On statistical Survival Analysis and its Applications in Medicine

Kardaun, Otto Jan Willem Frederik

Publication date:
1986

Citation for published version (APA):
Kardaun, O. J. W. F. (1986). *On statistical Survival Analysis and its Applications in Medicine*. s.n.

Summary

In this thesis some theoretical and practical aspects of survival-time analysis with censored data are considered. Survival-time analysis does appear in various contexts. Applications in the thesis, however, are restricted to the medical field. An observation is called censored if one knows that the true survival time, which cannot (yet) be observed, is larger than a certain lower bound. This happens for example in survival-time analysis of patients: If a patient is still alive at the closing date of the study, then the true survival time (measured, for example, from the date of diagnosis onwards) is not observable, but one knows that this time is larger than the time between diagnosis and the closing date of the study. Omitting the censored observations from the analysis is not to be advised, as this will select individuals with a relatively short lifetime from the available sample.

After the Second World War, various statistical techniques have been developed for correctly incorporating censored observations in the usual statistical situations. The existing literature is now extensive. Benchmark articles are KAPLAN & MEIER (1958) on the product-limit estimator, a distribution-free estimator for the survival function $S(t) = P\{T > t\}$ (which denotes the probability that the true survival time is larger than t), and COX (1972), which introduces the proportional hazards model

$$S(t \mid z_1, \ldots, z_k) = S(t \mid 0, \ldots, 0)^{exp(\beta_1 z_1 + \ldots + \beta_k z_k)}$$

to express the influence of risk factors (also called covariables) on the survival time. Here, $z_1, \ldots, z_k$ are the scores of the covariables, and $\beta_1, \ldots, \beta_k$ regression coefficients, which are to be estimated (as is also the base-line survival function $S(t \mid 0, \ldots, 0)$).

In the first part of the thesis (chapters I and II), survival-time analysis is used for the interpretation and evaluation of two medical data sets. In the last part (chapter III), some theoretical aspects are considered, which are motivated by the previous chapters.

The first chapter is concerned with the (remaining) survival time (measured from diagnosis onwards) of male larynx-cancer patients. The stage of the cancer (1 to 4) and the age of the patient are considered as covariables. The data are stratified with respect to stage, while for age Cox's proportional hazards model is used. For a patient with stage

i $(i = 1, \ldots, 4)$ of the cancer and of age z, a prognosis (which includes statistical uncertainties) is made of the remaining lifetime. Furthermore, a comparison is made with the lifetime of the normal population, taking into account the effects of age. Finally, the adequacy of the proportional hazards assumption to model the influence of age is investigated.

In chapter II a more complicated case-study is considered. The survival time (measured from the date of diagnosis until terminal renal failure) of patients with a certain kidney disease (IgA nephropathy) is investigated. Several correlated risk factors, such as the glomerular filtration rate of the kidneys, hypertension, proteinuria and hematuria (macroscopic as well as microscopic) seem to influence the survival time. Besides, there are kidney biopsy variables (with some missing observations), which are possibly related to the seriousness of the disease. The type of hypertension treatment changed rather drastically during the period of investigation. The problem of selecting a suitable set of covariables is addressed explicitly. A practical result of this chapter consists in the construction of two nomograms, from which one can read, for an arbitrary patient (with risk-factor scores $z_1, \ldots, z_k$), the estimated survival function $\hat{S}(t \mid z_1, \ldots, z_k)$ and (for a fixed time t) the involved statistical uncertainty. Moreover, a residual analysis is developed as a global model check of the proportional hazards assumption. This, however, leads to easily interpretable results only in the simple case that there are no censored observations. A general interactive computer program has been developed to support the analysis of this chapter.

In chapter III two theoretical developments, motivated by practice, are discussed. These are (1) the construction of simultaneous confidence bands for the survival function $S(t \mid \mathbf{z}) = S(t \mid z_1, \ldots, z_k)$ and (2) the implementation of the results of multivariable survival analysis, as described in chapter II, in decision analysis. This implementation is performed from the medical as well as from the mathematical side.

A (two-sided) simultaneous confidence band is a generalisation of a confidence interval for a fixed time t. It consists of a lower boundary $\underline{S}(t \mid, \mathbf{z})$ and an upper boundary $\overline{S}(t \mid \mathbf{z})$ such that, for a given interval $[t_{min}, t_{max}]$ and a given $\alpha \in [0, 1]$, the probability that the unknown survival function lies between $\underline{S}(t \mid \mathbf{z})$ and $\overline{S}(t \mid \mathbf{z})$ for every $t \in [t_{min}, t_{max}]$ is equal to $1 - \alpha$. The calculation of the width of a band with a realistic shape presents a non-trivial mathematical problem, which was solved

by using an integral equation approach. Such a confidence band can be used (among others) to test the null hypothesis H_0: $S(t \mid \mathbf{z}) = S_0(t)$, where $S_0(t)$ is a reference survival function (for example of the normal population at a certain age).

In the decision analysis section of chapter III, the following problem is considered. Suppose one is interested in choosing between two treatments, which we shall call A and B. (A may denote 'to treat' and B 'not to treat' according to some method.) We further assume that data are available from patients who have been treated according to A and from (other) patients who have been treated according to B. The question is whether for a new patient with covariables $z_1, \ldots, z_k$, on the basis of the available data, (1) treatment A has to be recommended, (2) treatment B has to be recommended, or (3) no recommendation should be made. Our analysis of this problem is based on the estimated survival functions of the 'effective lifetime' under each treatment. The effective lifetime is determined by subjective ratios, established in a discussion with the patient, of the qualities of life under various conditions. Sometimes, the difference between two 'effective' survival functions is well expressed by a one-dimensional parameter (such as the ratio of the mean, or median, effective survival times), the estimator of which is asymptotically normally distributed. For such a situation, the (two-sided) Neyman-Pearson testing theory is replaced by a loss-function formulation. This formulation leads to the same partition of the outcome space $\mathbb{R}$ as does some two-sided test. It adequately reflects, however, the three-decision character of the above mentioned problem, and allows a motivated choice of the level α of the corresponding (two-sided) test. In a traditional approach, this level is often set to 5% (or 1%).