Boosting First-Hitting-Time Regression Models

Vegard Stikbakke

February 9, 2018

Abstract

Empty.

Acknowledgements

Empty.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

sec:intro

In this thesis, we work with boosting for regression in the first-hitting-time model. First-hitting-time is a model in survival analysis which serves as an alternative to the proportional hazards model, typically known as Cox regression. Developments in FHT regression are relatively recent, and there has to our knowledge been no attempt at tackling it in the high-dimensional case, in which boosting is an appropriate choice of method.

CHAPTER 2

First hitting time regression models

2.1 Survival analysis and time-to-event models

sec:survival

In many fields, it is interesting to consider the lifetime of some entity. A lifetime ends when some event occurs. We are then interested in inferring things about this lifetime, and what it depends upon. In medical fields, this is the field of survival analysis, while in engineering fields, this is reliability analysis. In the first case, we consider e.g. the lifetime of patients with some chronic disease, or the length of a hospital stay after some treatment. In the latter, we consider e.g. the time before some component of a system breaks and must be replaced. It is in both cases natural to imagine that the event happens as some (underlying) process reaches some threshold (for the first time). We will use boundary and threshold interchangeably.

The time-to-event T is a continuous, non-negative random variable $T \sim f(t)$, $t > 0$, for some probability density function f.

Define hazard and survival functions.

Data structures

Let us say that we have observations from time points $t_0 = 0, \dots, t_T$. For each entity i, we have some characteristic of this entity, $\mathbf{x}_i$, which do not change over time. And we have observations at each time t_j, namely, δ_{ij}, which denotes if the event has occurred or not (1 if it has, 0 if not).

Proportional hazards

Write here.

Regression

sec:surv-reg

Lifetime regression models are usually fitted by maximum likelihood methods. The likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i|\mathbf{x}_i, \boldsymbol{\theta})^{\delta_i} S(t_i|\mathbf{x}_i, \boldsymbol{\theta})^{1-\delta_i} \tag{2.1}$$

{eq:surv-lik}

2.2 First hitting time models

sec:fht

The most widely used models in survival analysis are proportional hazards models. These consider hazard functions over time. Hazard functions say something about the chance of an event happening during some (small) interval. However, as mentioned, it seems natural to imagine the event happening due to some underlying process reaching some threshold. One way to model this is to in fact model the process itself, and look at the time it takes for the process to reach this threshold. We call this the first-hitting-time (FHT) of the process.

An FHT model has two main components.

1. A stochastic process $\{Y(t), t \in \mathcal{T}, y \in \mathcal{Y}\}$, with $Y(0) = y_0$.

2. A boundary set, $B \subset \mathcal{Y}$, where $y_0 \notin \mathcal{B}$

The first hitting time is the first time the process reaches the boundary set. Formally, the FHT is a stochastic variable S, which is defined as

$$S = \inf\{t \colon Y(t) \in \mathcal{B}\}$$

Note that it is possible that $P(S < \infty) < 1$.

Typically, one will consider a process with boundary $B = 0$. The event then occurs if and when the process $\{Ys\}$ reaches 0 at $y(T)$.

This framework makes for a flexible model, due to many possible choices for the process and the boundary. We now look at a common choice of the process.

Wiener process

sec:wiener

The Wiener process, also known as the standard Brownian motion process, is a process which is continuous in time and space, and has the properties (Caroni 2017, p. 61) that

- $Y(t)$ has independent increments, such that $Y(t_2)-Y(t_1)$ and $Y(t_4)-Y(t_3)$ are independent for any disjoint intervals, and

- for any interval (t_1, t_2),

$$Y(t_2) - Y(t-1) \sim N(\mu(t_2 - t_1), \sigma^2(t_2 - t_1)).$$

This process may both increase and decrease. However, if we want a monotonic restriction on the movement of the process, we may use a gamma process.

Gamma process

The gamma process is suitable for modelling something which we would require to be monotonic, typically a physical degradation, i.e. where the damage cannot mend itself. (Caroni 2017, p. 59)

Other choices of processes include Markov chain state models, the Bernoulli process, and the Ornstein-Uhlenbeck process.

2.3 First hitting time regression based on underlying Wiener process

The first hitting time of the Wiener process (section 2.2) follows an inverse Gaussian distribution (derivation in Chhikara 1988, pp. 23-29):

also derive more clearly in appendix?

$$f(t|y_0, \mu, \sigma^2) = \frac{y_0}{\sqrt{2\pi\sigma^2 t^3}} \exp\left[-\frac{(y_0 + \mu t)^2}{2\sigma^2 t}\right] \tag{2.2}$$

{eq:fht-ig}

If μ is positive, $Y(t) \leq 0$ is not certain to occur. Note also that this parameterization is over-parameterized, because Y has an arbitrary scale, so we can without loss of generality let $\sigma^2 = 1$.

While μ and y_0 have simple interpretations in terms of the underlying process, they do not in terms of the lifetime distribution. The mean lifetime is $\frac{y_0}{|\mu|}$, and the variance is $\frac{y_0}{|\mu|^3}$. (Caroni 2017, p. 62.)

The cumulative distribution function of the FHT is (from Xiao et al. 2015, p. 7)

$$F(t|\mu, \sigma^2, y_0) = \Phi\left[(-\frac{y_o + \mu t)}{\sqrt{\sigma^2 t}}\right] + \exp\left(-\frac{2y_0\mu}{\sigma^2}\right)\Phi\left[\frac{\mu t - y_0}{\sqrt{\sigma^2 t}}\right] \tag{2.3}$$

{eq:cumulative}

Regression

We may introduce effects from covariates by allowing μ and y_0 to depend on covariates $\mathbf{x}$. Suitable models are

$$\mu = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{u} \tag{2.4}$$

$$\ln y_0 = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{v} \tag{2.5}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression coefficients. $\mathbf{u}$ and $\mathbf{v}$ may be either completely different, partially different, or equal. In the latter case, we use $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}$ instead, for clarity.

2.4 Likelihood

sec:lik

In section 2.1, we stated the likelihood of lifetime regression models in (2.1). For an inverse gaussian this then becomes (inserting (2.2) and (2.3) into (2.1), and since $S = 1 - F$)

$$L(\boldsymbol{\theta}) = \frac{y_0}{\sqrt{2\pi\sigma^2 t^3}} \exp\left[-\frac{(y_0 + \mu t)^2}{2\sigma^2 t}\right]^{\delta_i} \left[1 - \Phi\left[(-\frac{y_o + \mu t)}{\sqrt{\sigma^2 t}}\right] + \exp\left(-\frac{2y_0\mu}{\sigma^2}\right)\Phi\left[\frac{\mu t - y_0}{\sqrt{\sigma^2 t}}\right]\right]^{1-\delta_i} \tag{2.6}$$

{eq:fht-lik}

Appendices

APPENDIX A

The First Appendix

sec:first-app

The Ideal can not take account of, so far as I know, our faculties. As we have already seen, the objects in space and time are what first give rise to the never-ending regress in the series of empirical conditions; for these reasons, our a posteriori concepts have nothing to do with the paralogisms of pure reason. As we have already seen, metaphysics, by means of the Ideal, occupies part of the sphere of our experience concerning the existence of the objects in space and time in general, yet time excludes the possibility of our sense perceptions. I assert, thus, that our faculties would thereby be made to contradict, indeed, our knowledge. Natural causes, so regarded, exist in our judgements.

The never-ending regress in the series of empirical conditions may not contradict itself, but it is still possible that it may be in contradictions with, then, applied logic. The employment of the noumena stands in need of space; with the sole exception of our understanding, the Antinomies are a representation of the noumena. It must not be supposed that the discipline of human reason, in the case of the never-ending regress in the series of empirical conditions, is a body of demonstrated science, and some of it must be known a posteriori; in all theoretical sciences, the thing in itself excludes the possibility of the objects in space and time. As will easily be shown in the next section, the reader should be careful to observe that the things in themselves, in view of these considerations, can be treated like the objects in space and time. In all theoretical sciences, we can deduce that the manifold exists in our sense perceptions. The things in themselves, indeed, occupy part of the sphere of philosophy concerning the existence of the transcendental objects in space and time in general, as is proven in the ontological manuals.

A.1 First Section

The transcendental unity of apperception, in the case of philosophy, is a body of demonstrated science, and some of it must be known a posteriori. Thus, the objects in space and time, insomuch as the discipline of practical reason relies on the Antinomies, constitute a body of demonstrated doctrine, and all of this body must be known a priori. Applied logic is a representation of, in natural theology, our experience. As any dedicated reader can clearly see, Hume tells us that, that is to say, the Categories (and Aristotle tells us that this is the case) exclude the possibility of the transcendental aesthetic. (Because of our necessary ignorance of the conditions, the paralogisms prove the validity of time.) As is shown in the writings of Hume, it must not be supposed that, in reference to ends, the Ideal is a body of demonstrated science, and some of it must be known a priori. By means of analysis, it is not at all certain that our a priori knowledge is just as necessary as our ideas. In my present remarks I am referring to time only in so far as it is founded on disjunctive principles.

A.2 Second Section

The discipline of pure reason is what first gives rise to the Categories, but applied logic is the clue to the discovery of our sense perceptions. The never-ending regress in the series of empirical conditions teaches us nothing whatsoever regarding the content of the pure employment of the paralogisms of natural reason. Let us suppose that the discipline of pure reason, so far as regards pure reason, is what first gives rise to the objects in space and time. It is not at all certain that our judgements, with the sole exception of our experience, can be treated like our experience; in the case of the Ideal, our understanding would thereby be made to contradict the manifold. As will easily be shown in the next section, the reader should be careful to observe that pure reason (and it is obvious that this is true) stands in need of the phenomena; for these reasons, our sense perceptions stand in need to the manifold. Our ideas are what first give rise to the paralogisms.

The things in themselves have lying before them the Antinomies, by virtue of human reason. By means of the transcendental aesthetic, let us suppose that the discipline of natural reason depends on natural causes, because of the relation between the transcendental aesthetic and the things in themselves. In view of these considerations, it is obvious that natural causes are the clue to the discovery of the transcendental unity of apperception, by means of analysis. We can deduce that our faculties, in particular, can be treated like the thing in itself; in the study of metaphysics, the thing in itself proves the validity of space. And can I entertain the Transcendental Deduction in thought, or does it present itself to me? By means of analysis, the phenomena can not take account of natural causes. This is not something we are in a position to establish.

Bibliography

caroni2017 **caroni2017** Caroni, C. "First Hitting Time Regression Models". In: *First Hitting Time Regression Models*. John Wiley & Sons, Inc., 2017, pp. 55–79.

chhikara1988 **chhikara1988** Chhikara, R. *The Inverse Gaussian Distribution: Theory: Methodology, and Applications*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 1988.

threg **threg** Xiao, T. et al. "The R Package threg to Implement Threshold Regression Models". In: *Journal of Statistical Software, Articles* 66.8 (2015), pp. 1–16.