Health & Medicine | Martin Nygård Johansen

Regression models using parametric pseudo-observations

The statistical analysis of survival data is the focus of research being carried out by Martin Nygård Johansen, a biostatistician at Aalborg University Hospital, Denmark, and his colleagues. Some patients' actual survival times are not observed during a clinical study and are said to be censored. This research offers a new way to transform such time-to-event data into a dataset without censoring. Pseudo-observations are generated using a flexible parametric model and replace the original censored observations. Extensive simulations revealed a substantial reduction in the uncertainty of the final regression estimates when compared with the traditional nonparametric method.

Medical research often involves the statistical analysis of survival data, also known as time-to-event data. Here, the outcome of interest is the expected duration of time that elapses before an event occurs or a task is completed. A unique characteristic of survival data is that the event, such as the recurrence of a tumour, discharge from hospital, or even death, will probably not have occurred for all patients by the end of the study or follow-up period. For some patients, therefore, the actual survival times remain unknown and the survival time is said to be censored. Censoring indicates that the observation period ended before the event occurred, so the researcher will not know when, or indeed if, the event occurred for these patients.

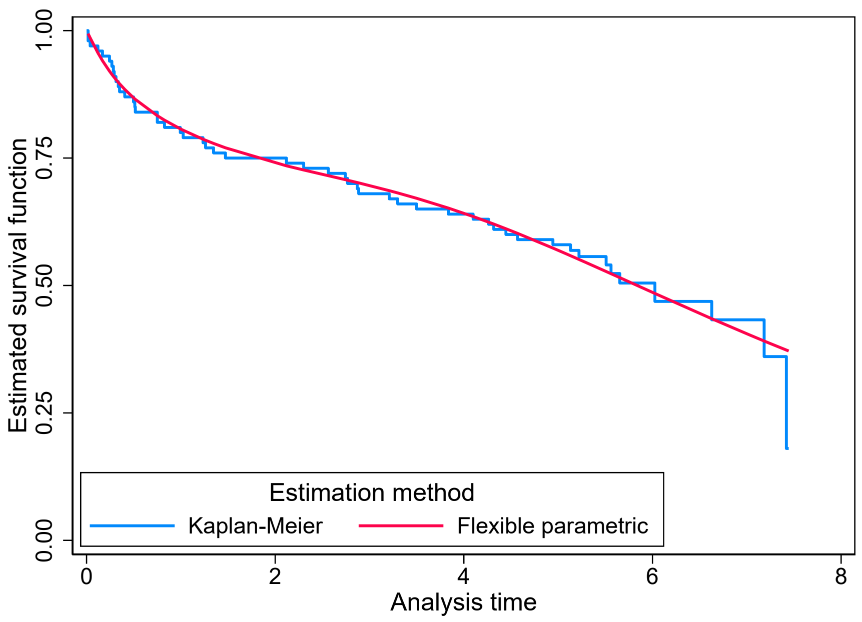
Traditional logistic and linear regression methods are not suitable for analysing time-to-event data as they cannot have both the event *and* time features as the modelled outcome. Furthermore, traditional regression methods are not equipped to handle the phenomenon of censoring, which must be accounted for in the analysis if the inferences are to be valid.

The analysis of such censored time-to-event data is the focus of research being carried out by Martin Nygård Johansen, a biostatistician at Aalborg University Hospital, Denmark, and his colleagues.

MODELLING CENSORED SURVIVAL DATA

For almost 50 years, the Cox proportional hazards model has been the predominant technique used in biostatistics to model censored survival data. The model is classified as a semiparametric model because the baseline hazard function remains unspecified. This can mean that little attention is paid to the baseline hazard.

A hazard function is used to model the distribution of data in survival analysis. It is used to model the time to an event for which the risk at any particular time point during follow-up can vary as a function of time (where time could be the patient's age). The baseline hazard is the hazard function with all of the covariates set to zero. In this application, baseline survival can almost be considered as the survival of an 'average' participant. The baseline hazard function is directly related to the time course of an illness and is likely to be of medical interest. Estimating it effectively requires a parametric model, that is a model that captures all the information required for its predictions from a finite set of parameters, which would provide precise estimates of survival probabilities. Moreover, a parametric method would provide a better understanding of the phenomenon being investigated.



The nonparametric Kaplan-Meier method estimates the unadjusted probability of surviving beyond a particular point in time by a step function. In contrast, Johansen's approach utilises a smooth parametric basis for pseudo-observations to generate more precise estimates.

PSEUDO-OBSERVATIONS

Johansen's research builds on an alternative approach that has been developed over the last 20 years. This involves transforming the censored dataset and applying a generalised linear model in order to estimate measures of association which quantify the relationship between two or more variables. He explains that the transformed data is referred to as pseudo-observations. The pseudo-observation approach involves transforming the time-to-event data into a dataset without censoring. This new dataset can then be used in place of the original censored observations.

The transformation usually takes the form of a nonparametric estimator of the cumulative incidence function. The cumulative incidence function gives the proportion of patients who have died from a particular cause at a specific time. This function also takes into account that patients can die from other competing causes during the study. The transformation is usually based on the Kaplan-Meier estimator for survival data or the Aalen-Johansen estimator in the presence of competing risks. The Kaplan-Meier method involves estimating the unadjusted probability of surviving beyond a particular point in time. The Aalen-Johansen estimator is an extension of the Kaplan-Meier method to handle competing risks.

Johansen used a typical clinical study comparing exposed and non-exposed individuals in order to evaluate the performance of nonparametric and parametric pseudo-observation approaches.

A NEW FLEXIBLE MODELLING APPROACH

Johansen has developed a different approach to calculating pseudo-observations underpinned by the theory behind the flexible parametric model for time-to-event data. Flexible parametric survival models can capture a wide range of hazard function shapes (i.e. when the hazard function is plotted) using splines to model the log-cumulative hazard function. Splines are function estimates used to smooth noisy data. Flexible parametric survival models can also include time-dependent effects for more flexibility.

> ***Combining flexible parametric modelling and pseudo-observations means that this novel modelling approach produces more precise pseudo-observations than the nonparametric approach.***

In this new approach, the baseline log cumulative hazard function is modelled using a spline function. The spline-based estimation of the hazard function is combined with the pseudo-observation approach, with parametric pseudo-observations being generated using an estimate of the cumulative incidence proportion (the proportion of participants experiencing the outcome of interest in a specific time interval). The cumulative incidence proportion is obtained from a flexible parametric model. Johansen describes how combining flexible parametric modelling and pseudo-observations means that this novel modelling approach produces more precise pseudo-observations than the nonparametric approach.

The resulting parametric pseudo-observations can be used in regression models to estimate absolute and relative association measures. While this innovative approach is fully parametric, it does not impose any distributional assumptions on the underlying time-to-event data.

SIMULATION STRATEGY

The researcher created seven different scenarios to evaluate the performance of both nonparametric and parametric pseudo-observation approaches. The general setup is a typical clinical study comparing exposed and non-exposed individuals. In each scenario, one aspect is varied. Individuals enter the study over a 6-year period. The follow-up continues for 13 years after the accrual period started, so individuals participate for between 7 and 13 years. A fixed sample size was set at n = 500. For each scenario, the overall event probability (the cumulative incidence proportion) was estimated at 10 years. The risk difference and risk ratio effect measures comparing the two exposure groups were also evaluated at 10 years. In each scenario, the researcher generated 5,000 replications of the datasets. For each sample, 50% of the subjects were considered exposed and 50% nonexposed. Then time-to-event data was simulated from exponential distributions. Uniform and independent loss-to-follow-up censoring was also included.

SIMULATION RESULTS

The simulation analysis revealed a reduction in the variability of the parameter estimates produced from the researcher's new method for calculating parametric pseudo-observations when compared with those pseudo-observations generated by the traditional nonparametric method. Johansen observed that the reduction in estimate variability depends on both the amount of additional follow-up time after the analysis time point and the size of the risk set at the analysis time point. Moreover, the observed reduction in variability translates to a reduction in the required sample size of up to 127%. This offers a significant gain in interventional studies, particularly when gathering data is expensive and/or time consuming.

MECHANISMS FOR GAINING EFFICIENCY

Johansen has discovered two mechanisms that contribute to the gain in efficiency of these new parametric pseudo-observations. Firstly, using additional information extending beyond the analysis time point; this is because information from events during the entire observed follow-up can be used to fit the spline in the parametric approach. In contrast, nonparametric pseudo-observations do not take into account any events that occur after the analysis time point. Secondly, the researcher observed that the nonparametric estimators were unstable when the risk set was very small as the size of the jumps in the nonparametric estimators depends on the size of the risk set at that particular time. This leads to greater uncertainty in the estimated cumulative incidence.

This extensive simulation study demonstrates the researcher's novel approach to calculating pseudo-observations compared to those of the traditional pseudo-observation method. The study revealed that this new method can reduce the uncertainty of the final regression estimates significantly. The researcher and his colleagues have published an example of Stata syntax demonstrating how to calculate parametric pseudo-observations and estimate cumulative incidence proportion, risk difference and risk ratio at a specific time point.

While this innovative approach is fully parametric, it does not impose any distributional assumptions on the underlying time-to-event data.

Behind the Research

Martin Nygård Johansen

E: martin.johansen@rn.dk **T:** +45 97 66 62 70 **W:** https://dk.linkedin.com/in/martinbergjohansen

Research Objectives

PhD student Johansen proposes a novel approach to calculating pseudo-observations.

Detail

Address
Sdr Skovvej 15
9000 Aalborg
Denmark

Bio
Martin Nygård Johansen is a biostatistician who has been working within medical research since 2006. He has handed in his PhD thesis concerning the use of pseudo-observations for interval censored time-to-event data, which will be defended in March 2021.

Funding
Aalborg University Hospital

Collaborators
- Erik Thorlund Parner, Aarhus University
- Søren Lundbye-Christensen, Aalborg University Hospital

References

Johansen, M., Lundbye-Christensen, S., Parner, E. (2020). Regression models using parametric pseudo-observations. *Statistics in Medicine*, 39, 2949–2961. Available at: https://doi.org/10.1002/sim.8586

Andersen, P. and Perme, M. (2009). Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19, 71–99. Available at: https://doi.org/10.1177/0962280209105020

Royston, P. and Parmar, M. (2002). Flexible Parametric Proportional-Hazards and Proportional-Odds Models for Censored Survival Data, with Application to Prognostic Modelling and Estimation of Treatment Effects. *Statistics in Medicine*, 21, 2175–2197. Available at: https://doi.org/10.1002/sim.1203

Personal Response

What inspired you to develop a new method for calculating pseudo-observations?

" The underlying motivation for this research is that we were looking for a way to calculate pseudo-observations for a more complicated data structure known as interval censored data where we only observe the event status at different time points during follow-up. In such a setting, all we know when we observe an event in a patient is that the event has occurred at some point since we last observed the patient without the event. This happens quite frequently if we study a group of patients using routine examinations after e.g. recovery from a disease or implantation of a medical device. "

AALBORG UNIVERSITY HOSPITAL