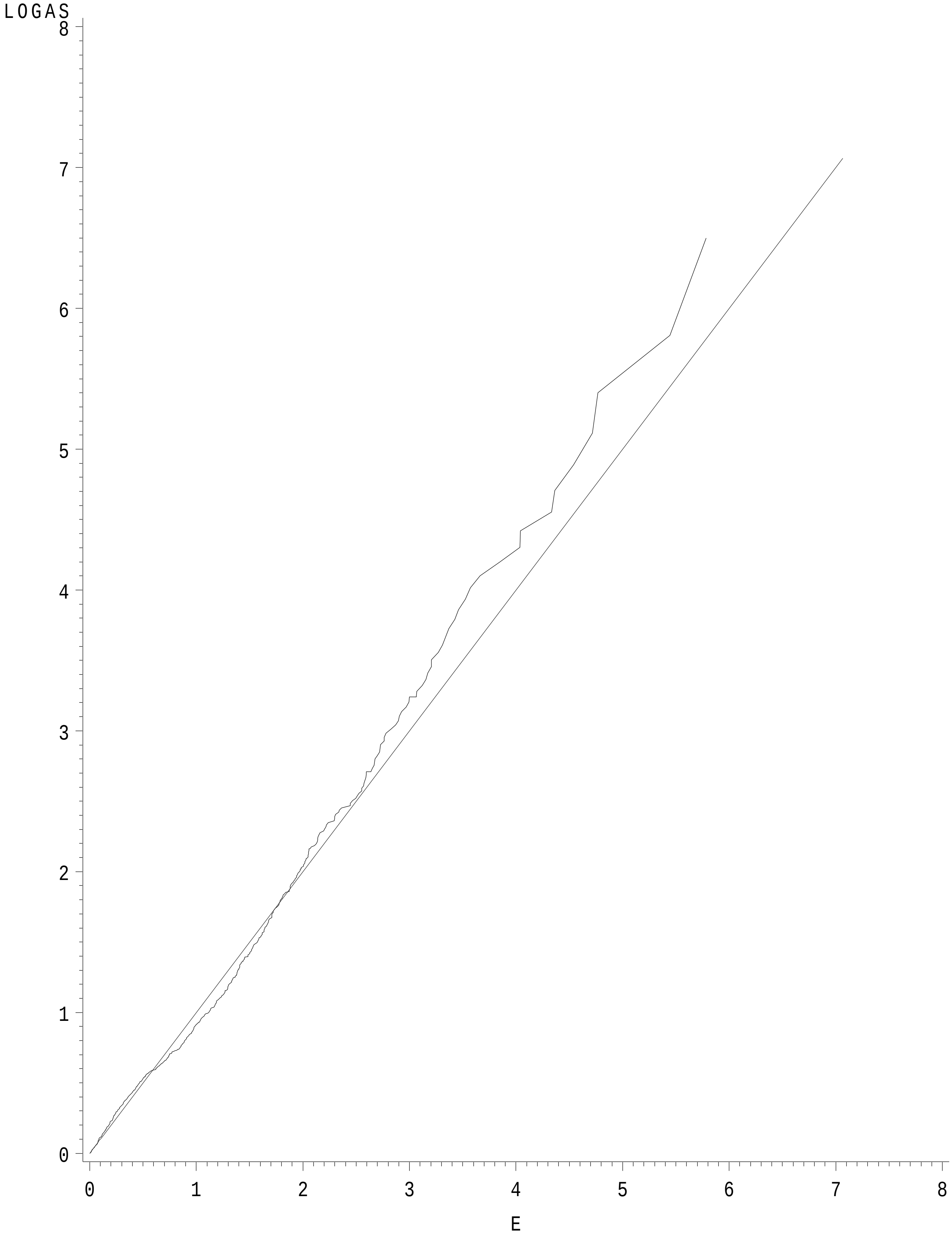


Figure 5.6: Residual plot for the log-normal model for the wage-earning duration

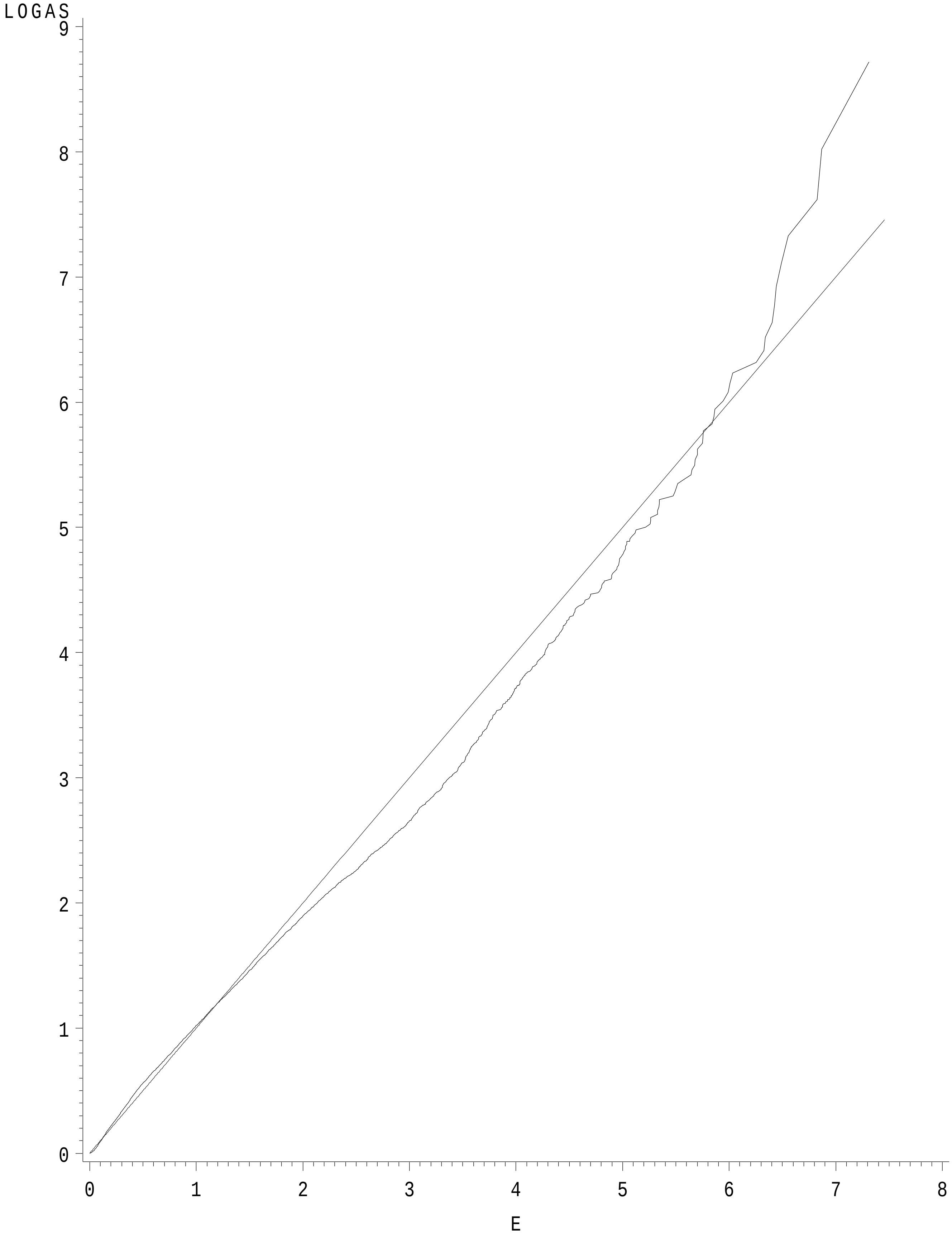
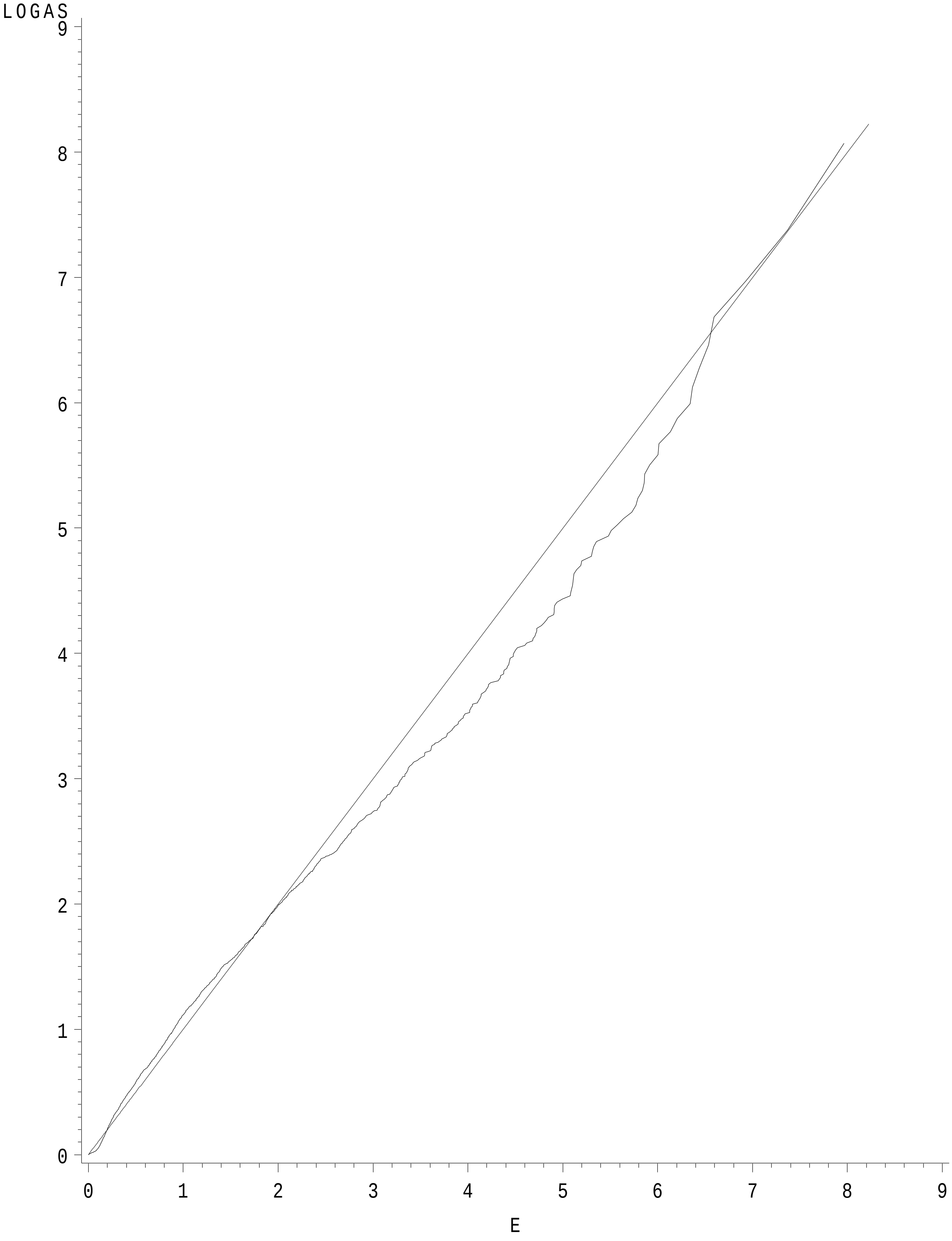


Figure 5.7: Residual plot for the Weibull model for the unemployment duration



5.10 Appendix E: Competing risks models for types of episodes

In the next tables there are the results of the estimates for the competing risks models introduced in Section 5.3.

5.10.1 Self-employment:

Table 5.18: Estimates for the transition self-employment to self-employment

variable	Estimate	Std Error	$p-$value
intercept	6.575	0.41	0.0001
sex	0.062	0.11	0.5815
age1	–0.217	0.20	0.2798
age2	–0.012	0.19	0.9504
age3	–0.073	0.21	0.7224
age4	–0.051	0.24	0.8277
age6	–0.406	0.38	0.7971
unemp1	–0.244	0.23	0.2980
unemp2	–0.154	0.18	0.3861
unemp3	–0.536	0.18	0.0026
unemp4	0.249	0.23	0.2864
unemp5	–0.060	0.20	0.7621
pre84	0.053	0.20	0.7971
empl=1	1.056	0.14	0.0001
empl=2	0.275	0.17	0.1022
nselfe	–0.433	0.06	0.0001
nw-ee	–0.081	0.02	0.0010
nnonce	0.087	0.04	0.0461
equarter1	0.223	0.16	0.1680
equarter2	–0.874	0.20	0.0001
equarter3	0.494	0.20	0.0130
squarter1	0.107	0.15	0.4849
squarter2	0.105	0.16	0.5238
squarter3	–0.010	0.16	0.9507
Scale	1.237	0.07	
Shape	0.345	0.19	

NOTE: The log-likelihood for the generalized gamma model is −1075.42

Table 5.19: Estimates for the transition self-employment to wage-earner

variable	Estimate	Std Error	p–value
intercept	8.245	1.07	0.0001
sex	0.331	0.23	0.1561
age1	–0.290	0.47	0.5348
age2	–0.475	0.44	0.2769
age3	–0.486	0.49	0.3254
age4	–0.291	0.51	0.5694
age6	–0.366	0.86	0.6715
unempl	–0.089	4.60	0.9845
pre84	0.199	0.32	0.5452
empl=1	–1.185	0.37	0.0014
empl=2	–0.998	0.37	0.0075
nselfe	1.128	0.27	0.0001
nw-ee	–0.076	0.04	0.0684
nnonce	0.023	0.08	0.7682
squarter1	0.723	0.26	0.0069
squarter2	0.624	0.27	0.0218
squarter3	–0.089	0.282	0.9845
Scale	1.1186	0.34	
Shape[a]	0.934	0.49	

[a] It is accepted a gamma model (shape=scale) with two parameters with shape=1.1

[a] NOTE: The log-likelihood for the generalized gamma model is −444.45. The fitted model does not accept the seasonal variables *equarter1-equarter3*

Table 5.20: Estimates for the transition self-employment to non-working

variable	Estimate	Std Error	p–value
intercept	9.609	0.82	0.0001
sex	–0.634	0.24	0.0088
age1	–1.545	0.49	0.0017
age2	–0.704	0.49	0.1476
age3	–0.331	0.53	0.5324
age4	–1.300	0.56	0.0196
age6	–1.229	0.76	0.1089
unemp1	–1.626	0.52	0.0020
unemp2	–1.750	0.40	0.0001
unemp3	–1.918	0.40	0.0001
unemp4	–1.342	0.51	0.0084
unemp5	–1.730	0.45	0.0001
pre84	0.546	0.43	0.2079
empl	0.063	0.15	0.6814
nselfe	–0.178	0.14	0.2034
nw-ee	0.087	0.07	0.2202
nnonce	–0.118	0.11	0.2861
equarter1	2.819	0.32	0.0001
equarter2	–0.017	0.36	0.9619
equarter3	0.598	0.32	0.0599
squarter1	1.081	0.35	0.0019
squarter2	0.763	0.35	0.0301
squarter3	0.199	0.34	0.5610
Scale	1.858	0.12	

NOTE: The log-likelihood for the log-normal model is -406.79

5.10.2 Wage-earner:

Table 5.21: Estimates for the transition wage-earner to self-employment

variable	Estimate	Std Error	$p-$value
intercept	12.152	0.94	0.0001
sex	0.716	0.19	0.0001
age1	–0.018	0.44	0.9677
age2	–0.052	0.43	0.9036
age3	–0.029	0.45	0.9479
age4	0.353	0.55	0.5182
age6	0.064	0.74	0.9307
unemp1	–0.599	0.41	0.1446
unemp2	–0.864	0.32	0.0077
unemp3	–0.736	0.33	0.0257
unemp4	–0.492	0.46	0.2904
unemp5	–0.528	0.37	0.1507
gdp	0.073	0.05	0.1653
pre84	0.535	0.36	0.1349
empl=1	–4.214	0.42	0.0001
empl=2	–0.327	0.20	0.0994
nselfe	0.511	0.46	0.2697
nw-ee	–0.012	0.03	0.7530
nnonce	–0.019	0.07	0.7784
lotimese	–0.320	0.17	0.0597
lotimewe	–0.167	0.09	0.0555
lotimenc	0.014	0.04	0.7480
equarter1	–0.062	0.23	0.7885
equarter2	0.146	0.24	0.5409
equarter3	1.329	0.23	0.0001
squarter1	0.332	0.24	0.1684
squarter2	0.168	0.25	0.4999
squarter3	0.044	0.23	0.8476
Scale	2.499	0.15	

NOTE: The log-likelihood for the log-normal model is −901.01

Table 5.22: Estimates for the transition wage-earner to wage-earner

variable	Estimate	Std Error	$p-$value
intercept	5.861	0.12	0.0001
sex	0.022	0.02	0.3233
age1	–0.174	0.05	0.0018
age2	–0.136	0.05	0.0132
age3	0.011	0.06	0.8507
age4	0.029	0.06	0.6588
age6	0.230	0.09	0.0143
unemp1	–0.101	0.05	0.0312
unemp2	–0.076	0.04	0.0394
unemp3	–0.078	0.04	0.0308
unemp4	0.053	0.06	0.3486
unemp5	–0.015	0.04	0.7305
gdp	–0.026	0.01	0.0001
pre84	–0.008	0.04	0.8504
empl=1	0.077	0.14	0.5687
empl=2	–0.579	0.03	0.0001
nselfe	–0.097	0.10	0.3521
nw-ee	–0.057	0.00	0.0001
nnonce	0.002	0.01	0.8184
lotimese	0.096	0.05	0.0617
lotimewe	0.099	0.01	0.0001
lotimenc	–0.015	0.01	0.0091
squarter1	–0.003	0.03	0.9193
squarter2	–0.012	0.03	0.7106
squarter3	–0.111	0.03	0.0004
Scale	1.434	0.01	
Shape	–1.138	0.23	

NOTE: The log-likelihood for the generalized gamma model is -19701.19

Table 5.23: Estimates for the transition wage-earner to non-working

variable	Estimate	Std Error	$p-$value
intercept	4.332	0.10	0.0001
sex	0.021	0.02	0.2845
age1	–0.288	0.05	0.0001
age2	–0.196	0.05	0.0001
age3	–0.095	0.05	0.0727
age4	–0.086	0.06	0.1407
age6	0.137	0.08	0.1047
unemp1	–0.077	0.04	0.0614
unemp2	–0.052	0.03	0.1045
unemp3	–0.044	0.03	0.1707
unemp4	–0.045	0.05	0.3621
unemp5	–0.093	0.04	0.0117
gdp	–0.010	0.01	0554
pre84	–0.031	0.04	0.3890
empl=1	0.158	0.12	0.1991
empl=2	–0.063	0.02	0.0084
nselfe	–0.083	0.08	0.2940
nw-ee	0.026	0.00	0.0001
nnonce	–0.118	0.01	0.0001
lotimese	0.048	0.04	0.1757
lotimewe	0.225	0.01	0.0001
lotimenc	–0.008	0.01	0.1414
equarter1	–0.215	0.03	0.0001
equarter2	–0.172	0.03	0.0001
equarter3	0.331	0.03	0.0001
squarter1	0.063	0.03	0.0392
squarter2	–0.171	0.03	0.0001
squarter3	–0.367	0.03	0.0001
Scale	1.286	0.01	
Shape	–1.618	0.04	

NOTE: The log-likelihood for the generalized gamma model is −20495.34

5.10.3 Non-working:

Table 5.24: Estimates for the transition non-working to self-employment

variable	Estimate	Std Error	$p-$value
intercept	8.356	2.29	0.0003
sex	0.919	0.23	0.0001
age1	–0.054	0.66	0.9340
age2	–0.465	0.63	0.4615
age3	–0.664	0.64	0.3031
age4	–0.464	0.68	0.4935
qual12	0.039	0.20	0.8472
qual34	–0.123	0.24	0.6094
pre84	0.738	0.26	0.0040
bempl3	0.477	0.59	0.4176
lodur	0.196	0.97	0.8402
lodur2	–0.020	0.09	0.8348
lotimewe	–0.238	0.13	0.0719
lotimenc	–0.010	0.04	0.7958
equarter1	0.296	0.26	0.2637
equarter2	0.120	0.24	0.6119
equarter3	1.747	0.28	0.0001
squarter1	–0.372	0.26	0.1572
squarter2	–0.524	0.27	0.0487
squarter3	–0.045	0.26	0.8656
Scale	0.860	0.06	

NOTE: The log-likelihood for the Weibull model is −432.83

Table 5.25: Estimates for the transition non-working to wage-earner

variable	Estimate	Std Error	$p-$value
intercept	7.066	0.50	0.0001
sex	–0.064	0.04	0.0884
age1	0.095	0.12	0.4147
age2	–0.021	0.11	0.8545
age3	0.051	0.12	0.6677
age4	0.109	0.12	0.3754
qual12	0.086	0.04	0.0323
qual34	–0.015	0.05	0.7515
unemp	3.239	0.82	0.0001
gdp	–0.035	0.01	0.0003
pre84	0.208	0.05	0.0001
bempl3	–0.001	0.17	0.9943
lodur	–0.805	0.20	0.0001
lodur2	0.077	0.02	0.0001
lotimewe	–0.154	0.03	0.0001
lotimenc	0.001	0.01	0.9477
equarter1	–0.040	0.06	0.5069
equarter2	0.389	0.06	0.0001
equarter3	0.850	0.06	0.0001
squarter1	–0.222	0.05	0.0001
squarter2	–0.510	0.06	0.0001
squarter3	–0.410	0.05	0.0001
Scale	1.140	0.02	
Shape	–1.156	0.07	

NOTE: The log-likelihood for the generalized gamma model is -5975.38

Conclusions and Future Research

This thesis has been structured in two parts with the common topic of the survival analysis. In the first part is proposed a consistent estimator for the regression coefficients of a censored linear model with measurement error on covariates. In the second part is analyzed a database about episodes of individual labor histories. The main goal has been the analysis of the duration of spells.

Results in Part I:

1. Using a Monte Carlo illustration we showed that ignoring the presence of measurement error on covariates lead to biases on the standard estimates of the regression coefficients of an accelerated failure time model. Moreover, even though only one of the covariates is affected by measurement error, the estimator of all coefficients are also biased.

2. We have proposed a methodology for obtaining consistent estimates for a linear model with a censored response and convariates contaminated with measurement error. It is named two-step estimator.

3. The two-step estimator modifies the standard procedures of estimation for linear measurement error models in order to account for censoring. In the first step is computing a consistent estimate, say $\hat{\boldsymbol{\kappa}}_{xy}$, of $\mathrm{E}(\mathbf{X}Y)$, being $\mathbf{X}$ the observed covariate and Y the complete response. The second step uses $\hat{\boldsymbol{\kappa}}_{xy}$ instead of $\mathrm{E}(\mathbf{X}Y)$ in the standard methods for estimating errors-in-variables models.

4. Standard errors have been computed using Bootstrap

5. The performance of the two-step estimator has been illustrated with a Monte

Carlo study varying sample sizes, amount of measurement error and percentage of censoring.

Results in Part II:

1. We have analyzed a database containing contributions to the social security, that is, sequences of individual labor spells.

2. Non-parametric estimates of the survival functions has been obtained for several sets of episodes: working versus non-working spells, first episodes compared with the next ones and according to the starting dates.

3. Comparison of estimated survival curves according to the transition at the end of the spell allow us to conclude that: the episodes of self-employment has higher survival than wage-earning, for the spells starting prior 1984 those of wage-earner has higher survival than non-working and lower survival than self-employment, and unemployment has higher survival than other causes of non-working spells.

4. Three statistical analysis have been carried out: for the first episodes, for the three type of episodes, and the five early spells taking into account for dependencies among observations.

5. The duration of the first episode, controlling for the other covariates, has been determined by: *age* where older individuals have longer duration, starting the episode prior 1984 lead to 81% larger durations, ending the spell in the third quarter of a year also increases the duration, and the transition to a non-working episode shorts the duration.

6. Differences have been found among type of episodes. Variables related with the previous history like number of previous episodes, total time spent on each type of episodes before the current one and the type of previous episode has been used as covariates. Controlling for the other covariates, the duration has been determined by: the starting date is significant for wage-earner and unemployment spells increasing the duration by 15%; previous experience on the same type of episodes increases the analyzed duration; the total number of previous episodes shorten the duration by 44% for the self-employment and

by 8% for the wage-earning; and the transition to non-working have longer durations than going to a working episode.

7. To analyze the labor histories with up to five spells we carried out three alternatives: Five separate analyses, one for each successive spell; a pooled analysis for all the observation; and pooled analysis controlling by the dependencies among observations.

8. From the corrected estimates for the five early spells, controlling for other variables, we conclude that: older people have longer durations; the risk of leaving a wage-earning spell is 14% greater than the risk of leaving from other episodes; the hazard of spells starting prior 1984 is only 70% of the hazard for those spells starting after 1984; and the risk to have a transition to wage-earner is about 85% of the risk to have a self-employment spells at the end of the current one.

Future research

The results and analyses presented in this thesis may be generalized on several aspects that will be the goals of our future research:

- Define the two-step estimator for the case of unknown covariance matrix of the measurement error.

- Generalize the two-step estimator to the case of multivariate regression model. That is to the case with several responses. In that case the second step could use the specialized methods of the factor analysis.

- Apply the generalized two-step estimator just mentioned to the analysis of the labor histories.

- Fitting competing risks models for the labor histories taking into account for the dependencies among observations.

Bibliography

[1] Aalen, O.O. (1978). *Nonparametric estimation of partial transitions probabilities in multiple decrement models.* The Annals of Statistics, 6, pp. 534-545.

[2] Adcock, R.J. (1877). *Note on the method of least squares.* Analyst, 4, pp. 183-184.

[3] Adcock, R.J. (1878). *A problem in least squares.* Analyst, 5, pp. 53-54.

[4] Allison, P.D. (1998). *Survival Analysis using the SAS system: A practical guide.* SAS Intitute Inc.

[5] Anh, N. & Ugidos-Olazabal, A. (1995). *Duration of unemployment in Spain: relative effects of unemployment benefit and family characteristics.* Oxford Bulletin of Economics and Statistics, 57, pp. 249-265.

[6] Andersen, P.K. (1988). *Statistical models for longitudinal labor market data based on counting processes.* In *longitudinal analysis of labor market data* by J.J. Heckman & B. Singer. Cambridge University Press.

[7] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical models based on counting processes.* New York: Springer-Verlag.

[8] Andrés, J. (1993). *La persistencia del desempleo agregado: una panorámica.* Moneda y Crédito, 197, pp: 91-127.

[9] Andrés, J., García, J. & Jiménez, S. (1989). *La incidencia y la duración del desempleo masculino en España.* Moneda y Crédito, 189, pp: 75-124.

[10] Antolín, P. (1995). *Movilidad laboral, flujos de desempleo, vacantes y comportamiento en la búsqueda de empleo en el mercaedo de trabajo español.* Moneda y Crédito, 201, pp: 1-19.

[11] Armitage, P. (1959). *The comparison of survival curves.* J.R. Stat. Soc. A, 122, pp. 279-292.

[12] Arranz, J.M. & Muro, J. (1999). *Recurrent unemployment and welfare system.* Universidad de Alcalá: preprint.

[13] Berkson, J. & Gage, R.P. (1952). *Survival curve for cancer patients following treatment.* JASA, 47, pp. 501-515.

[14] Blanco, J.M. (1995). *La duración del desempleo en España.* En *Estudios sobre el funcionamiento del mercado de trabajo español* de J.J. Dolado & F.J. Jimeno. FEDEA.

[15] Blossfeld, H., Hamerle, A. & Mayer, K. (1989). *Event history analysis. Statistical theory and application in the social sciences.* Lawrence Erlbaum Associates Inc.

[16] Bover, O., Arellano, M. & Bentolila, S. (1997). *Unemployment duration, benefit duration, and the business cycle.* CEMFI: WP-9717.

[17] Breiman, L., Tsur, Y. & Zemel, A. (1993). *On a simple estimation procedure for censored regression models with known error distributions.* The Annals of Statistics, 21, 4, pp. 1711-1720.

[18] Breslow, N. & Crowley, J. (1974). *A large sample study of life table and product limit estimates under random censorship.* The Annals of Statistics, 2, pp. 437-453.

[19] Breslow, N. & Day, N. (1987). *Statistical models in cancer research, 2.* The design and Analysis of Cohort Studies, IARC.

[20] Buckley, J. & James, I. (1979). *Linear regression with censored data.* Biometrika, 66, 3, pp. 429-436.

[21] Carrasco, R. (1997). *Transitions to and from self-employment in Spain: An empirical analysis.* CEMFI: WP-9710.

[22] Carrol, R., Rupert, D. & Stefanski, L. (1995). *Measurement error in nonlinear models.* New York: Chapman & Hall.

[23] Cebrián, I., García- Serrano, C., Muro, J., Toharia, L. & Villagómez, E. (1995). *Prestaciones por desempleo, duración y recurrencia del paro.* En *Estudios sobre el funcionamiento del mercado de trabajo espñol* de J.J. Dolado & F.J. Jimeno. FEDEA.

[24] Cheng, S. & Wang, N. (2001). *Linear transformation models for failure time data with covariate measurement error.* JASA, 96, 454, pp. 706-716.

[25] Cochran, W.G. (1968). *Errors of measurement in statistics.* Technometrics, 10, pp. 637-666.

[26] Collet, D. (1994). *Modelling survival data in medical research.* London: Chapman & Hall.

[27] Cox, D. (1972). *Regression models and life tables (with discussion).* J. R. Stat. Soc. B, 34, pp. 187-220.

[28] Cox, D. (1975). *Partial likelihood.* Biometrika, 62, pp. 269-276.

[29] Cox, D. (1979). *A note on the graphical analysis of survival data.* Biometrika, 66, pp. 188-190.

[30] Cox, D. & Oakes, D. (1984). *Analysis of survival data.* London: Chapman & Hall.

[31] Crowder, M. (2001). *Classical competing risks.* London: Chapman & Hall.

[32] Cutler, S.J. & Ederer, F. (1958). *Maximum utilization of the life tables method in analyzing survival.* J. of Chronic Diseases, 8, pp. 699-712.

[33] David, H.A. & Moeschberger, M.L. (1978). *The theory of competing risk.* London: Griffin.

[34] Drawoska, D.M. & Doksum, K.A. (1998). *Partial likelihood in transformation models with censored data.* Scandinavian Journal of Statistics, 15, pp. 1-23.

[35] Devine, T. & Kiefer, N. (1991). *Empirical labor economics.* New York: Oxford University Press.

[36] Dolton, P. & O'Neil, D. (1996). *Unemployment duration and the restart effect: some experimental evidence.* The Economic Journal, 435, pp. 387-400.

[37] Efron, B. (1967). *The two sample problem with censored data.* Proceedings of fifth Berkeley symposium in Mathematical Statistics, IV, New York: Prentice & Hall, pp. 831-855.

[38] Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

[39] Elbers, C. & Ridder, G. (1982). *True and spurious duration dependence: The identifiability of the proportional hazard model.* The Annals of Statistics, 7, pp. 1-26.

[40] Espinal, A. & Satorra, A. (1996). *Survival analysis with measurement error on covariates.* Proceedings in Computational Statistics, Physica-Verlag, pp. 253-258.

[41] Flinn, C.J. (1986). *Econometric analysis of CPS-Type unemployment data.* J. of Human Resources, pp. 456-484.

[42] Flinn, C.J. & Heckman, J.J. (1982). *Models for the analysis of labor force dynamics.* Advances in Econometrics, vol. 1, pp. 35-95.

[43] Florens, J.P., Ivaldi, M., Laffont, J.J. & Laisney, F. (1990). *Microeconometrics: Surveys and applications.* Basil Blackwell ltd.

[44] Florens, J.P., Gérard-Varet, L.A. & Werquin, P. (1990). *The duration of current and complete unemployment spells between 1984 and 1996 in France: Modelling and empirical evidence.* In *Microeconometrics: Surveys and applications* by J.P. Florens, et al. Basil Blackwell ltd.

[45] Fuller, W. (1987). *Measurement error models.* John Wiley & Sons, Inc.

[46] García-Fontes, W. & Hopenhayn, H. (1996). *Flexibilización y volatilidad del empleo.* Moneda y Crédito, 206 , pp. 205-227.

[47] García-Fontes, W. & Hopenhayn, H. (1996). *Creación y destrucción de empleo en la economía española.* En *La Economía Española: una visión diferente* de R. Marimon. Barcelona: Bosch ed., pp. 139-170.

[48] García-Fontes, W. & Hopenhayn, H. (1996). *Componentes cíclicos y cambio estructural en la destrucción del empleo.* En *La Economía Española: una visión diferente* de R. Marimon. Barcelona: Bosch ed., pp. 171-196.

[49] García-Pérez, J.J. (1997). *Las tasas de salida del empleo y el desempleo en España (1978-1993).* Investigaciones Económicas, XXI, pp. 29-53.

[50] Gehan, E.A. (1969). *Estimating survivor functions from life table.* J. of Chronic Diseases, 21 pp. 629-644.

[51] Gil, F., Martín, M.J. & Serrat, A. (1994). *Movilidad en el mercado de trabajo en España: un análisis econométrico de duración con riesgos en competencia.* Investigaciones Económicas, XVIII, pp. 517-537.

[52] Gill, R.D. (1980). *Censoring and stochastic integrals.* Mathematical Centre Tracts. Amsterdam: Mathematisch Centrum, 124.

[53] Glasser (1967). *Exponential survival with covariance.* JASA, 62 pp. 561-568.

[54] Gonzalo, R. (1998). *Duración del desempleo: estudio del caso español.* Tesis doctoral. Universidad de Valencia.

[55] Hamerle, A. (1989). *Multiple-spell regression models for duration data.* J. of Appl. Statistics, 38, pp. 127-138.

[56] Han, A. & Hausman, J.A. (1990). *Flexible parametric estimation of duration and competing risk models.* J. of Appl. Econometrics, 5, pp. 1-28.

[57] Heckman, J.J. & Borjas, G.J. (1980). *Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence.* Economica, 47, pp. 247-283.

[58] Heckman, J.J. & Singer, B. (1984). *Econometric duration analysis.* J. of Econometrics, 24, pp. 63-132.

[59] Heckman, J.J. & Singer, B. (1985). *Longitudinal analysis of labor market data.* In *Econometric Society monograph 10.* Cambridge University Press.

[60] Heckman, J.J. & Honoré, B.E. (1989). *The identifiability of the competing risks model.* Biometrika, 76, pp. 325-330.

[61] Holt, D., McDonald, J. & Skinner, C. (1991). *The effect of measurement error on event history analysis*, in *Measurement error in surveys* by Biemer, Groves, Lyberg, Mathiowetz & Sudman. John Wiley & Sons, Inc.

[62] Honoré, B.E. (1993). *Identification results for duration models with multiple spells.* Review of Economic Studies, 60, pp. 241-246.

[63] Hougaard, P. (1999). *Multi-state models: A review.* Lifetime Data Analysis, 5, pp. 239-264.

[64] Irwin, J.O. (1942). *The distribution of the logarithm of survival times when true law is exponential.* J. of Hyg., 42, pp. 328-333.

[65] James, I.R. & Smith, P.J. (1984). *Consistency results for linear regression with censored data.* The Annals of Statistics, 12, pp. 590-600.

[66] Jiang, W., Turnbull, B. & Clark, L. (1999). *Semiparametric regression models for repeated events with random effects and measurement error.* JASA, 94, 445, pp. 111-124.

[67] Jovanovic, B. (1979). *Job matching and the theory of turnover.* J. of Pol. Economy, 87, pp. 972-990.

[68] Kalbfleisch, J. & Prentice, R. (1980). *The statistical analysis of failure time data.* John Wiley & Sons, Inc.

[69] Kaplan, E.L. & Meier, P. (1958). *Nonparametric estimation from incomplete observations.* JASA, 53, pp. 457-481.

[70] Kiefer, N.M. (1988). *Economic duration data and hazard function.* J. of Economic Literature, 26, pp. 646-679.

[71] Klein, J. & Moeschberger, M. (1997). *Survival analysis. Techniques for censored and truncated data.* New York: Springer-Verlag Inc.

[72] Koul, H., Susarla, V. & Van Ryzin, J. (1981). *Regression analysis with randomly right-censored data.* The Annals of Statistics, 6, pp. 1276-1288.

[73] Kulich, M. & Lin, D.Y. (2000). *Additive hazards regression with covariate measurement error.* JASA, 95, pp. 238-248.

[74] Kummel, C.H. (1879). *Reduction of observed equations which contain more than one observed quantity.* Analyst, 6, pp. 97-105.

[75] Lai, T. & Ying, Z. (1991). *Large sample theory of a modified Buckley-James estimator for regression analysis with censored data.* The Annals of Statistics, 19, 3, pp. 1370-1402.

[76] Lancaster, T. (1979). *Econometric methods for duration of unemployment.* Econometrica, 47, pp. 939-956.

[77] Lancaster, T. (1990). *The Econometric Analysis of transition data.* Cambridge University Press.

[78] Lawless, J.F. (1982). *Statistical models and methods for lifetime data.* New York: Wiley

[79] Mantel, N. & Myers, M. (1971). *Problems of convergence of maximum likelihood iterative procedures in multiparameters situation.* JASA, 66, pp. 484-491.

[80] Mare, R.D. & Winship, C. (1988). *School enrollment, military enlistment, and the transition to work: implications for the age pattern of employment.* In *longitudinal analysis of labor market data* by J.J. Heckman & B. Singer. Cambridge University Press.

[81] MATLAB (1997). The mathWorks, Inc. Natic, MA.

[82] McCall, B.P. (1994). *Testing the proportional hazards model in the presence of unmeasured heterogeneity.* J. of Appl. Econometrics, 9, pp. 321-334.

[83] McCall, B.P. (1997). *The determinants of full-time versus part-time reemployment following job displacement.* J. of Labor Economics, 15, pp. 714-733.

[84] Mealli, F., Pudney, S. & Thomas, J. (1996). *Training duration and post-training outcomes: A duration-limited competing risks model.* The Economic Journal, 435, pp. 422-433.

[85] Miller, R. (1976). *Least squares regression with censored data.* Biometrika, 63, pp. 449-464.

[86] Miller, R. & Halpern, J. (1982). *Regression with censored data.* Biometrika, 69, pp. 521-531.

[87] Nakamura, T. (1992). *Proportional hazards model with covariates subject to measurement error.* Biometrics, 48, pp. 829-838.

[88] Narendranathan, W. & Stewart, M.B. (1993). *Modelling the probability of leaving unemployment: competing risks models with flexible base-line hazards.* Appl. Statistics, 42, pp. 63-83.

[89] Nickel, S., Narendranathan, W., Stern, J & García, J. (1989). *The nature of unemployment in Britain.* Oxford University Press.

[90] Nelson, W. (1972). *Theory and applications of hazard plotting for censored failure data.* Technometrics, 14, pp. 945-965.

[91] Petersen, T. (1995). *Analysis of event histories.* In *Handbook of Statistical Modelling for the Social and Behavioral Sciences* by G. Arminger et. al. Plenum Press, pp. 453-517.

[92] Peto, R. & Lee, P. (1973). *Weibull distributions for continuous carcinogenesis experiments.* Biometrics, 29, pp. 457-470.

[93] Pike, M. C. (1966). *A method of analysis of certain class of experiments in carcinogenesis.* Biometrics, 22, pp. 142-161.

[94] Prentice, R. (1974). *A log gamma model and its maximum likelihood estimation.* Biometrika, 61, pp. 539-544.

[95] Prentice, R. (1982). *Covariate measurement errors and parameters estimation in a failure time regression model.* Biometrika, 69, pp. 331-342.

[96] Prentice, R., Kalbfleisch, J.D., Peterson, A.V., Jr.,Flournoy, N.S., Farewell, V.T. & Breslow, N.E. (1978). *The analysis of failure times in the presence of competing risks.* Biometrics, 34, pp. 541-554.

[97] Rao, C. (1973). *Linear statistical inference and its applications.* New York: Wiley.

[98] Rabinowitz, A., Tsiatis, D. & Aragon, J. (1995). *Regression with interval-censored data.* Biometrika, 82 pp. 501-513.

[99] Ridder, G. (1990). *The nonparametric identification of generalized accelerated failure-time models.* The Review of Economic Studies, 57 pp. 167-181.

[100] Ritov, Y. (1990). *Estimation in linear regression model with censored data.* The Annals of Statistics, 18 pp. 303-328.

[101] Salant, S. (1977). *A search theory and duration data: a theory of sorts.* Quarterly journal of economics, 91, pp. 39-58.

[102] Satorra, A. (1992). *Asymptotic robust inferences in the analysis of mean and covariance structures.* Sociological Methodology, pp. 249-278. Blackwell.

[103] Segura, J., Durán, F., Toharia, L. & Bentolila, S. (1991). *Análisis de la contratación temporal en España.* Ministerio de Trabajo y Seguridad Social, Madrid.

[104] Schneider, H. & Weissfeld, L. (1986). *Estimation in linear models with censored data.* Biometrika, 73, 3, pp. 741-745.

[105] Stacey, E.W. (1962). *A generalization of the gamma distribution.* The Annals Math. Stat., 33, pp. 1187-1192.

[106] Sueyoshi, G.T. (1995). *A class of binary response models for grouped duration data.* J. of Appl. Econometrics, 10, pp. 411-431.

[107] Van den Berg, G.J. & Van Ours, J. (1996). *Unemployment dynamics and duration dependence.* J. of Labor Economics, 14, pp. 100-125.

[108] Tsatis, A. (1975). *A nonidentifiability aspect of the problem of competing risks.* Proceedings Nat. Acad. Sci. USA, 72, pp. 20-22.

[109] Tsiatis, A. (1990). *Estimating regression parameters using linear rank tests for censored data.* The Annals of Statistics, 18 pp. 354-372.

[110] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). *Regression analysis of multivariate incomplete failure time data by modelling marginal distributions.* JASA, 84 pp. 1065-1073.

[111] Zhang, Z. & Li, G. (1996). *A simple quantile approach to the 2-sample locatio-scale problem with random censorhip.* J. of Nonparamteric Statisics, 6, pp. 323-335.

[112] Zippin, C. & Armitage, P. (1966). *Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter.* Biometrics, 22, pp. 665-672.