TECHNISCHE
UNIVERSITÄT
WIEN

VIENNA
UNIVERSITY OF
TECHNOLOGY

DIPLOMARBEIT

Duration Analysis - Theory and Application to Austrian Unemployment Data

Ausgeführt am Institut für

WIRTSCHAFTSMATHEMATIK

der Technischen Universität Wien

unter der Anleitung von Ao.Univ.-Prof. Dr. Bernhard Böhm

durch

Irene Hofstetter

Pergkirchen 56
4320 Perg

20. März 2008
Datum

Unterschrift

TECHNISCHE
UNIVERSITÄT
WIEN

VIENNA
UNIVERSITY OF
TECHNOLOGY

DIPLOMA THESIS

Duration Analysis - Theory and Application to Austrian Unemployment Data

submitted for Partial Fulfillment
of the Requirements for the Degree

Master of Science (MSc) — Diplom-Ingenieur (Dipl.-Ing.)

at the

INSTITUTE FOR MATHEMATICAL METHODS IN ECONOMICS

Faculty for Mathematics and Geoinformation

VIENNA UNIVERSITY OF TECHNOLOGY
(TECHNISCHE UNIVERSITÄT WIEN)

Supervisor: Ao.Univ.-Prof. Dr. Bernhard Böhm

by

Irene Hofstetter

Vienna

March 20th, 2008
Date

Signature

Acknowledgements

I am grateful for the constructive comments and support of Prof. Dr. Bernhard Böhm and would also like to gratefully acknowledge the assistance in data preparation from Andreas Wolf. Moreover gratitude is extended to DI Albert Neumüller for his support in LaTeX-command-design matters and to Mr. Clapham and Mr. Breitkreutz for their proof-reading of the English phrasing in the paper. Additionally, I wish to thank my family and friends who supported me in various ways during the completion of this study. Without all these people this paper would not have enriched my academic and personal life in the same special way it has with their support.

Contents

1. Introduction

Unemployment is a relevant and recurring discussion topic in social and economic politics. Extensive research in the area of characteristics and effects of unemployment has been made by scientists of several countries. Some reference links to such unemployment related topics are given throughout this diploma thesis.

The Public Employment Service Austria (AMS), a service agency under public law, has been founded to assist unemployed in their job search and provide further labour-market related services. Pursuing the governmental policy of full employment it plays a significant role in prevention and reduction of unemployment in Austria. An unemployment benefit system has been established to guarantee financial support during periods of unemployment. To protect the work environment the government has passed laws that provide labour contract regulations. Furthermore, financial incentives for hiring underprivileged persons have been set. More recently the focus has been shifted to young people by increasing the amount of apprenticeship positions. The reduction of unemployment, embedded in the superior target of general improvement of economic welfare, has since been continuously given high priority from the state authorities. The reason behind this concentrated attention is that unemployment does not affect solely individuals but also the public budget by loss of tax revenues and increased expenditures in the form of benefits. Beside the budget concern unemployment reveals serious social problems that come along with losing a job or being unable to find one. It may lead to isolation and stigmatisation and in worst cases to loss of contact with family and friends. The prejudgement of society worsens the situation and negatively affects the self-esteem of unemployed. Finally this ends up in a vicious cycle, as long-term unemployment may result in loss of skills and required key qualifications for most job positions and thus further reduce the job opportunities. To render every assistance possible to help people to find a job is therefore of high social interest. To establish which individual properties may affect the time spent in unemployment, the duration time has to be analysed together with appropriate characteristics. Although regularly published statistics for unemployment data are available, the posted information usually does not contain additional background information concerning the *duration* of unemployment. The published information generally is represented only by partly categorised histograms. These statistics do however give interesting insights in how far the economy is affected by unemployment rates. Additional categorisation gives information about group effects, including which participants of the labour market are most concerned or threatened by the prospect of becoming unemployed. An overall development of the unemployment rates concerning the Austrian labour market on the basis of the microcensus data of Statistics Austria can be found on the webpage of Statistics Austria *(http://www.statistik-austria.at)*. The information is given in a way which provides a good starting point for deriving unemployment prevention measures. However, what is not included in this body of information is the treatment of individuals. We are interested in finding out more about the dynamics of the data, the in- and outflows of the state of unemployment of an observed individual. Our study therefore concentrates on unemployment duration and its determinants, which has so far been examined less intensively within available Austrian data. Our aim is to identify which characteristics may have an effect on the extension or shortening of unemployment duration.

The thesis is divided into several sections starting by a qualified statement about the theoretical technicalities concerning survival analysis (equivalently used for the term duration analysis) in the next chapter. This theoretical backup is an attempt to introduce the field of duration analysis in a simple way to catch the reader's interest for the topic and to provide a profound basis for better understanding of the subsequent application to our sample data. After presenting the mathematical tools comprising different methods, functions and distributions, a short outlook to our approach of investigating the data is given. In this context the topic of competing risks is re-emphasised in more detail following a short introduction in the theoretical section. The distinction between different cases regarding data observation takes centre stage in this section. What follows and precedes the actual application is the description of the data set in use. The main section, the exploration of our sample data, basically consists of three parts. It starts with a non-parametric analysis method, the Kaplan-Meier-estimation (*section* 3.4). All available covariates are screened for potential effects on duration of unemployment and plots are given to help visualize these effects. The data diagnostics are then carefully detailed thereafter. The second part in this chapter deals with parametric estimations. QQ-plots (*page* 79) are used to determine the appropriateness of certain parametric approaches. Four different parametric distributions, already touched on in the theory chapter, are examined for fitness. The third approach is the attempt to apply a semi-parametric model to the data. This part deals with Cox proportional hazard (CPH) modeling which provides the possibility for detection of influential factors on the length of unemployment, considering the assumption of multiplicative effects of covariates. The thesis is concluded with a summary of these findings and an outlook to further research in the field of unemployment duration.

2. Survival Analysis

There are numerous ways to try to get valid results from a given dataset, using different methods and approaches. Focusing on the question of how different circumstances and characteristics of the individuals affect the unemployment spell [1], we will concentrate on what is called Survival Analysis.

This method is favoured above other methods of data analysis commonly used, for several reasons.

The distribution of the time to leaving unemployment is usually not normally distributed. Very often it is exponentially distributed if the risks occur constantly over time, or the distribution might be bimodal. [Mario Cleves and Gutierrez, 2004, p1 f.]

Another reason is that we are dealing with duration data (unemployment spells) which are positive by their very nature, so distributions should be useful which take this restriction into account.
[Greene, 2003, p792]

A third aspect giving preference to Survival Analysis is the fact that it can deal with censoring. Simply defined, censoring means that some duration times are not observed until their end. This is the case if the survey is interrupted before the individuals have found another job and are still in a state of unemployment *(for more details on censoring see section 2.2)*. Censoring is a phenomenon that is frequently found as there are usually certain time frames for the realization of surveys.

Survival Analysis therefore qualifies itself to be an appropriate tool to evaluate the timing of events. In our study, this event is "leaving the state of unemployment". Note that the time until the occurrence of this event is called Survival time.
Generally, a survival analysis model is a simple example of a Markov process model with two states and a certain intensity of transition from one state to the other. (*Comprehensive explanations about Survival Analysis Models can be found in* [Crowley and Johnson, 1981] *amongst others*).

The following sections focus on the distinction between parametric and non-parametric survival analytic models, preceded by a brief description of required notions of the Hazard rate and Survival function. As primary reference literature the books of [Greene, 2003, p792 ff.], [Lancaster, 1994][p6 ff.] and [Hashem Pesaran and Schmidt, 1999, p301 ff.] have been used.

[1] 'A spell is an ordered triple {state, time of entry, time of exit}' [Winkelmann, 1994, p12]

2.1. Hazard

Consider a continuous non-negative random variable T, which describes the length of time until the required event of leaving unemployment occurs. T is therefore the "duration of time in the state of unemployment". The probability distribution of T can be specified as

$$F(t) = P(T < t), \qquad 0 < t < \infty, \tag{2.1}$$

and measures the unconditional probability of survival up to time t.
As the first derivative of this distribution we calculate the probability density function

$$f(t) = \frac{\partial F(t)}{\partial t} = lim_{dt \to \infty} \frac{P(t \leq T < t + dt)}{dt} \tag{2.2}$$

However, conditional probability rather than the unconditional approach is the relevant concept in the discussed statistical methods, focusing on the hazard function. We first define the survival function as

$$S(t) = 1 - F(t) = P(T \geq t) \tag{2.3}$$

which measures the probability that the random variable T will equal or exceed time t.

Functions (2.1) – (2.3) help us to define the hazard function

$$\lambda(t) = \lim_{\delta \to 0} \frac{P(t \leq T < t + \delta | T \geq t)}{\delta}. \tag{2.4}$$

which indicates the probability to leave unemployment in the short interval of length dt after time t, conditional on the state of unemployment being occupied in t, i.e. unemployment has not been left before t.

Additionally, from the basic notations above we can derive the following relations between them necessary for further calculations in the models mentioned later.

Considering the formula for conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$ of an event A given B, we get

$$\begin{aligned} \lambda(t) &= \lim_{\delta \to 0} \frac{P(t \leq T < t + \delta | T \geq t)}{\delta} = \lim_{\delta \to 0} \frac{\frac{P(t \leq T < t + \delta)}{P(T \geq t)}}{\delta} \\ &= \frac{1}{S(t)} \lim_{\delta \to 0} \frac{F(t + \delta) - F(t)}{\delta} = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)} \end{aligned}$$

Given the fact that $F'(t) = f(t) = -S'(t)$ and that $\frac{F'(x)}{F(x)} = [ln(F(x))]'$ further leads to

$$\lambda(t) = \frac{-S'(t)}{S(t)} = -\frac{dlnS(t)}{dt} \tag{2.5}$$

from which the integrated hazard function $\Lambda(t)$ can be derived by rewriting equation (2.5) in the form $\lambda(u)du = -dlnS(u)$ and subsequent integration.

$$-lnS(t) = \int_0^t \lambda(u)du = \Lambda(t) \tag{2.6}$$

The integrated hazard function is therefore a convenient mathematical term often used for validation of hypothesized models in residual analysis. The relation to the survivor function is $S(t) = e^{[-\Lambda(t)]}$. It is important to remember that the integrated hazard function cannot be viewed as probability. In addition, it does not have an equally simple interpretation regarding duration dependence as the hazard function itself, but has the advantage that it involves smoothing of the data.
If the hazard function increases, i.e. $\frac{d\lambda(t)}{dt} > 0$, positive duration dependence is revealed and the chances of emerging from the state of unemployment increases over time. The process would then be represented by a convex shaped integrated hazard function.
Conversely a concave integrated hazard is equivalent to a decreasing hazard function, i.e. $\frac{d\lambda(t)}{dt} < 0$, and points to a declining chance of emerging from unemployment.
The third case would be a memory-less system with constant hazard, i.e. $\frac{d\lambda(t)}{dt} = 0$, and is represented by an integrated hazard function as an ascending or descending straight line.
[Kavkler and Borsic, 2006, p11 f.], [Kiefer, 1988, Greene, 2003]

The notion, that the hazard rate does not vary over time, is a very basic one and gives us an appropriate introduction into analysing a duration process. It is for this reason that this is stated as a reference example (*compare* [Greene, 2003, p793]).

Ex.2.1 `Example:` Assume that the hazard rate is constant

$$\lambda(t) = \lambda$$

This implies that conditional probability of leaving unemployment is the same in any given short interval over time, no matter when the observation is made. We then arrive at the following differential equation by inserting λ into equation (2.5)

$$\lambda = -\frac{dlnS(t)}{dt}$$

Solving this equation we get

$$lnS(t) = -\lambda t + c$$

where c denotes the constant term. This expression can further be changed into

$$S(t) = C * e^{\lambda t}$$

where C again denotes the constant term of integration.
Since $P(T \geq 0) = 1$, our initial condition is $S(0) = 1$.
Putting these results into our general solution from above
$(S(0) = C * e^{\lambda 0} = 1 = C * 1)$
we calculate $C = 1$ and therefore get the exponential distribution as our final solution

$$S(t) = e^{\lambda t}$$

I should mention that next to the treated continuous case above with the continuous random variable T one could also be interested in the discrete case. This is dependent upon how the duration data has been observed in the survey. If it is obtained, for instance, on a monthly or seasonal basis one will presumably define a random variable T for the times measured $(t_1, t_2, \ldots, t_n)$ and consider the discrete version of the functions. These are expressed as follows

$$\begin{aligned} f(t_k) &= P(T = t_k) & S(t_k) &= \textstyle\sum_{j>k} f(t_j) \\ \lambda(t_k) &= \tfrac{f(t_k)}{S(t_k)} & \Lambda(t_k) &= \textstyle\sum_{i=0}^{k} \lambda(t_i) \end{aligned}$$

$$\text{for } k = 1, \ldots n.$$

[Kiefer, 1988, p652]

Having stated the basic equations of Survival Analysis, it is noted that they can be derived from each other. Once the distribution function $F(t)$ has been defined, you can easily derive $S(t)$ and $\lambda(t)$. Conversely, using the relationship between the functions, one can also define $\lambda(t)$ and simultaneously derive $S(t)$ and $F(t)$.

Other important characteristics in Survival Analysis that can be calculated from the above mentioned functions are:

- *the quantile function* $Q(p) = inf\{t : F(t) \geq p\}$ for $0 < p < 1$, the time at which a specific proportion p fails
- *the mean* $\mathbf{E} = \int_0^\infty S(t)dt$
- *the variance* $\mathbf{Var} = 2\int_0^\infty tS(t)dt - \{\mathbf{E}(t)\}^2$

[Bagdonavičius and Nikulin, 2002, p2]

2.2. Censoring

One of the previously mentioned advantages of Survival Analysis is that it can deal with censoring and for this reason I am including an extra section explaining the term and the ideas behind it.

Censoring is to be distinguished from truncation, which is a rather strong type of biased sampling and represents a minimum or maximum restriction. In truncated data, only the spells within a certain interval are observed. Information loss due to censoring is less than that lost due to truncation [Owen, 2001, p135]. The sample size is changed by truncation while with censoring the size of the sample remains unchanged. A spell is said to be censored, if the time of its completion is unknown. This is, for instance the case, if the event we are interested in has not occurred prior to the end of the study, which implies that the individual is still unemployed when the survey time has finished. This is termed right-censoring.
Denote the time until the event of interest by X and the right-censored time by C_r, then the duration time T is explained by $T = min\{X, C_r\}$. An auxiliary function is normally declared by the indicator δ.

$$\delta = \begin{cases} 1 & \text{if} \quad T = X \\ 0 & \text{if} \quad T = C_r \end{cases}$$

Conversely, left-censoring can be found in case of unknown original starting points of observed spells. The duration time is then defined by $T = max\{X, C_l\}$ with C_l representing the left-censored time. The indicator function η is then defined by

$$\eta = \begin{cases} 1 & \text{if} \quad T = X \\ 0 & \text{if} \quad T = C_l \end{cases}$$

According to [Hashem Pesaran and Schmidt, 1999, *section* 2.2] we can distinguish 3 other types of censoring:

- *type I censoring:* All spells not completed after a certain duration time are regarded as being censored. As a consequence the number of censored items is random while the end of the study has been fixed at the beginning.
- *type II censoring:* Sampling continues until the rth smallest failure time is observed. Thus r is predetermined in advance while the duration time of the study is random.
- *progressive type II censoring:* A given fraction of the sample may be censored after several observed failure times.

2.3. Non-parametric analysis

The idea behind an entirely non-parametric approach is to follow the philosophy of "letting the dataset speak for itself" and to leave assumptions about the distribution of failure times or how covariates serve to change or shift the survival experience aside [Mario Cleves and Gutierrez, 2004, p5].
Facing fewer restrictions is an advantage of non-parametric approaches. It can be favourable not to impose a certain shape to the hazard function, as is the case in parametric models, because of the possibility of a resulting bias in estimators.

Non-parametric estimators can be a useful tool of data exploration and provide relevant information. An example of an advanced approach to using non-parametric estimators in unemployment duration analysis can be found in the discussion paper of [Wichert and Wilke, 2007] where the distribution of regressors is allowed to be truncated by incorporating conditional quantile functions. But we will focus instead on the basic principles in non-parametric estimation and consider the conventional Kaplan-Meier estimator without further extensions.

2.3.1. Kaplan-Meier estimator

The Kaplan-Meier estimator, also called the product-limit estimator, is a non-parametric estimate of the survival function and was originally proposed by [Kaplan and Meier, 1958].
Expressed in a simple way, the Kaplan-Meier estimator is a step function which decreases by a step at each failure time, i.e. at the end of an unemployment period.
It has the advantage that it is not dependent upon the choice of intervals. Given the duration data, each observation at a certain time is a failure or is censored. *(for more details on censoring see section 2.2)*. Assume that the data of observed duration are sorted in ascending order ($t_1 < t_2 < \ldots < t_k$) and d_j denotes the departures from unemployment at time t_j

($j \in \{1, \ldots, k\}$). It is further assumed that c_j observations were censored in the interval $[t_{j-1}, t_j)$. Therefore we obtain

$$r_j = \sum_{i \geq j}(d_i + c_i) = n - \sum_{i<j}(d_i + c_i) \tag{2.7}$$

the number of individuals at risk just prior to t_j, those neither completed, nor censored until t_j. (Note: The term "risk" is to be interpreted as "chance" of getting out of unemployment). In other words, the number at risk r_j are those who have been in the state of unemployment in the preceding interval $[t_{j-1}, t_j)$. We can alternatively express r_j recursively by

$$r_{j-1} = r_j - d_j - c_j.$$

The initial condition is set to be $r_0 = n$, where n is the number of individuals observed, and indicates that all subjects start in the state of being unemployed.

We are interested in estimating the probability of survival $P(T \geq t_j)$, which can be expressed in terms of conditional probability as follows
(Considering the formula for conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$)

$$P(T \geq t_j) = P(T \geq t_j | T \geq t_{j-1}) * P(T \geq t_{j-1}).$$

$P(T \geq t_{j-1})$ can again be expressed in terms of conditional probability and we therefore derive by recursion

$$P(T \geq t_j) = \prod_{i=1}^{j} P(T \geq t_i | T \geq t_{i-1}) * P(T \geq t_0). \tag{2.8}$$

By our initial definition $P(T \geq t_0) = 1 - P(T < t_0) = 1 - F(t_0) = S(t_0) = 1$. Additionally, we can define the conditional probability as

$$P(T \geq t_i | T \geq t_{i-1}) = 1 - P(T < t_i | T \geq t_{i-1}) \tag{2.9}$$

where $P(T < t_i | T \geq t_{i-1})$ is the probability of completing a spell of unemployment in the interval $[t_{i-1}, t_i)$ which is the definition of the hazard function introduced in section 2.1. The hazard function can be estimated by

$$\hat{\lambda}(t_j) = \frac{d_j}{r_j} \tag{2.10}$$

which is the ratio of the completed spells of unemployment at t_j to the sum of all neither completed nor censored spells of unemployment until t_j, i.e. the number of individuals leaving unemployment at t_j divided by the number at risk at this time of interest.

We can now rewrite equation (2.9) as

$$P(T \geq t_i | T \geq t_{i-1}) = 1 - \hat{\lambda}(t) = 1 - \frac{d_i}{r_i} = \frac{r_i - d_i}{r_i}$$

and finally calculate the Kaplan-Meier Estimator of the survival function from equation (2.8)

$$P(T \geq t_j) = \prod_{i=1}^{j} P(T \geq t_i | T \geq t_{i-1}) * 1 = \prod_{i=1}^{j} \frac{r_i - d_i}{r_i}. \tag{2.11}$$

Another non-parametric estimator worth mentioning is the Nelson-Aalen estimator of the cumulative hazard function. A step function with vertical step size of $1/r_j$ defined by

$$\hat{\Lambda}_{NA}(t) = \sum_{j:t_j \leq t} \frac{d_j}{r_j}. \tag{2.12}$$

Often used in Cox proportional hazard models (*see section 2.5.1*) the Nelson-Aalen estimator is also known as Breslow estimator, due to the work by Breslow (1972), and represents a possible alternative to the Kaplan-Meier estimator. The function of interest is in general the Breslow-type estimate of the survival function

$$\hat{S}(t) = e^{-\hat{\Lambda}_{NA}(t)}.$$

More detailed information about the Nelson-Aalen estimator can be found in [Andersen et al., 1993] amongst others.

2.4. Parametric analysis

First I would like to point out, that it is good practice to start with an initially non-parametric estimation, which delivers the functional shape, before concentrating on parametric models.

The main reason for choosing the parametric method though is to handle the following, as stated in [Lancaster, 1994, p33 f.]

- ◇ The duration distributions of different people may differ because of varying inputs. Representation of this fact can be given by introducing a regression vector, $\mathbf{x}$, for each person to demonstrate the source of difference.

An advantage of parametric methods is that, due to the underlying distribution assumption, one can legitimately make predictions conditional on $\mathbf{x}$.

In the parametric approach, regarding increasing or decreasing hazard rates, one faces a choice of possible distribution models to fit to the observed data. In the following paragraph I will introduce some of the most common distributions chosen in parametric survival analysis. A basic description of them can be found in [Kiefer, 1988, Greene, 2003, p653 ff., p794 f.] and [Hashem Pesaran and Schmidt, 1999, p307] amongst others.

Ex.2.2 `Example:` (to illustrate the difference in shape of different hazard functions)
The hazard shape of the given data may be observed as constant, monotone decreasing or increasing. Others appear in form of a bell-shape or U-shape. Some example figures (Figure 2.1 - Figure 2.4) demonstrate the great variety in shape adjustment.

Depending on the suggested or presumed shape of the hazard function, one faces the decision of which distribution to take in order to support the model.

Some parametric specifications for duration data

distribution	*parameters*	*functions*
exponential	$\gamma > 0$	$F(t) = 1 - e^{-\gamma\, t}$ $S(t) = e^{-\gamma\, t}$ $f(t) = \gamma\, e^{-\gamma\, t}$ $\lambda(t) = \gamma$ $\Lambda(t) = \gamma\, t$
Weibull	$\gamma > 0,\, \alpha > 0$	$F(t) = 1 - e^{-(\gamma\, t)^{\alpha}}$ $S(t) = e^{-(\gamma\, t)^{\alpha}}$ $f(t) = \gamma\, \alpha\, (\gamma\, t)^{\alpha-1}\, e^{-(\gamma\, t)^{\alpha}}$ $\lambda(t) = \gamma\, \alpha\, (\gamma\, t)^{\alpha-1}$ $\Lambda(t) = (\gamma\, t)^{\alpha}$
log-logistic	$\gamma > 0,\, \alpha > 0$	$F(t) = 1 - [\frac{1}{(1+(\gamma\, t)^{\alpha})}]$ $S(t) = \frac{1}{(1+(\gamma\, t)^{\alpha})}$ $f(t) = \frac{\gamma\, \alpha\, (\gamma\, t)^{\alpha-1}}{(1+(\gamma\, t)^{\alpha})^2}$ $\lambda(t) = \frac{\gamma\, \alpha\, (\gamma\, t)^{\alpha-1}}{(1+(\gamma\, t)^{\alpha})}$ $\Lambda(t) = ln(1 + (\gamma\, t)^{\alpha})$
log-normal	$\gamma > 0,\, \alpha > 0$	$F(t) = \Phi[\alpha\, ln(\gamma\, t)]$ $S(t) = \Phi[-\alpha\, ln(\gamma\, t)] = 1 - \Phi[\alpha\, ln(\gamma\, t)]$ $f(t) = \frac{\alpha}{t}\phi[\alpha\, ln(\gamma\, t)] = \frac{\alpha}{t}\frac{1}{\sqrt{2\pi}} exp(\frac{-\alpha^2(ln(\gamma\, t))^2}{2})$ $\lambda(t) = \frac{\alpha}{t}\frac{\phi(\alpha\, ln(\gamma\, t))}{\Phi(-\alpha\, ln(\gamma\, t))}$ $\Lambda(t) = -ln(\Phi[-\alpha\, ln(\gamma\, t)])$

note: in the log-logistic distribution $ln(t)$ is assumed to be logistically distributed with location parameter $\mu = -ln(\gamma)$ and scale parameter $\sigma = \frac{1}{\alpha}$ and in the log-normal distribution $ln(t)$ is assumed to be normally distributed with mean $\mu = -ln(\gamma)$ and standard deviation $\sigma = \frac{1}{\alpha}$.

Table 2.1.: parametric functions for survival data

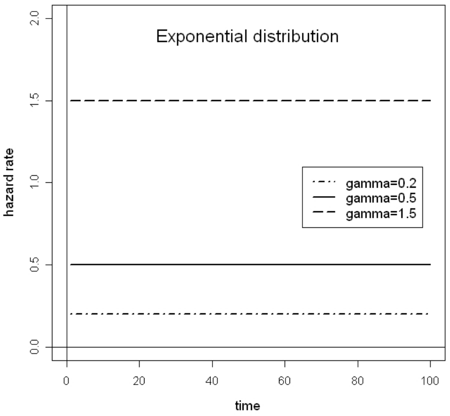


Figure 2.1.: exponential hazard functions

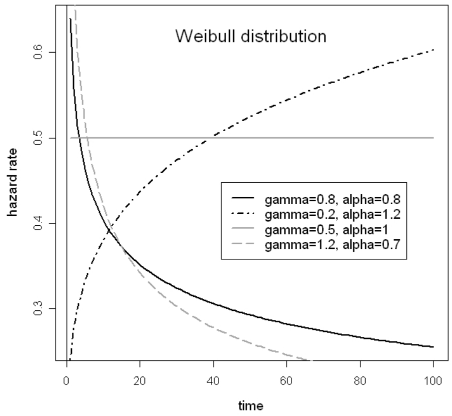


Figure 2.2.: Weibull hazard functions

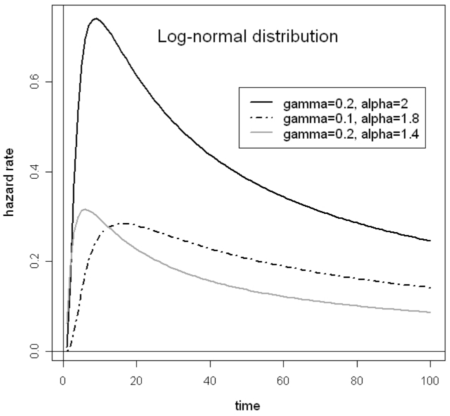


Figure 2.3.: log-normal hazard functions

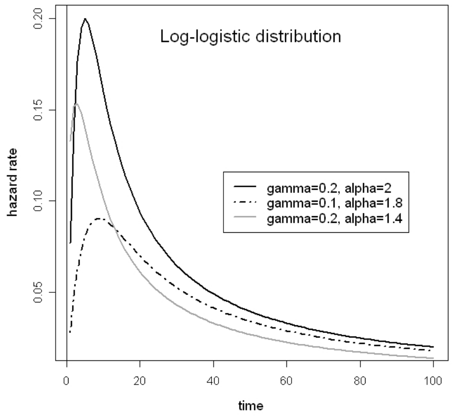


Figure 2.4.: log-logistic hazard functions

Beginning with the **exponential distribution**, I will briefly describe this distribution which is well known as a widely used base model for duration data. It is popular due to its simplicity and its characteristic of modeling a constant probability of leaving the state occupied. It seems to be adequate if there is not much variation exhibited because the constant hazard rate implies no duration dependence. For instance, it is adequate if approximately the same chance of emerging from unemployment after 2 months or 12 months is given. Due to the inherent constant hazard function, the exponential distribution is sometimes said to be ”memory-less”. Of further importance is the fact that the distribution depends on a single parameter (λ) and it can be uniquely characterized by the hazard function (*see* Example [Ex.2.1] *of section* 2.1). A disadvantage though is that distribution depends solely on one parameter and is therefore not very flexible. This can be seen in that both, the mean $E(T) = \frac{1}{\lambda}$ and the variance $var(T) = \frac{1}{\lambda}$ are equal. Another crucial fact is that the assumption of a ”memory-less” property inherent to the process is quite strong and not always appropriate. The exponential distribution with $\alpha = 1$ is a special case of the Weibull distribution stated below.

The **Weibull distribution**, a generalization of the exponential distribution, leads to a rescaling in the time axis compared with exponential distribution. If $\alpha > 1$, the time in the Weibull distribution is regarded to be faster and consequently results in an underestimation of the coefficients more likely. It is exactly opposite, increasing the probability of overestimation and slowing down the time, if $\alpha < 1$. [Kiefer, 1988, p665]
The shape of the hazard rate is characterized by an increasing slope if $\alpha > 1$, decreasing if $\alpha < 1$ and constant and thus representing the exponential case if $\alpha = 1$.

Just like the Weibull distribution, the **log-logistic distribution** is also explained by two parameters. Other than the two aforementioned distributions, the hazard shape can have a change in direction with duration. This is the case if $\alpha > 1$, showing a hazard that is first increasing and later decreasing. If $\alpha \in (0, 1]$ the hazard function decreases over time.

The **log-normal distribution** has, like those previously, a location parameter γ and a scale parameter α. The hazard function is reflected by a first increasing and then decreasing function when $\alpha > 1$ and solely decreasing function in the case of $\alpha \in (0, 1]$, and has therefore the same directions of motion as the previously mentioned Weibull distribution.

The exponential ($\alpha = m = 1$) and Weibull ($m = 1$) distributions are nested in the generalized gamma distribution and can therefore be discriminated amongst each other. The probability density function of the three parametric generalized gamma distributions is

$$f(t) = \frac{\alpha\, \lambda^{\alpha\, m}\, t^{\alpha\, m-1} e^{-(\lambda\, t)^{\alpha}}}{\Gamma(m)} \tag{2.13}$$

with shape parameters $\alpha > 0$, $m > 0$ and scale parameter $\lambda > 0$. Recall the gamma distribution

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

The hazard function is for $\alpha\, m > 1$ and $\alpha < 1$ an inverted U-shape (from 0 at $t = 0$ to 0 at $t \to \infty$) and for $\alpha\, m < 1$ and $\alpha > 1$ it shows a regular U-shape (from ∞ at $t \to 0$ to ∞ at $t \to \infty$). Outside these defined zones the hazards vary monotonically between 0 and ∞, are increasing if $\alpha > 1$ and decreasing if $\alpha < 1$. [Lancaster, 1994, p38 f.]

Attention should be paid to the fact that this is not intended to be a complete list of distributions used for parametric approaches in Survival Analysis but represents the distributions most commonly used.

Other failure time distribution functions are listed below:
- Gamma
- Gompertz-Makeham
- Generalized Weibull
- Exponential Weibull
- Inverse Gaussian
- Birnbaum and Saunders

More about these distribution functions can be found in the book ”Accelerated Life Models -

Modeling and Statistical Analysis" [Bagdonavičius and Nikulin, 2002, p2-17].

2.4.1. Parametric likelihood estimation

Having made a decision on the issue of specification of the distribution family, one then faces the necessity to estimate the related unknown parameters. Let us denote these parameters hereafter by the vector θ.

Recall the definition of the Likelihood-function: "*The **likelihood**, $L(\theta)$, of the observed data is a constant multiple of the joint distribution of the observed data. The **maximum likelihood estimator** $\hat{\theta}$ of θ is a function of the observed data which maximises L over values of θ in the parameter space of all possible values of θ*" (Definition 7.1) [Smith, 2002, p119]

$$L(\theta) = \prod_{i=1}^{n} f(t_i, \theta) \tag{2.14}$$

The log-likelihood function is then given by

$$ln(L(\theta)) = \sum_{i=1}^{n} ln(f(t_i, \theta)). \tag{2.15}$$

Integrating the idea of censoring into the likelihood approach, we consider that a censored observation at time t_j gives only information about the least time (t_j) the duration has lasted. Therefore it contributes via the survival function $S(t_j, \theta)$ to the log-likelihood function leading us to the following equation

$$ln(L(\theta)) = \sum_{i=1}^{n} \delta_i ln f(t_i, \theta) + \sum_{i=1}^{n} (1 - \delta_i) ln S(t_i, \theta) \tag{2.16}$$

where $\delta_i = \begin{cases} 0 \text{ if } i\text{-th spell is censored} \\ 1 \text{ if } i\text{-th spell is uncensored} \end{cases}$

Using the finding on page 4 that $f(t) = \lambda(t)S(t)$ and from equation (2.6) that $lnS(t) = -\Lambda(t)$, equation (2.16) can also be written as

$$ln(L(\theta)) = \sum_{i=1}^{n} \delta_i ln\lambda(t_i, \theta) - \sum_{i=1}^{n} \Lambda(t_i, \theta). \tag{2.17}$$

[Kiefer, 1988, p662]

Ex.2.3 `Example: Exponential model`

The hazard function and the integrated hazard of the exponential model are

$$\lambda(t) = \lambda \qquad \text{and} \qquad \Lambda(t) = \lambda t,$$

thus

$$L(\lambda) = \sum_{i=1}^{n} \delta_i ln\lambda - \lambda \sum_{i=1}^{n} t_i$$

and the first derivative is

$$\frac{\partial L(\lambda)}{\partial \lambda} = \lambda^{-1} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} t_i.$$

Maximising this equation by setting it to zero gives us the maximum-likelihood estimator as the solution of this calculation

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}.$$

Additionally note that ignoring the concept of censoring would lead to an exaggeration of probability that the unemployment spell would end after a certain duration time, i.e. the estimated hazard would face an upward bias.

2.5. Semi-parametric analysis

Whereas in the previous section the observations correspond with a predefined form up to certain parameters, the semi-parametric approach manages without exactly specifying a distribution family. The concentration on the data itself is regarded to be the principle advantage of this method.

The idea behind this modeling approach is to order the given duration data and analyze the probability of the first failure to occur. Then, excluding the first observation, examine the second and then the subsequent observations. As none of these separate analyses makes an assumption concerning the distribution of failure times, their combination does not either. Therefore time can be ignored except for an ordering of the observations, representing the non-parametric part. The parametric component is given by an assumption concerning the effect of the covariates. [Mario Cleves and Gutierrez, 2004, p3 ff.]

2.5.1. Proportional hazard

A semi-parametric model that has been in common use and well promoted in several life time analysis reports is the Cox's proportional hazard (CPH) model. It is also implemented in various programs that support mathematical functions (eg. Matlab, SPSS, R, Limdep).

The proportional hazard model is specified by

$$\lambda(t, x, \beta) = \phi(x, \beta)\lambda_0(t). \tag{2.18}$$

It is a product of two multiplicative terms, where λ_0 is called the "baseline" hazard and is that part of the function which is conditioned on all values related to the independent variables to be 0, thus representing the hazard function for an individual having $\phi(x, \beta) = 1$. Whereas $\phi(x, \beta)$ is the part that explains the influences of the covariates.
[Smith, 2002, p143 f.]

The model proposed by Cox (1972) is specified as follows

$$\lambda(t, x, \beta) = e^{\beta' x}\lambda_0(t). \tag{2.19}$$

Although the CPH itself is very well known, the diagnostic methods are not, but are essential for assessing the model. There has been an interesting approach pertaining to this issue in [Nikulin et al., 2004, p27 ff.]

A decisive benefit of PH-models can be achieved by taking advantage of the proportionality assumption. Leaving censoring aside, one gets the conditional probability that observation z finishes its spell at duration t_z, given N observations that could have ended at the same time t_z

$$\frac{\lambda(t_z, x_z, \beta)}{\sum_{i=z}^{N} \lambda(t_z, x_i, \beta)}$$

with $\lambda(t, x, \beta) = \phi(x, \beta)\lambda_0(t)$ this expression reduces to

$$\frac{\phi(x_z, \beta)}{\sum_{i=z}^{N} \phi(x_i, \beta)}. \tag{2.20}$$

This makes it possible to estimate β without specification of either the form or the family of the baseline hazard λ_0.
[Hashem Pesaran and Schmidt, 1999, p315 ff.]

Another option to incorporate the effect of the explanatory variable x on survival time, next to the ”proportional hazard”, is the modeling process called ”accelerated lifetimes”.

2.5.2. Accelerate failure time (AFT) model

In this kind of model it is assumed, that the covariates have direct influence on lifetime, slowing or speeding up its progress. The distinction between this and the PH model is consequently the rescaling of time.
In the following, $S(t, x, \beta)$ shall denote the survival function of duration $T(x)$, the observed response variable for the covariates x.

According to an AFT model the duration can be written as

$$T(x) = \frac{T_0}{\psi(x, \beta)}, \tag{2.21}$$

$\psi(.)$ being a positive function.

The survival function is given by

$$S(t, x, \beta) = S_0(t * \psi(x, \beta)), \tag{2.22}$$

where S_0, the baseline survival function, is representing the survival function for an individual having $\psi(x, \beta) = 1$.
This equation is derived from equation (2.21) by

$$S(t) = P(T > t) = P(\frac{T_0}{\psi(x, \beta)} > t) = P(T_0 > t * \psi(x, \beta)) = S_0(t * \psi(x, \beta)).$$

One gets an "accelerated duration" for $\psi(x,\beta) > 1$ and vice versa for $\psi(x,\beta) < 1$.

Choosing $\psi(x,\beta) = e^{\beta' x}$ and taking the logarithm of the above equation (2.21) we get

$$ln(T) = ln(T_0) - ln(\psi(x,\beta)) = ln(T_0) - (\beta' x), \tag{2.23}$$

a regression model for $ln(T)$ that relates $ln(T)$ linearly to the covariates.
[Smith, 2002, p148], [Lancaster, 1994, p40]

While the PH model requires an entire specification of the distribution of the error term and allows the duration time to be linearly related to $\beta' x$, as a transformation in some general way, it is the opposite with AFT models. The AFT model restricts the transformation of duration time, but allows the structure of the error time to be chosen arbitrarily. [Hashem Pesaran and Schmidt, 1999, p317]

Note: "*The proportional hazards model and the accelerated lifetime model coincide if and only if the lifetimes follow a Weibull distribution.*"
(Theorem 8.1) [Smith, 2002, p149]

2.6. Model extensions

At the end of this chapter some extensions to the basic concepts and equations of the previous sections are stated. The research and development in the field of survival analysis experiences steady modification as can be seen from the different topics chosen in other literature referred to throughout subsequent documentation. The following extensions are just a few basic aspects that the model can or should in some cases additionally allow for.

2.6.1. Ties

When there is more than a single observation at a time, then so-called ties occur. This is assumed not to be possible in a continuous case of modeling but rather common in discrete models. In practice one has to choose the smallest unit of time and therefore ties can occur in the measurements. If there are just a few ties, they can be ignored without drastic changes in the results. However, if there are a larger number of ties one should consider taking them into account by using discrete survival or failure time models which are then more consistent with the data. Often it is not clear which approach should be preferred among the choices available. Some of which worthy of note are the Breslow-, the Efron-method and the proportional odds model. The work of [Chalita et al., 2002] provides a helpful empirical guideline based on Monte Carlo simulations as well as on the amount of tied data present and the mean square error.

2.6.2. Time varying covariates

Giving attention to the covariates, x, that have been assumed to be invariant so far, one has to consider if some of them are time dependent by themselves. This implies that some variables would change over the length of the unemployment spell beside their function of influencing the duration. In this case we would use the expression $x = x(t, s)$ ($t \dots duration$, $s \dots calendertime$).
Time varying covariates can further be distinguished into *external* and *internal* covariates. A useful discussion regarding this issue is provided by [Lancaster, 1994, p323 ff.].
Briefly, we speak about external covariates if the complete time path of a covariate is predefined and known. Those covariates are conditioned on the entire path and can be treated as time invariant ones.
For some covariates, referred to as internal covariates, it is not relevant to condition on the complete path but just on the path to time t.
Examples for time varying covariates regarding unemployment data are for instance benefits received when reported unemployed, those payments may vary during the spell, or the job search intensity could possibly vary over time, too.

2.6.3. Heterogeneity

Heterogeneity occurs when different individuals exhibit different distributions of the dependent variable. Covariates are implemented into the model as a control for heterogeneity. Nevertheless, this control for the effects of explanatory variables on the dependent variable is often incomplete and one has to deal with remaining heterogeneity which can cause problems in the interpretation of results.
Heterogeneity is caused by mis-specification of the functional form and may result in misleading inferences about duration dependence or about the effects of the covariates. [Kiefer, 1988, p671]
In order to account for heterogeneity there are various possibilities to extend the duration model. The Kaplan-Meier Estimator is a simple approach to avoid heterogeneity problems but may not give a satisfactory amount of information. A more direct approach could be modeling heterogeneity in parametric models. In [Greene, 2003, p797] such an approach is given by the following common example.

Ex.2.4 `Example: Heterogeneity in the Weibull model`

Suppose that the survival function is conditioned on the individual specific effect ν_i and expressed by $S(t_i|\nu_i)$.
The unobserved heterogeneity is expressed by $f(\nu_i)$
Then

$$S(t) = E_\nu[S(t|\nu)] = \int_\nu S(t|\nu)f(\nu)d\nu.$$

Assume a gamma distribution for ν with mean $E[\nu] = 1$ and variance $Var[\nu] = \frac{1}{k} := \omega$, then

$$f(\nu) = \frac{k^k}{\Gamma(k)}e^{-k\nu}\nu^{k-1}$$

and

$$S(t|\nu) = e^{-(\nu\lambda t)^p}$$

resulting in the unconditional distribution to be

$$S(t) = \int_0^\infty S(t|\nu)f(\nu)d\nu = [1 + \omega(\lambda t)^p]^{-\frac{1}{\omega}},$$

and the corresponding hazard function to be

$$\lambda(t) = \lambda p(\lambda t)^{p-1}[S(t)]^\omega.$$

On the one hand, the problem that may arise using this approach, is over-parameterization of the survival distribution. This can lead to serious errors regarding inference of the explaining variables and is an aspect one should be aware of when including the issue of heterogeneity into the model building process.
On the other hand if the presence of unobserved heterogeneity is ignored (i.e. ignoring that $Var[\nu] > 0$) then the estimation of the duration dependence will show negative bias. This comes from the phenomenon called ”weeding out”, that individuals with high values of ν and thus higher hazards leave the state of unemployment quicker than those with lower values of ν. [Heckman and Leamer, 2001, p3407]

2.6.4. Competing risks

One may not always be solely interested in the transition from one state to another but would like to take into consideration several exit states. This is when we face the multiple destinations problem, which is dealt with by selection of a competing risk model. In the context of unemployment study, one could for instance consider two different final states when leaving unemployment, either finding a new job or withdrawing from the labour force. This last transition can have different reasons that vary from getting pregnant to becoming a retiree. Due to the apparent independence of those final states the hazard rate from unemployment is the sum of the two possible transition rates. When analyzing the transition process, the transition of every other state is considered as being censored respectively. Further information on this issue can be found in the books of [Lancaster, 1994, p99 ff.] and [Crowley and Johnson, 1981, p216 ff.] amongst others.

3. Investigating unemployment duration in Austria

3.1. Data

To investigate the determinants of unemployment duration, the data taken is a subset of the micro-census database of the Austrian statistical office (Statistics Austria).

It is a household-based survey of the year 2005, designed to measure the individual's labour market history, geographical mobility, history of studies and change in marital status amongst others. This data is based on questionnaires completed in the course of a face-to-face interview at initial contact and thereafter via telephone interviewing. Each interviewee remains in the survey sample for 5 data collections which are done at 3 monthly intervals. A unique personal identifying number is added to enable identification of repeated interviews.

In the tables displayed on the following pages all extracted and derived data from the micro-census sample set are listed.

The first part of the listed variables shows some given or derived data stratification and is the calculation base for our analysis. This data is comprised of information pertaining to the unemployment path of the observed individuals. Other parts represent all possible covariates which are examined later for significant influence on unemployment duration in the next chapter (chapter 3.2). For a better overview the variables are grouped into 'personal characteristics', 'education', 'job characteristics', 'regional variables', 'support and job search' and 'partner characteristics'. Further information according to the labelling, the grouping of categorical variables, the values, the range and the identification of missing values can be found in this descriptive table. Additionally counts of value 1 in dichotomous variables and two example data of each variable are reported to provide partial insight into the data structure. More detailed classification information according to the variables can be found later in the respective covariate examination during the course of our applications.

unemployment duration, censor- and miscellaneous variables 1/3

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
PNR	lfd. Nr.	consecutive number	-	[1,2824]	N	-	-	-
ASBPER	Personenkennzahl	personal code number	-	[10106015803, 90111286002]	derived number	-	-	-
AROT	Rotationskennzahl	rotation code number	number that determines the groups of different quarters the persons where asked in	[5,12]	N	-	-	-
Q1	Quartal1	1st quarter	1 if unemployed in this quarter	[0,1]	{0,1}	-	2044	-
Q2	Quartal2	2nd quarter	1 if unemployed in this quarter	[0,1]	{0,1}	-	2304	-
Q3	Quartal3	3rd quarter	1 if unemployed in this quarter	[0,1]	{0,1}	-	2297	-
Q4	Quartal4	4th quarter	1 if unemployed in this quarter	[0,1]	{0,1}	-	2059	-
XERWSTAT1	Erwerbsstatus Q1	employment status Q1	1=employed,2=unempl.,3=not in labour force, 4=civil-/military service	[1,9]	{1,2,3,4,9}	-1	-	-
XERWSTAT2	Erwerbsstatus Q2	employment status Q2	1=employed,2=unempl.,3=not in labour force, 4=civil-/military service	[1,9]	{1,2,3,4,9}	-1	-	-
XERWSTAT3	Erwerbsstatus Q3	employment status Q3	1=employed,2=unempl.,3=not in labour force, 4=civil-/military service	[1,9]	{1,2,3,4,9}	-1	-	-
XERWSTAT4	Erwerbsstatus Q4	employment status Q4	1=employed,2=unempl.,3=not in labour force, 4=civil-/military service	[1,9]	{1,2,3,4,9}	-1	-	-
XDAUER1	Arbeitslosendauer Q1	unemployment duration Q1	quotation of unempl. period in Q1	[0,201]	N	-1	-	-
XDAUER2	Arbeitslosendauer Q2	unemployment duration Q2	quotation of unempl. period in Q2	[0,182]	N	-1	-	-
XDAUER3	Arbeitslosendauer Q3	unemployment duration Q3	quotation of unempl. period in Q3	[0,182]	N	-1	-	-
XDAUER4	Arbeitslosendauer Q4	unemployment duration Q4	quotation of unempl. period in Q4	[0,284]	N	-1	-	-
ABSTAND12	Abstand zw. Q1 u. Q2	difference in duration inform. btw. Q1 and Q2	-	[0,55]	N	-99	-	-
ABSTAND13	Abstand zw. Q1 u. Q3	difference in duration inform. btw. Q1 and Q3	-	[0,91]	N	-99	-	-
ABSTAND14	Abstand zw. Q1 u. Q4	difference in duration inform. btw. Q1 and Q4	-	[0,93]	N	-99	-	-

unemployment duration, censor- and miscellaneous variables 2/3

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
ABSTAND24	Abstand zw. Q2 u. Q4	difference in duration inform. btw. Q2 and Q4	-	[0,127]	N	-99	-	-
ABSTAND34	Abstand zw. Q3 u. Q4	difference in duration inform. btw. Q3 and Q4	-	[0,79]	N	-99	-	-
HASDAU	Suchdauer	job seeking period	-	[0,284]	N	999	-	-
HANTR	voraussichtlicher Antritt neuer Stelle	future job prospective	-	[1]	{1}	-3	226	-
JLWA	Ende der letzten Tätigkeit	ending time of last job	-	[01.06.1962, 01.12.2005]	date	1.1.9993	-	-
DSEIT	neue Arbeit seit	starting time of new job	-	[01.01.1970, 01.12.2005]	date	1.1.9993	-	-
DISAB	Abstandsunpässlichkeit	discrepancy in duration inform. Difference	-	[0,1]	{0,1}	-	389	-
DISDS	Dseit und Jlwa-Ungereimtheit	discrepancy in starting and ending time of jobs	-	[0,1]	{0,1}	-	186	-
BEGINN	Beginn der Arbeitslosigkeit	start of unemployment period	-	[1,4]	{1,2,3,4}	-	-	-
PIVQUAR	Pivotquartal	pivot quarter	-	[1,4]	{1,2,3,4}	-	-	-
ENDEART	Art des Endes der Arbeitslosenphase	final state	0=employment, 1=censored, 2=out of labour force (inactivity), 3=civil-/military service	[0,3]	{0,1,2,3}	-	-	-
UEQUARTAL	Übertragungsquartal	active quarter	quarter after last unemployment state	[2,4]	{2,3,4}	-1	-	-
AREFWOUE	Referenzwoche der Befragung	reference week in active quarter	-	[10.04.2005, 01.01.2006]	date	no entry	-	-
AREFWOPIV	Befragungszeitpunkt im Pivotquartal	reference week in pivot quarter	-	[09.01.2005, 01.01.2006]	date	-	-	-
POTMGL	Potentiell Mögl.	possible	although defined as discrepant still possible- special cases	[0,1]	{0,1}	-	118	-
HANTR1	Antritt neuer Stelle im Q1	future job prospect in Q1	job offer within 3 months	[1]	{1}	{-3,-1}	51	-
HANTR2	Antritt neuer Stelle im Q2	future job prospect in Q2	job offer within 3 months	[1]	{1}	{-3,-1}	59	-

unemployment duration, censor- and miscellaneous variables 3/3

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
HANTR3	Antritt neuer Stelle im Q3	future job prospect in Q3	job offer within 3 months	[1]	{1}	{-3,-1}	75	-
HANTR4	Antritt neuer Stelle im Q4	future job prospect in Q4	job offer within 3 months	[1]	{1}	{-3,-1}	55	-
CENSOR	Zensuriert -comp. risks	censored data (competing risk relevant)	no longer observed - final state unknown	[0,1]	{0,1}	-	1613	dichotomy
CENSORB	Zensuriert -single-risk	censored data (single risk relevant)	no longer observed - final state unknown	[0,1]	{0,1}	-	1509	dichotomy
EFFDAUER	Effektive AL-Dauer für comp. risks	unemployment period (competing risk relevant)	-	[0,284]	N	neg. value	-	interval
EFFDAUVARB	Effektive AL-Dauer für single-risk	unemployment period (single risk relevant)	-	[0,284]	N	neg. value	-	interval

personal characteristics 1/1

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
VERW	verwitwet	widowed	-	[0,1]	{0,1}	-	32	dichotomy
GESCH	geschieden	divorced	-	[0,1]	{0,1}	-	288	dichotomy
XBSTAATO	Österreichischer Staatsbürger	Austrian citizenship	0=Austrian, 1=Foreigner	[0,1]	{0,1}	-	479	dichotomy
XBSTAAT	staatliche Zugehörigkeit	citizenship	1=Austria, 2=EU15, 3=EU25, 4=former Yugoslavia, 5=Turkey, 6=other countries	[1,6]	{1,2,3,4,5,6}	-	-	nominal
OE	Österreich	Austria	-	[0,1]	{0,1}	-	2345	dichotomy
EU15	15 EU-Mitgliedsstaaten	15 former European countries	-	[0,1]	{0,1}	-	57	dichotomy
EU25	neuen 25 EU-Mitgliedsstaaten	25 additional European countries	-	[0,1]	{0,1}	-	42	dichotomy
YU	ehem. Jugoslawien	former Yugoslavia	-	[0,1]	{0,1}	-	185	dichotomy
TURK	Türkei	Turkey	-	[0,1]	{0,1}	-	124	dichotomy
SONST	Sonstige	other countries	-	[0,1]	{0,1}	-	71	dichotomy
XBGEBLAO	Geburtsland Österreich	Austria as country of birth	0=Austria, 1=Foreign country	[0,1]	{0,1}	-	696	dichotomy
XANZKIND	Anzahl der Kinder	number of children	-	[0,6]	N	-3	-	interval

education 1/1

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
XKARTAB	Ausbildung	education	categories 1-8, see below	[1,8]	N		-	nominal
PFLSCH	Pflichtschule	compulsory education or none at all	1	[0,1]	{0,1}	-	947	dichotomy
LEHRAB	Lehrabschluss	apprenticeship	2	[0,1]	{0,1}	-	976	dichotomy
BMS	Berufsbildende Mittlere Schule	secondary vocational school	3	[0,1]	{0,1}	-	339	dichotomy
AHS	Allgemeinbildende Höhere Schule	high school	4	[0,1]	{0,1}	-	175	dichotomy
BHS	Berufsbildende Höhere Schule	vocational high school	5	[0,1]	{0,1}	-	187	dichotomy
KOLLEG	Kolleg	college	6	[0,1]	{0,1}	-	19	dichotomy
KURZSTUD	Kurzstudium	bachelor studies	7	[0,1]	{0,1}	-	34	dichotomy
UNI	Universität	university	8	[0,1]	{0,1}	-	147	dichotomy

job characteristics 1/4

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
JBERS	frühere berufliche Stellung	former job position	categories 0-8, see below	[0,8]	N	-	-	nominal
NIEGEARB	nie gearbeitet	never worked so far	0	[0,1]	{0,1}	-	435	dichotomy
ANG	Angestellter	white-collar worker (employee)	1	[0,1]	{0,1}	-	1007	dichotomy
ARBEIT	Arbeiter	blue-collar worker	2	[0,1]	{0,1}	-	1165	dichotomy
BEAMT	Beamter	official	3	[0,1]	{0,1}	-	13	dichotomy
VB	Vertragsbediensteter	contract agent	4	[0,1]	{0,1}	-	53	dichotomy
FREIER	Freier Dienstnehmer	freelancer	5	[0,1]	{0,1}	-	28	dichotomy
SELBO	Selbstständiger ohne Arbeitnehmer	self-employed person without further staff	6	[0,1]	{0,1}	-	74	dichotomy
SELBM	Selbstständiger mit Arbeitnehmer	self-employed person with staff	7	[0,1]	{0,1}	-	16	dichotomy
MITH	Mithelfender Familienangehöriger	"job supporting" family member	8	[0,1]	{0,1}	-	15	dichotomy
JTAET	Art der früheren Tätigkeit	type of former profession	categories 0-17, see *)	[0,17]	N	-	-	nominal
NIETAET	nie eine Tätigkeit ausgeübt	never had any kind of profession	0	[0,1]	{0,1}	-	435	dichotomy
LEHRLING	Lehrling	apprentice	1,6	[0,1]	{0,1}	-	105	dichotomy
HILFSTAET	Hilfstätigkeit	auxiliary work	2,7	[0,1]	{0,1}	-	425	dichotomy
LANDWIRT	Landwirtschaftl. Tätigkeit	farmer	12,13	[0,1]	{0,1}	-	17	dichotomy
HOCHNM	höhere Tätigkeit nicht manuell	high profession non manual	9,10,11	[0,1]	{0,1}	-	180	dichotomy
SELBSTST	Selbstständiger	self-employed	15,16,17	[0,1]	{0,1}	-	88	dichotomy
MITTNM	mittlere Tätigkeit nicht manuell	medium profession non manual	8	[0,1]	{0,1}	-	475	dichotomy
ANGM	Angestellter manuell	manual employee	3	[0,1]	{0,1}	-	649	dichotomy
HOCHM	höhere Tätigkeit manuell	high profession manual	4,5	[0,1]	{0,1}	-	433	dichotomy
DBERS	jetzige berufliche Stellung	current job position	categories 1-8, see next page	[1,8]	N	{-1,-3}	-	nominal

job characteristics 2/4

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
DANG	Angestellter	white-collar worker (employee)	1	[0,1]	{0,1}	-	303	dichotomy
DARB	Arbeiter	blue-collar worker	2	[0,1]	{0,1}	-	320	dichotomy
DBEAMT	Beamter	official	3	[0,1]	{0,1}	-	2	dichotomy
DVB	Vertragsbediensteter	contract agent	4	[0,1]	{0,1}	-	6	dichotomy
DFREIER	Freier Dienstnehmer	freelancer	5	[0,1]	{0,1}	-	22	dichotomy
DSELBO	Selbstständiger ohne Arbeitnehmer	self-employed person without further staff	6	[0,1]	{0,1}	-	21	dichotomy
DSELBM	Selbstständiger mit Arbeitnehmer	self-employed person with staff	7	[0,1]	{0,1}	-	4	dichotomy
DMITH	Mithelfender Familienangehöriger	"job supporting“ family member	8	[0,1]	{0,1}	-	2	dichotomy
DTAET	Art der Tätigkeit	type of current profession	categories 1-17, see **)	[1,17]	N	{-1,-3}	-	nominal
DLEHRLING	Lehrling	apprentice	1,6	[0,1]	{0,1}	-	61	dichotomy
DHILFSTAET	Hilfstätigkeit	auxiliary work	2,7	[0,1]	{0,1}	-	147	dichotomy
DLANDWIRT	Landwirtschaftl. Tätigkeit	farmer	12,13	[0,1]	{0,1}	-	1	dichotomy
DHOCHNM	höhere Tätigkeit nicht manuell	high profession non manual	9,10,11	[0,1]	{0,1}	-	45	dichotomy
DSELBSTST	Selbstständiger	self-employed	15,16,17	[0,1]	{0,1}	-	26	dichotomy
DMITTNM	mittlere Tätigkeit nicht manuell	medium profession non manual	8	[0,1]	{0,1}	-	108	dichotomy
DANGM	Angestellter manuell	manual employee	3	[0,1]	{0,1}	-	183	dichotomy
DHOCHM	höhere Tätigkeit manuell	high profession manual	4,5	[0,1]	{0,1}	-	109	dichotomy
DSTD	Wochenarbeitsstunden	weekly working hours	(restricted to 80 hours)	[0,80]	N	{-1,999}	-	interval
XDWZAB	Wirtschaftssektor (ÖNACE)	branch of industry	A-Q (no entry for B=fishing)	[1,17]	N	{-1,-3}	-	nominal
A	Land- u. Forstwirtschaft	agriculture and forestry	1	[0,1]	{0,1}	-	5	dichotomy
C	Bergbau	mining industry	3	[0,1]	{0,1}	-	0	dichotomy

job characteristics 3/4

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
D	Sachgütererzeugung	manufacture	4	[0,1]	{0,1}	-	113	dichotomy
E	Energie- u. Wasserversorgung	energy production and water supply	5	[0,1]	{0,1}	-	3	dichotomy
F	Bauwesen	civil engineering	6	[0,1]	{0,1}	-	87	dichotomy
G	Handel; Reparatur v. Kfz u. Gebrauchsgütern	trade and repairs	7	[0,1]	{0,1}	-	131	dichotomy
H	Beherberungs- u. Gaststättenwesen	catering trade	8	[0,1]	{0,1}	-	93	dichotomy
I	Verkehr- und Nachrichtenübermittlung	traffic and news transfer	9	[0,1]	{0,1}	-	32	dichotomy
J	Kredit- u. Versicherungswesen	bank and insurance	10	[0,1]	{0,1}	-	15	dichotomy
K	Realitätenwesen, Unternehmensdienstl.	realities and service sector	11	[0,1]	{0,1}	-	85	dichotomy
L	Öffentl. Verwaltung, Sozialversicherung	public sector and social insurance	12	[0,1]	{0,1}	-	17	dichotomy
M	Unterrichtswesen	teaching sector	13	[0,1]	{0,1}	-	20	dichotomy
N	Gesundheits-, Veterinär- und Sozialwesen	health, veterinary and social sector	14	[0,1]	{0,1}	-	43	dichotomy
O	Erbring. v. sonstigen öffentl. u. pers. Dienstl.	other public and personal services	15	[0,1]	{0,1}	-	32	dichotomy
P	Private Haushalte	private household	16	[0,1]	{0,1}	-	2	dichotomy
Q	Exterritoriale Organisationen	exterritory organizations	17	[0,1]	{0,1}	-	2	dichotomy
JLWI	wodurch wurde letze Arbeit beendet	reason for last job exit	categories 1-12, see below	[1,12]	N	-3	-	nominal
PENSION	Pensionierung	retirement	1,2,3 ***)	[0,1]	{0,1}	-	100	dichotomy
KUEND	Kündigung durch den Arbeitgeber	dismissal	4	[0,1]	{0,1}	-	830	dichotomy

job characteristics 4/4

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
KRANK	Krankheit od. Arbeitsunfähigkeit	illness or disability	5	[0,1]	{0,1}	-	113	dichotomy
ABLAUF	Ablauf eines befristeten Arbeitsvertrages	end of contract	6	[0,1]	{0,1}	-	275	dichotomy
PFLEGE	Betreuung von Kindern od. Pflegebedürftigen	caring for children or persons in need	7,10 ***)	[0,1]	{0,1}	-	203	dichotomy
ZIVI	Zivil- od. Präsenzdienst	civilian or military service	8	[0,1]	{0,1}	-	38	dichotomy
SCHULE	Schulische od. berufl. Ausbildung	education	9	[0,1]	{0,1}	-	40	dichotomy
RES	Selbstkündigung, einvernehmliche Lösung	resignation, mutually agreed working contract termination	11	[0,1]	{0,1}	-	531	dichotomy
SONSTIG	Sonstiger Beendigungsgrund	other exit reason	12	[0,1]	{0,1}	-	100	dichotomy

*)
0 .... nie gearbeitet=NIETAET (no profession)
1+6 .... Lehrvertrag(Lehrling) + Lehrvertrag(nicht manuell)=LEHRLING (apprenticeship)
2+7 .... Hilfstätigkeit(manuell) + Hilfstätigkeit(nicht manuell)=HILFSTAET (auxiliary work)
3 .... Angelernte Tätigkeit (manuell)=ANGM (employee (for manual work))
4+5 .... Facharbeiter, Vorarbeiter/Meister (manuell)=HOCHM (high profession- manual work)
8 .... Mittlere Tätigkeit (nicht manuell)=MITTNM (medium profession- non manual work)
9+10+11 .... Höhere, hochqualifizierte u. führende Tätigkeit (nicht manuell)=HOCHNM (high profession- non manual work)
12+13+15+16+17... Landwirtschaft klein + mittel=LANDWIRT (farmer), Freiberufler, neue Selbständige, Gewerbescheinbesitzer (freelancer)

**) same categories as listed above in *) but without class “0”.

***)
1+2+3 .... Mit dem Pensionsantritt gesetzlichen Pensionsalter + Frühpensionierung + Invaliditätspension (retirement)
7+10 .... Betreuung von Kindern oder pflegebedürftigen Erwachsenen + Andere persönliche oder familiäre Verpflichtungen (caring for children or other persons in need of help)

regional variable 1/1

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
XEINW	Größe der Gemeinde	size of municipality or town	categories 1-15, see ****)	[1,15]	N	-	-	ordinal
XNUTS2	Bundesland	federal states	categories, see below	[11,34]	{11,12,13, 21,22,31, 32,33,34}	-	-	nominal
BURGENLAND	Burgenland	Burgenland	11	[0,1]	{0,1}	-	273	dichotomy
KAERNTEN	Kärnten	Carinthia	21	[0,1]	{0,1}	-	328	dichotomy
NIEDEROE	Niederösterreich	Lower Austria	12	[0,1]	{0,1}	-	311	dichotomy
OBEROE	Oberösterreich	Upper Austria	31	[0,1]	{0,1}	-	269	dichotomy
SALZBURG	Salzburg	Salzburg	32	[0,1]	{0,1}	-	234	dichotomy
STEIERMARK	Steiermark	Styria	22	[0,1]	{0,1}	-	291	dichotomy
TIROL	Tirol	Tirol	33	[0,1]	{0,1}	-	260	dichotomy
WIEN	Wien	Vienna	13	[0,1]	{0,1}	-	523	dichotomy
VLBG	Vorarlberg	Vorarlberg	34	[0,1]	{0,1}	-	335	dichotomy

****)
1 .... bis 500
2 501-1000
3 1001-1500
4 1501-2000
5 2001-2500
6 2501-3000
7 3001-5000
8 5001-10000
9 10001-20000
10 20001-30000
11 30001-50000
12 50001-100000
13 100001-200000
14 200001-500000
15 .... Wien (Vienna)

support and job search 1/2

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
HAMS	Vormerkung beim AMS	notification at Austrian Labour Market Service	1 if notified	[0,1]	{0,1}	-	774	dichotomy
HAMSL	Leistungen vom AMS	receipt of services from the AMS	categories, see *****)	[0,630000]	37 diff. classes	-3	-	nominal
ALG	Arbeitslosengeld	unemployment benefit	<170000	[0,1]	{0,1}	-	2302	dichotomy
NOTSTAND	Nostandshilfe	minimum financial benefit	190000<x<270000	[0,1]	{0,1}	-	427	dichotomy
GELDSONST	sonstige Geldzuwendungen	other financial support	360000,500000,600000 or 630000	[0,1]	{0,1}	-	18	dichotomy
SCHULUNG	Schulungen	course of training	130000,230000,630000 or 290000<x<400000	[0,1]	{0,1}	-	139	dichotomy
HSART1	Kontakt mit AMS	in contact with AMS	-	[0,1]	{0,1}	-3	1987	dichotomy
HSART2	Jobangebot vom AMS	job offer placed via AMS	-	[0,1]	{0,1}	-3	1006	dichotomy
HSART3	Stellenangebot in Zeitungen studiert	looking for job offers in newspapers	-	[0,1]	{0,1}	-3	2335	dichotomy
HSART4	Freunde, Bekannte gefragt	verbal exchange with friends	-	[0,1]	{0,1}	-3	2125	dichotomy
HSART5	Bewerbung an Arb.geber geschickt	application by mail	-	[0,1]	{0,1}	-3	1716	dichotomy
HSART6	Stellenang. in Zeitungen aufgegeb. od. beworben	application via advertisement	-	[0,1]	{0,1}	-3	1128	dichotomy
HSART7	Warten auf Antwort auf Bewerbung	waiting for reply to application	-	[0,1]	{0,1}	-3	1475	dichotomy
HSART8	Bewerbungsgespräche geführt	job interview	-	[0,1]	{0,1}	-3	1486	dichotomy
HSART9	Warten auf Antwort vom AMS	waiting for placements through AMS	-	[0,1]	{0,1}	-3	1032	dichotomy
HSART10	Verbindung mit privater Stellenvermittlg. aufgen.	in contact with private employment agency	-	[0,1]	{0,1}	-3	386	dichotomy
HSART11	Räume od. Ausrüstung für Selbstst.igkeit ges.	preparation for self-employment	-	[0,1]	{0,1}	-3	130	dichotomy

support and job search 2/2

HSART12	Warten auf Ergebnisse v. Ausschreibungen	waiting for reply to advertisement	-	[0,1]	{0,1}	-3	118	dichotomy
HSART13	Bemühen um Genehmigungen u. Konzessionen	trying to get licences	-	[0,1]	{0,1}	-3	96	dichotomy
HSART14	Arbeitssuche auf andere Weise	other kinds of job seeking	-	[0,1]	{0,1}	-3	125	dichotomy

*****)

ALG= (unemployment benefit)
- 100000,00 Arbeitslosengeld
- 120000,00 Arbeitslosengeld+Notstandshilfe
- 130000,00 Arbeitslosengeld+Kurse,Schulungsmaßnahmen

NOTSTAND= (minimum financial benefit)
- 200000,00 Notstandshilfe
- 210000,00 Arbeitslosengeld+Notstandshilfe
- 230000,00 Notstandshilfe+Kurse,Schulungsmaßnahmen

SCHULUNG= (educational training)
- 130000,00 Arbeitslosengeld+Kurse,Schulungsmaßnahmen
- 230000,00 Notstandshilfe+Kurse,Schulungsmaßnahmen
- 300000,00 Kurse, Schulungsmaßnahmen
- 310000,00 Arbeitslosengeld+Kurse,Schulungsmaßnahmen
- 320000,00 Notstandshilfe+Kurse,Schulungsmaßnahmen

GELDSONST= (other (financial) support)
- 500000,00 Pensionsvorschuss
- 600000,00 Andere Leistungen

partner characteristics 1/2

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
PARTNERERW	Erwerbsstatus des Partners	employment status of partner	1=employed, 2=unempl., 3=not in labor force, 4=civil-/military service	[1,4]	{1,2,3,4}	{-1,0}	-	nominal
PARTERW1	Partner erwerbstätig	partner employed	-	[0,1]	{0,1}	-	922	dichotomy
PARTERW2	Partner arbeitslos	partner unemployed	-	[0,1]	{0,1}	-	111	dichotomy
PARTERW3	Partn. nicht erwerbstätig	partner inactive	-	[0,1]	{0,1}	-	262	dichotomy
PARTNERSEK	Wirtschaftssektor des Partners	branch of industry of partner	A-P (no entry for B=fishing and Q=exterritorial organisations)	[1,16]	N	{-3,-1}	-	nominal
PA	Land- u. Forstwirtschaft	agriculture and forestry	1	[0,1]	{0,1}	-	15	dichotomy
PC	Bergbau	mining industry	3	[0,1]	{0,1}	-	6	dichotomy
PD	Sachgütererzeugung	manufacture	4	[0,1]	{0,1}	-	220	dichotomy
PE	Energie- u. Wasserversorgung	energy production and water supply	5	[0,1]	{0,1}	-	8	dichotomy
PF	Bauwesen	civil engineering	6	[0,1]	{0,1}	-	99	dichotomy
PG	Handel; Reparatur v. Kfz u. Gebrauchsgütern	trade and repairs	7	[0,1]	{0,1}	-	126	dichotomy
PH	Beherberungs- u. Gaststättenwesen	catering trade	8	[0,1]	{0,1}	-	66	dichotomy
PAI	Verkehr- und Nachrichtenübermittlung	traffic and news transfer	9	[0,1]	{0,1}	-	69	dichotomy
PJ	Kredit- u. Versicherungswesen	bank and insurance	10	[0,1]	{0,1}	-	28	dichotomy
PK	Realitätenwesen, Unternehmensdienstl.	realities and service sector	11	[0,1]	{0,1}	-	85	dichotomy
PL	Öffentl. Verwaltung, Sozialversicherung	public sector and social insurance	12	[0,1]	{0,1}	-	53	dichotomy
PM	Unterrichtswesen	teaching sector	13	[0,1]	{0,1}	-	39	dichotomy
PN	Gesundheits-, Veterinär- und Sozialwesen	health, veterinary and social sector	14	[0,1]	{0,1}	-	72	dichotomy
PO	Erbring. v. sonstigen öffentl. u. pers. Dienstl.	other public and personal services	15	[0,1]	{0,1}	-	34	dichotomy

partner characteristics 2/2

label	variable name (German)	variable name (English)	description	range	values	default values	counts	type of scale
PP	Private Haushalte	private household	16	[0,1]	{0,1}	-	2	dichotomy
PARTDBERS	Berufliche Stellung des Partners	current job position of partner	categories 1-8, see below, (2110 missing values)	[0,8]	N	{-3,-1,0}	-	nominal
PANG	Angestellter	white-collar worker (employee)	1	[0,1]	{0,1}	-	359	dichotomy
PARB	Arbeiter	blue-collar worker	2	[0,1]	{0,1}	-	414	dichotomy
PBEAMT	Beamter	official	3	[0,1]	{0,1}	-	41	dichotomy
PVB	Vertragsbediensteter	contract agent	4	[0,1]	{0,1}	-	30	dichotomy
PFREIER	Freier Dienstnehmer	freelancer	5	[0,1]	{0,1}	-	6	dichotomy
PSELBO	Selbstständiger ohne Arbeitnehmer	self-employed person without further staff	6	[0,1]	{0,1}	-	41	dichotomy
PSELBM	Selbstständiger mit Arbeitnehmer	self-employed person with staff	7	[0,1]	{0,1}	-	31	dichotomy
PMITH	Mithelfender Familienangehöriger	"job supporting“ family member	8	[0,1]	{0,1}	-	0	dichotomy
PARTDTAET	Tätigkeit des Partners	type of current profession of partner	categories 1-17, see *), (2110 missing values)	[0,17]	N	{-3,-1,0}	-	nominal
PLEHRLING	Lehrling	apprenticeship	1,6	[0,1]	{0,1}	-	2	dichotomy
PHILFSTAET	Hilfstätigkeit	auxiliary work	2,7	[0,1]	{0,1}	-	127	dichotomy
PLANDWIRT	Landwirtschaftl. Tätigkeit	farmer	12,13	[0,1]	{0,1}	-	11	dichotomy
PHOCHNM	höhere Tätigkeit nicht manuell	high profession- non manual	9,10,11	[0,1]	{0,1}	-	158	dichotomy
PSELBSTST	Selbstständiger	self-employed	15,16,17	[0,1]	{0,1}	-	61	dichotomy
PMITTNM	mittlere Tätigkeit nicht manuell	medium profession- non manual	8	[0,1]	{0,1}	-	162	dichotomy
PANGM	Angestellter manuell	employee- manual	3	[0,1]	{0,1}	-	216	dichotomy
PHOCHM	höhere Tätigkeit manuell	high profession- manual	4,5	[0,1]	{0,1}	-	185	dichotomy

3.2. Data exploration

Initial data filtering was carried out using the program 'Foxpro'. Records were to be examined for each individual and therefore all the data pertaining to an observed individual had first to be merged from the whole data set entries. From this the subset with those individuals having been unemployed at some time over the year has been extracted for further calculations. In our data selection, the duration variable labelled as `EFFDAUER` (competing risk) or `EFFDAUVARB` (single risk) represents the response variable to be explained by ascertained covariates (note: the variable listing is recorded in the previous section 3.1). The duration under consideration is the one of the last occurred unemployment period of a person in the selected subset. This guarantees independence of the data under observation as an individual does not recur in the final selected sample set due to multiple unemployment spells. However, such temporary layoffs could be integrated in the model to give additional information in form of multiple phase duration models. A paper focusing on this area of expertise is provided by [Jensen and Svarer, 2003].

The definition of unemployment considered is the one according to the methodology of the International Labour Organisation (ILO). Figure 3.1 provides a graphical illustration.

The individuals surveyed are of working age ($\geq$ 15 years) and are divided into three mutually exclusive and exhaustive groups, namely "persons in employment", "unemployed persons" and "inactive persons" (i.e. persons considered as having left the labour force). Through the survey questionnaire the interviewees are treated as being in one of these groups depending on the stated actual activity within a particular reference week. The groups are defined as follows:

1 *'**Employed persons** are persons aged 15 year and over, who during the reference week performed work, even for just one hour a week, for pay, profit or family gain or were not at work but had a job or business from which they were temporarily absent because of, e.g., illness, holidays, industrial dispute and education and training.'*
— [ILO-statement].

2 *'**Unemployed persons** are persons aged 15-74, who were without work during the reference week, were currently available for work and were either actively seeking work in the past four weeks or had already found a job to start within the next three months.'*
— [ILO-statement].

3 *'**Inactive persons** are those who neither classified as employed nor as unemployed.'*
— [ILO-statement].

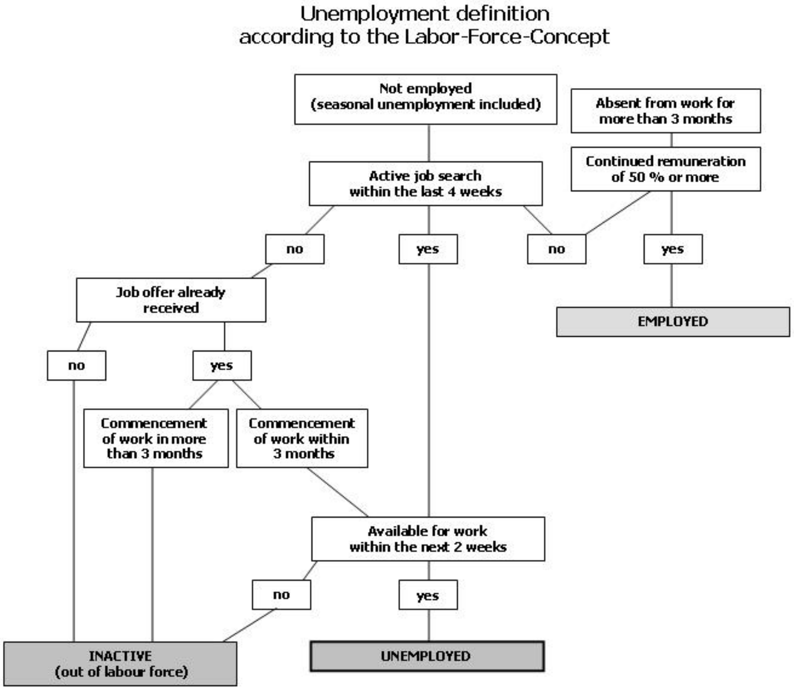

Unemployment definition
according to the Labor-Force-Concept
Not employed
(seasonal unemployment included)
Absent from work for more than 3 months
Continued remuneration of 50 % or more
Active job search within the last 4 weeks
no
yes
no
yes
EMPLOYED
Job offer already received
no
yes
Commencement of work in more than 3 months
Commencement of work within 3 months
Available for work within the next 2 weeks
no
yes
INACTIVE
(out of labour force)
UNEMPLOYED

Figure 3.1.: unemployment definition - Labour Force Concept

Another important definition that has to be stated at the beginning of this section is the one for duration of unemployment [ILO-statement]:
*'**Duration of unemployment** is defined as:*

- *the duration of search for a job, or*
- *the length of the period since the last job was held (if this period is shorter than the duration of search for a job).'*

Given the repeated interviewing data from the questionnaires over time, the change of the individuals' status, which group they were currently assigned to, was observed. The period of unemployment could be evaluated in this way for approximately half of all observed individuals while the other half exhibits an uncertain exit from the unemployment period as the survey ended without exit information. Those data have to be regarded as censored *(for more details to censoring see section 2.2)*.
In this way transition rates are obtained by identifying the unemployed individuals and the elapsed unemployment period where the exit from unemployment status could result in employment or inactivity. These two different destinations need to be considered and this matter will subsequently be satisfied by implementing competing risk models *(section 3.3.2)*.
What also has to be borne in mind is that different individuals have different time origins for the unemployment duration they experience. Unemployment spells in our investigation can begin at any date (before 2005 or within this year) which is then defined to be the time origin for the spell. The duration of a spell is its length. Those spell lengths represent our dependent variables under study. [Kiefer, 1988, p650]
The duration of staying unemployed shows notable variation. Some individuals leave the

status of unemployment rather quickly, after a few weeks, while others remain unemployed for several years. Nevertheless, given the structure of the records, the most appropriate unit of analysis for these estimations seems to be a monthly one. Furthermore, it is reasonable and necessary for logarithmic calculations to draw a cutting line, which is defined to be at 0.25 months, which is in accordance with 7.5 days or is approximately one week. This is a plausible assumption and represents a subtler specification than rounding off to 0 or 1. The resulting range in our data for the lengths of duration spells is [0.25,284].

After scaling into quarters and observing the respective group status of the individual, the last quarter, in which the person occurred as unemployed before exiting in any way, was defined and named as the "pivot quarter". When the exit was due to a new job within the next quarter, then the duration of the unemployment spell was calculated as follows

```
effdauer = xdauer(pivquar) + [month(dseit)-month(arefwopiv)],
```

where `xdauer` is the minimum of the searching period `(hasdau)` and the time difference to the last job exit `(jlwaz)`.

Further, `month(dseit)` is the month in which the new job was entered and `month(arefwopiv)` the month in which the interviewee was still unemployed and the interview took place (reference week in pivot quarter). For other exits, either quitting the labour market or being censored, the spells have to be calculated as

```
effdauer = xdauer(pivquar)
```

Unfortunately the actual exit point here cannot be determined as precisely as in the former case.

An additional calculation for the expected value of actually getting work was conducted for the case that the data entry had been censored and showed a value of 1 for `hantr` before. Note, `hantr` equal to 1 indicates that the observed individual was expected to start work within 3 months from the time of the questioning. Derived from the observed values from datasets which were not censored, the expected value resulted in 0.94 which is rounded up to 1 month and added to `effdauer` in the cases concerned (i.e. where `hantr==1`) resulting in `effdauvarb`. This amendment refers to single risk models while in competing risk models it is of no avail and `effdauer` is regarded as being the dependent variable in that situation. The intention of this slight amendment was to increase the volume of uncensored data.

Finally the whole dataset, including spell durations and censoring indicator, was screened for detection of discrepancies in the answers of the interviewees. It is essential to check the data for the presence of non-sense or inconsistent variables. Those entries with discrepancies according to spell duration information (`DISAB`) and to end- and restart-date of jobs (`DISDS`) have been eliminated from the sample dataset (`DISAB+DISDS-POTMGL`).

Comprehensive analysis can begin utilising the prepared calculation base of the original data including the demographic and other individual characteristics. Values for those characteristics are taken from the corresponding pivot quarter. These include personal, regional, educational and job information as well as information about partners (the partner information has been extracted from the data base, given the personal identification number `asbper`). More details on the structure and significance of the characteristics are identified in sections 3.4, 2.4 and 3.6.1. Further information about software in use for the applications can be found in the appendix A.1.

3.3. Single versus competing risks

3.3.1. Case distinction

A general approach of analysing duration data is to consider only a single type of exit event which represents the end of an unemployment spell by whatever reason. This restriction gives us a single-risk model. What is emphasized again is that the term ”risk” in our context is actually the ”chance” to leave the unemployment status. Despite the possible lack of information concerning the exit distinction a single-risk model often reveals sufficiently satisfactory results. However, the end of the unemployment period in our sample set is not caused by a single event but can be divided into exits by different routes. One of these routes is the exit from unemployment to a job. The other one that is taken into account in our study is the exit to economic inactivity. As those two exit reasons are mutually exclusive, that means one cannot occupy the state of economic inactivity and be employed at the same time, we speak of competing risks. Consequently, the term 'competing-risk models' is used to refer to models of such kind. A possible extension to this competing risk approach would be further distinction between full time (re)employment or start of part-time work after the unemployment spell. Referring to this risk distinction and considering frailties and smoothing effects is given in a paper of [Kauermann and Khomski, 2006] or a paper using Canadian data of [McCall, 1997]

In our study we do not distinguish further between the two exits into employment or inactivity. But we incorporate the working hours as potential covariate of the model in the case of (re)employment.
The final choice of structuring the data for the estimation procedure is done as follows.

First we regard all exits as one event of termination in total and do not further distinguish. Including the exit into inactivity in the cause of termination prevents the results to be biased due to censoring. Let this be explained in more detail. The Kaplan-Meier estimator assumes underlying independence of the censoring distribution. This implies that at the discretized time points the hazard of the event of interest, i.e. return to employment, is the same for individuals that have not failed until then as for those having experienced a competing event (exit to inactivity). This would ignore the fact that an individual being censored because of failure from the competing risk is not able to experience the event of interest. [Putter et al., 2007]
To avoid this overestimated probability of failure the first attempt is to consider a single event of interest, namely 'either exit' (case 1) out of unemployment as the only event of interest.

In the second treatment of data exploration, exit distinction is taken into consideration. A division is made into exit into 'employment' (case 2) or into 'economic inactivity' (case 3). The intention is to detect any special influence structures for both exit possibilities. The situation of an individual appears different in both exit routes which is likely to be represented in different influencing factors. It is intended to give this fact its emphasis by fitting a separate model for each transition. In order to prevent any bias in the estimation, only uncensored sample data is taken for the observation of effects on unemployment duration. The intention behind this restriction is similar to that mentioned in the previous paragraph. The disadvantage of this approach is that censored data is ignored but would incorporate additional information concerning the probability that the censored individual is still at risk.

This gives reason to consider a third approach of data exploration, to include the idea of competing risks of the second approach but still observing the censoring in the data. This is achieved by the implementation of the sub-distribution function, called 'cumulative incidence function'. The competing risks approach is further applied to CPH-models. The basic issue in competing risks models and the introduction and explanation of the functions involved is succinctly stated in the next section 3.3.2.

For further reading on the topic of comparison and benefits of competing risk approaches one is referred to [Pintilie, 2006].

3.3.2. Competing risks approach

In general there may be several reasons to exit the state of unemployment. Retirement, child or family care and working disability are stated to mention a few of them. The main and most interesting case in respect of unemployment prevention policy is to leave unemployment because of a job offer, thus to return to or start working life. For simplicity we combine all other reasons of exit to a single event which represents the exit to *economic inactivity* (subsequently referred to as case 3). This exit into inactivity prevents the occurrence of the event of interest, the exit to *employment* (case 2). Therefore we speak of competing risks, as both transition states are independent from each other. Hence the time to an exit of the state of unemployment, defined as time to first departure, is given by the minimum of time to event of either exit case 2 (T_2) or 3 (T_3). Though the interest lies in estimating the probability of case 1 departure by a certain time (t), it can only be observed for an individual if the time for the competing departure is lower. The mathematical term for this estimable probability is

$$P(T_2 \leq t, T_3 > T_2)$$

This expression is called subdistribution function or cumulative incidence function (CIF) Another expression for the CIF can be derived from the cause-specific hazard and is stated below in equation 3.5. The cause of departure is denoted $C = 2, 3$. Thus the cause-specific hazard can be expressed as

$$\lambda_2(t) = \lim_{\delta \to 0} \frac{P(t \leq T < t + \delta, C = 2 | T \geq t)}{\delta} \tag{3.1}$$

or equivalently for the second type of exit as

$$\lambda_3(t) = \lim_{\delta \to 0} \frac{P(t \leq T < t + \delta, C = 3 | T \geq t)}{\delta} \tag{3.2}$$

(Compare with the general hazard definition *(equation* (2.4) *on page* 4*)*. The resulting cumulative hazard is then defined by

$$\Lambda_k(t) = \int_0^t \lambda_k(u) du \qquad for: \ k \in \{2, 3\}$$

This formula is part of the cause-specific Survival function

$$S_k(t) = exp(-\Lambda_k(t)) \qquad for: \ k \in \{2, 3\} \tag{3.3}$$

The survival function as probability for no departure of either case is then given by

$$S(t) = exp(-\Lambda_2(t) - \Lambda_3(t)) \tag{3.4}$$

The finally resulting CIF (cumulative incidence function) from the above equations is expressed as

$$I_k(t) = \int_0^t \lambda_k(u)S(u)du \qquad for: \ k \in \{2,3\} \tag{3.5}$$

[Putter et al., 2007]

The non-parametric cumulative incidence estimator is defined as the sum of the unconditional probability of departure from cause 2 or 3 at time t_j

$$\hat{I}_k(t) = \sum_{j:t_j \leq t} \hat{\lambda}_k(t_j)\hat{S}(t_j) \tag{3.6}$$

where $\hat{\lambda}_k(t_j) = \frac{d_{kj}}{r_j}$ (*compare equation* (2.10))

The estimator $\hat{I}_k(t)$ according to the definition of Kalbfleisch and Prentice (1980, p169) [Tableman and Kim, 2004, p199] is derived as follows. Similar to formula (2.7) for the single-risk consideration, in case of competing risks we get

$$r_j = \sum_{i \geq j}(d_{2i} + d_{3i} + c_i) = n - \sum_{i<j}(d_{2i} + d_{3i} + c_i) \tag{3.7}$$

where

r_j	...	number of individuals at "risk" (i.e. have not experienced any exit) just before t_j
n	...	total number of individuals under study
d_{ki}	...	number of individuals who experience exit-cause k at t_i
c_i	...	censored at t_i

The Kaplan-Meier estimate of survival regarding to exit cause 2 is given as

$$\hat{S}_2(t) = \prod_{t_j \leq t}(\frac{r_j - d_{2j}}{r_j}) \tag{3.8}$$

The individuals who depart from unemployment due to the competing exit cause 3 are treated as censored. In similar manner we calculate the survival estimate for exit cause 3 as

$$\hat{S}_3(t) = \prod_{t_j \leq t}(\frac{r_j - d_{3j}}{r_j}) \tag{3.9}$$

and regard departures of cause 2 as being censored.
The probability of "surviving" all causes, thus staying unemployed beyond time t ($P(min(T_2, T_3) > t)$ is defined as the product

$$\hat{S}(t) = \hat{S}_2(t) \times \hat{S}_3(t) \tag{3.10}$$

Therefrom we derive the cumulative incidence estimate of equation (3.6).

The non-parametric estimation of cumulative incidence curves is an easy and reasonable approach for investigation of the effect of a categorical or dichotomous covariate. For testing if the curves differ by covariate value a log-rank test is performed. The application to our data is found in chapter 3.4.

In order to account for a higher number of covariates in the evaluation of their importance for duration (survival) prediction, we will also investigate competing risks incorporated in the CPH-model (*section* 3.6). Similar to the single-risk approach introduced in section 2.5.1 (*compare formula* (2.18)) the model for the cause-specific hazard for the two distinct cases 2 and 3 is defined as

$$\lambda_k(t, x, \beta_k) = e^{\beta'_k x} \lambda_{k,0}(t) \tag{3.11}$$

where x represents the covariate vector and β_k the covariate effects on cause $k \in \{2, 3\}$. Remember that $\lambda_{k,0}(t)$ is the baseline cause-specific hazard.
We investigate each time point at which case 2 occurs. The covariate values of a person moving to the state of employment are then compared to other individuals who appear event-free at that time. Persons who move to another state imply censoring [Putter et al., 2007]. The same investigation is carried out for departure of case 3. A regression on the cause-specific hazards concerning the Austrian unemployment data set is given in section (3.6.1).

3.4. Non-parametric analysis

In the empirical calculations we discriminate between the following cases

1. take only uncensored data, regard the exit state to be non-relevant and thus consider all exits in total
2. take solely data related to exit state: ”employment”
3. take solely data related to exit state: ”inactivity”

ad 2.

```
> datawork<-data[which(ENDEART==0),] #exit state is new working place
```

ad 3.

```
> dataoutofl<-data[which(ENDEART==2),] #exit state is out of labour force
```

In the consecutive calculations to all code-names related to exit 2 (”employment”), number 2, for those related to exit 3 (”inactivity”), number 3 will be added at the end of the object-name.

In addition to this case discrimination we apply the cumulative incidence functions in order to take account of competing risks and censoring. A short introduction to this approach was given in section 3.3.2.

The books ”Survival Analysis Using S” by [Tableman and Kim, 2004] and ”Analysis of Failure and Survival Data” by [Smith, 2002] have been primary literature resources for the subsequent analyses.

Our data presentation and analysis begins with a closer look at the uncensored spell durations. An exploratory data analysis of our dependent variable - the length of unemployment (`efftime`) - is conducted and reveals the following. The range of the observed data runs from 0.25 to 105 months for exit ’employment’. For exit ’inactivity’ the highest observed duration time is 182 months. The mean values according to the case distinction 1-3 are mean1= 8.826046, mean2= 6.581618 and mean3= 12.669962 respectively. We recognize that the mean-value of the individuals who retire or leave the working environment by some other route is almost double the value of those who have found a job. This factor substantiates the allegation that consideration of competing departures via the competing risk model is an essential extension to standard single risk models. The median for either exit cause is 4 months. Being the same for all groups, this median emphasizes its characteristic as a robust estimator. The first group (sample set of case 1) can be regarded as the combination of the latter two. The difference between group two and group three is clearly seen in the far higher variance and thus higher standard deviation of group three.

```
 variance1    variance2    variance3
254.286355    71.895207   509.791223

 std. dev1    std. dev2    std. dev3
 15.946359     8.479104    22.578557
```

To conclude this attempt of giving transparency to the data structure we additionally state the quantiles.

```
25%quantile1   25%quantile2   25%quantile3
           2              2              1
75%quantile1   75%quantile2   75%quantile3
           9              8             12
```

The boxplot in figure 3.2 provides the corresponding graphical overview.

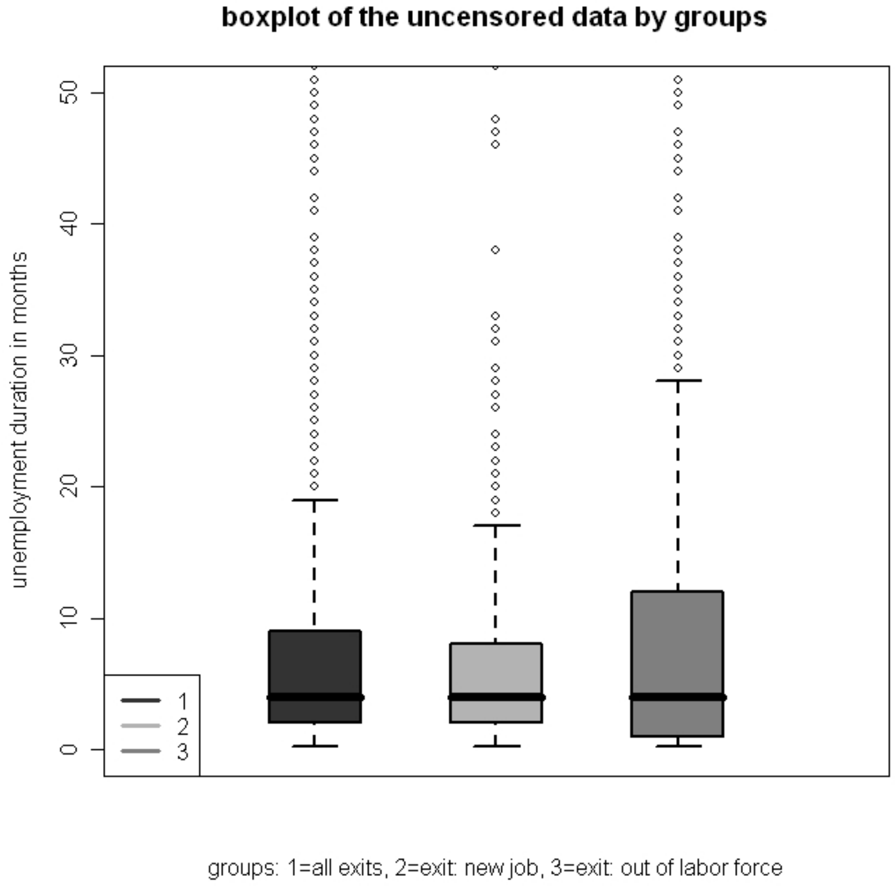


Figure 3.2.: Box-plots

To move on from the general data exploration, one among the first steps in analysing survival data is to use the given data for an a-priori estimation of the survival function, an apparent and recommended starting point. No assumptions are initially made but the data itself is the focus of interest in estimating the survival function. This is achieved by calculating the **empirical survivor function**. The simple method behind this estimation is the use of an indicator function which leads to a totalling of the number of data points larger than the observed survival point t. The formula for the empirical survivor function is therefore

$$S_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{t,\infty}(t_{(i)})$$

[Smith, 2002, p5]

A basic fact that has to be considered here is that this estimation is a method that cannot treat censoring and therefore the data has to be manipulated or treated as if censoring were absent.
The R code of the empirical survivor function (ESF) is given below.

```
> empsurvfunc<-function(n,y,z)  {
+   i<-as.integer
+   j=0
+   for (i in 1:length(z)) {
+   if (z[i]>y) {j=j+1}
+   }
+   i=i+1;
+   res=1/n*j
+   return(res)
+ }
```

where n...*length of dataset* and y...*fixed time point at which the ESF is evaluated.*
The survivor function (S) is a monotone decreasing function that satisfies $S(0) = 1$ and $S(\infty) = 0$, which is also valid for the empirical estimation.

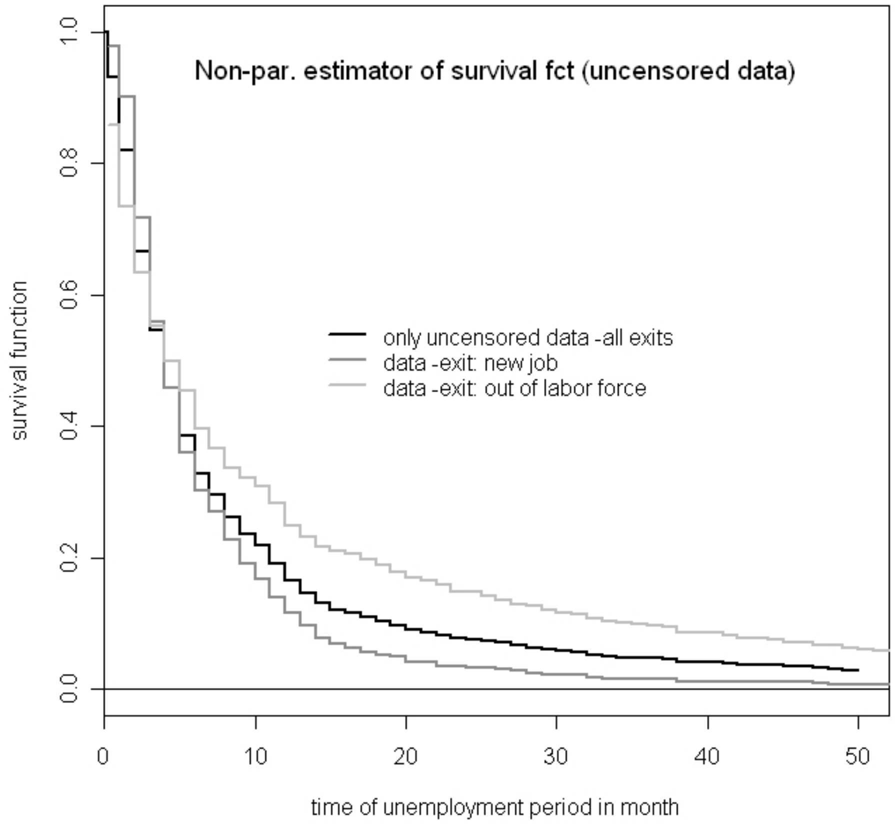


Figure 3.3.: empirical survivor function

Figure 3.3 shows a significant difference between the curve representing those who leave the labour force and the one that considers only unemployment spells of those who either return or commence work after unemployment.
A closer look at the curves and their summaries gives further information (table-appendix A.2). One column there represents the number at risk (`n.risk`), those individuals who are still in the state of unemployment just before the observed point of time. The number of events (`n.event`) indicates how many individuals have left unemployment at each observed time point. From these two sets of information ($n_i \ldots$`n.risk` and $d_i \ldots$`n.event`) the estimate of the **hazard risk** can be derived.

$$\hat{h}(t_i) = \frac{d_i}{n_i}.$$

The estimation of the hazard and survivor function is shown in the appendix (Appendix A.3). The hazard function (*described earlier on page* 4) is the probability of leaving the state of unemployment within an indefinitely small time interval after the observed point of time (`efftimeuc`), given the individual is still unemployed at that point.

In the Appendix A.3, analogous to the one described above, the table for the case of exit to employment and exit from the labour force can be found.

A better overview can be achieved by looking at the graph showing the corresponding hazard rates of our three defined case distinctions (figure 3.4). This figure also shows very well the differences among the exit states.

Another interesting aspect worthy of consideration is the **mean residual life time** (`MRLT`) - How much time of unemployment remains on average at time t_i. Specifying the actual survival time by the random variable Y_i, the discrepancy $Y_i - t_i$ in case of $Y_i > t_i$ is called the residual life time at time t_i. This is therefore defined as

$$MRLT(t_i) = E(Y_i - t_i | Y_i > t_i)$$

```
> MRLT<-function(f)  {
+ Mrl<-vector(mode = "integer", length = (length(f)-1))
+ Meanreslt<-vector(mode = "integer", length = (length(f)))
+ j<-as.integer
+ for (j in 1:(length(f)-1)) {
+      Mrl[j]<-sum(f[j:length(f)]);
+      Meanreslt[j]<-Mrl[j]/f[j];
+      }
+      j=j+1;
+      Meanreslt[length(f)]=0;  #ending point - no residual lifetime
+      return(Meanreslt)
+ }
```

Note that the first entry (at time 0) is equivalent to the mean time of failure (`MTTF`) which is the expected value $E(Y)$. Related to the residual lifetime the **mean life expectancy** (`MLEX`) at time t_i shall be outlined here as well. The `MLEX` is obtained by adding the time of passed unemployment to the `MRLT`.

```
> MLEX<-function(f)  {
+ Mle<-vector(mode = "integer", length = (length(f)-1))
```

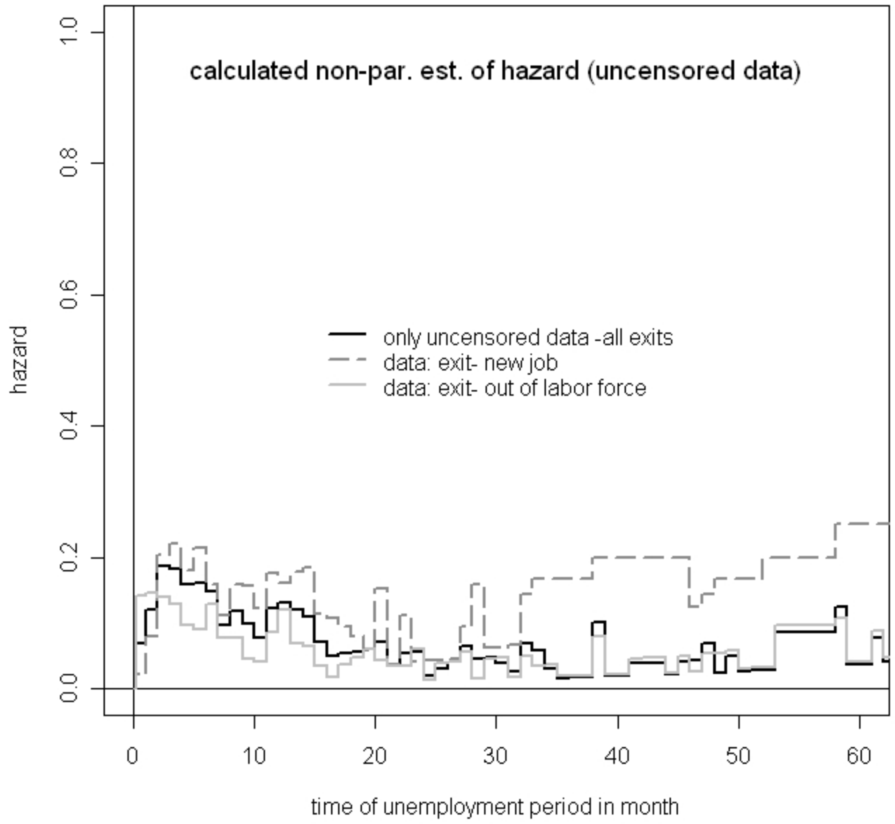


Figure 3.4.: non-param. hazard function

```
+ Meanlexp<-vector(mode = "integer", length = (length(f)))
+ j<-as.integer
+ for (j in 1:(length(f)-1)) {
+     Mle[j]<-sum(f[j:length(f)]);
+     Meanlexp[j]<-j+Mle[j]/f[j];
+     }
+     j=j+1;
+     Meanlexp[length(f)]=length(f);  #ending point
+     return(Meanlexp)
+ }
```

These calculations could be done for different groupings to detect different behaviour, e.g. men versus women (*examples can be found in* [Smith, 2002, p8 ff.]). The reason for not discussing such differences at this point is that the effect of covariates is examined in more detail in the following paragraphs (*section* 3.4.1).

Another option in data examination that takes censoring into account and therefore the next step in this data analysis is the **Kaplan-Meier-estimator** KM (*equation* (2.8)). Note,

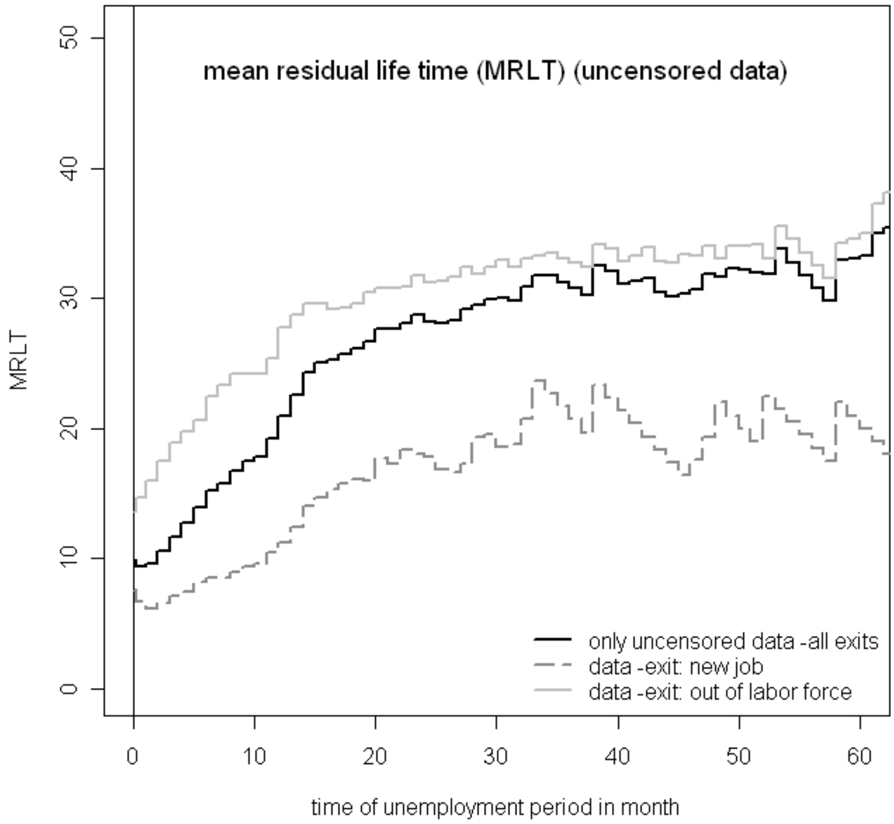


Figure 3.5.: mean residual life time

when there are no censored data values, KM reduces to the `ESF`. The KM-curve is a right continuous step function which steps down only at an uncensored observation.
A two-sided 95%-confidence interval is included in the graph below (*figure* 3.7) using broken lines. A comparison of the KM curves with the empirical survivor functions can be found in figure 3.8. This provides an indication of the crucial importance of taking censoring into account. The next step in examining the data involves visualization and filtering of covariates. Codes have been used to designate data that has not been available (e.g. $-1, -3, \ldots$). Related data records to those filter-codes have been excluded. Histograms, survival curves and tests have been created to demonstrate the dependence of unemployment duration on particular covariates.
In the following characterisation the covariates shall be examined according to their order in the tables in section 3.1.

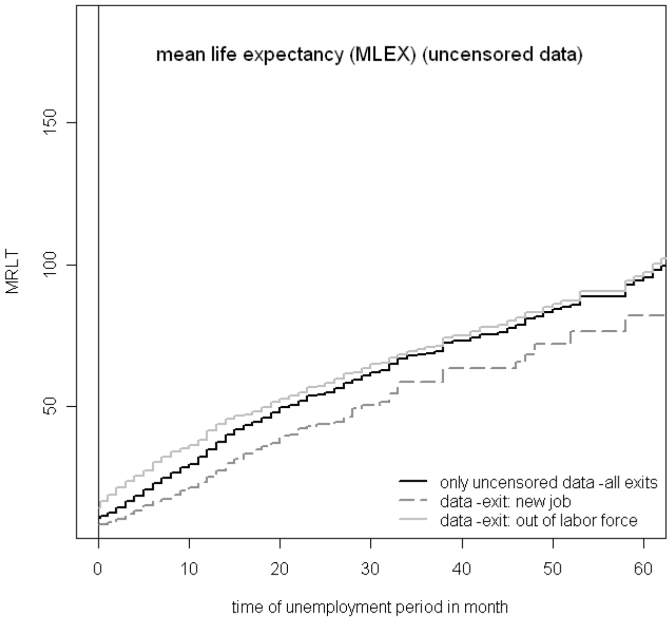


Figure 3.6.: mean life expectancy

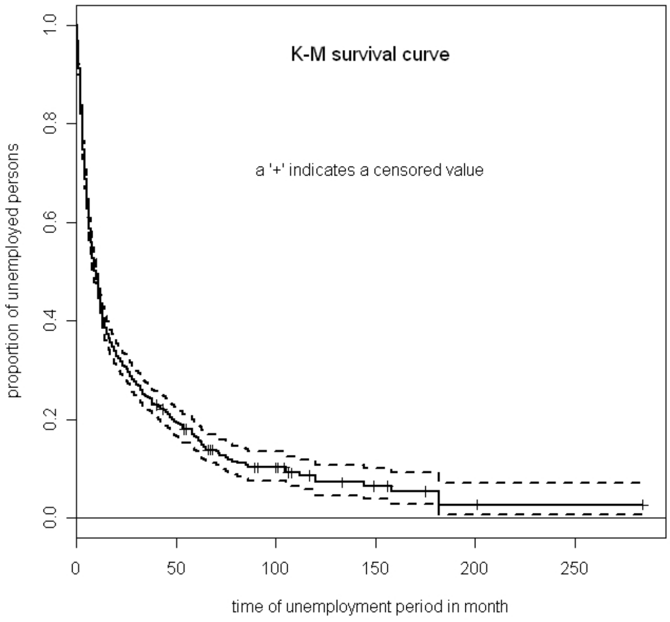


Figure 3.7.: Kaplan-Meier survival curve

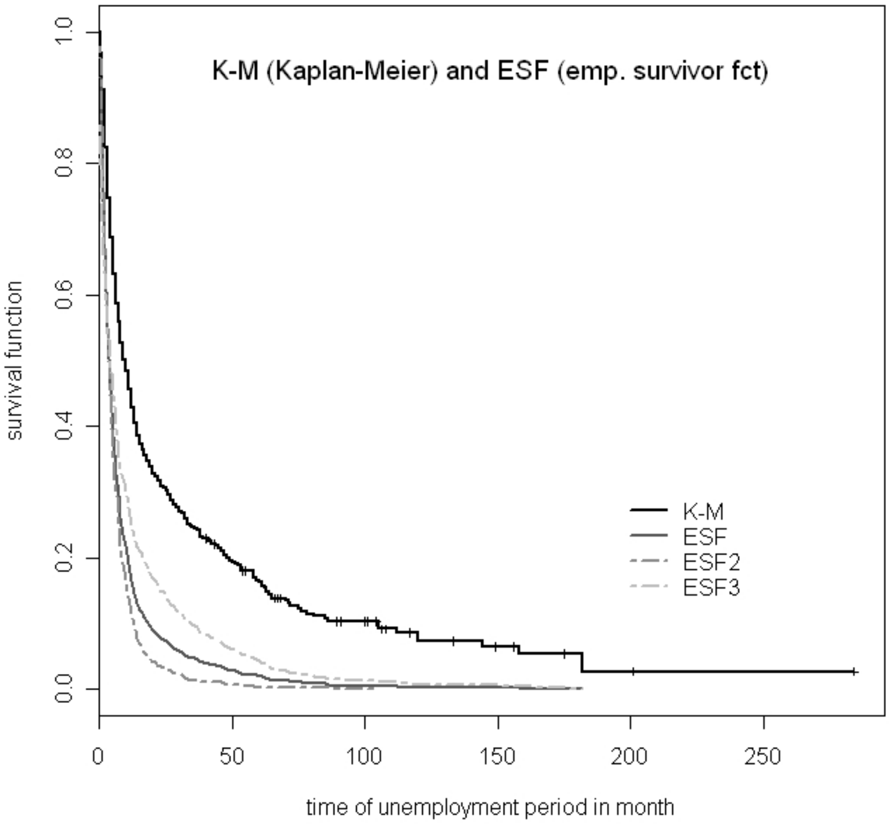


Figure 3.8.: Kaplan-Meier and empirical survival curve

3.4.1. Covariate examination

PERSONAL CHARACTERISTICS

age

To begin with BALT, the variable indicating the age of the observed individual, we subdivide this covariate into 6 age-groups. The intervals for those groups are $[15, 20)$, $[20, 25)$, $[25, 35)$, $[35, 45)$, $[45, 55)$ and $[55, 72)$. The group shares are apportioned as listed in column 'n' of the following R-Output (km.fit.age) which calculates the Kaplan-Meier curves according to their age-groups. These curves are plotted in figure 3.9.

```
> km.fit.age
Call: survfit(formula = Surv(efftime, status) ~ (BALTGROUP))
              n events median 0.95LCL 0.95UCL
BALTGROUP=1 456    239      6       5       8
BALTGROUP=2 462    225      5       4       6
BALTGROUP=3 644    274      9       7      11
BALTGROUP=4 654    290     12       9      14
BALTGROUP=5 449    209     14      11      22
BALTGROUP=6 159     78     23      20      42
```

Note that there is a natural bias occurring due to the fact that those individuals in the lower age group have an upper limit of their potential unemployment period. Their starting point of an employment period cannot commence before a certain age.
For a more balanced view of the data, a closer look at the short-time unemployment (< 2 years) shall be made and the lowest age group consisting of 15 and 16 year old boys and girls (new "group 1") shall be omitted. Furthermore, a distinction between the exit states is also provided.

```
> km.fit.age1
Call: survfit(formula = Surv(efftime[which(efftime < 2 years)], status[which(efftime <
    24)]) ~ (BALTGROUP1[which(efftime < 2 years)]))
                                          n events median 0.95LCL 0.95UCL
BALTGROUP1[which(efftime < 2 years)]=1 159     91      6       5       9
BALTGROUP1[which(efftime < 2 years)]=2 281    143      5       5       7
BALTGROUP1[which(efftime < 2 years)]=3 442    218      4       3       5
BALTGROUP1[which(efftime < 2 years)]=4 606    262      7       6       9
BALTGROUP1[which(efftime < 2 years)]=5 595    269      9       8      11
BALTGROUP1[which(efftime < 2 years)]=6 384    173      9       8      11
BALTGROUP1[which(efftime < 2 years)]=7 116     57     13      11      16
```

The results still show a greater likelihood of leaving unemployment for young people, especially those of age between 20 and 25. Those of age under 20 perform slightly worse which leads to the presumption of an existing correlation between education and age. Another clear statement of the results is that the older working generation is far more limited in their chances of finding a job.
Next, the same partitioning as before is used for examination of the two different exits. Note, given the information about the exit, the censoring aspect is consequently ignored, as the censored data has an unknown exit status. The corresponding results, showing the median along with 95%-confidence levels, are given in table 3.1. Once more, greater flexibility of young people can be seen as is indicated not only by this group finding a job faster than their older

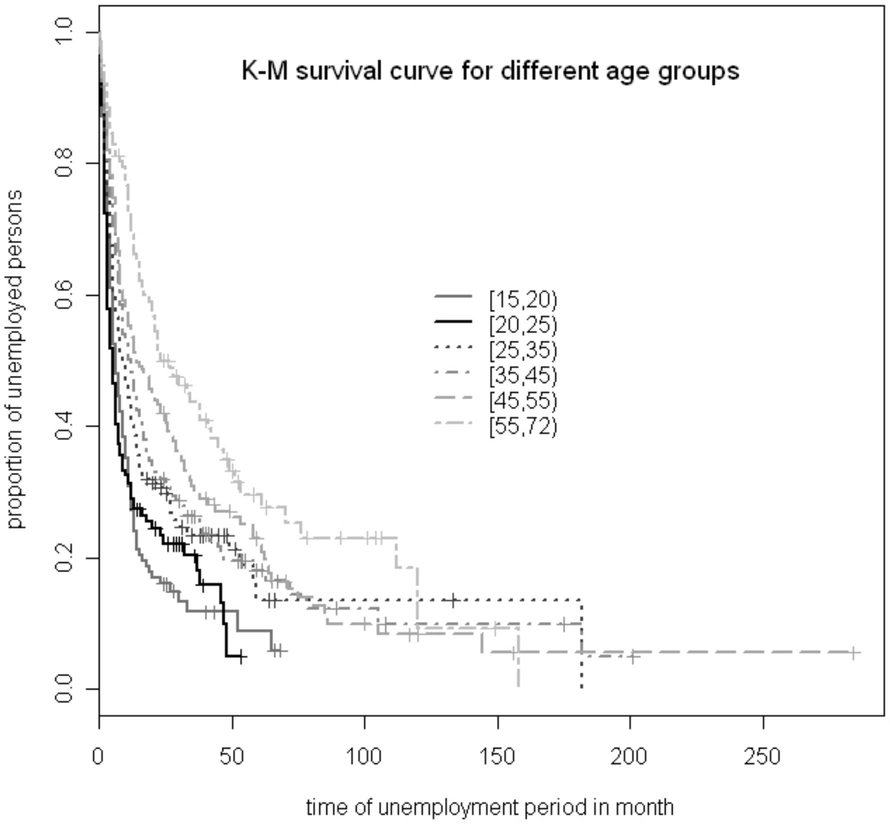


Figure 3.9.: Kaplan-Meier - age

counterparts but also by performing better in the case of faster exit to 'out of labour force'. This exit may be, amongst others, a result of further educational training or marriage and maternity, while for the oldest age-group it may predominantly result from early retirement.

However, better insights concerning the different exits might be given in our competing risks approach. After the first case consideration, that regarded all exits as one event of termination in total and the second, in which we have split the data into the two considered exit cases 2 and 3, we now calculate the cumulative incidence function (CIF) in respect of competing risks for each covariate. Therefore, we use the implemented function `cuminc` of the R software. This function estimates the CIF from competing risks data and tests for equality of the subdistributions across the groups of a chosen covariate. Recall the short introduction of non-parametric competing risks in section 3.3.2. The CIF for the different age-groups has been examined and yields the following results. Testing for equality across the age-groups according to Gray [Gray, 1988] reveals significance for both exit causes. People of age 20 to 25 have the highest probability of leaving unemployment into employment but one of the lowest probabilities of leaving into inactivity. Second, and with similar results, are those being in the age-group below 20. People from the two middle-age groups perform slightly worse than

exit	'EMPLOYMENT'			
AGEGROUP	observ.	median	95%-LCL	95%-UCL
1	44	4	3	6
2	82	4	3	4
3	132	3	3	3
4	141	4	3	5
5	153	5	4	6
6	86	5	4	6
7	19	11	4	16
exit	'INACTIVITY'			
AGEGROUP	observ.	median	95%-LCL	95%-UCL
1	40	2.0	2	6
2	51	3.0	1	5
3	65	2.0	2	3
4	96	3.0	2	4
5	86	4.0	3	6
6	76	5.0	3	6
7	38	7.5	4	13

Table 3.1.: age

the younger ones but still have a higher probability to find work than to leave into inactivity while for those of the older age groups it turns out to be vice versa. Far from the others, showing the lowest probability of finding work, are the ones in the age group of 55 or more years. However, this age-group shows the highest probability curve regarding exit into economic inactivity, most likely through retirement.
Observing the age-groups in case of the competing risk 'leaving the labour force', we find the ordering of the age-groups reversed compared to the event of interest, case 2 ('employment'). Hence, unemployed elder people are expected to stay longer in the state of unemployment before getting retired or leave the labour force in any other way than younger unemployed.

Mantel-Haenszel or log-rank test

Comparison of two possible influencing factors (covariates) on survival curves can be made by performing a Mantel-Haenszel or log-rank test ([Tableman and Kim, 2004, p41 ff.]). The procedure of the test is shortly described below. Allowing for differences among the observed objects (interviewed persons) the Null-hypothesis for the test is:

$$H_0 = p_{11} = p_{12}, \ldots, p_{k1} = p_{k2} \qquad \text{for } k \ldots \textit{number of observations}$$

where
$p_{i1} = P(\textit{end of unempl.-period} | \textit{influence 1, observation } i)$
$p_{i2} = P(\textit{end of unempl.-period} | \textit{influence 2, observation } i)$
$\vdots$

The Null-hypothesis is therefore that the distribution of 'time to event' or in other words

'the survival rate' is equal for all groups.
The Mantel-Haenszel statistic (1959) is ([Tableman and Kim, 2004, p42])

$$MH = \frac{\sum_{i=1}^{k}(a_i - E_0(A_i))}{\sqrt{\sum_{i=1}^{k} Var_0(A_i)}}$$

where $a_i \ldots$ *observed event of interviewed person i.* And for each observation

$$E_0(A) = \frac{d_1 * r_1}{n}$$
and
$$Var_0(A) = \frac{d_1(n-d_1)}{n-1} * \frac{r_1}{n}(1 - \frac{r_1}{n})$$

where
$d_1 \ldots$ *number of unempl. periods that ended at the respective observed time*
$r_1 \ldots$ *number at risk of those of group 1*
$n \ldots$ *number of all observations.*

In case of independence, the *MH test statistic* is approximately standard normal distributed $MH \overset{a}{\sim} N(0,1)$.
The function `survdiff` used in the calculations (R-code), which executes the log-rank test provides us the square of the *MH test statistic.*
Note: $MH^2 \overset{a}{\sim} \chi^2_{df}$
where *df ... degrees of freedom* are the number of groups minus 1.

The χ^2-statistic corresponds to a two-tailed test and therefore the p-value in the calculation output has to be considered as being twice that of the *MH-statistic* stated above. A rule that has become generally accepted is that a p-value (one-sided) less than 0.05 indicates that the Null-hypothesis should be rejected, with stronger evidence against the Null-hypothesis the nearer the p-value comes to zero.
Applying this test to the covariate `BALT` (age) showed evidence for a difference among the age-groups, as the p-value was very small (p= 1.64e-14).
In the subsequent paragraphs, an analogous examination by the log-rank test is provided for other covariates. The next, in turn of the listing in section 3.1, is the gender of an unemployed person.

sex

We observed a balanced sample regarding gender with only slightly more women in the total number of observations (1525) compared to men (1299).

We test if there is a difference between the survival curves of men and women using the Mantel-Haenszel (log-rank) test. The corresponding R-Output with the test results is found below.

```
survdiff(formula = Surv(efftime, status) ~ BSEX, rho = 0)
            N Observed Expected (O-E)^2/E (O-E)^2/V
BSEX=0 1299      612      619    0.0822     0.166
BSEX=1 1525      703      696    0.0731     0.166
 Chisq= 0.2  on 1 degrees of freedom, p= 0.683
```

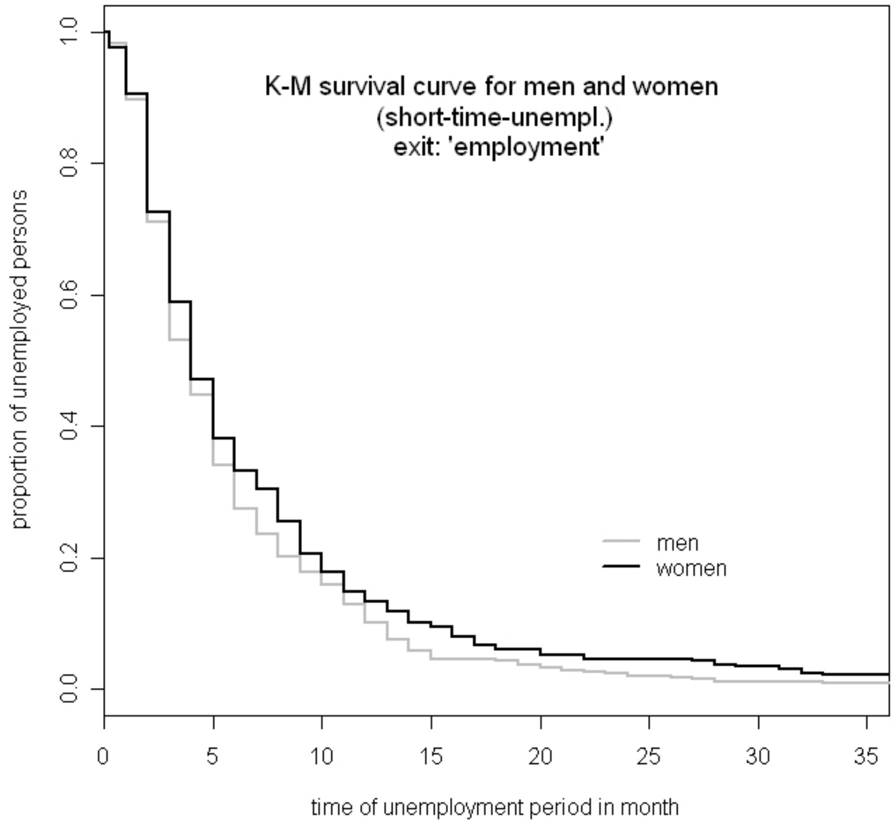


Figure 3.10.: Kaplan-Meier - sex

The resulting p-value suggests that the Null-hypothesis of equality among the survival curves is not rejected. Therefore there is no strong difference in survival rates regarding the sex-criterion. Considering solely short-time unemployment gives similar results. Including the distinction between the competing exit states the one-sided p-values are 0.054 and 0.207 respectively. There is evidence for sex discrimination regarding the time the person spent being jobless before entering into the labour market again. As the p-value of 0.054 is close to our defined rejection level of 0.05 (*subsection* 3.4.1), the possibility of sex discrimination in case of exit to employment cannot be completely excluded. The corresponding Kaplan-Meier curves are plotted in figure 3.10. We recognise a marginally worse performance for women but should recall that the results do not give any information about gender discrimination in jobs or earnings and stress the point that only the duration of unemployment is examined for differences. Although there has not been found strong evidence for a different duration distribution between male and female job seekers, the covariate BSEX can also play a significant role e.g. due to interaction with other covariates such as the one appearing next in our listing displayed in section 3.1, the marital status (BFST).

Considering the cumulative incidence functions for the gender specification according to exit

behaviour, the result of Gray's test [Gray, 1988] for equality of the functions across men and women shows significant p-values for both, exit to employment (p-value = 0.0025812) and exit to inactivity (p-value= 0.0001788).
Other than in the previous approaches a different performance between genders becomes obvious and is shown in the plotting of the cumulative incidence curves in figure 3.11. We see for instance that the 3-year cumulative incidence of men was 45% concerning exit to employment, but was 26% for exit into an economic inactive life while for women the results were 39% and 36% respectively. All in all women show a higher probability of transition to inactivity and a lower probability in finding a job than their male counterparts.

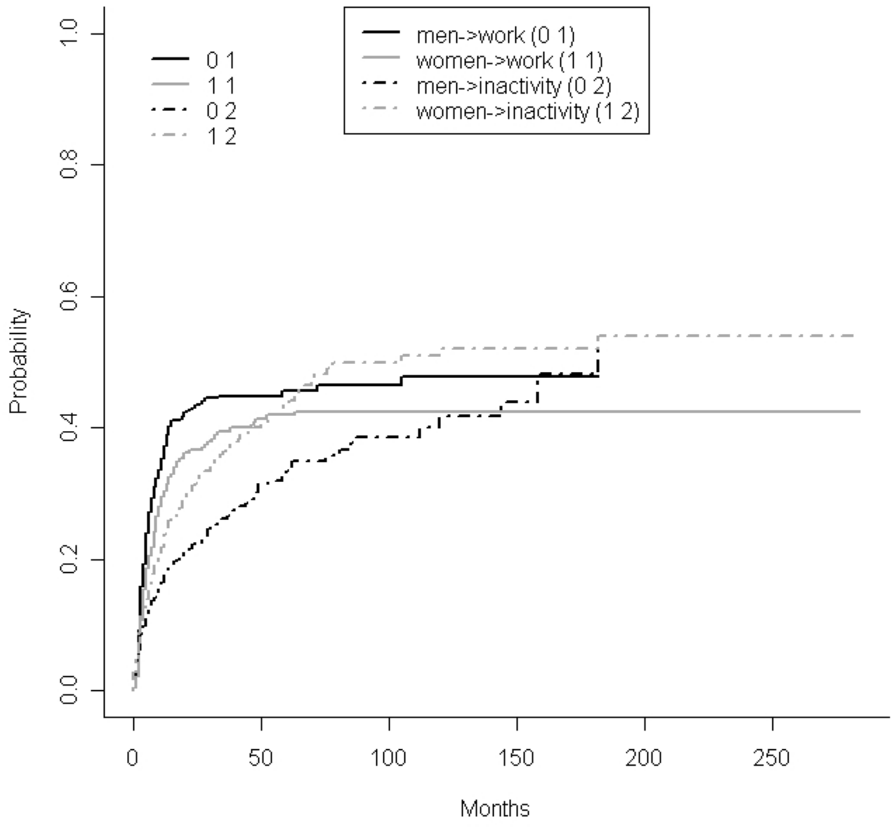


Figure 3.11.: CIF - sex

marital status

What is considered next in our examination is the marital status of a person. A listing of the corresponding group counts is provided below.

marital status:	single	married	widowed	divorced
observations:	1396	1108	32	288

As there are very few widowed persons, this category is added to the group of divorced people. The log-rank test produced a p-value near zero (p= $2.47e-07$) and therefore possible differences among singles, married or divorced persons for case 1 examination (*see* figure 3.12). The test proves to be significant in case 2 (p= 0.077 (two-sided)) and case 3 (p= $2.05e-07$). Exiting unemployment by entering into a new job seems to be slightly easier for singles. Singles are also the category showing the highest probability of leaving unemployment into inactivity while those divorced or widowed show the longest expected survival in the state of unemployment before exit from the labour market. In this context it has to be taken into consideration that mostly young people are singles which is made clear by the strong correlation coefficient between age and the status of being single. As shown before, young people seem to be favoured in being able to end their period of unemployment. This is partly reflected in

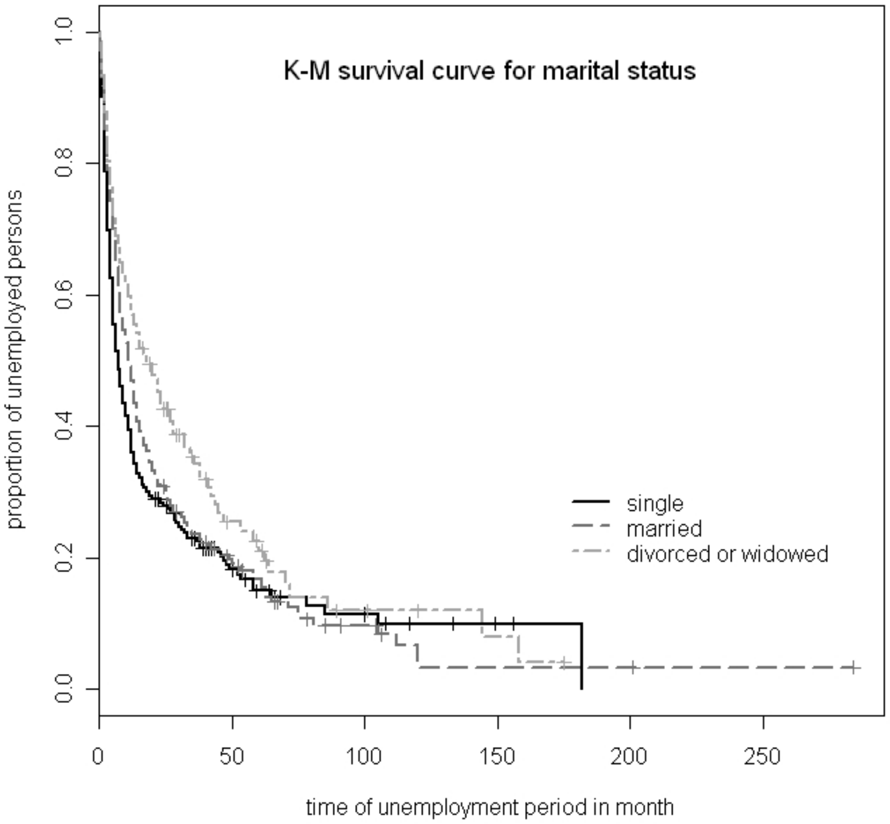


Figure 3.12.: Kaplan-Meier - marital status

the survivor curve for singles (LEDIG) due to the high negative correlation of -0.6584788. An interesting and possibly unexpected development is that divorced persons are seen to have a higher median of unemployment durations than married ones. Adding the gender aspect does not provide further information for significant changes in the differentiation of unemployment length.

The cumulative incidence functions plotted in figure 3.13 confirm the better chances for singles to leave unemployment and start to work but show a reduced likelihood of them quitting the labour market early (figure 3.13).

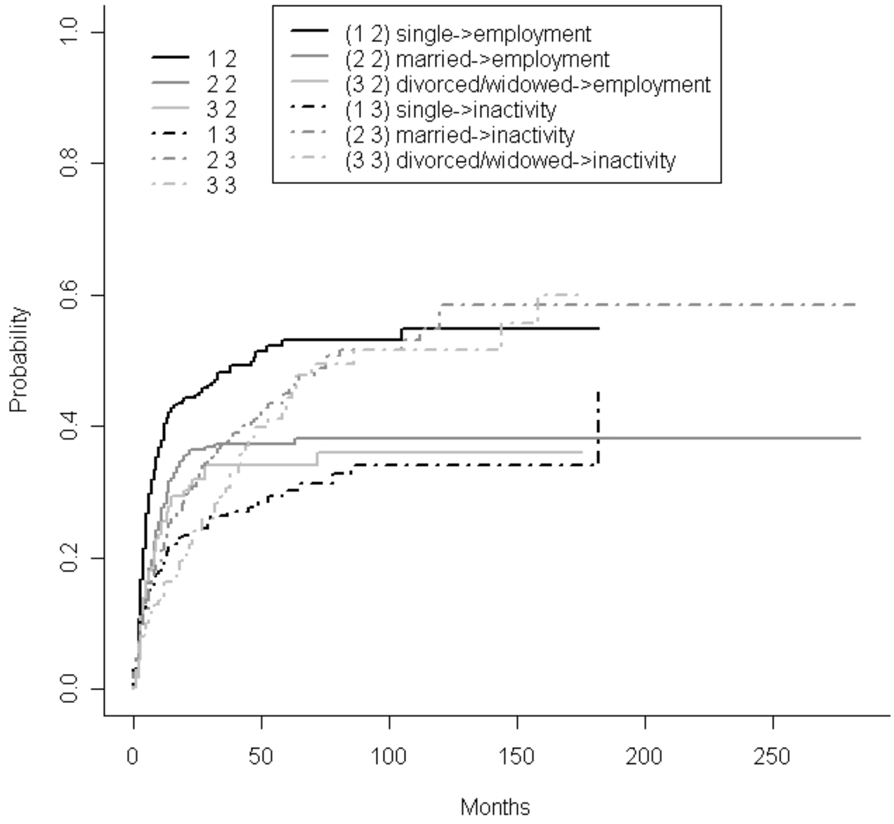


Figure 3.13.: CIF - marital status

For distinction of effects due to nationality we distinguish between those who are born in Austria (XBGEBLAO) or have got the Austrian citizenship (XBSTAATO) and those who don't. Further, we separate into different groups regarding the country of citizenship by the variable XBSTAAT described in table 3.2.

The log-rank test for the first covariate (XBSTAATO) furnishes no proof for differences among Austrian and Foreign job seekers in either case.

Similarly, there are no further indicators for differences when splitting Foreigners into the groups listed in table 3.2.

Unlike the two previous covariates, the test result for foreign-born persons gives evidence for

label		name	observations
`OE`	...	Austria	2345
`EU15`	...	EU members (membership started before 2004)	57
`EU25`	...	new EU members (membership in or past 2004)	42
`YU`	...	Former Yugoslavia	185
`TURK`	...	Turkey	124
`Others`	...	Other citizenship	71

Table 3.2.: citizenship

inequalities in unemployment durations (p= 0.00242). However, restricting the data to the exit 'employment' and 'inactivity' categories respectively, no longer provides the same results and therefore does not verify the differences observed in case 1 (p2= 0.479, p3= 0.86).

We must now judge if consideration of the probability of experiencing different events by a given time, including the aspect of censoring, confirms these results. Considering the competing risks by applying the cumulative incidence function for the question of Foreign or Austrian origin (`XBSTAATO`) provides the following information.
For exit cause 3, ending up in no further attempt of job search, we get a p-value of 0.049617. This gives mild evidence for difference among Austrians and Foreigners while for exit cause 2 the p-value was not low enough to reject the Null-hypothesis of equality between Austrian and Foreign citizens.
Thus, the binary covariate solely indicates an increased probability for Foreigners to leave unemployment into inactivity than for Austrians.
Further distinction into 6 categories of citizenship classification does not give any reason for assuming inherent differences among the groups indicated by high p-values in both cases.
The effect of the covariate `XBGEBLAO`, retaining information whether Austria is the individuals' country of birth or not is different. Testing for this covariate provides a highly significant p-value for exit to employment as well as remaining significance for exit of type 3. Plotted in figure 3.14 we find the people not born in Austria to be discriminated by a lower chance of finding work but a higher exit rate to inactivity.

children

Another personal characteristic considered in the covariate examination is the number of children. This variable has been divided into 5 groups:

marital status:	no child	1 child	2 children	3 children	4 or more
observations:	372	740	786	359	139

Figure 3.15 shows that those with 2 or 3 children do best ending the period of unemployment while the parent of 1 or more than 3 children has slightly worse chances to exiting unemployment. Those who do not have children perform worst. For specification reason the different exit states are taken into consideration when performing the log-rank test. In this tests the exits of finding a job or returning to work do not provide evidence for differences due to the number of children a person has. The obtained differentiation in the duration distribution is therefore based on the exit 'out of labour force' and other covariates, possibly interacting with

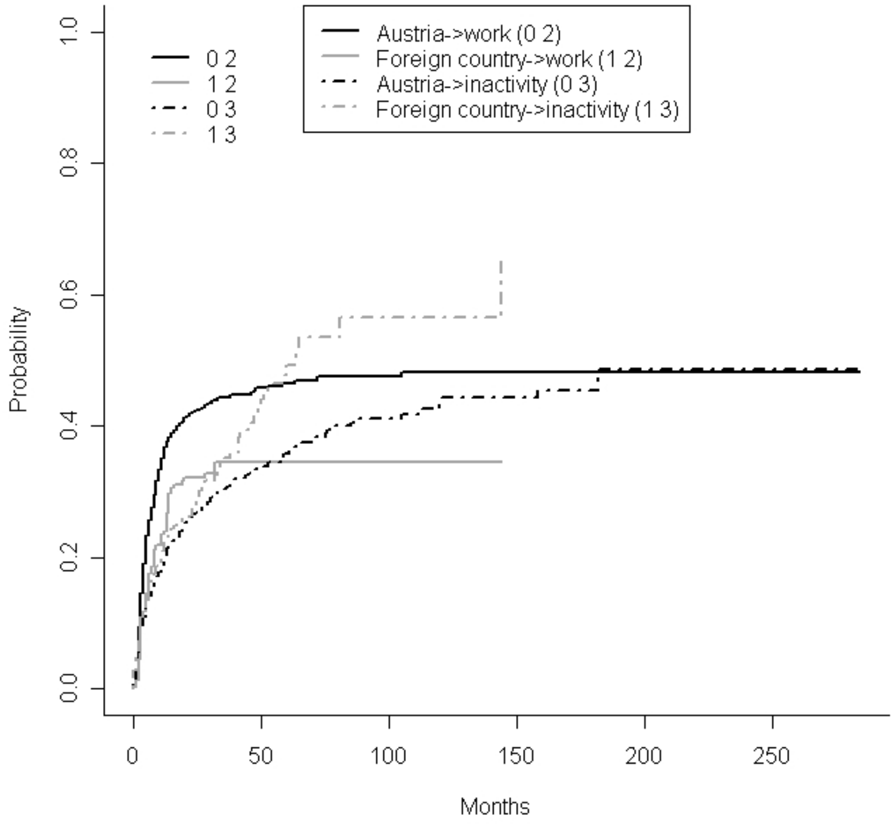


Figure 3.14.: CIF - country of birth

XANZKIND, like sex (BSEX) or marriage (VERH) or age (BALT) should be considered. The only eye catching result emanating from observation for this kind of correlation was that a female parent having 4 or more children is on average longer unemployed than the male counterparts of the sample.

Given this preliminary result, we will observe the cumulative incidence function next.

In the sample test for comparing the cumulative incidence for competing risks the number of children seems to play a significant role only for exit into employment, demonstrated by a p-value of 0.005071. This is different to the suggested significance of the competing risk observed above. Best chances of exiting unemployment are again evident for those with 2 or 3 children. This seems to be the socially most stable family constellation. The lowest probability to exit unemployment is given for those who do not have children.

EDUCATION

Another covariate considered to be of relevance for our aim of finding duration influence factors, hence often used in unemployment duration analyses, is the one comprising the levels of

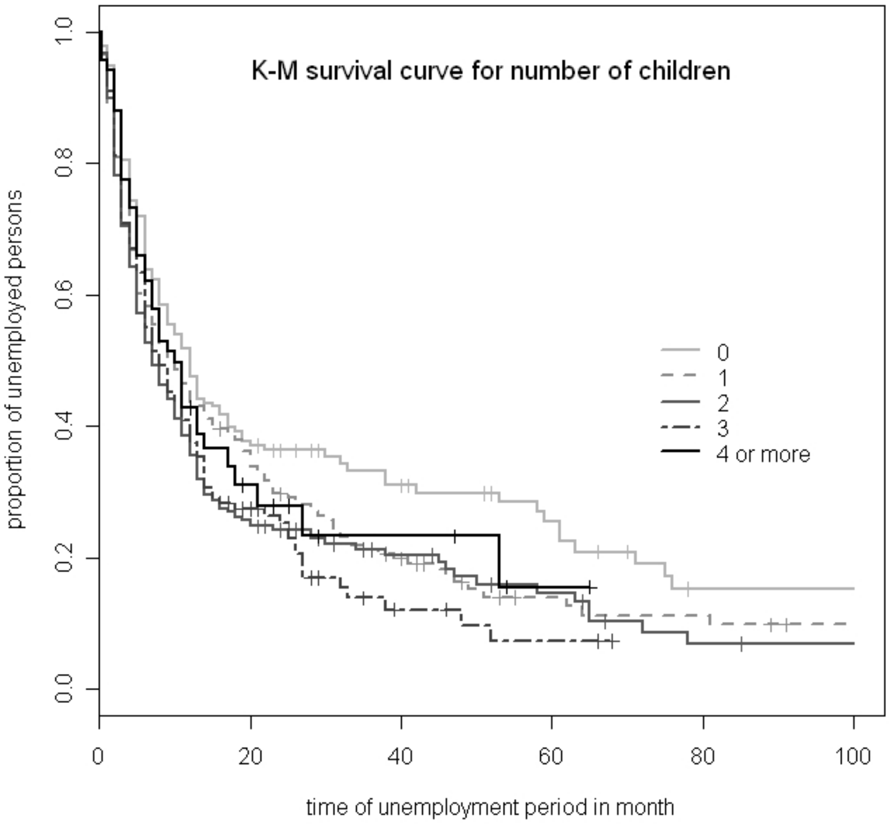


Figure 3.15.: Kaplan-Meier - number of children

education. In this micro-census survey it is labelled as `XKARTAB`. The partitioning has been made as listed in table 3.3.

For better optical interpretation the graph (*in figure* 3.16) is constrained in a way to show short-time unemployment only. The test for differences shows mild evidence (p= 0.0696 (two-sided) - *see subsection 3.4.1*) of differences among the levels of education. Additionally, restriction of the data to exit 'employment', having solely uncensored data where it is known that the individuals found a job after the unemployment period, lays emphasis on this thesis of inherent differences. The achieved p-value in the log-rank test is much lower (p= 0.00712). The corresponding survival curves are visualized in figure 3.17. Still, those with the lowest education level (`PFLSCH`), and especially those having attended high school, perform worst in finding a job. This is not surprising as this school type is primarily intended for continuing education at universities.
The examination of case 3, the transition to exit from the labour force, reveals that a longer duration is most probable for the lowest education level. Further, in this case, those who attended a general high school or a vocational high school seem to be quickest in exiting into inactivity from the state of unemployment.

PFLSCH...compulsory education or none
LEHRAB...apprenticeship
BMS...secondary vocational school
AHS...high school
BHS...vocational high school
UNI...university and colleges

Table 3.3.: educational levels

Considering marginal failure probabilities via the cumulative incidence function reveals slightly different results. In case 2, transition to employment, the covariate test appears significant with a p-value of 0.001622679 and we get the following results. The highest probability to leave unemployment is observed for persons who attended a University or experienced other high education. Next in the probability ranking, with similar probability appear people whose educational training has ended with an apprenticeship or general high school (AHS). People who attended the vocational school or vocational high school appear next in the probability ranking. As expected, people with the lowest or no education at all have the worst probability to leave the state of unemployment.
For case 3, there is only mild evidence of group differences indicated by a p-value of 0.047690622. The ranking in this transition to economic inactivity is reversed compared to the ranking regarding to exit case 2.

JOB CHARACTERISTICS

What comes next in the variable-listing in section 3.1 are the covariates that concern the job properties of the observed persons. A list of p-values resulting from the respective log-rank tests are given in table 3.4. Note that the number added to the labels indicate the observed cases described on page 41. Due to the fact that some covariates are derived from questions concerning the job found after the unemployment period, those data are available solely for case 2 examination (DBERS, DTAET, DSTD, XDWZAB). A detailed description is provided in section 3.1.

former job position (JBERS)

Lower unemployment duration is shown for those who have not had a job before. This may result from the correlation to younger people who, as discovered earlier, are favoured in the labour market. Another well-known fact is that changing the working position or sector one has started in is not easy and therefore those who have not worked before are usually more flexible in their decisions. The restriction to exit 'employment' shows no more evidence for differences. This leads to the interpretation that the lower unemployment duration of those who had no job before could be due to short periods in between educational phases. This hypothesis might be confirmed by the fact that the job situation seems to play a role for exit of cause 3, the leaving into inactivity. Without having ever been in employment an

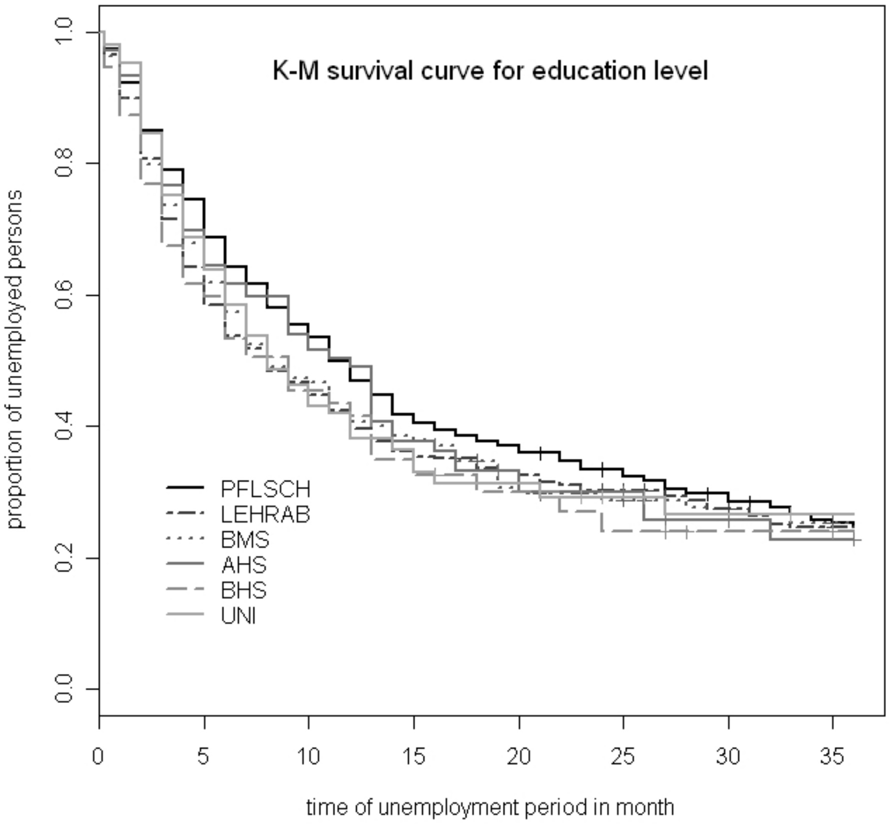


Figure 3.16.: Kaplan-Meier - education

employed person is more likely to withdraw from the labour market than otherwise.

Considering the competing risks comparison according to Gray's test [Gray, 1988], no evidence for rejecting the equality assumption is given on basis of a 5%-level test, showing p-values of p2= 0.45613718 and p3=0.08822283 for the respective causes 2 and 3.

type of former profession (JTAET)

The only profound statement that can be made from examining this covariate is that those who are 'high professional manual' have a significantly lower median of unemployment duration regarding total exit consideration as well as the distinct causes 2 and 3. All other categories show considerably large variation.

Considering transition to employment, the cumulative incidence test provides a p-value of 0.0002610973. Conversely, consideration of case 'exit from labour force' does not result in rejection of the Null-hypothesis of equality among the groups. This is based on a p-value of

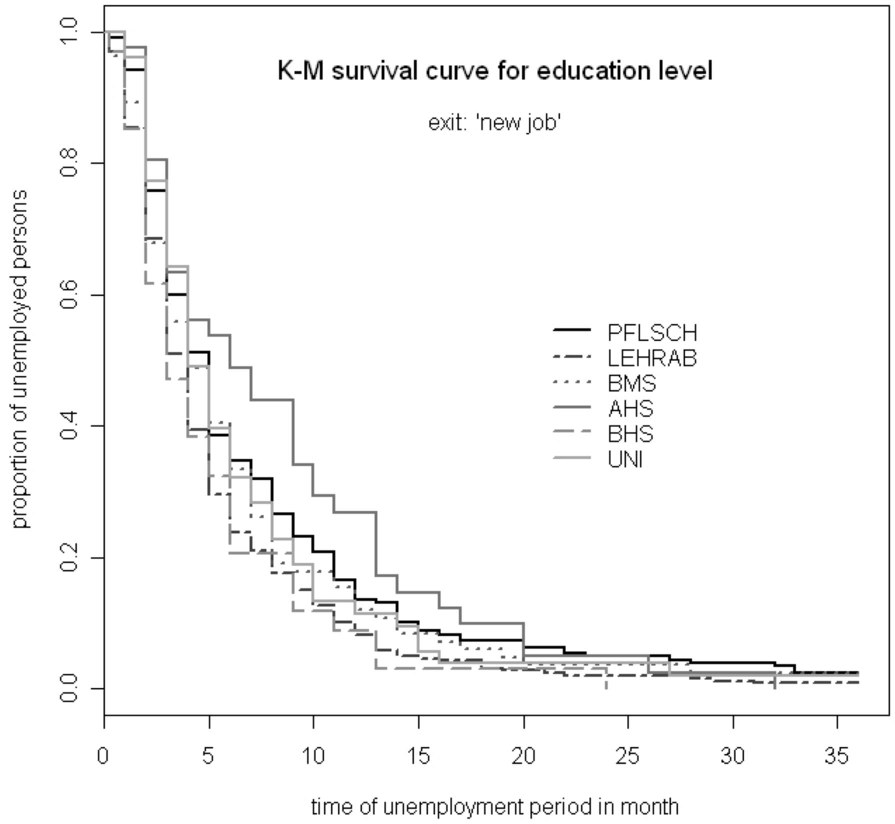


Figure 3.17.: Kaplan-Meier - education (exit 'employment')

0.0561110305 being outside of our level of acceptance of 5%. Therefore we observe the covariate solely for our event of interest, case 2. After two years time, the cumulative incidence is highest for 'high professionals doing manual work' with a probability of 0.4704240 to leave unemployment. 'Manual employees' show an estimated probability of 0.4526700 and those having been in apprenticeship show a value of 0.4154889 after two years. Falling last in the ranking made after 2 years of unemployment are those who had done auxiliary work before, showing a corresponding probability to leave unemployment of 0.3097932.

For the next four covariates no competing risk, i.e. exit into inactivity (case 3), is given as it concerns the job situation after the unemployment period, hence the covariate is investigated only for departure to (re)employment. Therefore we cannot calculate cumulative incidence curves but take the opportunity to look at log-rank tests for our cause of interest, the return to employment.

covariate	label	two-sided p-value of log-rank test
former job position (case 1)	JBERS1	p= 0.13
former job position (case 2)	JBERS2	p= 0.94
former job position (case 3)	JBERS3	p= 0.00736
type of former profession (case 1)	JTAET1	p= 2.99e-06
type of former profession (case 2)	JTAET2	p= 0.00382
type of former profession (case 3)	JTAET3	p= 0.00188
current job position (case 2)	DBERS2	p= 0.204
current type of profession (case 2)	DTAET2	p= 0.000226
weekly working hours (case 2)	DSTD2	p= 0.0384
branch of industry (case 2)	XDWZAB2	p= 0.249
reason for quitting last job (case 1)	JLWI1	p= 4.72e-09
reason for quitting last job (case 2)	JLWI2	p= 0.023
reason for quitting last job (case 3)	JLWI3	p= 0.000338

Table 3.4.: covariates - job

current job position (DBERS)

This covariate is automatically restricted to exit 'employment' as it is the job position after the unemployment period that is asked for. As the p-value of the log-rank test is not low enough, there is no evidence for different distributions.

current type of profession (DTAET)

As expected, this covariate provides similar results to JTAET, the type of profession the individual had before the occuring work interruption. Note that the group 'no profession' is no longer contained in the classification and again the profession of 'high professional manual' shows increased chances of leaving unemployment. Additionally, 'Manually working employees' and 'Non manual medium professionals' have slightly better chances of leaving unemployment than people with other professions.

weekly working hours (DSTD)

As there have been some unreasonably high entries for the amount of working hours per week in our sample set, the data was reordered into the following 5 intervals: $[1, 15)$, $[15, 25)$, $[25, 35)$, $[35, 45)$ and $[45, 85)$, hence, excluding entries above 85. The histogram, displayed in figure 3.18 gives an overview about the binning.

Figure 3.19 shows the survival curves according to the working hours per week observed in the reference week and has been terminated after 3 years of unemployment for better visualization. The log-rank test for this covariate provides the following results. Overall, there is mild evidence that the chances of leaving unemployment are better for the last two groups that contain those professionals with a larger number of working hours per week.

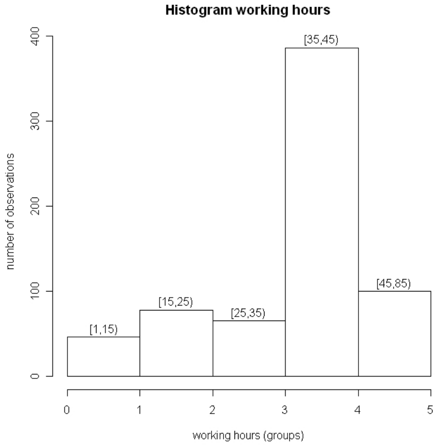


Figure 3.18.: histogram - working hours

Those with less than 15 working hours per week perform a little better on average than those who work between 15 and 35 hours (*defined as 'part-time'*). A reason for these better odds of (re)employment compared to part-time jobs might be a greater number of jobs offered that claim less working hours. It further became apparent that there is no correlation between the amount of working hours and the age of a person due to a correlation coefficient of 0.006314875.
For the examined covariates `DBERS, DTAET, XKARTAB, BFST` and `BSEX` no significant positive or negative correlation to weekly working hours was detected either.

branch of industry (`XDWZAB`)

With a p-value of p= 0.249 this covariate does not give evidence of any duration dependence.

reason for quitting last job (`JLWI`)

For all individuals who became or more precisely have been unemployed in the year of our observation (2005) but had a job before, we investigate possible influences concerning their reason for quitting the previous job (`JLWI`). As expected, the reason for the last job-exit seems to play a significant role for the succeeding unemployment duration (*see* figure 3.20). Considering all exits in total as departure event from unemployment, those persons who have already retired but start searching for a job thereafter have the longest expected duration, a median of 15 months, followed by dismissal and illness or disability. The possibility that a retired person can be regarded as unemployed is justified by our particular definition of unemployment (*this definition is given in section* 3.2). It is emphasized once more that in this visualization (figure 3.20) of survival curves no exit distinction is made. The shortest expected unemployment duration is given for those individuals who left their job for civil or military service and it is to be noted that these are mostly young male people. Also good chances of early rehabilitation into working life or for definitely quitting from the labour force,

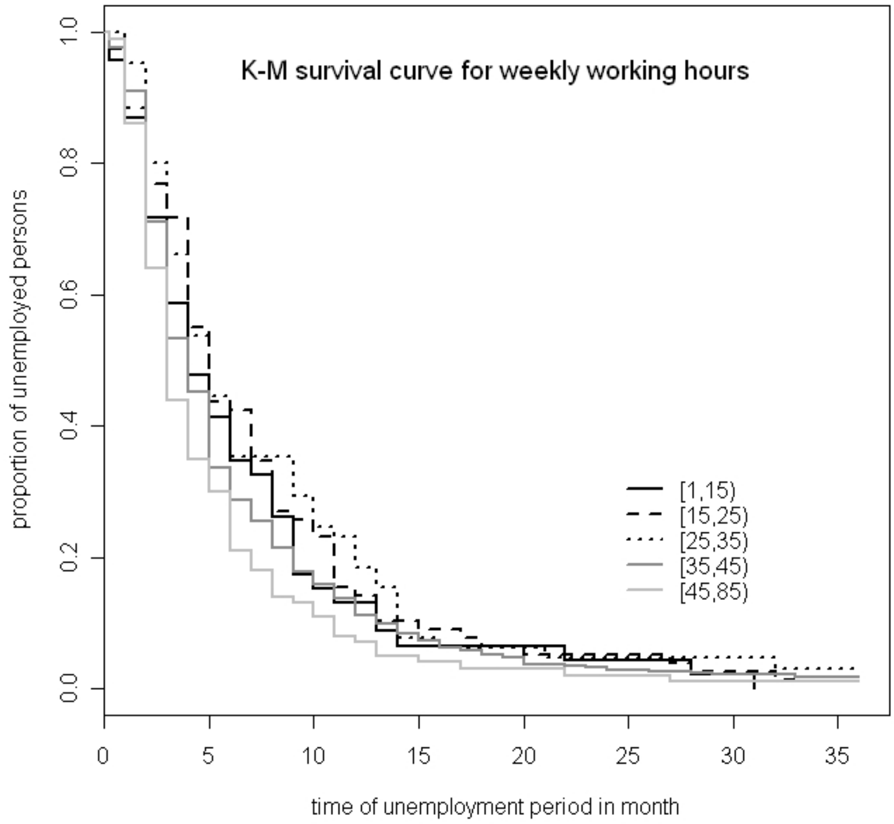


Figure 3.19.: Kaplan-Meier - working hours

which is not separated in this first investigation, face those who experienced unemployment due to the fact that their working period was regulated by a fixed-term contract or the case that they resigned themselves (including mutually agreed termination of a work contract).

As the p-values for the log-rank test for both exit distinctions indicated significance, the resulting survival curves for the distinct reasons were examined in both cases. Let us first consider the transition to employment. Here one reason for ending the prior employment contract that predominantly seems to elongate a successive unemployment period is illness or disability with an expected duration time (median) of 7 months. All other duration expectations lie between 2.5 and 5 months. Surprisingly, with a median value of 4 months, dismissal from the last job does not show strong evidence for adversely affecting the chances of finding a new job compared to other reasons. To mention others, 'end of contract' shows a median value of 3 months and 'resignation or mutual agreement on terminating the work contract' and 'caring for others' show a median duration of 4 months.

However, regarding the case of transition to inactivity after the state of unemployment, dismissal indeed plays a crucial role, adversely affecting the duration time. This is indicated by

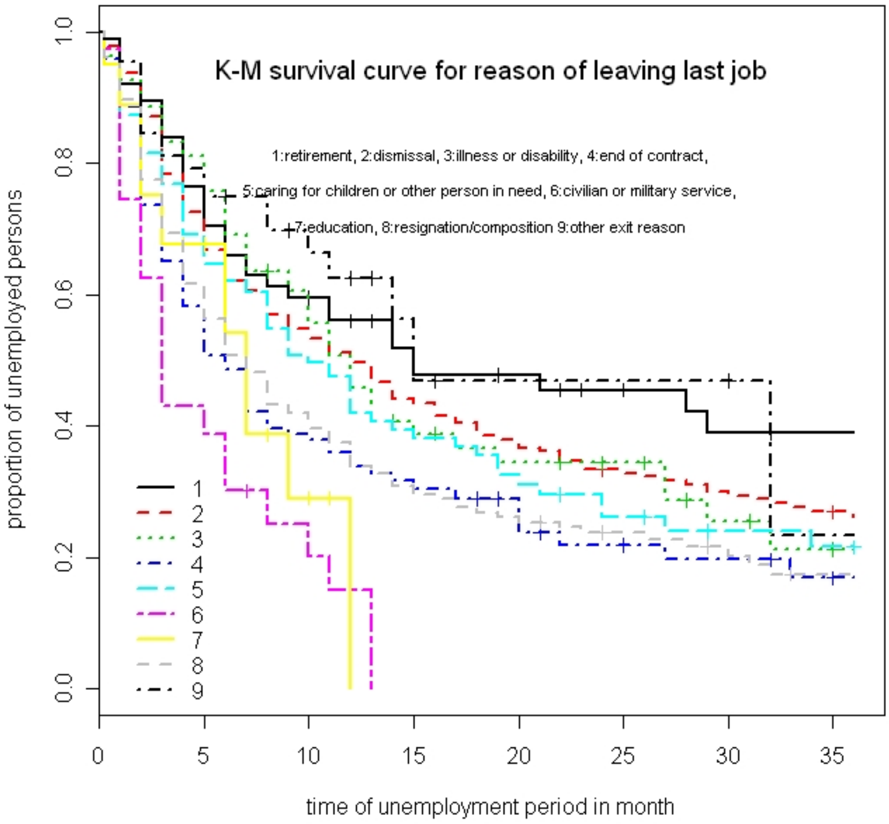


Figure 3.20.: Kaplan-Meier - reason for exit

one of the longest expected duration times in this case, a median duration time of 6 months, experienced before quitting the labour market. The same median of 6 months is observed for already retired persons. Illness or disability comes next in the ranking for longest duration times before exit to inactivity (median = 5.5 months). For comparison, all other exit reasons have a median between 2 and 4 months except the exits caused by further educational training or interruption due to military or civil service. The median of those two exit reasons is not significant due to underrepresentation in entries in the sample selection of case 3. There are only 7 and 2 counts respectively for these categories.

Combining the two exit causes in a competing risks approach helps gaining a larger pool of observed individuals to be used for the calculation. The estimation of the cumulative incidence functions for the covariate `JLWI` provides the following significant results. Comparing the probabilities of departure from employment after a chosen time point of 12 months confirms the findings for this covariate that have been stated in the preceding paragraphs. After one year of unemployment those who had further education or civilian or military service as reason for their last job exit show a probability of more than 50% to return to employment. On the other hand poor expectations result for already retired persons (15%) and for those

who where quitting their job due to illness or disability (28%). Child or family care also worsens the chances of early re-employment with this group having a probability of approximately 25% to leave unemployment. Dismissal shows a cumulative incidence of approximately 34%. The curves of end of contract and resignation lie close together and show a probability to leave after a year near to 40%.
Considering the time period of 12 months again but transition to the competing event, which is transition to inactivity, the worst chances with a cumulative incidence of 0.14 are given for those who experienced dismissal from their last job. Note that due to the exit to inactivity, which is for the most part retirement, correlation to the elder working generation may be inherent. The second lowest probability to leave into inactivity is given for those whose previous job contract has ended because of predefined termination (18%). All other exit reasons show a probability between 22% and 34% to leave unemployment by experiencing the competing event.

REGIONAL VARIABLES

inhabitants

A closer look at the regional component of an observed individual visualized in figure 3.21 reveals that those who live in the biggest cities of Austria are affected by a longer median unemployment duration. The graph visualizes this supposition. A division into the defined two exit states 'employment' and 'inactivity' though does not confirm this finding. A p-value from the log-rank test above 0.5 is obtained in both cases. For comparing the cumulative incidence of competing risks for this covariate, Gray's test [Gray, 1988] is performed. The covariate labelled `XEINW` gives information about the size of town the observed person lives in. The results of the test show significance for our cause of interest, which is exit to employment, but not for the competing risk to exit from the labour force. The results reveal that the probability to leave unemployment for (re)employment is far lower for those living in one of the biggest cities in Austria with more than 100.000 inhabitants, namely Innsbruck, Salzburg, Linz, Graz and Vienna. Individuals living in other cities or towns with less than 100.000 inhabitants all show better probabilities to leave unemployment. The probabilities among people living in one of these towns however, are not remarkably differing from each other.

federal states

Beginning with case 1 investigation, which does not distinguish between any cause of unemployment termination, a glance at the federal states provides similar results as those obtained above. Vienna appears to be the living place with highest unemployment duration. Specifying case 2, the consideration of uncensored observation with exit to employment, people from Vienna and Lower Austria together with Vorarlberg show a slightly decreased chance of finding a job compared to the other federal states. For exit 'out of labour force' no evidence for differences was detected.
The same results are confirmed by the cumulative incidence curves, where only the exit to employment reveals significance and Vienna has the lowest cumulative incidence curve, sig-

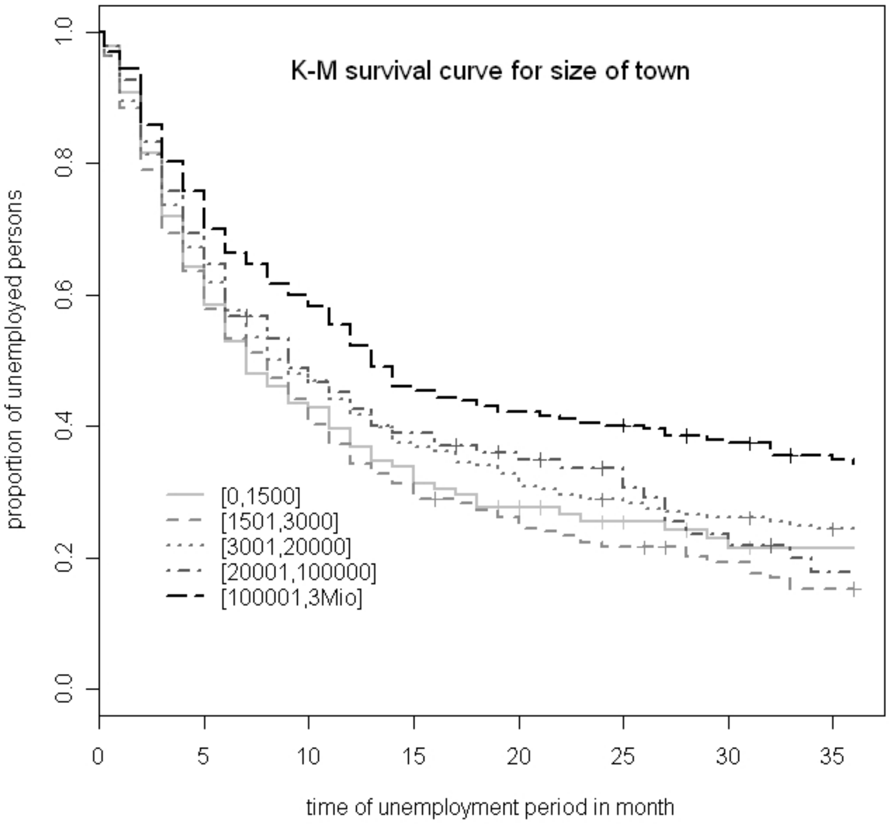


Figure 3.21.: Kaplan-Meier - number of inhabitants

nificantly behind the others. Tirol, Salzburg, Upper Austria and Carinthia are the federal states with the highest cumulative incidence curves. Unemployed persons who live there show a higher probability to exit from unemployment than those from Burgenland, Lower Austria and Styria.

SUPPORT AND JOB SEARCH

support from AMS

As far as support form the Public Employment Service Austria (AMS) is concerned, support which is received during the period of unemployment, the situation presents itself as follows. There is strong evidence for differences in either case of exit-type distinction. Those who receive unemployment benefits are in all cases more likely to end their unemployment period than those who receive other or no support. A graph showing the survival curves for the case without exit distinction is given as an example in figure 3.22. While for the category of 'other financial support' the counts are too few to make clear statements (3 and 4 entries respectively to the exit cause distinction 2 and 3), the group of those receiving minimum financial benefit seem to perform worse compared to others showing longer unemployment duration expectations.

Performing the cumulative incidence test reveals significance for both causes, the return to working life and the quitting from labour force. In a high p-value for both causes we find evidence to reject the null-hypothesis of equality among the classifications. Let us consider the cause of interest, exit to employment, first. Showing a probability to leave unemployment of more than 50% after a chosen observation time of 12 months confirms better chances for those receiving regular unemployment benefits. This has already been discovered in the preceding case observations of this covariate. Quite far behind with a probability to leave after one year of unemployment of 30% are those who attend a course offered from the AMS during their unemployment period. There might be two possible reasons for the reduction in chances of this group. One could be that further educational training is offered especially to those with difficulties finding a new job and the other could be that the unemployed person who attends a course is less eager to abort the vocational training for early job entry and/or has less time for intensive search. The same percentage for the probability to leave unemployment after one year is shown for those who do not get any support. With 11%, the bad performing of minimum financial benefit receivers may originate in their social position which causes this type of benefit classification.
Concerning exit 'inactivity' the probability curve (CI-curve) for those receiving no support is always above the curve for those who receive unemployment benefit. The unemployed who receive minimum financial benefits are represented by a constantly increasing cumulative incidence function.

job search

Regarding the kind of job searching activity, indicated by the covariate `HSART`, two types are demonstrated to be not favourable in raising the chances of exiting unemployment. One of such search criteria that is leading to a higher expected duration of unemployment of more than 3 months, is when the person has been in contact with the Public Employment Service Austria (AMS) (`HSART1`). The other one which raises the average duration is the search type defined as waiting for placements provided from the AMS, which is obviously correlated to the first one. Further correlation may be assumed with the education level, as mostly well educated persons are rarely placed by the AMS but search by other means. Significant, and increasing the chance of leaving unemployment, is the indicator variable for performed job interviews (`HSART8`). All other search activities (*find them listed in section* 3.1) do not show strong enough differences concerning the expected duration times to give any reason for assuming influence on the unemployment duration. However, it should be noted that only those who do perform active search are regarded as unemployed according to our chosen definition (*see section* 3.2) and only those persons form part of our sample set.

Distinction between the exit cases 2 and 3 yields the following additional insights. For exit to employment 'job interviews' no longer appear to be significant. To the previously listed unfavourable indicator of 'contact to AMS' (which indicates slightly worsening chances regarding this exit), looking for jobs in newspapers, applications sent by mail and contact to

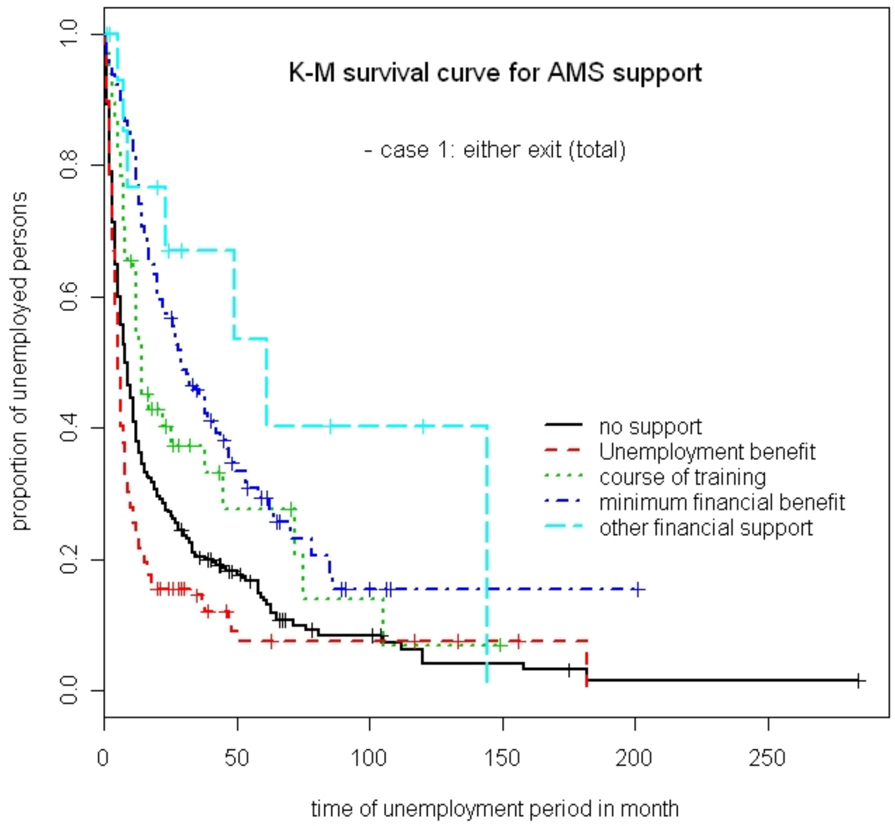


Figure 3.22.: Kaplan-Meier - AMS support

private employment agencies join in giving mild evidence for being significant factors that accompany marginally reduced chances of finding a job.

Regarding the exit to economic inactivity, the only search characteristic that gives mild evidence (p-value $p = 0.0455$) for being a significant influence factor on the preceding unemployment duration is the verbal exchange with friends. However, no good reason is found for explaining the possibly inherent elongating influence.

Next in our covariate study, the cumulative incidence curves for the search activities that turned out to be significant are described below. The first one, the indication of contact to AMS, is plotted in figure 3.23.
We observe higher probability to leave unemployment by our cause of interest, (re)employment, for those who are in contact with AMS. Remember that the above results from the cases of observing uncensored data revealed a longer median in each case for those who had been in contact with AMS. It is therefore interesting to see that the competing risks approach suggests higher probability of getting out of unemployment by finding a job for individuals who are in contact with the labour market service. But note that for the first 3 months the chances

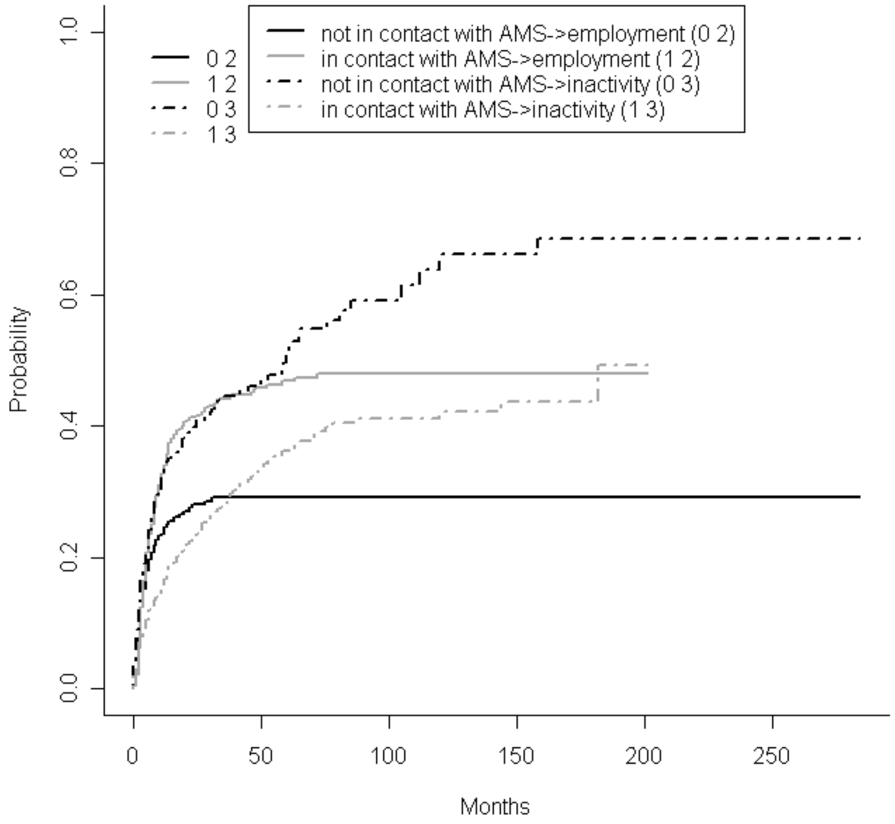


Figure 3.23.: CIF - contact to AMS

appear to be equal to those searching by other means. On the contrary, the competing exit cause shows reversed probabilities.

Consideration of the fact whether a job offer has been placed by the AMS or not provides similar results to the question of whether contact has been made with AMS or not. This is explained by the high certainty of getting a job offer from the AMS once being registered there as someone who is looking for a job.

The criterion whether the unemployed avail themselves of the job offers in newspapers or not appears significant for statistical difference for the competing cause of exit to inactivity. Those who use newspapers for job search have a slightly lower probability of ending the unemployment period by all observed time points compared to others who are not using this medium. Also significant in the competing exit cause only, appears the variable of verbal exchange with friends about job opportunities, indicating marginally lower probability of earlier exit over time.

Statistically different in probability to leave the status of unemployment for either exit cause seems to be whether a person has sent an application via mail to potential employers. For finding a job the probabilities are equal at the beginning but increase less for those who did not sent an application. Note that in the questionnaire the number of applications sent is not asked for and thus no information about the intensity of job search is given for this degree of searching activity. Reversed probability-curve-ordering is given considering the competing exit cause. The curve for application via mail being drawn below the other one. If the application is not sent by mail but the applicant has put an advertisement concerning his job search in a newspaper or other similar medium, the tendencies apply the same as the ones for applications per mail.

In conjunction with those two search activities is the question whether a person is waiting for a reply to an application or not and therefore the same structure of the cumulative incidence curves is given. The probability of an individual becoming employed without waiting times for an answer to an application is reduced by time more rapidly than with such waiting times.

Those persons who had already job interviews during their time of unemployment have an increased probability to find a job. Reversed ordering of the probability curves is again given considering the competing cause. This is shown in a form that those without previous job interviews have a higher probability to end up in inactivity at all considered time points compared to those who have been interviewed for a job during their unemployment period. As the cumulative incidence curves highly differ, they are plotted in figure 3.24 for illustration.

Waiting for placements provided by the AMS is significant for the competing exit cause only. A slightly increased probability to end unemployment earlier for exit to inactivity is given for individuals not waiting for a placement.

If the unemployed person is in contact with private employment agencies the probability for departure from the state of unemployment into a new working contract appears to be advantageous. At the very beginning (first 3 months), the probabilities are almost equal to those not in contact with private agencies, but the positive effect is strengthening with the duration of unemployment. For the competing cause the probabilities are lower for those being in contact with private agencies than those without such connections to end the unemployment period throughout all observed duration times.

No significant difference could be detected for the individuals that prepared for self-employment and those who stated to await for reply to an advertisement compared to those who did not experience either of these activities. For the variable of either trying to get working licences or not, no relevance could be detected either.

As seasonal unemployment is included in the considered labour force definition (section 3.2) we are interested in finding out whether the month in which the last job came to an end plays a role in the subsequent duration of unemployment. Therefore we applied the log-rank test for case 1, 2 and 3, which are 'no exit distinction', 'exit: employment' and 'exit: inactivity' respectively. While for case 3 the month of job exit does not appear relevant, case 1 shows mild evidence for statistical differences and case 2, our event of interest, allows for interpretation

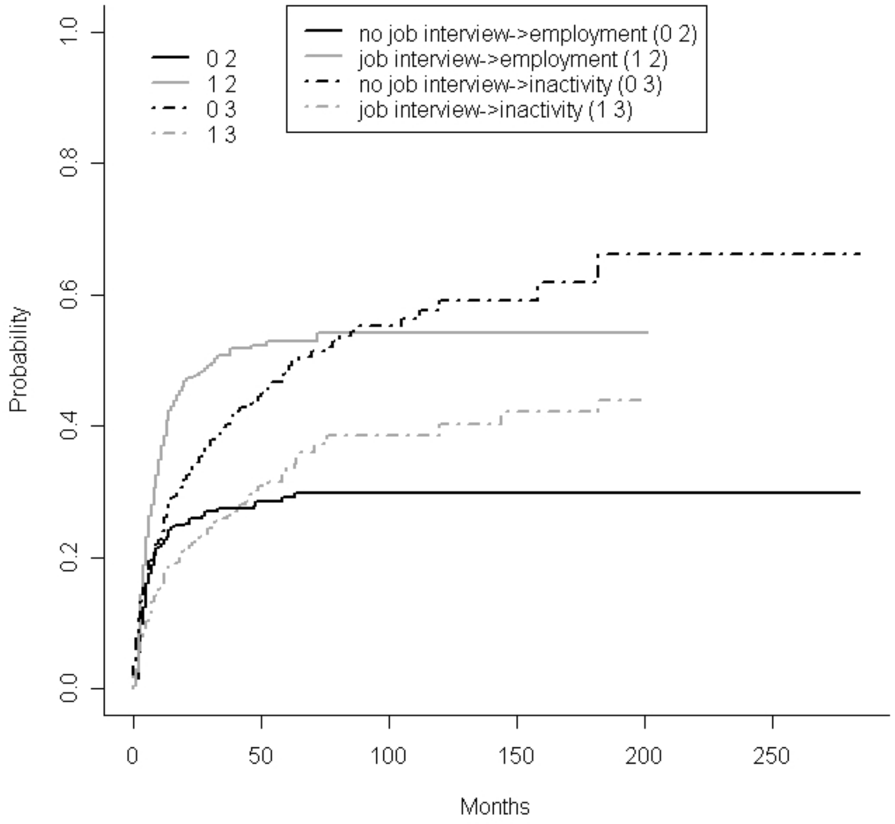


Figure 3.24.: CIF - job interview

with a p-value of 0.054 in the log-rank test. In the general consideration of case 1, the median for unemployment duration is apparently higher for the months from May until October with a value between 11 and 13 months. All other months of the year show a median-value below 9 months. When considering the exit cause of finding the way (back) to employment, the average median for unemployment duration is 4 months, exceeded by median-values of 8 months in September, 6 months in October and 5 months in November. Below the average are the median of March and April with 3 and 2 months respectively. This gives reason for the assumption of better job opportunities in summer compared to the winter period.

The competing risks approach via the visualisation of cumulative incidence curves shows better probabilities to exit into employment when the last job ended in winter or spring and least chances for those who had to leave their last job in autumn. These findings confirm seasonal unemployment and could in further research be examined for correlation to the branch of industry.

PARTNER CHARACTERISTICS

According to influences on unemployment duration, the last section in our preliminary covariate selection stated in section 3.1 concerns the characteristics of partners. Clearly, this selection criterion reduces our observable sample set to those being in a relationship with another person at the time of elicitation of responses by the questionnaire. The variables in question are listed in table 3.5, where the number added to their label indicates the case to which it belongs (the case distinction is described in section 3.4). The resulting p-values from the log-rank tests are listed in the third column of the table.

covariate	label	two-sided p-value (log-rank test)
Employment status of partner (case 1)	PARTNERERW1	p= 1.15e-06
Employment status of partner (case 2)	PARTNERERW2	p= 0.218
Employment status of partner (case 3)	PARTNERERW3	p= 0.00118
Branch of industry of partner (case 1)	PARTNERSEK1	p= 0.0431
Branch of industry of partner (case 2)	PARTNERSEK2	p= 0.965
Branch of industry of partner (case 3)	PARTNERSEK3	p= 0.901
Current job position of partner (case 1)	PARTDBERS1	p= 0.85
Current job position of partner (case 2)	PARTDBERS2	p= 0.618
Current job position of partner (case 3)	PARTDBERS3	p= 0.923
Type of current profession of partner (case 1)	PARTDTAET1	p= 0.74
Type of current profession of partner (case 2)	PARTDTAET2	p= 0.786
Type of current profession of partner (case 3)	PARTDTAET2	p= 0.682

Table 3.5.: partner

Covariate 'Employment status of partner' (PARTNERERW) shows evidence for significant differences between the distinction of whether the partner has a job or is unemployed or out of the labour force in course of case 1 and 3 investigation, but not for uncensored data according to our exit cause of interest (case 2). The Kaplan-Meier curves for the first case, when the cause differentiation is ignored, are shown in figure 3.25. What has to be kept in mind though is that the confidence intervals for the later two states are larger as there are fewer entries in these categories.

Focusing on exit 'inactivity', the following statement qualifies for this third case of observation. Inactivity of the partner increases the individuals expected unemployment duration before finally departing from the labour force themselves.

Other than the log-rank test for uncensored data, Gray's test for comparison of cumulative incidences of competing risks, where censoring is also considered, identifies solely the event of interest, exit to employment, to be significant, indicated by a p-value of 0.004216253. The highest cumulative incidence is given for those being in a relationship with an employed partner. An inactive partner seems to prolong the duration of unemployment and if the partner is unemployed as well, the situation presents itself at its worst concerning the probability of leaving the state of unemployment.

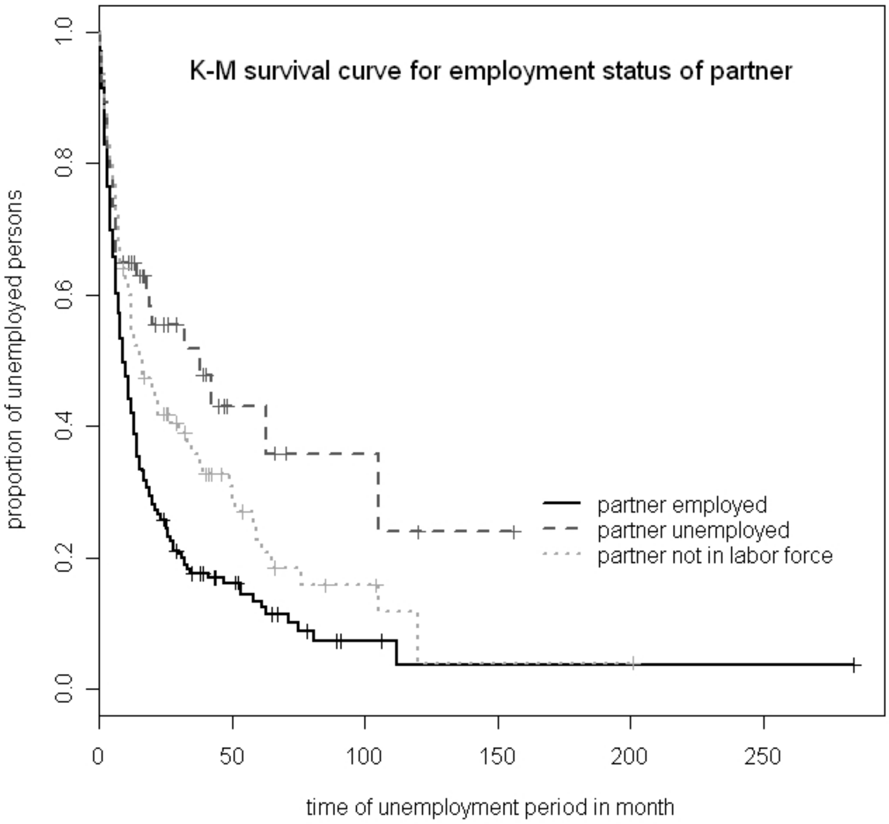


Figure 3.25.: Kaplan-Meier - employment status of partner

Branch of industry of partner (PARTNERSEK): Although mild evidence for differences due to the working branch of one's partner is given in the general case 1, this theory is not confirmed by the subsequently realised exit distinction. For neither type of exit cause 2 and 3 the log-rank test suggested a rejection of equal distributions.

According to Gray's test the only eye-catching finding is that partners who work in the sector of 'catering trade' seem to negatively influence the individuals' probability of leaving unemployment to work in a new job. The estimated cumulative incidence therefore never exceeds 0.2138. A pragmatic possibility is that in this sector a helping hand is commonly desireable which could be provided by the unemployed partner.

Current job position (PARTDBERS) and type of profession of partner (PARTDTAET): The Null-hypothesis of none of the performed tests can be rejected in either case for these covariates. Therefore the exact specification of a partner's job does not seem to play a significant role for the unemployment duration of the observed individuals themselves.

3.4.2. Nelson-Aalen estimation

As an extension to the non-parametric estimator of the hazard in the previous section the Nelson-Aalen estimator of our data is calculated. The Nelson-Aalen estimator provides a non-parametric estimator of the *cumulative* hazard function based on right censored data. A short theoretical description has already been given in section (2.3.1). To repeat, the Nelson-Aalen estimator is a staircase function with

- the location of the steps to be at each observed unemployment-period
- the vertical size of the steps to be $\frac{1}{r}$ ($r \ldots$*number at risk*).

The purpose of the hazard estimation is to find out which type of distribution might be adequate for a duration model. A short overview of the hazard shapes of the relative distributions can be found in section 2.4.
First the Kaplan-Meier type hazard estimate is plotted in figure 3.26. Note that the hazard is evaluated at the mid-points of the observed unemployment ending time points.

```
> hazest<-kphaz.fit(efftime,status,q=1,method="nelson")
> #where q...number of failure times combined for estimating the hazard at their
> #midpoint. Default is 1.
> kphaz.plot(hazest, lwd=2)
> mtext("Kaplan-Meier type hazard estimate",side=3,line=-3,cex=1.2)
```

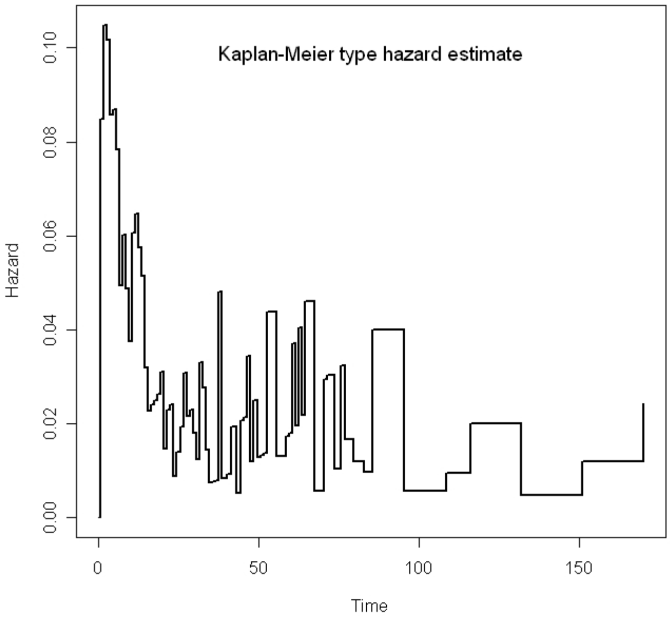


Figure 3.26.: parametric hazard function

The Nelson-Aalen estimator according to the 'single-exit assumption' is plotted in figure 3.27.

```
> fit <- survfit(Surv(efftime, status) ~ 1)
> nelscumhaz<-cumsum(fit$n.event/fit$n.risk)
```

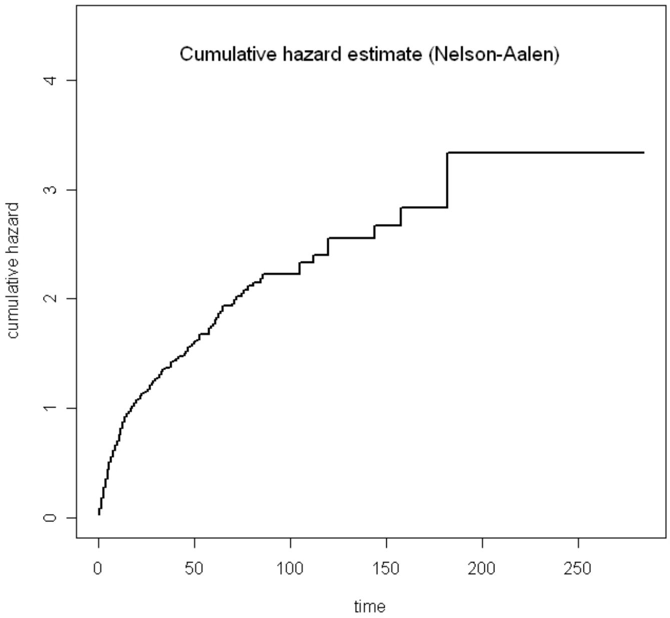


Figure 3.27.: single risk cumulative hazard fct.

Also of major interest for us is the comparison to 'competing-risk' models. Therefore the Nelson-Aalen estimator is also calculated for the exit differentiation between exit 'employment' (*left graph of* figure 3.28) and exit 'inactivity' (*graph on the right hand side of* figure 3.28). The curves take the possible transitions into account and show, for the respective exit, how the cumulative hazard increases over time. While the curve for exit into inactivity shows a constant increase, exit into employment is characterized by an increased probability of exit at the beginning. This curve is flattening for longer experienced unemployment periods.

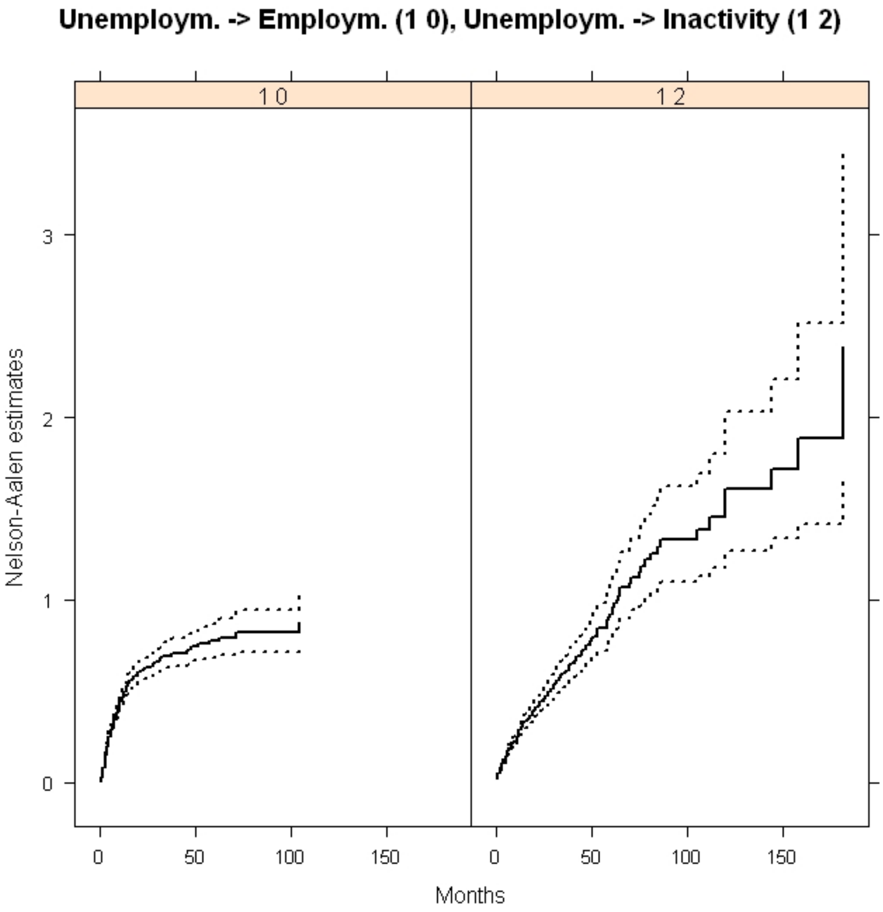


Figure 3.28.: competing risk cumulative hazard fct.

3.5. Parametric analysis

Given the above non-parametric results the adequacy of parametric distributions for a model are considered next. (*a short theoretical introduction of parametric models in duration analyses has been given in section* 2.4).
This section is an estimation attempt for uncensored data observed in 3 different ways. Firstly without unemployment exit distinction, then for the transition to employment (case 2) and as third case the transition to economic inactivity (*compare case distinction in section* 3.4).

It is of high relevance to find out about how likely a probable model assesses the data before making the model decision definite. QQ-plots are used as an assisting diagnostic tool to check model adequacy. For better understanding and revision a few remarks on quantiles are made. Therefore we recall the definition for quantiles for a certain distribution from page 6. The *sample* quantile function to which it is going to be in comparison with is defined as

$$Q_n(p) = inf\{t : S_n(t) \geq 1 - p\} \quad \text{for } 0 < p < 1$$

Note that $Q_n(p) = S_n^{-1}(1-p) = F^{-1}(p)$ for $F = 1 - S_n$. [Smith, 2002, p57]
Within the range $\frac{j-1}{n} \leq p \leq \frac{j}{n}$ the $Q_n(p) = t_{(j)}$. The evaluation points are

$$p_i = \frac{i - \frac{1}{2}}{n} = \frac{(2i-1)}{(2n)} \quad \text{for } i \in \{1, \dots, n\}$$

where i stands for the according ith-position in the ranked duration data which is given by $t_{(1)}, t_{(2)}, \dots, t_{(n)}$. And n is the amount of data points. Thus t_i is the $100(\frac{i-\frac{1}{2}}{n})$th sample percentile or $\frac{i-\frac{1}{2}}{n}$th sample quantile as $Q_n(p_i) = t_{(i)}$.
Then for plotting reasons of all p_i pertaining to an occurring duration-time-point ($t_{(j)}$) the mean value is taken and defined as p_j.
The QQ-plot is represented by a graph having the ranked data plotted on the vertical axis and the theoretical percentiles of a comparison distribution on the horizontal axis. The 45°-line through the origin (0,0) represents equality of both percentiles. A good description by a chosen distribution is given, if the points lie close to this line.
The points are given as

$$(Q(p_j), Q_n(p_j)) = (F^{-1}(p_j), t_{(j)})$$

First, the data is tested for the basic parametric model, the exponential model. As the exponential distribution is a special case of the Weibull distribution with scale 1, it is at the same time a check whether the data alternatively comes from a Weibull distribution with scale different from 1.
To repeat the distributional attributes, the related functions of the exponential and Weibull fit (*from section* 2.4) are stated below

distribution	*parameters*	*functions*
exponential	$\gamma > 0$	$F(t) = 1 - e^{-\gamma\, t}$ $S(t) = e^{-\gamma\, t}$ $f(t) = \gamma\, e^{-\gamma\, t}$ $\lambda(t) = \gamma$ $\Lambda(t) = \gamma\, t$
Weibull	$\gamma > 0,\ \alpha > 0$	$F(t) = 1 - e^{-(\gamma\, t)^{\alpha}}$ $S(t) = e^{-(\gamma\, t)^{\alpha}}$ $f(t) = \gamma\, \alpha\, (\gamma\, t)^{\alpha-1}\, e^{-(\gamma\, t)^{\alpha}}$ $\lambda(t) = \gamma\, \alpha\, (\gamma\, t)^{\alpha-1}$ $\Lambda(t) = (\gamma\, t)^{\alpha}$

The relationship between $\Lambda(t)$ (the integrated hazard) and $S(t)$ (the survivor function) for the exponential distribution is

$$ln(\Lambda(t)) = ln(-ln(S(t)) = ln(\gamma) + ln(t).$$

And for the Weibull distribution the relationship is

$$ln(\Lambda(t)) = ln(-ln(S(t)) = \alpha(ln(\gamma) + ln(t)).$$

This equation can be used for a graphical test of goodness of fit. Using the fact that the transformed equation

$$ln(t) = -ln(\gamma) + \sigma * ln(-ln(S(t)) \tag{3.12}$$

is a straight line with slope σ (*note that in case of exponential fit* $\sigma = \frac{1}{\alpha} = \alpha = 1$) and intercept $-ln(\gamma)$ in a plot with $ln(t)$ on the x-axis and $ln(-ln(S(t))$ on the y-axis.
The first thing to do before constructing the plot is substituting the continuous time t in the survivor function $S(t) = e^{-\gamma\, t}$ with the observed $t_{(j)}$.
In case of good fit of the data $S(t_{(j)}) \approx 1 - (p_j)$ and

$$ln(t_{(j)}) \approx -ln(\gamma) + \sigma * ln(-ln(1 - (p_j)) \tag{3.13}$$

We use figure 3.29 to figure 3.31 to demonstrate the parameter estimation.

For getting the y-intercept, $ln(-ln(1 - (p_j)))$ is set to be zero. Then $p_j = 1 - \frac{1}{e} = 0.6321$. For the case of no exit distinction t_j turns out to be approximately 5.834169. For exit 'employment' it is 5.534754 and for exit 'inactivity' 7.42942.
Then $ln(5.834169)$, $ln(5.534754)$ or $ln(7.42942)$ is multiplied by (-1) and the exponentiated interim result provides a rough estimate for γ which is $\hat{\gamma} = 0.171404$, $\hat{\gamma}_2 = 0.1806765$ and $\hat{\gamma}_3 = 0.1346$ respectively.
Further, the shape parameter α is estimated from the inverse slope of an estimated straight line through the quantile points. We obtain this line by standard linear regression using the *R*-function `lm()`. If the slope is close or equal to 1, the exponential fit is deemed to be adequate. According to the cases of 'no exit distinction (either exit)' (1), 'employment' (2) and 'inactivity' (3) the α-estimates are $\hat{\alpha} = 0.7450815$, $\hat{\alpha}_2 = 1.094604$ and $\hat{\alpha}_3 = 0.6234616$.
The regression line additionally provides another way of estimating γ. γ can be derived from the intercept of this line, which equals $-ln(\gamma)$. The corresponding estimates are $\hat{\gamma} =$

0.1207328, $\hat{\gamma}_2 = 0.1222323$ and $\hat{\gamma}_3 = 0.1112761$.
In case of exponential fit, the mean duration time θ is $\frac{1}{\gamma}$. A third way of estimating the parameters α and γ is via the `survreg`-function in R. The resulting intercept-value is $-ln(\hat{\gamma}) = ln(\hat{\theta}) = 2.177707$. The estimates in this case are based on a MLE (maximum likelihood estimation) approach and are $\hat{\gamma} = 0.1133010$, $\hat{\gamma}_2 = 0.1519383$ and $\hat{\gamma}_3 = 0.07892683$ regarding to the exponential fit. For the Weibull fit this approach leads to estimates $\hat{\alpha} = 0.7948045$, $\hat{\gamma} = 0.1332505$, $\hat{\alpha}_2 = 1.040121$, $\hat{\gamma}_2 = 0.1490851$, $\hat{\alpha}_3 = 0.6613934$ and $\hat{\gamma}_3 = 0.1102422$.
Note that a higher value of the parameter γ results in a quicker downward slope of the survivor function indicating that there is an increased probability of exit from the state of unemployment.
Recapitulating the different estimation approaches, the above estimates are listed in table 3.6.

case distinction 'either exit'	'exit employment'	'exit inactivity'
manually $\hat{\gamma} = 0.171404$	$\hat{\gamma}_2 = 0.1806765$	$\hat{\gamma}_3 = 0.1346$
linear regression $\hat{\alpha} = 0.7450815$ $\hat{\gamma} = 0.1207328$	$\hat{\alpha}_2 = 1.094604$ $\hat{\gamma}_2 = 0.1222323$	$\hat{\alpha}_3 = 0.6234616$ $\hat{\gamma}_3 = 0.1112761$
MLE approach exponential $\hat{\gamma} = 0.1133010$ Weibull $\hat{\alpha} = 0.7948045$ $\hat{\gamma} = 0.1332505$	$\hat{\gamma}_2 = 0.1519383$ $\hat{\alpha}_2 = 1.040121$ $\hat{\gamma}_2 = 0.1490851$	$\hat{\gamma}_3 = 0.07892683$ $\hat{\alpha}_3 = 0.6613934$ $\hat{\gamma}_3 = 0.1102422$

Table 3.6.: parametric estimation results - expon./Weibull

For further use in plots and calculations we take the MLE-estimators (*recall equation* (2.14)). The MLE-estimators provide the best fit of the estimated parametric survivor functions to the empirical survivor functions.

Since the empirical hazard plot in case (2) (exit employment) shows that the hazard decreases for larger duration times, the log-normal distribution is chosen for another parametric data observation. The log-normal distribution is an alternative parametric distribution that seems suitable when large values of t are given less weight as the hazard increases early and decreases later.
For the same reason the log-logistic distribution is selected as parametric model for survival estimation.
Both distributions are additionally considered in our parametric estimations and their attributes already listed in section 2.4 are tabulated a second time for better comprehension.

distribution	*parameters*	*functions*
log-normal	$\gamma > 0,\, \alpha > 0$	$F(t) = \Phi[\alpha\, ln(\gamma\, t)]$ $S(t) = \Phi[-\alpha\, ln(\gamma\, t)] = 1 - \Phi[\alpha\, ln(\gamma\, t)]$ $f(t) = \frac{\alpha}{t}\phi[\alpha\, ln(\gamma\, t)] = \frac{\alpha}{t}\frac{1}{\sqrt{2\pi}}exp(\frac{-\alpha^2(ln(\gamma\, t))^2}{2})$ $\lambda(t) = \frac{\alpha}{t}\frac{\phi(\alpha\, ln(\gamma\, t))}{\Phi(-\alpha\, ln(\gamma\, t))}$ $\Lambda(t) = -ln(\Phi[-\alpha\, ln(\gamma\, t)])$
log-logistic	$\gamma > 0,\, \alpha > 0$	$F(t) = 1 - [\frac{1}{(1+(\gamma\, t)^{\alpha})}]$ $S(t) = \frac{1}{(1+(\gamma\, t)^{\alpha})}$ $f(t) = \frac{\gamma\, \alpha\, (\gamma\, t)^{\alpha-1}}{(1+(\gamma\, t)^{\alpha})^2}$ $\lambda(t) = \frac{\gamma\, \alpha\, (\gamma\, t)^{\alpha-1}}{(1+(\gamma\, t)^{\alpha})}$ $\Lambda(t) = ln(1 + (\gamma\, t)^{\alpha})$

note: in the log-normal distribution $ln(t)$ is assumed to be normally distributed with mean $\mu = -ln(\gamma)$ and standard deviation $\sigma = \frac{1}{\alpha}$ and in the log-logistic distribution $ln(t)$ is assumed to be logistically distributed with location parameter $\mu = -ln(\gamma)$ and scale parameter $\sigma = \frac{1}{\alpha}$.

If the data fits the model well $S(t_j) \approx 1 - p_j$, the theoretical quantile $Q(p_j)$ is close to the sample quantile $Q_n(p_j)$. This is equivalent to $p_j \approx F(t_j)$, thus for the Log-normal fit

$$p_j = \Phi[\alpha\, ln(\gamma\, t_j)] = \Phi\left[\frac{ln(t_j - \mu}{\sigma}\right]$$

This equation can be transformed by first taking Φ^{-1} on each side and subsequently bringing it into linear shape

$$ln(t_j) = \mu + \sigma\Phi^{-1}(p_j) \tag{3.14}$$

With $ln(t_j)$ on the y-axis and $\Phi^{-1}(p_j)$ on the x-axis the intercept μ and the slope σ can again be estimated via linear regression.

Similarly, for the log-logistic fit a linear function can be derived from the survivor function

$$S(t) = \frac{1}{(1 + (\gamma\, t)^{\alpha})}$$

which can be rewritten as

$$(\gamma\, t)^{-\alpha} = \frac{S(t)}{1 - S(t)} \approx \frac{1 - p_j}{p_j}. \tag{3.15}$$

By making use of the invariance property of the log-function the equation can be brought into linear relationship.

$$ln(t_j) = \mu + \sigma(-ln(\frac{1 - p_j}{p_j})) \tag{3.16}$$

In the plots of figur 3.32 to figure 3.37 the above explained relations are graphically exemplified.

The resulting estimates from the previous equations are listed in table 3.7

case distinction 'either exit'	'exit employment'	'exit inactivity'
linear regression		
log-normal		
$\hat{\alpha} = 0.7631507$	$\hat{\alpha}_2 = 0.9751516$	$\hat{\alpha}_3 = 0.6777508$
$\hat{\gamma} = 0.2623127$	$\hat{\gamma}_2 = 0.1207328$	$\hat{\gamma}_3 = 0.1207328$
log-logistic		
$\hat{\alpha} = 1.564896$	$\hat{\alpha}_2 = 1.958049$	$\hat{\alpha}_3 = 1.350597$
$\hat{\gamma} = 0.2217837$	$\hat{\gamma}_2 = 0.2429339$	$\hat{\gamma}_3 = 0.1832742$
MLE approach		
log-normal		
$\hat{\alpha} = 0.7908107$	$\hat{\alpha}_2 = 1.063532$	$\hat{\alpha}_3 = 0.6124891$
$\hat{\gamma} = 0.2498771$	$\hat{\gamma}_2 = 0.2382537$	$\hat{\gamma}_3 = 0.2466502$
log-logistic		
$\hat{\alpha} = 1.419571$	$\hat{\alpha}_2 = 1.916303$	$\hat{\alpha}_3 = 1.047737$
$\hat{\gamma} = 0.2467708$	$\hat{\gamma}_2 = 0.2392512$	$\hat{\gamma}_3 = 0.2381536$

Table 3.7.: parametric estimation results - log-norm./log-logis.

Another graphical method of checking the distribution in the data is to plot the empirical and the estimated parametric survival functions on a single graph. Then both curves are compared and if they are close, the particular distribution model is appropriate. Again, the MLE-estimates are taken for this graphical visualization. Figure 3.38 to 3.40 show this comparison for the 3 exit distinctions and respective parametric model fit.

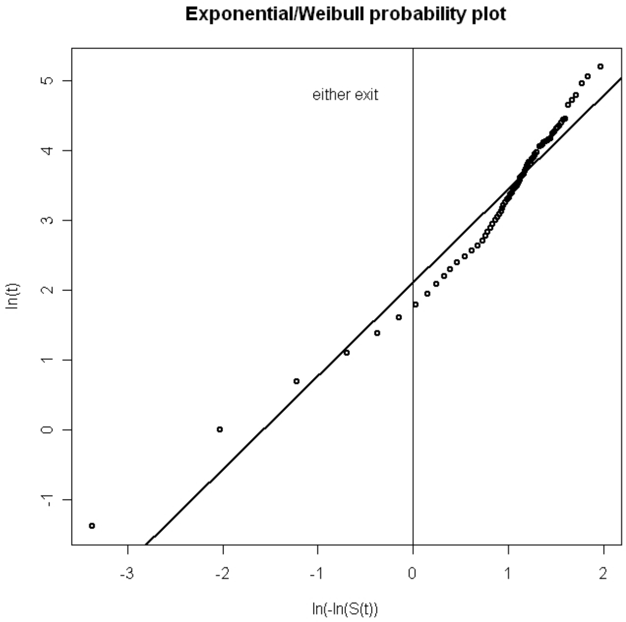
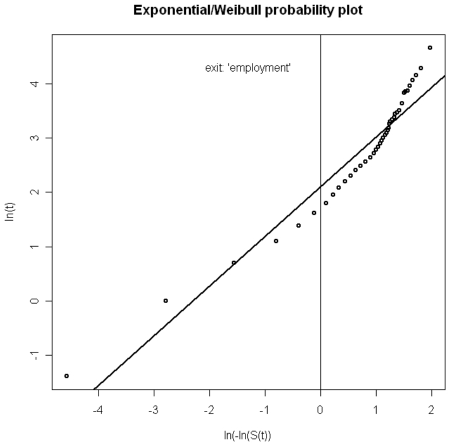


Figure 3.29.: prob.plot: expon./weib.-'total'



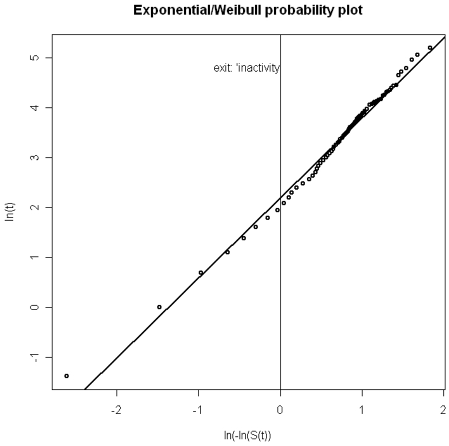


Figure 3.30.: prob.plot: expon.-'employment'

Figure 3.31.: prob.plot: expon.-'inactivity'

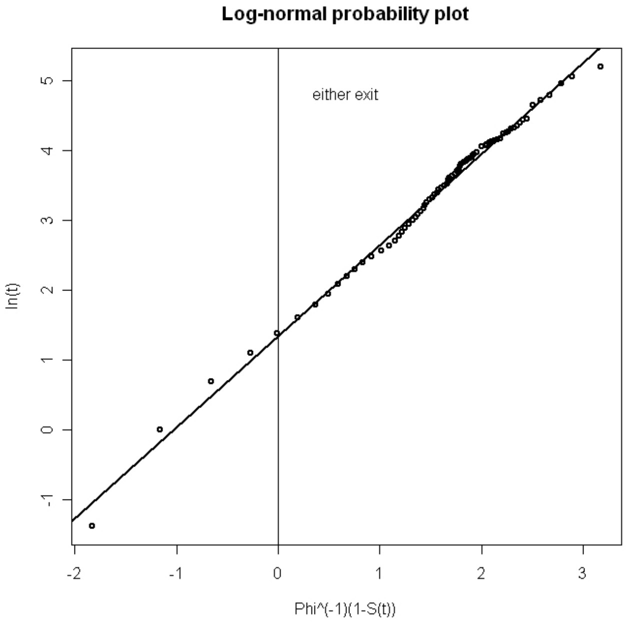


Figure 3.32.: prob.plot: log-norm.-'total'

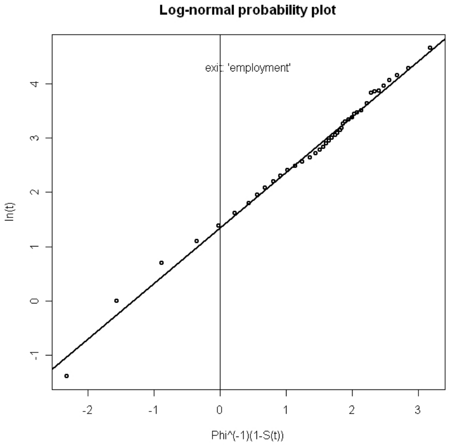


Figure 3.33.: prob.plot: log-norm.-'employment'

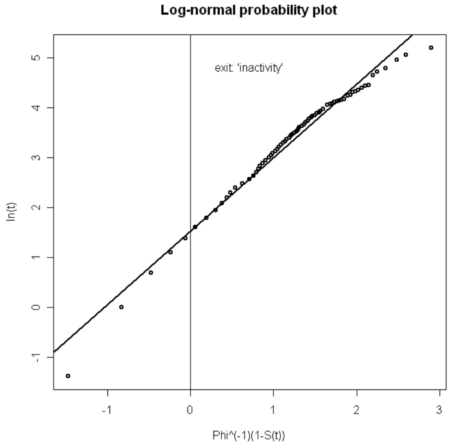


Figure 3.34.: prob.plot: log-norm.-'inactivity'

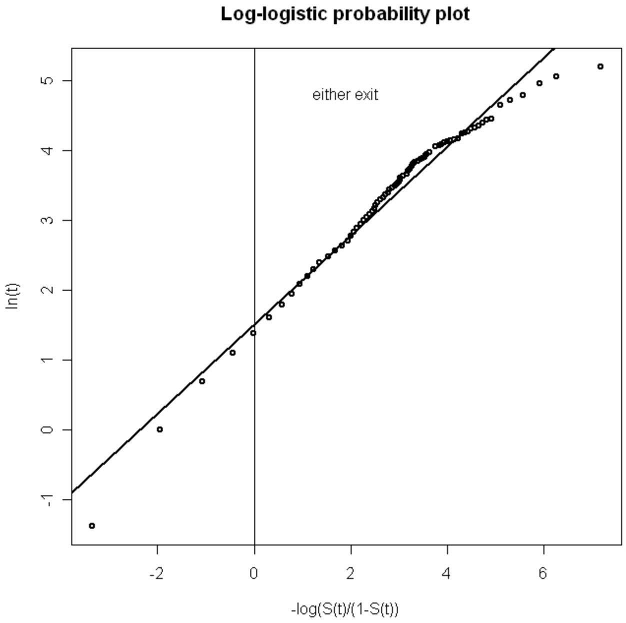


Figure 3.35.: prob.plot: log-log.-'total'

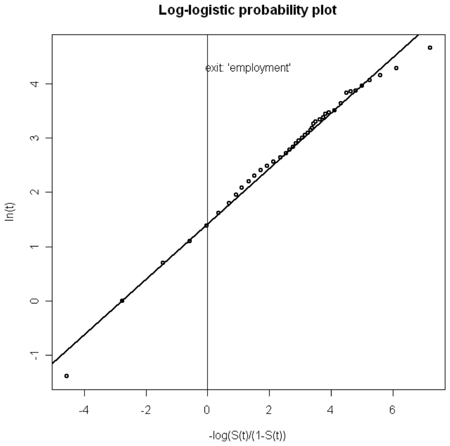


Figure 3.36.: prob.plot: log-log.-'employment'

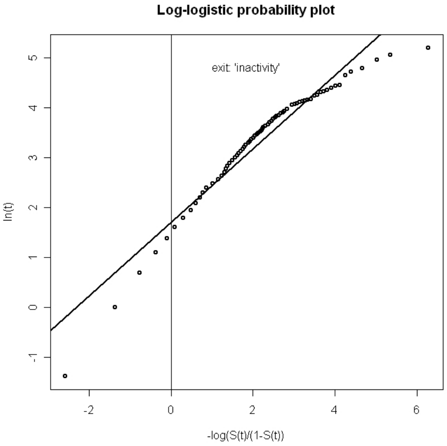


Figure 3.37.: prob.plot: log-log.-'inactivity'

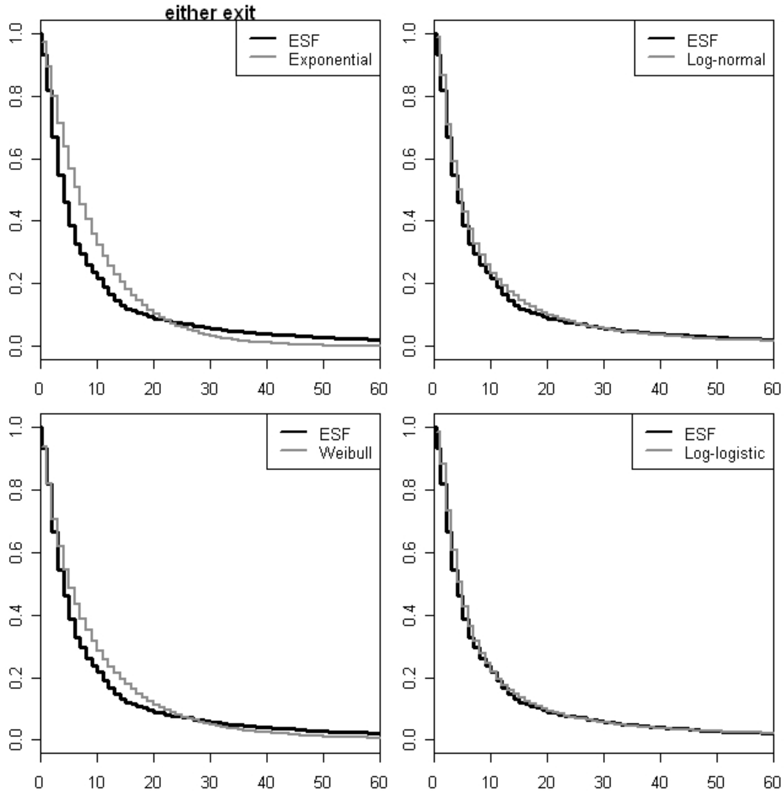


Figure 3.38.: empirical and parametric survivor curves - either exit

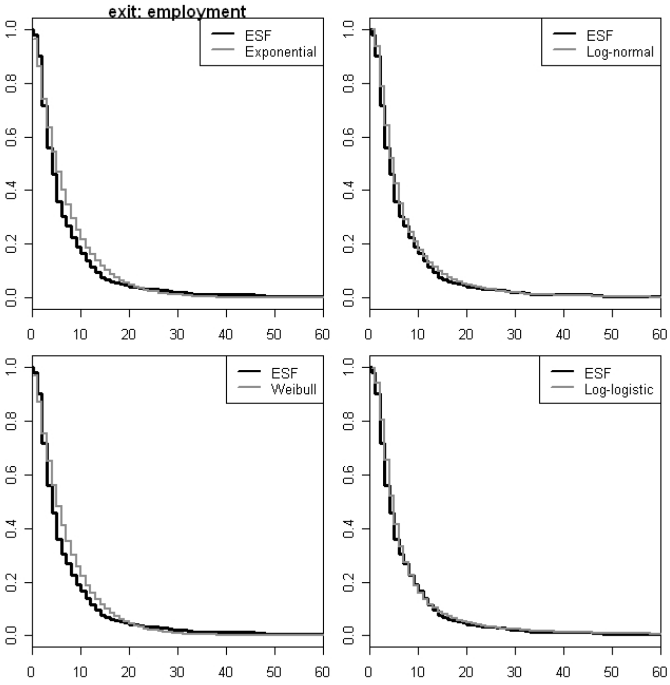


Figure 3.39.: empirical and parametric survivor curves - exit 'employment'

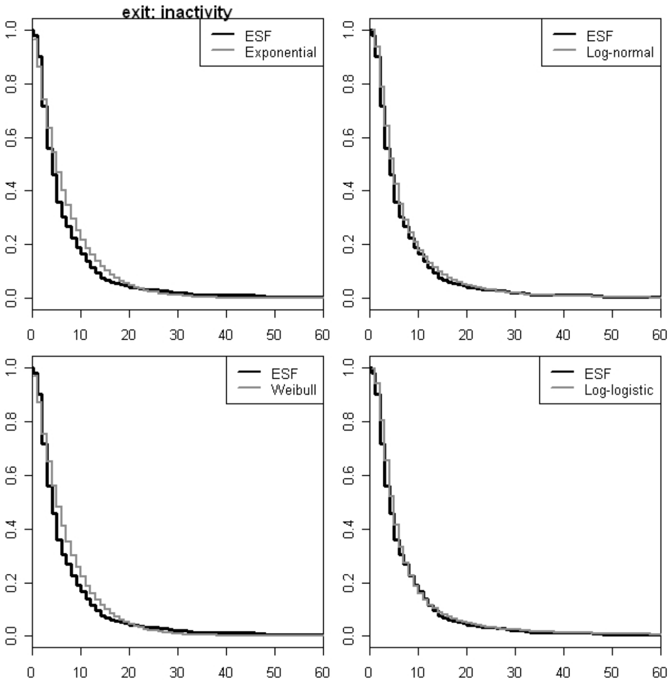


Figure 3.40.: empirical and parametric survivor curves - exit 'inactivity'

In order to compare the 4 chosen parametric models we summarize the estimated median, mean, variance and log-likelihood values in table 3.8.

EITHER EXIT			
distribution	*median*	*mean*	*log-likelihood*
sample	4	8.826046	—
exponential	1.811194	8.826046	-4178.685
Weibull	1.554389	8.542838	-4097.763
log-normal	1.386786	8.902197	-3998.154
log-logistic	1.399295	11.19968	-3992.62
EXIT: EMPLOYMENT			
distribution	*median*	*mean*	*log-likelihood*
sample	4	6.581618	—
Exponential	1.517768	6.581618	-1961.311
Weibull	1.550863	6.60227	-1960.243
log-normal	1.434419	6.530392	-1898.399
log-logistic	1.430241	6.868383	-1891.306
EXIT: INACTIVITY			
distribution	*median*	*mean*	*log-likelihood*
sample	4	12.66996	—
exponential	2.172721	12.66996	-1879.333
Weibull	1.650923	12.16062	-1776.623
log-normal	1.399784	15.37295	-1757.051
log-logistic	1.434840	88.26073	-1770.103

Table 3.8.: summary - parametric estimation

From table 3.8 we see that none of the estimated medians is close to the K-M-estimated median of 4. The log-logistic distribution performs worst in estimating the mean value. This listing of estimates reconfirms the importance of additional graphical considerations. Including the appearance of the QQ-plots, we derive the following suggestions. In case of exit to employment a log-logistic model deems to be adequate while for inactivity exit the Weibull model provides a better fit. For the estimation without exit distinction the log-normal distribution is regarded to describe the data best of all.

Having attained the general parametrically specified models in this section, we are further interested in testing the covariate influences. We will focus on the issue regarding the relationship between the unemployment duration and the considered covariates. The question is whether and which covariates help to explain the duration time of unemployment and if such an influence is given, how strong it affects the duration times. This effect is measured by the estimated coefficients of the according covariate in the model.

In course of this section we have derived linear forms for the different parametric models without covariate consideration (*equations* (3.13), (3.14) *and* (3.16)). Now, we formulate a generalized linear relationship for the case of integration of covariates.

$$ln(T) = \tilde{\mu} + \sigma Z = \beta_0^* + x^{'}\beta^* + \sigma Z \tag{3.17}$$

where Z denotes either an extreme value, a standard normal or a standard logistic random variable, depending on the distributions stated in table 3.8. x is the covariate vector, σ the scale parameter and $\tilde{\mu}$ the linear predictor. Turning to consider the dependent duration variable T instead of $ln(T)$, $\tilde{\gamma}$ is defined as $\tilde{\gamma} = exp(-\tilde{\mu})$ (*compare with equation* (3.18) *from* Example [Ex.3.1] *below*) and the shape parameter $\alpha = \frac{1}{\sigma}$.

Ex.3.1 `Example: Weibull model`

As an example for the inclusion of the covariate vector x into a parametric model, we choose the Weibull distribution for a short demonstration. Recall the hazard function of the Weibull distribution from table 2.1 ($\lambda(t) = \gamma\,\alpha\,(\gamma\,t)^{\alpha-1}$). The hazard function for a given covariate vector x is then defined as

$$\lambda(t|x) = \lambda_0(t) * exp(x^{'}\beta) = \gamma\,\alpha\,(\gamma\,t)^{\alpha-1}exp(x^{'}\beta) =$$

$$= \alpha\,\gamma^{\alpha}\,t^{\alpha-1}exp(x^{'}\beta) = \alpha(\gamma(exp(x^{'}\beta))^{\frac{1}{\alpha}})^{\alpha}\,t^{\alpha-1} =$$

$$= \alpha\,\tilde{\gamma}^{\alpha}\,t^{\alpha-1} = \tilde{\gamma}\,\alpha\,(\tilde{\gamma}\,t)^{\alpha-1}$$

where $\tilde{\gamma} = \gamma(exp(x^{'}\beta))^{\frac{1}{\alpha}}$.
Further, we get

$$\tilde{\mu} = -ln(\tilde{\gamma}) = -ln(\gamma(exp(x^{'}\beta))^{\frac{1}{\alpha}}) = -ln(\gamma) - \frac{1}{\alpha}x^{'}\beta. \tag{3.18}$$

β_0^* and β^* from equation (3.17) can therefore be derived from the parameters γ and β as follows

$$\beta_0^* = -ln(\gamma),$$

and

$$\beta^* = -\sigma\,\beta.$$

Using the R-function `survreg` gives us estimates for the parameters β_0^*, β^* and σ.

To test for goodness of fit of a model the AIC (Akaike's information criterion) is used to give a starting selection criterion. This automated process is implemented in R via the `stepAIC` function. Alternatively to the AIC a slightly different statistical criterion for model selection is given by the Bayesian information criterion (BIC). The intention is to find significant predictor variables. We do not rely solely on the results provided by the AIC or the BIC criterion, but rather use them as an initial selection method. Additionally, this test methods allow for inclusion of interaction terms. The idea behind the criterion is to adjust the goodness of fit, which is measured by the log-likelihood, by penalizing for complexity in case of a high number of parameters. This penalty-term is intended to avoid 'overfitting' of the model. For a chosen predetermined constant of 2 being the multiplying value of the parameters, the AIC is given by

$$AIC = -2 * log(L_{max}) + 2 * p \tag{3.19}$$

where p is the *number of parameters* involved and L_{max} the maximized value of the likelihood function for the estimated model.

The formula for the BIC is written as

$$BIC = -2 * log(L_{max}) + ln(n) * p \tag{3.20}$$

where n is the *number of observations* used in the model estimation.
The smaller the value of the AIC or BIC, the better the fit of the tested model to the given data.
[Tableman and Kim, 2004, p106]

The BIC is based on the Bayesian idea of learning from observations, given an a-priori distribution, resulting in an a-posteriori conditional probability. The penalty term for the number of parameters included in the model is weighted by the logarithm of the number of observations. Therefore the criticism of the BIC criterion lies in the crucial setting of a-priori information and the stronger penalty for a higher number of parameters. Thus the model selected via the BIC inherits a less complex structure which could be disadvantageous in prediction matters. Detailed information about comparison of these two selection criterions can be found in [Kuha, 2004].

In our analysis we applied both methods using the implemented pre-programmed R procedures. Recall, they are intended to provide solely an initial model selection.
In course of detailed non-parametric covariate analysis in section 3.4, we have already eliminated some covariates that appeared to be non-relevant. For the remaining ones, stated in table 3.9, potential for duration influence has been detected. This preselection represents the third important pillar for our final model choice next to the application of the AIC and the BIC criterion.

In course of the model selection procedure it is possible to consider the effect of three-way or even higher interaction terms, too. However, this would result in high covariate dimension besides difficult interpretation. Therefore we consider two-way interaction only.
The selection process is a quite long procedure with many intermediate results. It goes beyond the scope of manageable presentation to particularize all of them. Consequently only the final results are presented. Note that for our selection process the data is taken as equally weighted. This seems to be a plausible assumption as our main interest is to find the covariates effect on *individuals'* unemployment durations.

As the cases of exit distinction (1-3) are already examined separately in the previous non-parametric analyses and our main interest is exit cause 'employment' (case 2), we will focus on this event in the following parametric-model estimations. The other cases could be examined in similar manner. However, to demonstrate all 3 cases would go beyond the scope of our aim to give a short introduction in the usage of parametric models in survival analysis. For consideration of exit cause distinction, we rather concentrate on the more general Cox proportional hazard model and the competing risks regression . The corresponding model estimates are presented in the following section (*section* 3.6).

		observations
covariate	label	exit 'employment'
Age	BALT	680
Gender	BSEX	680
Marital status	BFST	680
Born in Austria	XBGEBLAO	680
Number of children	XANZKIND	589
Level of education	XKARTAB	680
Former job position	JBERS	680
Type of former profession	JTAET	680
Working hours	DSTD	679
Branch of industry	XDWZAB	680
Reason for last job exit	JLWI	543
Inhabitants	XEINW	680
Federal States	XNUTS	680
End of last job	START	564
Support of AMS	HAMSL	680
Job search activity	HSART1,3,5,8,10	608
Employment status of partner	PARTNERERW	286

Table 3.9.: selected covariates

The parametric model estimation starts using already attained information from the non-parametric analysis in section 2.3.1. In a stepwise search the model is reduced to a form including significant covariates only. In the process of this reduction attention should be given to the correct covariate elimination. If covariates are left aside that actually turned out to be significant, the attained model would yield wrong coefficients for the remaining variables. Another fact that has to be paid attention to is the possible interaction between certain terms. Ignoring this interaction effects also results in different estimates.

Bearing this in mind the following variables (listed in table 3.9) are chosen as initial covariate-selection.

The third column of the table shows the number of observations available related to the case of exit to 'employment' (2). This number of available data entries evinces a crucial point relevant in model estimation. Missing values in some data sets would result in excluding them from the sample set. To be able to consider a large number of covariates for coefficient estimation, a large sample set is needed to retain good estimation accuracy.

We start our analysis giving attention to the covariate that indicates the reason for the last job exit (`JLWI`). This variable is relevant only for unemployed individuals who have had a job before. The same applies to the variable that indicates the time of termination of the last job (`START`). Therefore we have to split the data set into the group who have never worked before *(no job experience)* and those who have had a work relationship *(job experience).*
First we will treat the group of persons who have been in employment before but lost their job.
We choose the Log-logistic model according to our suggestion after the examination without covariate inclusion. Note that the log-logistic regression model is an example for an accelerated duration time model (*compare section 2.5.2 AFT Model*). We then apply the BIC and AIC-step criterion.
Note that possible influences due to the employment status of a partner are not considered in this first attempt as consideration would result in further reduction of the sample set to persons who have lived with a partner in the period of data collection. This group of people will be examined later in a separate model (*table* 3.12).
Assuming that the gender aspect is the only covariate interacting with several others, we do not consider further interaction terms but split the sample into male and female observations and similarly build an appropriate model for each subset. The variable grouping of the covariates (*all of them are categorical or binary*) is the same as in the non-parametric section 3.4 documented on page 49 ff. We consider all covariates as indicator variables for better interpretation and consideration of the contribution of certain groups. Further information about the categorization is given in section 3.1. Any changes from this previously defined classifications are pointed out separately. Such a covariate whose category has been redefined is the one giving information about the number of children (`XANZKIND`). The last two categories are combined to a single one (3 or more children). Some classes of former profession are also combined. The new groups are 1:'apprentice', 2:'auxiliary work', 3:'manual employee', 4:'high professional manual', 5:'medium profession non manual', 6:'higher profession non manual' and 7:'farmer & freelancer'. Further, the last category of the covariate indicating the support of AMS ('other financial support') was added to the group of unemployment benefit receivers. The new classes are 0:'no support', 1:'unemployment benefit', 2:'minimum financial benefit' and 3:'course of vocational training'. Moreover, we consider only a few of the searching activities. We take those that appeared significant in the non-parametric analysis or seem to be indicators for intensive search. The BIC-step method is applied to the remaining covariates. First we consider the subset of female unemployed. The BIC criterion reduces the covariates suggested to two predictor variables, the support from AMS (`HAMSL`) and the contact to private agencies (`HSART10`). This drastic reduction from the initial covariate choice confirms the above mentioned doubts concerning the strong penalty term in this method. Performing the AIC step procedure by contrast suggests a longer list of covariates. Together with the significant covariates from a model approach with the plain initial setting we derive the final model stated in table 3.10. Note that further selection was done by a step-by-step procedure, removing non-significant predictor variables indicated by high p-values ('p-value method') or adding and removing covariates via the LRT (likelihood ratio test). The LRT statistic is a method for testing two nested models and its statistic is given by

$$-2 * ln\, L(\theta_0) + 2 * ln\, L(\hat{\theta}) \approx \chi^2_{df}$$

where df, the *degrees of freedom* equal the difference between the parameters of the compared models. Note that the second model is always the simpler one being a reduced special case of the first one. As the 'p-value method' and the LRT are asymptotically equivalent both are suitable for one-variable at a time reduction. To repeat, only women who have been in an employment relationship before are considered in this model estimation.

Recall that the log-logistic model is an AFT (accelerated failure time) model. Given the fact that the duration times are positive by nature it is possible to take logarithm of T. Compare to the previously stated formula (3.17). We can rewrite this formula in simpler manner as

$$ln(T) = x^{'}\beta^* + Z^*. \tag{3.21}$$

Then

$$T = exp(x^{'}\beta^*)\, exp(Z^*) := exp(x^{'}\beta^*)\, T^*. \tag{3.22}$$

Suppose that the log-logistic time T^* is a transformed random variable $T^* = g^{-1}(T)$ and thus $t^* = g^{-1}(t) = exp(-x'\beta^*)t$. The according baseline hazard function is given by

$$\lambda_0^*(t^*) = \frac{\alpha\gamma^{\alpha}(t^*)^{\alpha-1}}{1+\gamma^{\alpha}(t^*)^{\alpha}} \tag{3.23}$$

where $\sigma = \frac{1}{\alpha}$ (*recall equation* (3.17))
The hazard function can be written in terms of this baseline function as

$$\lambda(t,x) = \frac{\alpha\tilde{\gamma}^{\alpha}(t)^{\alpha-1}}{1+\tilde{\gamma}^{\alpha}(t)^{\alpha}} \tag{3.24}$$

where $\gamma = \gamma(exp(-x'\beta^*))$.
The resulting log-logistic survivor function is

$$S(t,x,\beta_0^*,\beta^*,\alpha) = \frac{1}{1+exp(\alpha(ln(t)-\beta_0^*-x'\beta^*))} \tag{3.25}$$

where $\beta_0^* = -ln(\gamma)$. As demonstrated in formula (3.15) the odds of survival beyond time t is given by

$$\frac{S(t,x,\beta_0^*,\beta^*,\alpha)}{1-S(t,x,\beta_0^*,\beta^*,\alpha)} = exp(-\alpha(ln(t)-\beta_0^*-x'\beta^*)) \tag{3.26}$$

Recall that the negative logarithm of these odds is used to define a linear relationship explaining the variable $ln(t)$ (*compare equation* (3.16)). Evaluation at two different covariate combinations x_1 and x_2 gives us the odds-ratio

$$OR(t,x_1,x_2) = exp(\alpha(x_2-x_1)'\beta^*) \tag{3.27}$$

We see, that this ratio does not depend on time. Therefore the log-logistic regression model is a model with a proportional odds property. The odds-ratio is used to measure the effect of the model's predictor variables. But before explaining the results of the model we first test the goodness of fit of the model.

Comparing the maximum log-likelihood of the attained log-logistic model to the exponential, Weibull and log-normal model with the defined covariate combination gives information about the goodness of fit. With the lowest maximum log-likelihood value among the four tested models, the log-logistic model appears to be the distribution of best fit.

Log-logistic Model - Women - exit: ’employment’

Intercept	description	obs.	Intercept	description	obs.
BALTGROUP1	15-19 years	22	BFST1	single	81
XDWZAB1	see table 3.11	22	JBERS1	employee	110
XANZKIND1	no child	45	HAMSL0	no support	76
XEINW1	[0,1500] inhabitants	25			

covariate	description	obs.	Value	Std. Error	z	p
(Intercept)	see above		0.9647	0.309	3.1201	1.81e-03
BALTGR.2	20-24 years	31	-0.1188	0.215	-0.5513	5.81e-01
BALTGR.3	25-34 years	50	0.1749	0.212	0.8266	4.08e-01
BALTGR.4	35-44 years	66	0.4336	0.219	1.9774	4.80e-02
BALTGR.5	45-54 years	35	0.6319	0.252	2.5046	1.23e-02
BALTGR.6	$\geq$ 55	5	1.3187	0.361	3.6547	2.57e-04
BFST2	married	105	-0.0675	0.145	-0.4646	6.42e-01
BFST3	divorced/widowed	23	-0.4041	0.192	-2.1095	3.49e-02
XDWZAB2	see table 3.11	6	0.6332	0.311	2.0352	4.18e-02
XDWZAB3	see table 3.11	58	0.0743	0.192	0.3879	6.98e-01
XDWZAB4	see table 3.11	39	0.1597	0.202	0.7920	4.28e-01
XDWZAB5	see table 3.11	43	0.0142	0.206	0.0688	9.45e-01
XDWZAB6	see table 3.11	26	0.4476	0.221	2.0263	4.27e-02
XDWZAB7	see table 3.11	15	-0.2541	0.240	-1.0572	2.90e-01
JBERS2	blue-collar worker	86	0.0781	0.111	0.7029	4.82e-01
JBERS3	public servant	7	-0.8051	0.293	-2.7485	5.99e-03
JBERS4	freelancer	6	-0.6708	0.307	-2.1841	2.90e-02
XANZKIND2	1 child	64	-0.1991	0.147	-1.3534	1.76e-01
XANZKIND3	2 children	68	-0.0795	0.142	-0.5609	5.75e-01
XANZKIND4	3 or more childr.	32	-0.6125	0.185	-3.3031	9.56e-04
HAMSL1	unempl. benefit	100	-0.1689	0.119	-1.4160	1.57e-01
HAMSL3	training	10	0.6107	0.235	2.6021	9.26e-03
HAMSL2	min. fin. benefit	23	0.8044	0.187	4.3076	1.65e-05
XEINW2	[1501,3000]	41	0.3005	0.200	1.5022	1.33e-01
XEINW3	[3001,20000]	82	0.4201	0.171	2.4569	1.40e-02
XEINW4	[20001,100000]	26	0.2922	0.214	1.3636	1.73e-01
XEINW5	[100001,3Mio]	35	0.5468	0.198	2.7577	5.82e-03
Log(scale)			-0.8797	0.059	-14.9127	2.73e-50

Model Information:
Scale=$\hat{\sigma}$=0.415
Loglik(model)= -547, Loglik(intercept only)= -593.1
Chisq= -2*(-593.1)+2*(-547)=92.23 on 26 degrees of freedom, p= 2.4e-09
n= 209…number of female individuals observed in this modelling approach

Table 3.10.: log-logistic model -women

XDWZAB1:	agriculture and forestry, manufacture, energy and water supply
XDWZAB2:	civil engineering
XDWZAB3:	trade and repairs
XDWZAB4:	catering trade
XDWZAB5:	traffic and news transfer, bank and insurance, realities and service sector
XDWZAB6:	public sector and social insurance, teaching sector, other public and personal services, private household, exterritory organisations
XDWZAB7:	health, veterinary and social sector

Table 3.11.: branch of industry

Additional to the comparison of maximum log-likelihoods we further assess the goodness of fit via residual plots. Therefore we consider Cox-Snell residuals which are widely used in survival analysis. Recall from equation 2.3 and equation 2.6 that the random variable $F(T)$ is distributed uniformly on (0,1). Hence, the random variable $\Lambda(T)$ is distributed exponentially with hazard rate 1 ($\lambda = 1$).

For a given covariate combination x, when the proportional hazards model reveals good fit, the true cumulative hazard function is

$$\Lambda(T, x) = \Lambda_0(T) * exp(x^{'}\beta) \quad \sim exp(1) \tag{3.28}$$

Considering right censored data (*compare section* 2.2), for $T = min\{X, C_r\}$, the Cox-Snell residuals are defined as

$$r_C = \hat{\Lambda}_0(T) * exp(x^{'}\hat{\beta}) \tag{3.29}$$

The vector $\hat{\beta}$ contains the estimates obtained from maximizing Cox's partial likelihood (*equation* (3.35)). And the $\hat{\Lambda}_0(t)$ is the empirical Nelson-Aalen estimate of the cumulative hazard (*compare equation* (2.12)). If the chosen proportional hazards model is correct, the Cox-Snell residuals should have a unit exponential distribution. We plot the residuals (r_C against the non-parametric estimate $-ln\hat{S}(r_C)$, which is equivalent to the cumulative hazard function $\hat{\Lambda}(r_C)$ (*equation* (2.6)). If the points are close to a 45°-line through the origin, the model fits the data well. The according plot to our chosen model is given in figure 3.41. We find it to be appropriate.

Hence, the log-logistic model for women given in table 3.10 gives a good model approximation. The scale parameter of 0.415 provides a measure of the steepness of the incidence curve. Since $\hat{\sigma}$ is less than one, the risk function increases to a maximum and then declines. Thus, the hazard function has an inverted U-shaped form.

The estimated coefficients reflect the importance of the predictor variable on unemployment duration time. A positive regression parameter means that an increase in the regressor leads to an increase in duration time. The intercept β_0 indicates either a higher median duration time when it reveals a large positive value or a shorter median duration time in case of a large negative value. In our model $\beta_0 = 0.9647$. Thus the median duration time for the log-logistic distribution with all baseline variables is $exp(0.9647) = 2.624$ months. The list of the baseline variables is given in the intercept-section on top of table 3.10. The median duration time value is compared to the model dealing with the subgroup of male individuals in table 3.13. In that model $\beta_0 = 1.6873$, hence, the median duration time for the baseline covariate combination is 5.405 months. What has to be kept in mind is that the baseline is different for these two model approaches. Therefore β_0 cannot be interpreted the same way.

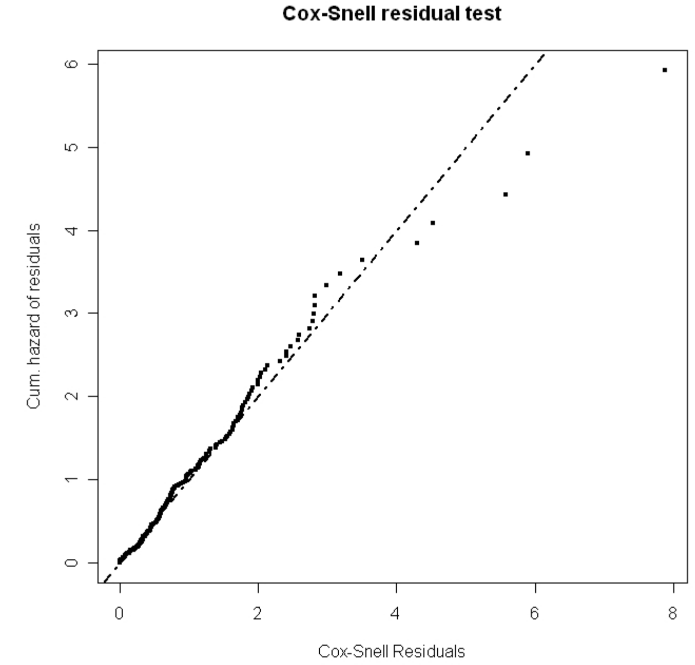


Figure 3.41.: Cox-Snell residual test -women

It is not correct to claim that men have longer median duration in general.
But first, we remain focused on the interpretation of the model for the female subgroup. As highly significant among the predictor variables appears the covariate age-group (`BALTGROUP`) of the observed women. The older a female person is, the worse are the chances of getting out of the state of unemployment. With a coefficient of 0.4336 the age group of 35-44 year old women has an odds-ratio of survival beyond time t of 2.843546 ($= (exp((1-0)*0.4336))^{\frac{1}{\sigma}}$). This means that the odds of duration beyond time t among women of age 35-44 is almost 3 times that of very young women of age less than 20 years. The odds apparently increase the higher the age of women. What is surprising is the resulting coefficient for divorced or widowed women. They seem to have lower odds (approx. one third) compared to singles. Hence they have good chances of finding employment. Having a closer look at the sample data reveals that the "duration-time-range" for 'divorced or widowed' is one to eleven month, while for single or married women the range varies from 0.25 to 48 and 63 months respectively. Therefore no long-term unemployment is detected for the group in question. Concerning the former job of the observed women, the female civil engineers as well as those who worked in public sectors show increased duration times indicated by positive regression coefficients. Furthermore, female public servants and freelancer seem to be better positioned regarding to finding work than employees. Favourable as well seems to be the fact of having children. Significant at a 5%-level appears the group of women with three or more children. The negative coefficient indicates lower odds for remaining in unemployment for this group than for childless women. Similar to the results obtained in the non-parametric analysis, the women who attend courses for vocational training have longer median duration times than those who do not get any support from the Public Employment Service Austria (AMS). The ones who do not receive regular unemployment benefit but minimum financial benefit have even

worse chances to find a job, indicated by a regression coefficient of 0.8044. What is further interesting to see is that women who live in smaller towns or villages have better chances compared to the bigger towns and cities. Those who come from Innsbruck, Salzburg, Graz, Linz or Vienna have about 4 times the odd of duration beyond time t compared to the reference group. Note in particular that the educational level does not appear to be significant for women's unemployment duration and is therefore not among the predictor variables.

Additionally as second sample in the course of parametric estimations, we consider the restriction to women who live with a partner. This allows for the inclusion of the variable 'employment status of partner', labelled `PARTNERERW`. The estimated model is given in table 3.12. Those women whose partner has left the labour force, have an odd of duration beyond time t that is about 10 times that of women with an employed partner. What appears additionally to be relevant compared to the model estimation in table 3.10 are the working hours and the reason for the last job exit. Not surprisingly willingness to work for less hours (1-14) as well and the exit due to the expiration of a fixed working contract have a positive effect. Indicated by negative coefficients women with this attributes are more likely to return faster to employment than others.

Log-logistic Model - Women (with partner) - exit: 'employment'

Intercept	description	obs.	Intercept	description	obs.
BFST1	single	23	XDWZAB1	see table 3.11	18
XANZKIND1	no child	45	DSTD3	25-34 hours/week	29
JLWI2	dismissal	56	HAMSL0	no support from AMS	53
HSART50	no application sent	34	PARTNERERW1	employed	115

covariate	description	obs.	Value	Std. Error	z	p
(Intercept)			1.1674	0.3317	3.520	4.32e-04
BFST2	married	103	0.4004	0.1651	2.425	1.53e-02
BFST3	widowed/divorced	9	-0.2839	0.2776	-1.023	3.06e-01
XDWZAB2	see table 3.11	2	0.6887	0.4763	1.446	1.48e-01
XDWZAB3	see table 3.11	37	-0.1049	0.2059	-0.509	6.11e-01
XDWZAB4	see table 3.11	20	-0.0534	0.2513	-0.212	8.32e-01
XDWZAB5	see table 3.11	30	-0.1299	0.2302	-0.564	5.72e-01
XDWZAB6	see table 3.11	18	0.5024	0.2550	1.970	4.88e-02
XDWZAB7	see table 3.11	10	-0.6292	0.2792	-2.254	2.42e-02
XANZKIND2	1 child	34	-0.0475	0.1741	-0.273	7.85e-01
XANZKIND3	2 children	42	-0.1305	0.1656	-0.788	4.31e-01
XANZKIND4	3 or more childr.	14	-0.7727	0.2419	-3.194	1.40e-03
DSTD2	15-24 hours/week	35	0.2435	0.1838	1.325	1.85e-01
DSTD4	35-44 hours/week	47	0.0507	0.1640	0.309	7.57e-01
DSTD5	$\geq$ 45 hours/week	7	0.1933	0.3324	0.582	5.61e-01
DSTD1	1-14 hours/week	17	-0.4559	0.2316	-1.969	4.90e-02
JLWI3	illness/disability	6	-0.3664	0.3120	-1.175	2.40e-01
JLWI4	end of contract	14	-0.9177	0.2184	-4.202	2.64e-05
JLWI5	caring for children/others	18	0.1562	0.2051	0.762	4.46e-01
JLWI8	resignation or composition	35	-0.0748	0.1509	-0.496	6.20e-01
JLWI7	education/other reason	7	-0.1137	0.4485	-0.254	8.00e-01
JLWI10	retirement	2	0.2111	0.4211	0.501	6.16e-01
HAMSL1	unemployment benefit	62	-0.1070	0.1430	-0.748	4.54e-01
HAMSL3	training	6	0.5907	0.2840	2.079	3.76e-02
HAMSL2	minimum fin. benefit	14	1.1849	0.2288	5.179	2.24e-07
HSART51	sent application	101	0.2978	0.1657	1.797	7.23e-02
PARTNERERW2	unemployed	8	-0.3486	0.3056	-1.141	2.54e-01
PARTNERERW3	inactive	12	0.8867	0.2138	4.148	3.36e-05
Log(scale)			-0.9618	0.0732	-13.140	1.95e-39

Model Information:
Scale=$\hat{\sigma}$=0.382
Loglik(model)= -356, Loglik(intercept only)= -392.7
Chisq= -2*(-392.7)+2*(-356)=73.49 on 27 degrees of freedom, p= 3.5e-06
n= 135...number of female individuals with partner observed in this modelling approach

Table 3.12.: log-logistic model -women with partner

The same examination procedure as for women is applied to the subgroup of male unemployed. But for the male subgroup influence of any kind from partners did not appear to be of relevance for unemployment duration. The coefficient estimates for the model suggested in this case are listed in table 3.13 on page 101. The maximum log-likelihood comparison as well as the Cox-Snell residual test (figure 3.46) confirmed the choice of a log-logistic model.

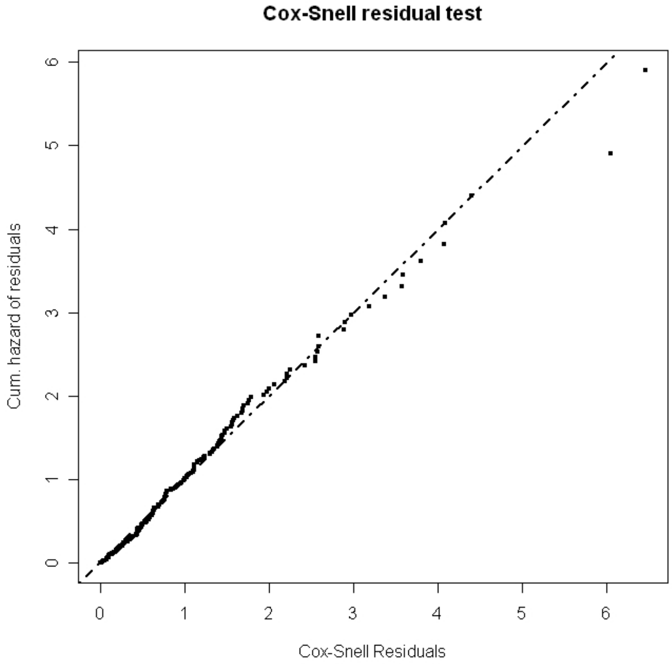


Figure 3.42.: Cox-Snell residual test -men

Other than for women, for men age does not seem to play the same crucial role on unemployment duration. The only age group appearing to be significant at a 5%-level is the 2nd one, that emcompasses the years between 20 and 24. This age group seems to be advantageous compared to the group of very young men aged under 20.

Among the branch of industry, the catering and trade sector is highly significant, showing a p-value of 0.00424. Men from this branch are more likely to leave unemployment. This might result from a great variety as well as from high fluctuation in this branch of industry.

An exceptional role of freelancers can be detected, too. This profession inherits clearly higher chances for men to leave the state of unemployment. Please recall our definition of unemployment, explained on page 34. For freelancers it seems to be easier to find work for at least a couple of hours a week. But it might also be that they find themselves in the state of 'unemployment' more often in between the states of 'employment' due to lacks of orders from clients.

Further, male parents with two children have almost half the odds of duration beyond time t than men without children.

What is among the factors of highest relevance for prediction are the support of AMS and the time the last job has come to an end, which is usually equal to the starting point of unemployment duration. Similar to the results for female persons, the unemployed men who receive minimum financial benefit show a considerable large coefficient (0.9427) indicating

Log-logistic Model - Men - exit: 'employment'

Intercept	description	obs.	Intercept	description	obs.
BALTGR.1	15-19 years	24	XANZKIND1	no child	25
XDWZAB1	see table 3.11	65	HAMSL0	no support	41
JBERS1	employee	51	START1	winter	70

covariate	description	obs.	Value	Std. Error	z	p
(Intercept)			1.6873	0.2351	7.177	7.10e-13
BALTGR.2	20-24 years	51	-0.3125	0.1586	-1.971	4.87e-02
BALTGR.3	25-34 years	38	0.0539	0.1702	0.317	7.51e-01
BALTGR.4	35-44 years	52	-0.0880	0.1620	-0.543	5.87e-01
BALTGR.5	45-54 years	32	-0.2024	0.1842	-1.099	2.72e-01
BALTGR.6	$\geq$ 55	9	0.3626	0.2884	1.257	2.09e-01
XDWZAB2	see table 3.11	42	-0.0581	0.1240	-0.469	6.39e-01
XDWZAB3	see table 3.11	29	-0.1614	0.1460	-1.105	2.69e-01
XDWZAB4	see table 3.11	17	-0.4941	0.1728	-2.860	4.24e-03
XDWZAB5	see table 3.11	40	0.1407	0.1291	1.090	2.76e-01
XDWZAB6	see table 3.11	13	0.1733	0.1759	0.985	3.25e-01
JBERS2	blue-collar worker	140	-0.1893	0.1065	-1.778	7.55e-02
JBERS3	public servant	5	-0.2516	0.3114	-0.808	4.19e-01
JBERS4	freelancer	10	-0.7003	0.2146	-3.263	1.10e-03
XANZKIND2	1 child	63	-0.1057	0.1442	-0.733	4.64e-01
XANZKIND3	2 children	70	-0.3054	0.1416	-2.157	3.10e-02
XANZKIND4	3 or more childr.	48	0.1018	0.1568	0.649	5.16e-01
HAMSL1	unempl. benefit	126	-0.1868	0.1246	-1.500	1.34e-01
HAMSL3	training	16	0.3188	0.2100	1.518	1.29e-01
HAMSL2	min. fin. benefit	23	0.9427	0.1721	5.479	4.29e-08
START2	spring	59	0.0527	0.1116	0.472	6.37e-01
START3	summer	39	0.2653	0.1385	1.916	5.54e-02
START4	autumn	38	0.4480	0.1206	3.716	2.03e-04
Log(scale)			-1.0506	0.0593	-17.731	2.41e-70

Model Information:
Scale=$\hat{\sigma}$=0.35
Loglik(model)= -491.4, Loglik(intercept only)= -549.2
Chisq= -2*(-549.2)+2*(-491.4)=115.65 on 22 degrees of freedom, p= 1.1e-14
n= 206…number of male individuals observed in this modelling approach

Table 3.13.: log-logistic model -men

longer duration times.
Also affected from longer unemployment duration are those whose last job ended in autumn or in summer. This reflects a seasonal unemployment effect.
As in the previous model for the female subgroup, in this model for man education does not seem to be an influential factor either. The reason for this insignificance probably lies in the chosen subset that considers only those individuals who had a job before getting unemployed. The previous career might show more influence than formal education.

Having excluded the group of persons without any job experience in the previous models, we now want to consider this subgroup separately in the next model approach. To maintain a sufficiently large data set, we do not further distinguish between male and female individuals. As for the circumstance of not having worked so far, mostly young people are concerned. We therefore redefine the age-groups and combine the last ones to a group of age 24 and older. What is very interesting to see is that even though mostly young people are in this subset, only two of them are childless. Moreover, more than 40% of all considered individuals experienced solely the lowest education level (38 out of 89). Due to very few entries concerning the different kind of support from AMS, we redefine the former groups to a new dichotomous variable, and label it HAMSL again. This indicator variable takes the value 0 for no support and 1 otherwise, i.e. any kind of support from the AMS. The variable of gender distinction (BSEX) has been included in the covariate testing, too. Both, however, the support of AMS and the gender specification did not turn out to be significant for this model approach. Covariates that do play a significant role are the ones listed in table 3.14. Other than in the two previous parametric models, two levels of education appear to be significant, indicated by low p-values. One of them is the vocational high school that has about a third of the odds of remaining unemployed than those with compulsory or no education. Comparing the odds of those having attended a college or university to the reference level suggests that they have odds being almost 5 times that of the other. Thus the highest educational level is also affected by long unemployment durations.
As far as the number of children is concerned, the following statement can be made. The more children, the higher the odds of staying unemployed. What joins in as predictor variable that has not been significant in the previous two model estimations is the type of job seeking. A person being in contact with the Public Employment Service Austria (AMS) has usually longer expected unemployment duration. Those who search by their own means might be more motivated and educated. Further, those who had job interviews already have higher odds to stay unemployed than those who did not have an interview.

Log-logistic Model - exit: 'employment' (no job before)

Intercept	description	obs.	Intercept	description	obs.
XANZKIND2	one child	20	XKARTAB1	no/compulsory education	38
HSART10	no contact to AMS	34	HSART80	no job interview	25

covariate	description	obs.	Value	Std. Error	z	p
(Intercept)			0.1631	0.3135	0.520	6.03e-01
XANZKIND3	2 children	42	0.4855	0.2395	2.028	4.26e-02
XANZKIND4	3 or more childr.	25	0.7838	0.2755	2.845	4.45e-03
XANZKIND1	no child	2	0.2012	0.5190	0.388	6.98e-01
XKARTAB2	apprenticeship	6	-0.5707	0.3837	-1.487	1.37e-01
XKARTAB3	second. vocational school	19	-0.0475	0.2341	-0.203	8.39e-01
XKARTAB4	high school	7	0.5240	0.3629	1.444	1.49e-01
XKARTAB5	vocational high school	9	-0.4960	0.2802	-1.770	7.67e-02
XKARTAB6	college/university	10	0.7206	0.3064	2.352	1.87e-02
HSART11	in contact with AMS	55	0.7562	0.1797	4.207	2.59e-05
HSART81	job interview	64	0.5052	0.1965	2.571	1.01e-02
Log(scale)			-0.7969	0.0898	-8.878	6.83e-19

Model Information:
Scale=$\hat{\sigma}$=0.451
Loglik(model)= -238.3, Loglik(intercept only)= -252.9
Chisq= -2*(-252.9)+2*(-238.3)=29.22 on 10 degrees of freedom, p= 0.0011
n= 89…number of male & female individuals observed in this modelling approach

Table 3.14.: log-logistic model -no job experience

3.6. Semi-parametric analysis

3.6.1. Cox-PH calculations

Having discussed parametric analysis, this section is devoted to semi-parametric models. Recall that the considerable advantage of the studied Cox model is that it is a distribution-free approach (*compare section* (2.5)). There is no necessity for a parametric form of the baseline hazard in the Cox PH model [Smith, 2002, p167]. The hazard function of the Cox PH model has the form (*compare equation* (2.19))

$$\lambda(t,x,\beta) = e^{\beta' x}\lambda_0(t) = e^{\beta_1 x^{(1)}+\beta_2 x^{(2)}+\cdots+\beta_k x^{(k)}}\lambda_0(t) = e^{\beta_1 x^{(1)}} * e^{\beta_2 x^{(2)}} * \cdots * e^{\beta_k x^{(k)}}\lambda_0(t) \quad (3.30)$$

Further recall the proportional characteristic

$$\frac{\lambda(t,x_1,\beta)}{\lambda(t,x_2,\beta)} = \frac{e^{\beta_1 x_1^{(1)}} * e^{\beta_2 x_1^{(2)}} * \cdots * e^{\beta_k x_1^{(k)}}\lambda_0(t)}{e^{\beta_1 x_2^{(1)}} * e^{\beta_2 x_2^{(2)}} * \cdots * e^{\beta_k x_2^{(k)}}\lambda_0(t)} \quad (3.31)$$

which eliminates the need for specification of the baseline hazard as $\lambda_0(t)$ cancels out in the equation. Since the respective hazard itself is regarded as chance to leave unemployment, this ratio can be interpreted as relative risk. First, the coefficient β is estimated while the estimation of the baseline hazard can be deferred ([Smith, 2002, p180]). However, we do not try to estimate the baseline hazard but concentrate on the coefficient estimation of the covariates. The CPH-function available in R having the needed procedures pre-programmed is called `coxph`. This function would also allow for time dependent covariates, but our approach deals with non time-dependent covariates. As literature about the topic of extension to time dependence, a paper of [Petersen, 1986], with the title 'Estimating Fully Parametric Hazard Rate Models with Time-dependent Covariates', shall be mentioned for background reading on this topic. The function `coxph` also incorporates the possibility of considering competing risks via cause-specific hazard functions (*compare section* 3.3.2). Moreover, the function requires no restriction on the baseline hazard $\lambda_0(t)$.

Regarding to a certain covariate combination the following assumption underlying the proportional hazards regression is made. All groups regarding to a certain covariate combination have constant relative risk over time. The hazard ratio (*see equation* (3.31)), expressed in terms of an exponential of the regression coefficients does not involve t. Thus, this assumption is called the proportional hazards property. It implies that the survivor functions of each group do not cross.

To get back to the baseline hazard $\lambda_0(t)$, note that this function is the same for all individuals. The baseline hazard is a function of time but independent of explanatory variables and observations. Therefore it does not necessarily need to be specified to obtain estimates for β. Recall that the hazard function of each group is a multiple of the baseline hazard. The resulting hazard ratio, defined above in equation (3.31), is used to measure the effect of a predictor variable on the time to departure of unemployment. A hazards ratio of one ($\frac{\lambda(t,x_1,\beta)}{\lambda(t,x_2,\beta)} = 1$) implies no effect. If the hazard ratio is less than one, the group put in the numerator has larger probability of staying in the state of unemployment at any given time, after adjusting for other covariates. The same is valid vice versa. This effect comes from the relationship between hazard and survival, which is an adverse up- and downwards movement.

A lower hazard implies higher probability of unemployment duration while a higher hazard implies lower duration probability.

For estimation, we recognize that the likelihood function cannot be fully specified in the CPH-model. For this reason, Cox defined a partial likelihood , which is based on conditional probabilities, free of the baseline hazard.
Let x_j denote the vector of covariates of an individual departing from unemployment at time t_j. And let r_j be the number of individuals that are still unemployed and not censored just before t_j. Finally, if we have k distinct ordered exit times, the partial likelihood $L_C(\beta)$ is defined as a product over these k uncensored departure times. Let $L_j(\beta)$ denote the following conditional probability:

$$L_j(\beta) = \frac{P(\text{individual with } x_j \text{ departs at } t_j| \text{ individual is in } r_j)}{P(\text{one departure at } t_j| \ r_j)} \tag{3.32}$$

$$L_j(\beta) = \frac{exp(x_j^{'}\beta)\lambda_0(t_j)}{\sum_{l \in r_j} exp(x_l^{'}\beta)\lambda_0(t_j)} \tag{3.33}$$

Therefrom the partial likelihood function is derived as

$$L_C(\beta) = \prod_{j=1}^{k} L_j(\beta) = \prod_{j=1}^{k} \frac{exp(x_j^{'}\beta)}{\sum_{l \in r_j} exp(x_l^{'}\beta)} \tag{3.34}$$

Including censoring via the indicators $\delta_1, \ldots, \delta_n$, being one if uncensored and zero otherwise, the partial likelihood function for all n observed times can also be expressed as

$$L_C(\beta) = \prod_{j=1}^{k} \left(\frac{exp(x_j^{'}\beta)}{\sum_{l \in r_j} exp(x_l^{'}\beta)} \right)^{\delta_i} \tag{3.35}$$

Dealing with competing risks (*compare section* (3.3)) we perform two different types of analysis. In the course of this section we model a cause-specific hazard regression as well as competing risks regression on subdistribution hazards.

3.6.2. Cause-specific CPH-model (CS-CPH)

We start with a cause-specific hazards model. A short introduction on this issue has been given on page 40. The selection procedure to define an appropriate model has been performed in similar manner to the one in the parametric approach. Again we estimate distinct models for the subgroups of men and women. The selection process has been made via the AIC criterion as well as the LRT-test. A short description of both methods has been given on page 90.

Recall for the following interpretations that the smaller the hazard rate, the risk of leaving the state in question, the larger the probability of staying in this state of being unemployed. The exponentiated coefficients in the output of the estimated CPH-models are interpretable as multiplicative effects on the hazard. We test the null hypothesis that all coefficients are zero. A positive coefficient in the model represents a risk increase and therefore a reduction in the expected time of unemployment duration. Adverse effect is given for negative coefficients.

We use the cause-specific hazard to examine the effect of predictor variables on the exit rate of unemployment due to a particular cause, conditional on not having exited from any other cause before.

Our first model for women, restricted to those with job experience, is given in table 3.15. Variables concerning the job a person got after the unemployment period are not considered in this model as they would reduce the subset of data to solely uncensored observations. Excluded from the initial covariate list for the examination are therefore the covariates 'working hours per week' (`DSTD`) and 'branch of industry' (`XDWZAB`).

A test for the proportional hazard assumption (according to Grampsch and Therneau) shows that the proportional hazards assumption is valid for all covariates except for the support of AMS (`HAMSL`). Hence, in a study comparing no support to unemployment benefit, we would expect the hazard ratio to vary over time. Model extensions that account for this issue are as follows: Application of different CPH-models for different time periods, extended Cox models that consider time-dependent variables, quantile regression approaches and stratification approaches. Stratification for example would lead to estimation of separate baseline hazards for each level of the stratified covariate. Applying such a model, allows to observe the obtained survival curves for the different levels. As those curves turned out to be close to being proportional, we decide to use the CPH model without stratification. A disadvantage when stratifying would also be that we could not obtain the estimated coefficient of the categorical variable effect of 'support of AMS'. Therefore we choose the simplified model without stratification and apply the cause-specific proportional hazards model.
The goodness of fit of the estimated model is validated by a Cox-Snell residual plot given in figure 3.43. A definition for Cox-Snell residuals has been given on page 96. The base case of our model is assumed to be women of the middle age-group (34-44 years), living in Burgenland, who have an employed partner and have neither sent applications nor done job interviews by the time of observation.

The corresponding survival function along with the 95% confidence bands represented by broken lines is displayed in figure 3.44. The cumulative hazard together with a smoothing line for the model in table 3.15 is plotted in figure 3.45.

Apparently, older women of age 45 plus are at a disadvantage in the labour market. Finding work in our rapidly changing economic environment appears to be more difficult for them. With other covariates fixed, older women have a smaller hazard, and, hence, have longer expected duration time than those of lower age groups. Decreased hazard is also given for women living in Vienna. Recall that in the parametric model estimation on page 95 we detected already that living in bigger cities has negative influence on the length of unemployment spells. This is consistent with the derived statement from this model of decreased chances for those living in Vienna.
Further, the receipt of unemployment benefit appears to be highly significant and the large corresponding coefficient implies higher probability of leaving unemployment. As previously mentioned, this estimate has to be considered with caution as the receipt of benefits might be correlated with time. Long-term unemployed are more likely to receive minimum financial

CS-CPH model- Women (with partner) -> exit: 'employment'

Baseline	description	obs.	Baseline	description	obs.
BALTGR.3	35-44 years	218	XNUTS211	Burgenland	71
HAMSL0	no support from AMS	265	PARTNERERW1	employed	505
HSART50	no application sent	244	HSART80	no job interview	291

covariate	description	obs.	coef	exp(coef)	se(coef)	z	p
BALTGR.1	15-24 years	49	0.42102	1.524	0.329	1.28079	2.0e-01
BALTGR.2	25-34 years	194	0.04364	1.045	0.225	0.19353	8.5e-01
BALTGR.4	45-54 years	132	-0.60705	0.545	0.248	-2.45109	1.4e-02
BALTGR.5	$\geq$ 55	30	-0.39326	0.675	0.557	-0.70662	4.8e-01
XNUTS212	Lower Austria	68	-0.02809	0.972	0.375	-0.07488	9.4e-01
XNUTS213	Vienna	83	-0.76659	0.465	0.404	-1.89694	5.8e-02
XNUTS221	Carinthia	77	0.60524	1.832	0.337	1.79467	7.3e-02
XNUTS222	Styria	67	0.19044	1.210	0.384	0.49651	6.2e-01
XNUTS231	Upper Austria	73	0.48914	1.631	0.367	1.33212	1.8e-01
XNUTS232	Salzburg	45	0.26399	1.302	0.412	0.64035	5.2e-01
XNUTS233	Tirol	56	0.62574	1.870	0.401	1.56130	1.2e-01
XNUTS234	Vorarlberg	83	-0.00222	0.998	0.346	-0.00641	9.9e-01
HAMSL1	unemploym. benefit	239	1.04473	2.843	0.218	4.79065	1.7e-06
HAMSL3	training	36	-0.28641	0.751	0.447	-0.64050	5.2e-01
HAMSL2	minim. fin. benefit	83	-0.58178	0.559	0.317	-1.83320	6.7e-02
PART.ERW2	employed	47	-0.19372	0.824	0.390	-0.49718	6.2e-01
PART.ERW3	inactive	71	-0.88707	0.412	0.396	-2.23984	2.5e-02
HSART51	sent application	379	0.47562	1.609	0.219	2.16962	3.0e-02
HSART81	job interview	332	0.40623	1.501	0.202	2.01318	4.4e-02

Model Information:
Rsquare= 0.136 (max possible= 0.902)
(Remember, the R^2-value can be interpreted as the proportion of explained variance)
Likelihood ratio test= 91.3, p=1.96e-11
n= 623 . . . number of female individuals observed in this modelling approach

Table 3.15.: cause-specific CPH model -women (with partner)

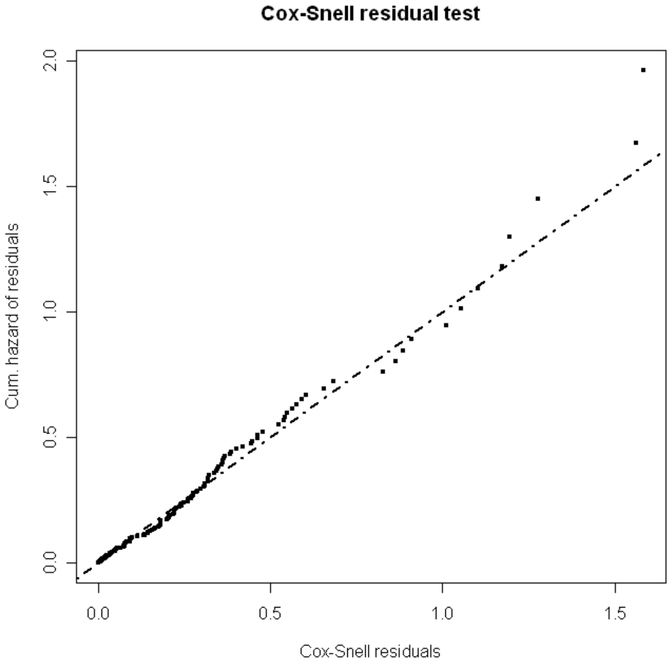


Figure 3.43.: Cox-Snell residual test -women

benefits than regular unemployment benefit.

What can further be observed from the model is that for women an inactive partner reduces the probability of early reemployment. And not surprisingly active search via applications and interviews lead to increased hazards, hence, shorter expected unemployment duration.

Note, all variables that do not appear in the list of covariates for this model but have been in the initially defined covariable set (table 3.9), have turned out to be not significant according to the LRT-test with a 5%-threshold.

Recall that for the first consideration of cause-specific hazards focus has been given to women living with partners. Next, we consider an extended model, leaving the restriction of a partnership aside. The corresponding model estimation is presented in table 3.16.

As in the previous model, we see similar negative, prolonging effect on unemployment duration when women are of higher age.

As far as the educational level is concerned only apprenticeship and vocational high school attendants appear significant in this model approach, indicated by low p-values. Woman who have entered the labour market after an apprenticeship are more likely to leave unemployment to employment.

CS-CPH model - Women - exit: 'employment'

Baseline	description	obs.	Baseline	description	obs.
BALTGR.3	35-44 years	280	XKARTAB1	no/compulsory education	322
XEINW3	[3001,20000] inhabitants	328	HAMSL0	no support from AMS	390
HSART50	no application sent	338	HSART80	no job interview	426
JLWI2	dismissal	333	JTAET1	apprentice/auxiliary work	194

covariate	description	obs.	coef	exp(coef)	se(coef)	z	p
BALTGR.1	15-24 years	215	0.5911	1.806	0.202	2.9275	0.00340
BALTGR.2	25-34 years	264	0.2457	1.279	0.200	1.2311	0.22000
BALTGR.4	45-54 years	169	-0.4415	0.643	0.216	-2.0487	0.04000
BALTGR.5	≥ 55	33	-0.7695	0.463	0.480	-1.6039	0.11000
XKARTAB2	apprenticeship	330	0.3729	1.452	0.190	1.9657	0.04900
XKARTAB3	second. vocation. school	142	0.2466	1.280	0.228	1.0795	0.28000
XKARTAB4	high school	55	-0.3010	0.740	0.362	-0.8319	0.41000
XKARTAB5	vocational high school	62	-1.1853	0.306	0.537	-2.2081	0.02700
XKARTAB6	college/university	50	-0.0246	0.976	0.373	-0.0659	0.95000
XEINW1	[0,1500]	124	-0.1044	0.901	0.238	-0.4396	0.66000
XEINW2	[1501,3000]	203	-0.3403	0.712	0.199	-1.7132	0.08700
XEINW4	[20001,100000]	105	0.0477	1.049	0.233	0.2050	0.84000
XEINW5	[100001,3Mio]	201	-0.5223	0.593	0.211	-2.4750	0.01300
HAMSL1	unemploym. benefit	381	0.6156	1.851	0.173	3.5676	0.00036
HAMSL3	training	58	-0.7271	0.483	0.347	-2.0979	0.03600
HAMSL2	minim. fin. benefit	132	-0.9088	0.403	0.247	-3.6849	0.00023
HSART51	sent application	623	0.4613	1.586	0.185	2.4890	0.01300
HSART81	job interview	535	0.4084	1.504	0.162	2.5269	0.01200
JLWI1	retirement	29	-0.6261	0.535	0.615	-1.0181	0.31000
JLWI3	illness/disability	46	-0.5930	0.553	0.415	-1.4300	0.15000
JLWI4	end of contract	119	0.6643	1.943	0.224	2.9692	0.00300
JLWI5	caring f. children/others	148	-0.5919	0.553	0.276	-2.1420	0.03200
JLWI6	education/other reason	67	-0.2429	0.784	0.346	-0.7016	0.48000
JLWI7	resignation/agreement	219	0.1367	1.147	0.184	0.7413	0.46000
JTAET3	manual employee	282	0.5243	1.689	0.220	2.3856	0.01700
JTAET4	high profess. manual	122	0.2153	1.240	0.272	0.7920	0.43000
JTAET5	med. prof. non man.	25	-1.0769	0.341	0.744	-1.4479	0.15000
JTAET6	high prof. non man.	298	0.2503	1.284	0.243	1.0285	0.30000
JTAET7	farmer/freelancer	40	0.8706	2.388	0.409	2.1283	0.03300

Model Information:
Rsquare= 0.148 (max possible= 0.922)
Likelihood ratio test= 154, p=0
n= 961 …number of female individuals observed in this modelling approach

Table 3.16.: CS-CPH model -women

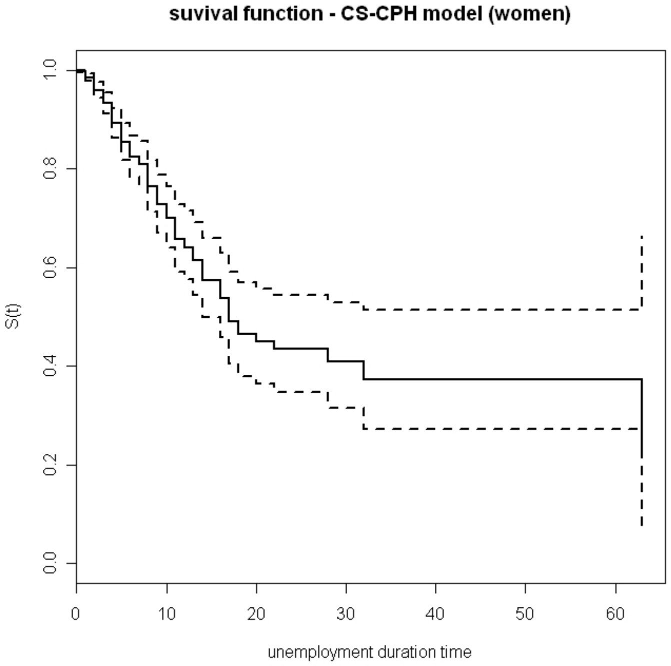


Figure 3.44.: survival function (women)

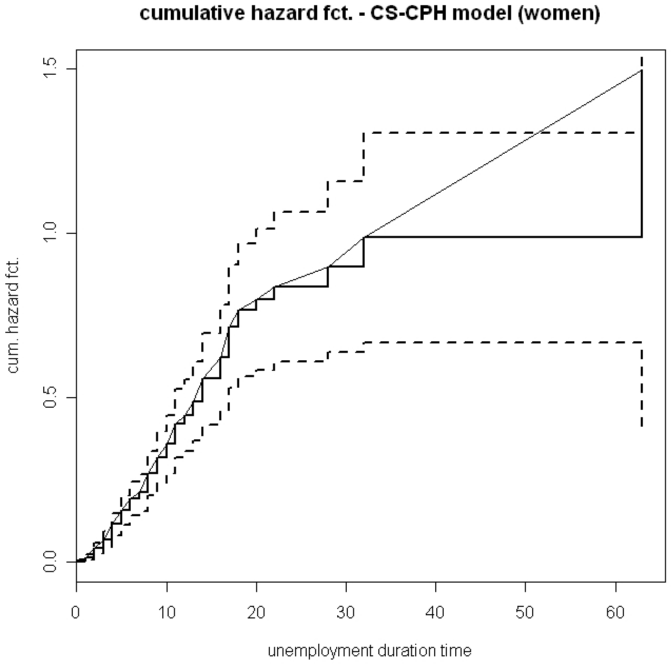


Figure 3.45.: cumulative hazard function (women)

For men the employment status of their partners did not turn out to be relevant. The corresponding covariate is therefore left aside in the model (table 3.17) in order to gain a larger subset for the estimation.
Different from the parametric model suggestion on page 3.13, the last two age-groups appear to be significant and negatively correlated to the duration of unemployment spells. Older men have a smaller hazard rate to employment than both prime-age and younger workers. Health problems or some other attributes may lead to reduced job search intensity or discouragement in this age groups.
As expected, support of the AMS does play a significant role as predictor variable. Those receiving regular unemployment benefits have a larger hazard rate to employment than individuals who do not get this kind of financial support. Conversely, men who receive minimum financial benefits show reduced hazard rates.
Further interpretation that can be made is that men being parent of two children have the best chances to leave the state of unemployment compared to others. This suggests that social and family life are determinants of unemployment duration as well.
Moreover, for men, the starting season of unemployment seems to be important for the determining of unemployment duration before reemployment.

Again, for this model estimation, a Cox-Snell residual test as well as the corresponding survival function and cumulative hazard function are plotted in figures 3.46 - figure 3.48.

cause-specific CPH-model - Men - exit cause: ’employment’

Baseline	description	obs.	Baseline	description	obs.
BALTGR.3	35-44 years	170	XANZKIND3	2 children	214
HAMSL0	no support from AMS	177	HSART100	no contact to private empl. agency (PEA)	629
START1	winter	214			

covariate	description	obs.	coef	exp(coef)	se(coef)	z	p
BALTGR.1	15-24 years	239	0.303	1.354	0.189	1.602	1.1e-01
BALTGR.2	25-34 years	153	-0.297	0.743	0.217	-1.369	1.7e-01
BALTGR.4	45-54 years	133	-0.515	0.598	0.231	-2.227	2.6e-02
BALTGR.5	$\geq$ 55	69	-1.334	0.263	0.384	-3.475	5.1e-04
XANZKIND1	no child	133	-0.568	0.567	0.250	-2.270	2.3e-02
XANZKIND2	one child	258	-0.478	0.620	0.179	-2.667	7.7e-03
XANZKIND4	$\geq$ 3 childr.	159	-0.346	0.707	0.191	-1.809	7.0e-02
HAMSL11	unemploym. benefit	395	0.754	2.126	0.186	4.054	5.0e-05
HAMSL13	training	41	0.152	1.164	0.329	0.461	6.4e-01
HAMSL12	minim. fin. benefit	151	-1.024	0.359	0.261	-3.924	8.7e-05
HSART101	contact to PEA	135	0.424	1.529	0.182	2.335	2.0e-02
START2	spring	161	0.266	1.305	0.180	1.475	1.4e-01
START3	summer	185	-0.509	0.601	0.206	-2.477	1.3e-02
START4	autumn	204	-0.556	0.573	0.204	-2.722	6.5e-03

Model Information:
Rsquare= 0.182 (max possible= 0.956)
Likelihood ratio test= 154, p=0
n= 764 . . . number of male individuals observed in this modelling approach

Table 3.17.: CS-CPH model -men

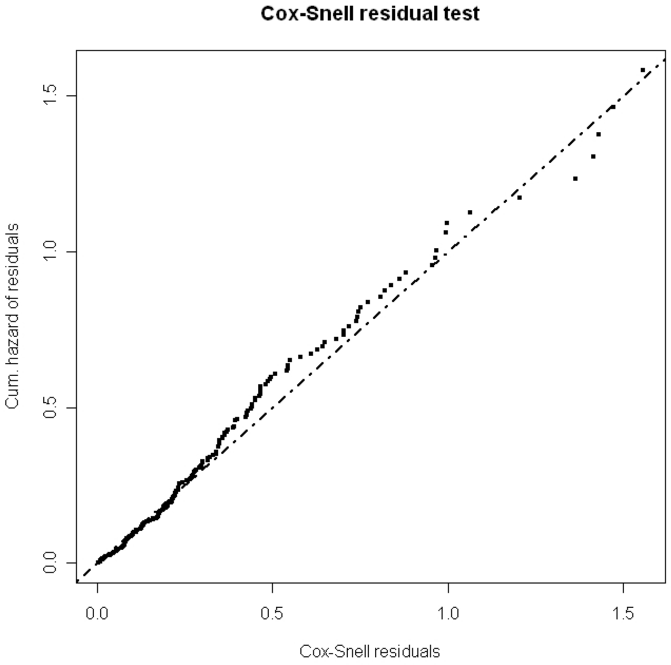


Figure 3.46.: Cox-Snell residual test -men

Considering the exit cause 'inactivity', i.e. leaving the labour force, two separate cause-specific hazard models are estimated. One for women with a subgroup restricted to those with partners and one for men without this restriction. For both models individuals, who had been in a work relationship before unemployment, are examined. The respective tables for these models are table 3.18 and table 3.18.

Turning to this different exit cause, several primary significant covariates emerge. As still of importance for determining unemployment duration appears the kind of support or whether or not an individual gets support from the AMS. The results obtained in both examined model distinctions imply that the existing unemployment benefit system is not contributing to longer unemployment spells.
For women the employment status of a partner stays relevant. Inactive or unemployed partners decrease the hazard.
Two covariates ('application sent' and 'job interview'), partly reflecting the search intensity, are significant and have negative coefficient estimates. Therefore, unsurprisingly, active search reduces the probability to exit to an inactive working life.
For men the type of profession seems to play a significant role. The combined classes of apprenticeship and auxiliary work show lower expected unemployment duration times before exit to 'inactivity' compared to all other groups.

cause-specific CPH-model - Women (with partner) - exit: 'inactivity'

Baseline	description	obs.	Baseline	description	obs.
HAMSL0	no support from AMS	265	PARTNERERW1	employed	505
HSART50	no application sent	244	HSART80	no job interview	291

covariate	description	obs.	coef	exp(coef)	se(coef)	z	p
HAMSL1	unemploym. benefit	239	0.434	1.544	0.200	2.167	0.03000
HAMSL3	training	36	-0.283	0.753	0.411	-0.689	0.49000
HAMSL2	minim. fin. benefit	83	-0.653	0.520	0.282	-2.319	0.02000
PARTNERERW2	unemployed	47	-0.681	0.506	0.367	-1.856	0.06300
PARTNERERW3	inactive	71	-0.629	0.533	0.257	-2.446	0.01400
HSART51	sent application	379	-0.647	0.523	0.185	-3.507	0.00045
HSART81	job interview	332	-0.457	0.633	0.187	-2.444	0.01500

Model Information:
Rsquare= 0.068 (max possible= 0.908)
Likelihood ratio test= 43.7, p=2.46e-07
n= 623 ...number of women individuals observed in this modelling approach

Table 3.18.: CS-CPH model -women (exit: 'inactivity')

cause-specific CPH-model - Men - exit cause: 'inactivity'

Baseline	description	obs.	Baseline	description	obs.
HAMSL0	no support from AMS	177	JTAET1	apprentice/auxiliary work	189

covariate	description	obs.	coef	exp(coef)	se(coef)	z	p
HAMSL1	unemploym. benefit	395	0.0064	1.006	0.232	0.0277	0.98000
HAMSL3	training	41	-0.6270	0.534	0.449	-1.3961	0.16000
HAMSL2	minim. fin. benefit	151	-0.9513	0.386	0.274	-3.4751	0.00051
JTAET3	manual employee	201	-0.7769	0.460	0.263	-2.9488	0.00320
JTAET4	high profess. manual	187	-0.4542	0.635	0.247	-1.8354	0.06600
JTAET5	med. prof. non man.	91	-0.7260	0.484	0.329	-2.2073	0.02700
JTAET6	high prof. non man.	51	-0.2675	0.765	0.369	-0.7255	0.47000
JTAET7	farmer/freelancer	45	-0.7880	0.455	0.426	-1.8480	0.06500

Model Information:
Rsquare= 0.039 (max possible= 0.826)
Likelihood ratio test= 30.6, p=0.000162
n= 764 ...number of men individuals observed in this modelling approach

Table 3.19.: CS-CPH model -men (exit: 'inactivity')

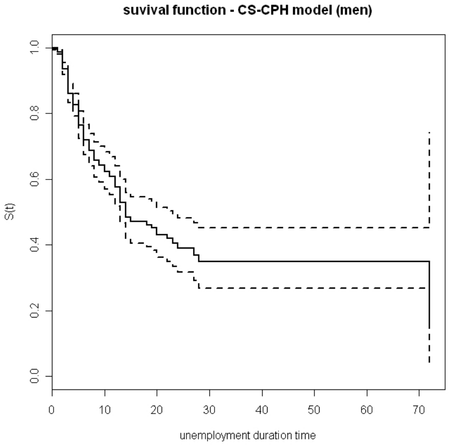


Figure 3.47.: survival function (men)

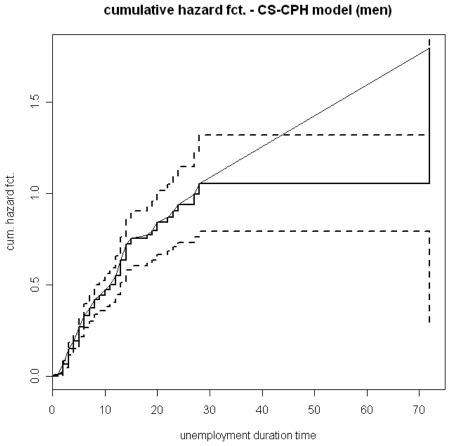


Figure 3.48.: cumulative hazard function (men)

3.6.3. Cumulative incidence regression

Finally for comparison, we regress on cumulative incidence functions. Recall that other than in cause-specific hazard estimation, the CIF for cause 2 does not only depend on the hazard of this cause 2 ('employment'), but also on hazards of cause 3 ('inactivity').

The difference to the cause-specific hazard is that individuals who depart from other cause than the one of interest remain in the risk set.

The related subdistribution hazard is given as

$$\lambda_k(t) = -\frac{dln(1 - I_k(t))}{dt} \qquad for\ k = 2, 3$$

(a *definition for* $I_k(t)$ *has been given in equation* (3.5))
[Putter et al., 2007, p11]

For modelling the hazard of subdistributions the partial likelihood

$$L_C(\beta) = \prod_{j=1}^{k} L_j(\beta) = \prod_{j=1}^{k} \frac{exp(x_j^{'}\beta)}{\sum_{l \in R_j} w_{jl} exp(x_l^{'}\beta)} \tag{3.36}$$

is maximised. [Fine and Gray, 1999]

There are two alterations in this formula compared to the partial likelhood function of Cox (3.36). Weights are included in the denominator (w_{jl}) and the risk set is defined differently. As mentioned above, the observation according to the competing event remains in the risk set (R_j) at all times. Hence, the risk set is divided into the following groups. First, the individuals who participate fully in the partial likelihood with weight $w_{jl} = 1$ where $t_l \geq t_j$. Second, those who experienced a competing event before t_j, thus $t_l \leq t_j$. This group is given a weight of less than one, becoming smaller than one the further t_l is from t_j. [Pintilie, 2006, p1365] Details on the topic of weight calculations can be found in [Fine and Gray, 1999]. A function available to apply the competing risk regression in R is named `crr`. This function is contained in the `cmprsk` package. It returns estimated coefficients along with their standard errors and the according two-sided p-value.

The division into female and male subgroups is done the same way as in the previous cause-specific observation. But other than before we consider solely dichotomous or numerical variables as initial covariate set. E.g. marital status has not been included as categorical variable with three classifications, but each class, 'married', 'single' or 'divorced or widowed' is included as separate dummy variable. As selection procedure to find all the significant predictor variables we choose a backward process, the 'p-value-method'. The initial variable choice has been included in the first model, then the variables with the highest p-values, i.e. the least significant variables, have been successively removed. To verify that the previously excluded variables do not add a significant contribution to the model, each covariate has been re-added separately at the end. The inclusion and exclusion of variables has been performed according to a 'p-value-threshold' of 0.05. The resulting final model estimations are presented in table 3.20 - 3.23. The considered subgroups consist of 623 and 764 observations respectively.

The estimated subdistribution hazards ratio (SDHR) is calculated as the exponential of the estimated coefficients. Women who receive unemployment benefit have a SDHR of $2.6491 = exp(0.97421)$. They have 2.6 times the chance (or hazard) of ending their unemployment duration of those who do not receive this financial support. An age over 40 implies lower chances for women to get reemployed. Additionally negative influence factors are caring for children or other persons in need, living in Vienna, not being native or if a women has worked as a high professional doing non manual work. Positive effect on the duration for exit cause 'employment' is revealed for active job search. Opposite effect is shown when exit cause 'inactivity' is considered for the corresponding job search covariates.

For men once more the reduced chances for leaving unemployment for elder unemployed is evident. Further, having a family with two or more children seems to increase the chance of early reemployment. A former type of profession of manual work in a middle or higher position also seems to come along with increased chances of finding a job.
For exit by the cause of leaving the labour force, dismissal from the last job appears to increase the subsequent expected unemployment duration. As for women, intensive job search has a prolonging effect for the stay in the state of unemployment. Adverse effect, hence, increased risk to exit from unemployment, have empoyees, auxiliary workers or apprentices.
All in all the results appear to be consistent with previously estimated models.

CRR - Women (with partner) - exit: 'employment'

covariate	description	obs.	coef	std. errors	two-sided p-val.
XBGEBLAO	not borne in Austria	184	-0.40220	0.1992	0.04400
BALTGROUP	≥40 years	248	-0.45900	0.1767	0.00940
HOCHNM	high profession -non manual	46	-0.80280	0.4806	0.09500
PFLEGE	caring for children/others	115	-0.63770	0.2544	0.01200
XEINW	inhabitants	623	0.05217	0.0305	0.08700
KAERNTEN	Carinthia	77	0.36040	0.2155	0.09500
WIEN	Vienna	83	-1.19200	0.3768	0.00160
ALG	unemployment benefit	514	0.97421	0.2578	0.00016
PARTERW1	employed partner	505	0.46280	0.2443	0.05800
HSART1	in contact with AMS	462	0.66080	0.2415	0.00620
HSART8	job interview	332	0.55060	0.1950	0.00470
HSART5	application sent	379	0.56680	0.2173	0.00910

Table 3.20.: CRR-model -women ->'employment'

CRR - Women - exit: 'inactivity'

covariate	description	obs.	coef	std. errors	two-sided p-val.
PENSION	retirement	19	0.6899	0.3567	0.05300
PFLEGE	caring for children/others	115	0.5141	0.1889	0.00650
NOTSTAND	minimum fin. benefit	89	-0.6476	0.2582	0.01200
HSART8	job interview	332	-0.5498	0.1802	0.00230
HSART5	application sent	379	-0.6585	0.1745	0.00016
START	summer,autumn	340	-0.3731	0.1665	0.02500

Table 3.21.: CRR-model -women ->'inactivity'

CRR - Men - exit: 'employment'

covariate	description	obs.	coef	std. errors	two-sided p-values
XANZKIND	≥2 children	373	0.4207	0.1435	3.4e-03
BALTGROUP	≥40 years	269	-0.3314	0.1513	2.8e-02
ANGM	manual employee	201	0.4923	0.1609	2.2e-03
WIEN	Vienna	145	-0.4658	0.2062	2.4e-02
ALG	unemployment benefit	586	1.2070	0.1809	2.5e-11
START	summer,autumn	389	-0.4785	0.1399	6.2e-04
HSART10	contact with PEA	135	0.3475	0.1822	5.7e-02
PFLSCH	compulsory/no education	232	-0.2999	0.1623	6.5e-02
ABLAUF	end of contract	91	0.3624	0.1794	4.3e-02
SCHULUNG	vocational training	47	0.4306	0.2174	4.8e-02
HOCHM	high professional -manual	187	0.2881	0.1725	9.5e-02

Table 3.22.: CRR-model -men ->'employment'

CRR - Men - exit: 'inactivity'

covariate	description	obs.	coef	std. errors	two-sided p-values
ANG	employee	215	0.3935	0.2080	5.9e-02
WORK	apprentice &auxiliary work	189	0.7785	0.1938	5.9e-05
KUEND	dismissal	314	-0.3977	0.1873	3.4e-02
HSART1	in contact with AMS	660	-0.6464	0.2144	2.6e-03
HSART3	newspaper job search	674	-0.5486	0.2499	2.8e-02

Table 3.23.: CRR-model -men ->'inactivity'

4. Conclusion

A non-parametric, a parametric and a semi-parametric estimation method for Austrian unemployment data is designed to incorporate the available information about predictor variables on unemployment duration. The non-parametric estimation is a fast and powerful tool for data exploration that provides good information about the basic structure and influential factors. Results therefrom are subsequently used in the parametric analysis, determining initial covariate selection for the parametric model estimation. Given the fact of different exit reasons from the state of unemployment we discriminate between two mutually exclusive causes. Firstly and the one of main interest, being the reemployment of an individual. Secondly and as a competing risk, the exit of an observed individual from the labour force. The strong interest for this case distinction is verified by an estimation of the corresponding cumulative hazard functions. They appear to be of different shape and therefore suggest treatment by different models. To account for the importance of competing risks we consider a non-parametric estimation of cumulative incidence in the first part of our data analysis. In the parametric approach a log-logistic model for our case of interest is fitted to uncensored data. And in the semi-parametric section, models based on cause-specific and subdistribution hazards are estimated. The log-logistic model allows for good interpretation of the significant coefficients of predictor variables. A model for female and male subgroups has been estimated in order that we take the interaction between gender and other covariates into account. For women, the employment status of their partner seems to be of relevance while for men influences attributed by partners appears not significant. To compare the results of the log-logistic models to others and to also consider censored data, we complete the analysis with semi-parametric modelling. From the estimates of the various model approaches one can clearly observe the different influences of the covariates. The result of the effect of age on unemployment duration is in sound agreement with common perceptions whereby elderly persons face more difficulties of reemployment than younger ones. Receipt of regular unemployment benefit also appears as highly significant as this kind of financial support increases the chances to leave the state of unemployment. Furthermore, a seasonal effect is detected in the models for unemployed men as it seems to be harder for them to get work during the winter months than in spring or summer. Except for the lowest educational level, the level of education does not seem to affect unemployment duration in a substantial way when reemployment is observed. The educational level, however, seems to be more relevant for people who enter their first employment relationship. In this case vocational high school attendants seem to have good chances for employment. The consideration of competing risks has been found to be very important. It is emphasised by the fact that different issues appear to influence unemployment duration very differently depending on the cause of leaving unemployment. Furthermore, reversed effects on unemployment duration are detected for covariates retaining information about job search intensity. Hence, intensive job search helps to find a job quicker but also extends the time spent unemployed if departure to an inactive working life follows.

4.1. Outlook

Having considered the covariates being independent of time an extension to our approach could be to allow and test for time-dependence. In a basic test (Grambsch and Therneau's test for PH assumption) carried out we found some evidence that the financial support of the Public Employment Agency is likely to be time-dependent. This seems to be a plausible assumption for further modelling approaches. Extended CPH models or quantile regression models are two of such possible model extensions.
In our analysis focus is given on reemployment, thus on people with job experience, differentiated by gender. This research could be continued by concentrating on other target groups or model assumptions. A combined data set of men and women together with inclusion of specific interaction terms accounting for the gender differences could be chosen as sample base for other models.
Finally the application of the Cox-Snell residual test could be amplified by other methods of model checks. Other residuals used for data diagnostics in duration analysis are for example Martingale, Deviance or Schönfeld residuals.

Bibliography

[Andersen et al., 1993] Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer.

[Bagdonavičius and Nikulin, 2002] Bagdonavičius, V. and Nikulin, M. (2002). *Accelerated life models : modeling and statistical analysis.* Chapman and Hall/CRC Press LLC.

[Chalita et al., 2002] Chalita, L. V. A. S., Colosimo, E. A., and Demétrio, C. G. B. (2002). *Likelihood approximations and discrete models for tied survival data*, volume Vol. 31, Issue 7. Journal Communications in Statistics - Theory and Methods.

[Crowley and Johnson, 1981] Crowley, J. and Johnson, R. A. (1981). *Survival Analysis*, volume II. Lecture notes-monograph series.

[Fine and Gray, 1999] Fine, J. P. and Gray, R. J. (1999). *A Proportional Hazards Model for the Subdistribution of a Competing Risk*, volume Vol. 94, No. 446. Journal of the American Statistical Association.

[Gray, 1988] Gray, R. J. (1988). *A Class of k-Sample Tests for Comparting the Cumulative Incidence of a Competing Risk*, volume Vol. XXVI. The Annals of Statistics.

[Greene, 2003] Greene, W. H. (2003). *Econometric Analysis*, chapter "'Models for duration data"', pages 790–801. Prentice Hall, 5th edition.

[Hashem Pesaran and Schmidt, 1999] Hashem Pesaran, M. and Schmidt, P. (1997, 1999). *Handbook of Applied Econometrics*, volume II: Microeconomics, chapter "'Search Models and Duration Data"', pages 300–351. Blackwell Publishers Ltd.

[Heckman and Leamer, 2001] Heckman, J. and Leamer, E. (2001). *Handbook of Econometrics*, volume Vol. 5, chapter "Duration Models: Specification, Identification and Multiple Durations", pages 3381–3460. Elsevier.

[Jensen and Svarer, 2003] Jensen, P. and Svarer, M. (2003). *Short- and long-term unemployment: How do temporary layoffs affect this distinction?*, volume Vol. 28. Empirical Economics.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). *Nonparametric Estimation from Incomplete Observations*, volume Vol. 53, No. 282. Journal of the American Statistical Association.

[Kauermann and Khomski, 2006] Kauermann, G. and Khomski, P. (2006). *Full Time or Part Time Reemployment: A Competing Risk Model with Frailties and Smooth Effects using a Penalty based Approach.* University of Bielefeld.

[Kavkler and Borsic, 2006] Kavkler, A. and Borsic, D. (2006). *The Main Characteristics of the Unemployed in Slovenia.* Original Scientific Papers.

[Kiefer, 1988] Kiefer, N. M. (1988). *Economic Duration Data and Hazard Functions*, volume Vol. XXVI. Journal of Economic Literature.

[Kuha, 2004] Kuha, J. (2004). *AIC and BIC: Comparisons of Assumptions and Performance*, volume Vol. 33, No. 2. Sociological Methods and Research.

[Lancaster, 1994] Lancaster, T. (1990, 1992, 1994). *The Econometric Analysis of Transition Data.* Cambridge University Press.

[Mario Cleves and Gutierrez, 2004] Mario Cleves, W. W. G. and Gutierrez, R. G. (2004). *An Introduction to Survival Analysis Using Stata - Revised Edition*, chapter "'The problem of survival analysis"', pages 1–6. Stata Press (StataCorp LP), Texas.

[McCall, 1997] McCall, B. P. (1997). *The Determinants of Full-Time versus Part-Time Reemployment following Job Displacement*, volume Vol. 15, No. 2. Journal of Labor Economics.

[Nikulin et al., 2004] Nikulin, M. S., Balakrishnan, N., Mesbah, M., and N., L. (2004). *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*, chapter "Diagnostics for Cox's Proportional Hazards Model", pages 27–38. Birkhäuser.

[Owen, 2001] Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall/CRC Press LLC.

[Petersen, 1986] Petersen, T. (1986). *Estimating Fully Parametric Hazard Rate Models with Time-Dependent Covariates: Use of Maximum Likelihood*, volume Vol. 14, No. 3. Sociological Methods and Research.

[Pintilie, 2006] Pintilie, M. (2006). *Analysing and interpreting competing risk data*, volume Vol. 26. STATISTICS IN MEDICINE - Published online in Wiley InterScience (www.interscience.wiley.com).

[Putter et al., 2007] Putter, H., Fiocco, M., and Geskus, R. B. (2007). *Competing risks and multi-state models*, volume Vol. 26, No. 11. STATISTICS IN MEDICINE - Published online in Wiley InterScience (www.interscience.wiley.com).

[Smith, 2002] Smith, P. J. (2002). *Analysis of Failure and Survival Data.* Chapman and Hall/CRC Press LLC.

[Tableman and Kim, 2004] Tableman, M. and Kim, J. S. (2004). *Survival Analysis Using S.* Chapman and Hall/CRC Press LLC.

[Wichert and Wilke, 2007] Wichert, L. and Wilke, R. (2007). *Application of a Simple Nonparametric Conditional Quantile Function Estimator in Unemployment Duration Analysis.* ZEW (Center for European Economic Research), Discussion Paper No. 05-67.

[Winkelmann, 1994] Winkelmann, R. (1994). *Count Data Models - Econometric Theory and an Application to Labor Mobility.* Springer Verlag.

List of Figures

List of Tables

A. Appendix

A.1. Software applications

An initial approach to study the data is to look at, demonstrate and visualize the structure and type of data. Ranges, histograms and densities as well as other graphical visualization tools can help to get a good overview of the data.
The following procedures to analyze the data are carried out by using the program R. Versions are available for Windows, Unix, Linux and Macintosh. The R software, together with documentation, can be obtained from the Comprehensive R Archive Network (CRAN) (homepage: http://cran.r-project.org). Aside from the fact that R is free of charge, another advantage is that this software provides a wide variety of statistical techniques that qualifies R to be a useful statistical computing tool. Another detail worth mentioning is that R is permanently changing software due to new extensions provided by programmers. Additionally extensions via user-defined functions are possible as well. R is quite similar to the widely used commercial statistical program called S-plus. It is a command-line driven package which is why I would also recommend, if Windows is the version in use, to install the code editor Tinn-R. This editor allows for syntax highlighting of the R language and can be downloaded, again free of charge (https://sourceforge.net/projects/tinn-r). Throughout this document the following convention is used:

- R commands and objects will appear in a different font and commands have parentheses after their name, eg.
  `summary()`
- typed in commands will appear preceded by the symbol ">", eg.
  `> range(efftime)`
- R output will be indented and appears mostly with a preceding number in squared brackets, eg.
  `[1] 0.25 284.00`

The packages where the subsequent commands stem from are:

```
> library(survival)
> library(stats)
> library(MASS)
> library(base)
> library(muhaz)
> library(tools) #sweave
> library(boot) #cor
> library(mvna)
> library(car) #qq.plot
> library(actuar) #rllogis
> library(cmprsk)
```

A description of these libraries and the content of these packages can be found in the built-in help software of R.
The calculation base for the data, edited as described in detail in section 3.2, also needs to be read into the program for further usage:

```
> data<-read.table("D:/Diploma/rdata.txt",
+ header=T,sep="\t",dec=",")
```

```
> dim(data)
```

```
[1] 2824  171
```

The dimension of the data set shows the number of observed individuals of the data set as row counts, with the number of columns (171) representing the number of the individuals' characteristics next to other extracted information. The data is attached in 'R' by the command

```
> attach(data)
```

to ensure direct access to the information by variable-name. This encodes the documentation for further use.

A.2. Empirical survivor function

```
> esf.fit


Call: survfit(formula = Surv(efftimeuc, status[which(status == 1)]))

      n  events  median 0.95LCL 0.95UCL
   1315    1315       4       4       4


> esf.fit2


Call: survfit(formula = Surv(efftimeuc2, status2[which(status2 == 1)]))

      n  events  median 0.95LCL 0.95UCL
    680     680       4       4       4


> esf.fit3


Call: survfit(formula = Surv(efftimeuc3, status3[status3 == 1]))

      n  events  median 0.95LCL 0.95UCL
    531     531       4       4       5


> summary(esf.fit)


Call: survfit(formula = Surv(efftimeuc, status[which(status == 1)]))

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0.25   1315      89  0.93232 0.00693     0.918841      0.94600
   1.00   1226     148  0.81977 0.01060     0.799258      0.84081
   2.00   1078     201  0.66692 0.01300     0.641927      0.69289
   3.00    877     160  0.54525 0.01373     0.518987      0.57284
   4.00    717     113  0.45932 0.01374     0.433155      0.48706
   5.00    604       97  0.38555 0.01342     0.360122      0.41278
   6.00    507       75  0.32852 0.01295     0.304088      0.35491
   7.00    432       42  0.29658 0.01260     0.272891      0.32232
   8.00    390       46  0.26160 0.01212     0.238889      0.28646
   9.00    344       34  0.23574 0.01171     0.213881      0.25984
  10.00    310       24  0.21749 0.01138     0.196298      0.24097
  11.00    286       35  0.19087 0.01084     0.170773      0.21334
  12.00    251       33  0.16578 0.01026     0.146850      0.18715
  13.00    218       26  0.14601 0.00974     0.128117      0.16640
  14.00    192       21  0.13004 0.00928     0.113072      0.14955
  15.00    171       12  0.12091 0.00899     0.104515      0.13988
  16.00    159        8  0.11483 0.00879     0.098828      0.13342
  17.00    151        8  0.10875 0.00859     0.093156      0.12694
  18.00    143        8  0.10266 0.00837     0.087501      0.12045
  19.00    135        8  0.09658 0.00815     0.081863      0.11394
  20.00    127        9  0.08973 0.00788     0.075543      0.10659
  21.00    118        4  0.08669 0.00776     0.072743      0.10332
  22.00    114        6  0.08213 0.00757     0.068553      0.09839
  23.00    108        6  0.07757 0.00738     0.064377      0.09346
  24.00    102        2  0.07605 0.00731     0.062988      0.09181
  25.00    100        3  0.07376 0.00721     0.060907      0.08934
  26.00     97        4  0.07072 0.00707     0.058139      0.08603
  27.00     93        6  0.06616 0.00685     0.054001      0.08106
  28.00     87        4  0.06312 0.00671     0.051253      0.07773
  29.00     83        4  0.06008 0.00655     0.048513      0.07440
  30.00     79        3  0.05779 0.00644     0.046463      0.07189
  31.00     76        2  0.05627 0.00635     0.045100      0.07022
  32.00     74        5  0.05247 0.00615     0.041704      0.06602
  33.00     69        4  0.04943 0.00598     0.038999      0.06265
  34.00     65        2  0.04791 0.00589     0.037651      0.06096
  35.00     63        1  0.04715 0.00584     0.036978      0.06012
  36.00     62        1  0.04639 0.00580     0.036306      0.05927
  37.00     61        1  0.04563 0.00575     0.035635      0.05842
  38.00     60        6  0.04106 0.00547     0.031626      0.05332
  39.00     54        1  0.04030 0.00542     0.030961      0.05247
  41.00     53        2  0.03878 0.00532     0.029634      0.05076
  42.00     51        2  0.03726 0.00522     0.028311      0.04904
  44.00     49        1  0.03650 0.00517     0.027651      0.04819
  45.00     48        2  0.03498 0.00507     0.026336      0.04646
  46.00     46        2  0.03346 0.00496     0.025025      0.04474
  47.00     44        3  0.03118 0.00479     0.023068      0.04214
  48.00     41        1  0.03042 0.00474     0.022419      0.04127
  49.00     40        2  0.02890 0.00462     0.021124      0.03953
  50.00     38        1  0.02814 0.00456     0.020480      0.03866
  51.00     37        1  0.02738 0.00450     0.019837      0.03778
  52.00     36        1  0.02662 0.00444     0.019195      0.03691
  53.00     35        3  0.02433 0.00425     0.017282      0.03427
  58.00     32        4  0.02129 0.00398     0.014760      0.03072
  59.00     28        1  0.02053 0.00391     0.014136      0.02982
  60.00     27        1  0.01977 0.00384     0.013514      0.02893
  61.00     26        2  0.01825 0.00369     0.012278      0.02713
  62.00     24        1  0.01749 0.00361     0.011665      0.02623
  63.00     23        2  0.01597 0.00346     0.010448      0.02441
  64.00     21        1  0.01521 0.00337     0.009845      0.02350
  65.00     20        2  0.01369 0.00320     0.008652      0.02166
  70.00     18        1  0.01293 0.00312     0.008061      0.02073
  71.00     17        1  0.01217 0.00302     0.007476      0.01980
  72.00     16        1  0.01141 0.00293     0.006897      0.01887
  75.00     15        1  0.01065 0.00283     0.006323      0.01793
  76.00     14        1  0.00989 0.00273     0.005756      0.01698
  78.00     13        1  0.00913 0.00262     0.005196      0.01603
  81.00     12        1  0.00837 0.00251     0.004644      0.01507
  85.00     11        1  0.00760 0.00240     0.004101      0.01410
  86.00     10        1  0.00684 0.00227     0.003569      0.01312
 105.00      9        2  0.00532 0.00201     0.002543      0.01114
 112.00      7        1  0.00456 0.00186     0.002054      0.01014
 120.00      6        2  0.00304 0.00152     0.001143      0.00809
 144.00      4        1  0.00228 0.00132     0.000737      0.00706
 158.00      3        1  0.00152 0.00107     0.000381      0.00607
 182.00      2        2  0.00000      NA           NA           NA
```

```
> summary(esf.fit2)

Call: survfit(formula = Surv(efftimeuc2, status2[which(status2 == 1)]))

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0.25    680      14  0.97941 0.00545     0.968797       0.9901
   1.00    666      53  0.90147 0.01143     0.879346       0.9242
   2.00    613     125  0.71765 0.01726     0.684599       0.7523
   3.00    488     108  0.55882 0.01904     0.522723       0.5974
   4.00    380      68  0.45882 0.01911     0.422858       0.4978
   5.00    312      67  0.36029 0.01841     0.325958       0.3982
   6.00    245      39  0.30294 0.01762     0.270299       0.3395
   7.00    206      23  0.26912 0.01701     0.237765       0.3046
   8.00    183      29  0.22647 0.01605     0.197099       0.2602
   9.00    154      24  0.19118 0.01508     0.163792       0.2231
  10.00    130      16  0.16765 0.01433     0.141796       0.1982
  11.00    114      20  0.13824 0.01324     0.114582       0.1668
  12.00     94      15  0.11618 0.01229     0.094425       0.1429
  13.00     79      14  0.09559 0.01128     0.075858       0.1205
  14.00     65      12  0.07794 0.01028     0.060186       0.1009
  15.00     53       6  0.06912 0.00973     0.052456       0.0911
  16.00     47       5  0.06176 0.00923     0.046081       0.0828
  17.00     42       4  0.05588 0.00881     0.041030       0.0761
  18.00     38       3  0.05147 0.00847     0.037276       0.0711
  19.00     35       2  0.04853 0.00824     0.034791       0.0677
  20.00     33       5  0.04118 0.00762     0.028651       0.0592
  21.00     28       1  0.03971 0.00749     0.027436       0.0575
  22.00     27       3  0.03529 0.00708     0.023826       0.0523
  23.00     24       1  0.03382 0.00693     0.022634       0.0505
  24.00     23       1  0.03235 0.00679     0.021448       0.0488
  26.00     22       1  0.03088 0.00663     0.020270       0.0471
  27.00     21       2  0.02794 0.00632     0.017935       0.0435
  28.00     19       3  0.02353 0.00581     0.014499       0.0382
  29.00     16       1  0.02206 0.00563     0.013373       0.0364
  31.00     15       1  0.02059 0.00545     0.012260       0.0346
  32.00     14       2  0.01765 0.00505     0.010072       0.0309
  33.00     12       2  0.01471 0.00462     0.007949       0.0272
  38.00     10       2  0.01176 0.00413     0.005908       0.0234
  46.00      8       1  0.01029 0.00387     0.004926       0.0215
  47.00      7       1  0.00882 0.00359     0.003978       0.0196
  48.00      6       1  0.00735 0.00328     0.003070       0.0176
  52.00      5       1  0.00588 0.00293     0.002214       0.0156
  58.00      4       1  0.00441 0.00254     0.001426       0.0136
  63.00      3       1  0.00294 0.00208     0.000737       0.0117
  72.00      2       1  0.00147 0.00147     0.000207       0.0104
 105.00      1       1  0.00000      NA           NA           NA

> summary(esf.fit3)

Call: survfit(formula = Surv(efftimeuc3, status3[status3 == 1]))

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0.25    531      75  0.85876 0.01511     0.829640       0.8889
   1.00    456      66  0.73446 0.01916     0.697846       0.7730
   2.00    390      54  0.63277 0.02092     0.593068       0.6751
   3.00    336      43  0.55179 0.02158     0.511071       0.5958
   4.00    293      28  0.49906 0.02170     0.458292       0.5435
   5.00    265      24  0.45386 0.02161     0.413430       0.4982
   6.00    241      31  0.39548 0.02122     0.356004       0.4393
   7.00    210      16  0.36535 0.02090     0.326604       0.4087
   8.00    194      15  0.33710 0.02051     0.299198       0.3798
   9.00    179       8  0.32203 0.02028     0.284646       0.3643
  10.00    171       7  0.30885 0.02005     0.271951       0.3508
  11.00    164      14  0.28249 0.01954     0.246675       0.3235
  12.00    150      18  0.24859 0.01876     0.214416       0.2882
  13.00    132       9  0.23164 0.01831     0.198397       0.2704
  14.00    123       8  0.21657 0.01788     0.184225       0.2546
  15.00    115       4  0.20904 0.01765     0.177164       0.2467
  16.00    111       2  0.20527 0.01753     0.173640       0.2427
  17.00    109       4  0.19774 0.01728     0.166606       0.2347
  18.00    105       5  0.18832 0.01697     0.157840       0.2247
  19.00    100       6  0.17702 0.01656     0.147363       0.2127
  20.00     94       4  0.16949 0.01628     0.140404       0.2046
  21.00     90       3  0.16384 0.01606     0.135200       0.1986
  22.00     87       3  0.15819 0.01584     0.130009       0.1925
  23.00     84       5  0.14878 0.01544     0.121388       0.1823
  24.00     79       1  0.14689 0.01536     0.119669       0.1803
  25.00     78       3  0.14124 0.01511     0.114521       0.1742
  26.00     75       3  0.13559 0.01486     0.109388       0.1681
  27.00     72       4  0.12806 0.01450     0.102572       0.1599
  28.00     68       1  0.12618 0.01441     0.100872       0.1578
  29.00     67       3  0.12053 0.01413     0.095786       0.1517
  30.00     64       3  0.11488 0.01384     0.090720       0.1455
  31.00     61       1  0.11299 0.01374     0.089035       0.1434
  32.00     60       3  0.10734 0.01343     0.083996       0.1372
  33.00     57       2  0.10358 0.01322     0.080649       0.1330
  34.00     55       2  0.09981 0.01301     0.077312       0.1289
  35.00     53       1  0.09793 0.01290     0.075648       0.1268
  36.00     52       1  0.09605 0.01279     0.073986       0.1247
  37.00     51       1  0.09416 0.01267     0.072328       0.1226
  38.00     50       4  0.08663 0.01221     0.065723       0.1142
  39.00     46       1  0.08475 0.01209     0.064080       0.1121
  41.00     45       2  0.08098 0.01184     0.060804       0.1078
  42.00     43       2  0.07721 0.01158     0.057543       0.1036
  44.00     41       1  0.07533 0.01145     0.055917       0.1015
  45.00     40       2  0.07156 0.01119     0.052679       0.0972
  46.00     38       1  0.06968 0.01105     0.051066       0.0951
  47.00     37       2  0.06591 0.01077     0.047854       0.0908
  49.00     35       2  0.06215 0.01048     0.044660       0.0865
  50.00     33       1  0.06026 0.01033     0.043071       0.0843
  51.00     32       1  0.05838 0.01017     0.041487       0.0822
  53.00     31       3  0.05273 0.00970     0.036771       0.0756
  58.00     28       3  0.04708 0.00919     0.032112       0.0690
  59.00     25       1  0.04520 0.00902     0.030573       0.0668
  60.00     24       1  0.04331 0.00883     0.029042       0.0646
  61.00     23       2  0.03955 0.00846     0.026007       0.0601
  62.00     21       1  0.03766 0.00826     0.024503       0.0579
  63.00     20       1  0.03578 0.00806     0.023009       0.0556
  64.00     19       1  0.03390 0.00785     0.021527       0.0534
  65.00     18       2  0.03013 0.00742     0.018598       0.0488
  70.00     16       1  0.02825 0.00719     0.017153       0.0465
  71.00     15       1  0.02637 0.00695     0.015724       0.0442
  75.00     14       1  0.02448 0.00671     0.014311       0.0419
  76.00     13       1  0.02260 0.00645     0.012917       0.0395
  78.00     12       1  0.02072 0.00618     0.011543       0.0372
  81.00     11       1  0.01883 0.00590     0.010192       0.0348
  85.00     10       1  0.01695 0.00560     0.008868       0.0324
  86.00      9       1  0.01507 0.00529     0.007574       0.0300
 105.00      8       1  0.01318 0.00495     0.006315       0.0275
 112.00      7       1  0.01130 0.00459     0.005099       0.0250
 120.00      6       2  0.00753 0.00375     0.002838       0.0200
 144.00      4       1  0.00565 0.00325     0.001828       0.0175
 158.00      3       1  0.00377 0.00266     0.000944       0.0150
 182.00      2       2  0.00000      NA           NA           NA
```

A.3. Estimated hazard and survivor function

```
> table.estim<-cbind(efftimuc,d1,n1,estimhaz,estimsurvf)
> table.estim
```

```
      efftimuc  d1   n1   estimhaz  estimsurvf
 [1,]     0.00   0 1315 0.00000000 1.000000000
 [2,]     0.25  89 1315 0.06768061 0.932319392
 [3,]     1.00 148 1226 0.12071778 0.819771863
 [4,]     2.00 201 1078 0.18645640 0.666920152
 [5,]     3.00 160  877 0.18244014 0.545247148
 [6,]     4.00 113  717 0.15760112 0.459315589
 [7,]     5.00  97  604 0.16059603 0.385551331
 [8,]     6.00  75  507 0.14792899 0.328517110
 [9,]     7.00  42  432 0.09722222 0.296577947
[10,]     8.00  46  390 0.11794872 0.261596958
[11,]     9.00  34  344 0.09883721 0.235741445
[12,]    10.00  24  310 0.07741935 0.217490494
[13,]    11.00  35  286 0.12237762 0.190874525
[14,]    12.00  33  251 0.13147410 0.165779468
[15,]    13.00  26  218 0.11926606 0.146007605
[16,]    14.00  21  192 0.10937500 0.130038023
[17,]    15.00  12  171 0.07017544 0.120912548
[18,]    16.00   8  159 0.05031447 0.114828897
[19,]    17.00   8  151 0.05298013 0.108745247
[20,]    18.00   8  143 0.05594406 0.102661597
[21,]    19.00   8  135 0.05925926 0.096577947
[22,]    20.00   9  127 0.07086614 0.089733840
[23,]    21.00   4  118 0.03389831 0.086692015
[24,]    22.00   6  114 0.05263158 0.082129278
[25,]    23.00   6  108 0.05555556 0.077566540
[26,]    24.00   2  102 0.01960784 0.076045627
[27,]    25.00   3  100 0.03000000 0.073764259
[28,]    26.00   4   97 0.04123711 0.070722433
[29,]    27.00   6   93 0.06451613 0.066159696
[30,]    28.00   4   87 0.04597701 0.063117871
[31,]    29.00   4   83 0.04819277 0.060076046
[32,]    30.00   3   79 0.03797468 0.057794677
[33,]    31.00   2   76 0.02631579 0.056273764
[34,]    32.00   5   74 0.06756757 0.052471483
[35,]    33.00   4   69 0.05797101 0.049429658
[36,]    34.00   2   65 0.03076923 0.047908745
[37,]    35.00   1   63 0.01587302 0.047148289
[38,]    36.00   1   62 0.01612903 0.046387833
[39,]    37.00   1   61 0.01639344 0.045627376
[40,]    38.00   6   60 0.10000000 0.041064639
[41,]    39.00   1   54 0.01851852 0.040304183
[42,]    41.00   2   53 0.03773585 0.038783270
[43,]    42.00   2   51 0.03921569 0.037262357
[44,]    44.00   1   49 0.02040816 0.036501901
[45,]    45.00   2   48 0.04166667 0.034980989
[46,]    46.00   2   46 0.04347826 0.033460076
[47,]    47.00   3   44 0.06818182 0.031178707
[48,]    48.00   1   41 0.02439024 0.030418251
[49,]    49.00   2   40 0.05000000 0.028897338
[50,]    50.00   1   38 0.02631579 0.028136882
[51,]    51.00   1   37 0.02702703 0.027376426
[52,]    52.00   1   36 0.02777778 0.026615970
[53,]    53.00   3   35 0.08571429 0.024334601
[54,]    58.00   4   32 0.12500000 0.021292776
[55,]    59.00   1   28 0.03571429 0.020532319
[56,]    60.00   1   27 0.03703704 0.019771863
[57,]    61.00   2   26 0.07692308 0.018250951
[58,]    62.00   1   24 0.04166667 0.017490494
[59,]    63.00   2   23 0.08695652 0.015969582
[60,]    64.00   1   21 0.04761905 0.015209125
```

```
[61,]    65.00   2   20 0.10000000 0.013688213
[62,]    70.00   1   18 0.05555556 0.012927757
[63,]    71.00   1   17 0.05882353 0.012167300
[64,]    72.00   1   16 0.06250000 0.011406844
[65,]    75.00   1   15 0.06666667 0.010646388
[66,]    76.00   1   14 0.07142857 0.009885932
[67,]    78.00   1   13 0.07692308 0.009125475
[68,]    81.00   1   12 0.08333333 0.008365019
[69,]    85.00   1   11 0.09090909 0.007604563
[70,]    86.00   1   10 0.10000000 0.006844106
[71,]   105.00   2    9 0.22222222 0.005323194
[72,]   112.00   1    7 0.14285714 0.004562738
[73,]   120.00   2    6 0.33333333 0.003041825
[74,]   144.00   1    4 0.25000000 0.002281369
[75,]   158.00   1    3 0.33333333 0.001520913
[76,]   182.00   2    2 1.00000000 0.000000000

> table.estim2<-cbind(efftimuc2,d12,n12,estimhaz2,estimsurvf2)
> table.estim2

      efftimuc2 d12 n12  estimhaz2 estimsurvf2
 [1,]      0.00   0 680 0.00000000 1.000000000
 [2,]      0.25  14 680 0.02058824 0.979411765
 [3,]      1.00  53 666 0.07957958 0.901470588
 [4,]      2.00 125 613 0.20391517 0.717647059
 [5,]      3.00 108 488 0.22131148 0.558823529
 [6,]      4.00  68 380 0.17894737 0.458823529
 [7,]      5.00  67 312 0.21474359 0.360294118
 [8,]      6.00  39 245 0.15918367 0.302941176
 [9,]      7.00  23 206 0.11165049 0.269117647
[10,]      8.00  29 183 0.15846995 0.226470588
[11,]      9.00  24 154 0.15584416 0.191176471
[12,]     10.00  16 130 0.12307692 0.167647059
[13,]     11.00  20 114 0.17543860 0.138235294
[14,]     12.00  15  94 0.15957447 0.116176471
[15,]     13.00  14  79 0.17721519 0.095588235
[16,]     14.00  12  65 0.18461538 0.077941176
[17,]     15.00   6  53 0.11320755 0.069117647
[18,]     16.00   5  47 0.10638298 0.061764706
[19,]     17.00   4  42 0.09523810 0.055882353
[20,]     18.00   3  38 0.07894737 0.051470588
[21,]     19.00   2  35 0.05714286 0.048529412
[22,]     20.00   5  33 0.15151515 0.041176471
[23,]     21.00   1  28 0.03571429 0.039705882
[24,]     22.00   3  27 0.11111111 0.035294118
[25,]     23.00   1  24 0.04166667 0.033823529
[26,]     24.00   1  23 0.04347826 0.032352941
[27,]     26.00   1  22 0.04545455 0.030882353
[28,]     27.00   2  21 0.09523810 0.027941176
[29,]     28.00   3  19 0.15789474 0.023529412
[30,]     29.00   1  16 0.06250000 0.022058824
[31,]     31.00   1  15 0.06666667 0.020588235
[32,]     32.00   2  14 0.14285714 0.017647059
[33,]     33.00   2  12 0.16666667 0.014705882
[34,]     38.00   2  10 0.20000000 0.011764706
[35,]     46.00   1   8 0.12500000 0.010294118
[36,]     47.00   1   7 0.14285714 0.008823529
[37,]     48.00   1   6 0.16666667 0.007352941
[38,]     52.00   1   5 0.20000000 0.005882353
[39,]     58.00   1   4 0.25000000 0.004411765
[40,]     63.00   1   3 0.33333333 0.002941176
[41,]     72.00   1   2 0.50000000 0.001470588
[42,]    105.00   1   1 1.00000000 0.000000000

> table.estim3<-cbind(efftimuc3,d13,n13,estimhaz3,estimsurvf3)
> table.estim3

      efftimuc3 d13 n13  estimhaz3 estimsurvf3
 [1,]      0.00   0 531 0.00000000 1.000000000
 [2,]      0.25  75 531 0.14124294 0.858757062
 [3,]      1.00  66 456 0.14473684 0.734463277
 [4,]      2.00  54 390 0.13846154 0.632768362
 [5,]      3.00  43 336 0.12797619 0.551789077
 [6,]      4.00  28 293 0.09556314 0.499058380
 [7,]      5.00  24 265 0.09056604 0.453860640
 [8,]      6.00  31 241 0.12863071 0.395480226
 [9,]      7.00  16 210 0.07619048 0.365348399
[10,]      8.00  15 194 0.07731959 0.337099812
[11,]      9.00   8 179 0.04469274 0.322033898
[12,]     10.00   7 171 0.04093567 0.308851224
[13,]     11.00  14 164 0.08536585 0.282485876
[14,]     12.00  18 150 0.12000000 0.248587571
[15,]     13.00   9 132 0.06818182 0.231638418
[16,]     14.00   8 123 0.06504065 0.216572505
[17,]     15.00   4 115 0.03478261 0.209039548
[18,]     16.00   2 111 0.01801802 0.205273070
[19,]     17.00   4 109 0.03669725 0.197740113
[20,]     18.00   5 105 0.04761905 0.188323917
```

```
[21,]    19.00   6 100 0.06000000 0.177024482
[22,]    20.00   4  94 0.04255319 0.169491525
[23,]    21.00   3  90 0.03333333 0.163841808
[24,]    22.00   3  87 0.03448276 0.158192090
[25,]    23.00   5  84 0.05952381 0.148775895
[26,]    24.00   1  79 0.01265823 0.146892655
[27,]    25.00   3  78 0.03846154 0.141242938
[28,]    26.00   3  75 0.04000000 0.135593220
[29,]    27.00   4  72 0.05555556 0.128060264
[30,]    28.00   1  68 0.01470588 0.126177024
[31,]    29.00   3  67 0.04477612 0.120527307
[32,]    30.00   3  64 0.04687500 0.114877589
[33,]    31.00   1  61 0.01639344 0.112994350
[34,]    32.00   3  60 0.05000000 0.107344633
[35,]    33.00   2  57 0.03508772 0.103578154
[36,]    34.00   2  55 0.03636364 0.099811676
[37,]    35.00   1  53 0.01886792 0.097928437
[38,]    36.00   1  52 0.01923077 0.096045198
[39,]    37.00   1  51 0.01960784 0.094161959
[40,]    38.00   4  50 0.08000000 0.086629002
[41,]    39.00   1  46 0.02173913 0.084745763
[42,]    41.00   2  45 0.04444444 0.080979284
[43,]    42.00   2  43 0.04651163 0.077212806
[44,]    44.00   1  41 0.02439024 0.075329567
[45,]    45.00   2  40 0.05000000 0.071563089
[46,]    46.00   1  38 0.02631579 0.069679849
[47,]    47.00   2  37 0.05405405 0.065913371
[48,]    49.00   2  35 0.05714286 0.062146893
[49,]    50.00   1  33 0.03030303 0.060263653
[50,]    51.00   1  32 0.03125000 0.058380414
[51,]    53.00   3  31 0.09677419 0.052730697
[52,]    58.00   3  28 0.10714286 0.047080979
[53,]    59.00   1  25 0.04000000 0.045197740
[54,]    60.00   1  24 0.04166667 0.043314501
[55,]    61.00   2  23 0.08695652 0.039548023
[56,]    62.00   1  21 0.04761905 0.037664783
[57,]    63.00   1  20 0.05000000 0.035781544
[58,]    64.00   1  19 0.05263158 0.033898305
[59,]    65.00   2  18 0.11111111 0.030131827
[60,]    70.00   1  16 0.06250000 0.028248588
[61,]    71.00   1  15 0.06666667 0.026365348
[62,]    75.00   1  14 0.07142857 0.024482109
[63,]    76.00   1  13 0.07692308 0.022598870
[64,]    78.00   1  12 0.08333333 0.020715631
[65,]    81.00   1  11 0.09090909 0.018832392
[66,]    85.00   1  10 0.10000000 0.016949153
[67,]    86.00   1   9 0.11111111 0.015065913
[68,]   105.00   1   8 0.12500000 0.013182674
[69,]   112.00   1   7 0.14285714 0.011299435
[70,]   120.00   2   6 0.33333333 0.007532957
[71,]   144.00   1   4 0.25000000 0.005649718
[72,]   158.00   1   3 0.33333333 0.003766478
[73,]   182.00   2   2 1.00000000 0.000000000
```

A.4. Kaplan-Meier survivor function

```
> km.fit

Call: survfit(formula = Surv(efftime, status), type = "kaplan-meier")

      n  events  median 0.95LCL 0.95UCL
   2824    1315      10       9      11


> summary(km.fit)

Call: survfit(formula = Surv(efftime, status), type = "kaplan-meier")

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0.25   2824      89   0.9685 0.00329      0.96206       0.9749
   1.00   2504     148   0.9112 0.00551      0.90050       0.9221
   2.00   2090     201   0.8236 0.00771      0.80864       0.8388
   3.00   1723     160   0.7471 0.00906      0.72958       0.7651
   4.00   1420     113   0.6877 0.00991      0.66851       0.7074
   5.00   1207      97   0.6324 0.01059      0.61199       0.6535
   6.00   1024      75   0.5861 0.01108      0.56477       0.6082
   7.00    887      42   0.5583 0.01135      0.53652       0.5810
   8.00    807      46   0.5265 0.01164      0.50419       0.5498
   9.00    725      34   0.5018 0.01183      0.47915       0.5256
  10.00    664      24   0.4837 0.01197      0.46078       0.5077
  11.00    614      35   0.4561 0.01216      0.43288       0.4806
  12.00    542      33   0.4283 0.01235      0.40481       0.4532
  13.00    473      26   0.4048 0.01250      0.38102       0.4301
  14.00    427      21   0.3849 0.01262      0.36093       0.4104
  15.00    390      12   0.3730 0.01268      0.34899       0.3988
  16.00    360       8   0.3648 0.01274      0.34062       0.3906
  17.00    341       8   0.3562 0.01279      0.33198       0.3822
  18.00    327       8   0.3475 0.01285      0.32320       0.3736
  19.00    314       8   0.3386 0.01289      0.31428       0.3649
  20.00    296       9   0.3283 0.01295      0.30391       0.3547
  21.00    281       4   0.3237 0.01298      0.29920       0.3501
  22.00    268       6   0.3164 0.01302      0.29190       0.3430
  23.00    255       6   0.3090 0.01306      0.28440       0.3357
  24.00    241       2   0.3064 0.01308      0.28181       0.3331
  25.00    222       3   0.3023 0.01312      0.27762       0.3291
  26.00    212       4   0.2966 0.01318      0.27183       0.3235
  27.00    199       6   0.2876 0.01328      0.26274       0.3149
  28.00    190       4   0.2816 0.01334      0.25660       0.3090
  29.00    180       4   0.2753 0.01340      0.25025       0.3029
  30.00    169       3   0.2704 0.01346      0.24529       0.2981
  31.00    163       2   0.2671 0.01350      0.24192       0.2949
  32.00    158       5   0.2586 0.01359      0.23334       0.2867
  33.00    146       4   0.2516 0.01367      0.22615       0.2798
  34.00    140       2   0.2480 0.01371      0.22250       0.2763
  35.00    138       1   0.2462 0.01373      0.22069       0.2746
  36.00    133       1   0.2443 0.01375      0.21881       0.2728
  37.00    129       1   0.2424 0.01377      0.21688       0.2710
  38.00    128       6   0.2311 0.01388      0.20539       0.2599
  39.00    120       1   0.2291 0.01390      0.20345       0.2581
  41.00    110       2   0.2250 0.01396      0.19921       0.2541
  42.00    105       2   0.2207 0.01402      0.19485       0.2499
  44.00     99       1   0.2185 0.01405      0.19258       0.2478
  45.00     97       2   0.2140 0.01412      0.18800       0.2435
  46.00     94       2   0.2094 0.01418      0.18337       0.2391
  47.00     90       3   0.2024 0.01427      0.17630       0.2324
  48.00     85       1   0.2000 0.01430      0.17389       0.2301
  49.00     82       2   0.1952 0.01436      0.16895       0.2254
  50.00     79       1   0.1927 0.01439      0.16646       0.2231
  51.00     77       1   0.1902 0.01442      0.16393       0.2207
  52.00     75       1   0.1877 0.01445      0.16137       0.2182
  53.00     72       3   0.1798 0.01453      0.15349       0.2107
  58.00     64       4   0.1686 0.01467      0.14216       0.1999
  59.00     59       1   0.1657 0.01470      0.13929       0.1972
  60.00     56       1   0.1628 0.01473      0.13632       0.1944
  61.00     55       2   0.1569 0.01478      0.13041       0.1887
  62.00     51       1   0.1538 0.01480      0.12734       0.1857
  63.00     50       2   0.1476 0.01484      0.12123       0.1798
  64.00     47       1   0.1445 0.01485      0.11813       0.1767
  65.00     45       2   0.1381 0.01487      0.11180       0.1705
  70.00     36       1   0.1342 0.01494      0.10792       0.1670
  71.00     34       1   0.1303 0.01501      0.10394       0.1633
  72.00     33       1   0.1263 0.01507      0.10000       0.1596
  75.00     32       1   0.1224 0.01511      0.09609       0.1559
  76.00     31       1   0.1184 0.01513      0.09221       0.1521
  78.00     30       1   0.1145 0.01513      0.08837       0.1483
  81.00     28       1   0.1104 0.01513      0.08440       0.1444
  85.00     27       1   0.1063 0.01511      0.08046       0.1405
  86.00     25       1   0.1021 0.01510      0.07638       0.1364
 105.00     19       2   0.0913 0.01530      0.06576       0.1268
 112.00     15       1   0.0852 0.01544      0.05975       0.1216
 120.00     13       2   0.0721 0.01560      0.04719       0.1102
 144.00      9       1   0.0641 0.01579      0.03955       0.1039
 158.00      6       1   0.0534 0.01638      0.02929       0.0974
 182.00      4       2   0.0267 0.01567      0.00846       0.0843
```

Index