Robert Tibshirani is professor of statistics and biomedical data science at Stanford University, and winner of the International Statistical Institute's 2021 Founders of Statistics Prize (bit.ly/3GX774R).

What is Cox's proportional hazards model?

Robert Tibshirani gives an overview of one of David Cox's most widely applied ideas, for which he was awarded the International Prize in Statistics in 2017

The proportional hazards model developed by David Cox[14] is widely used for a type of problem known as *survival analysis*. Such problems concern estimating the time until a particular event occurs, such as the death of a patient being treated for a disease, or the failure of an engine part in a vehicle.

Cox's 1972 paper, which sets out his idea, is one of the most cited statistics papers of all time. The aim of this article is to explain the proportional hazards model, which as we will see is closely related to the Kaplan–Meier survival curve, from another path-breaking paper.[36]

Censored survival data

Our focus here is on the analysis of data with a special kind of outcome variable: *the time until an event occurs*.

For example, suppose that we have conducted a 10-year medical study, in which patients have been treated for cancer. We would like to fit a model to predict patient survival time, using features such as their health information at the start of the study, or the type of treatment they received. At first pass, this may sound like a regression problem. But there is an important complication: hopefully some or many of the patients have survived until the end of the study. Such a patient's survival time is said to be *censored*: we know that it is at least 10 years, but we do not know its true value. We do not want to discard this subset of surviving patients, as the fact that they have survived at least 10 years amounts to valuable information. However, it is not clear how to make use of this information.

Though the phrase "survival analysis" suggests a medical study, the applications of survival analysis extend beyond medicine. For example, an automobile company may want to model the time until failure of an engine part. The company might collect data on engine parts over some time period, in order to model each part's failure as a function of the age of the vehicle, number of miles driven and other factors. However, presumably not all parts will have failed by the end of this time period; for such parts, the time to failure is censored.

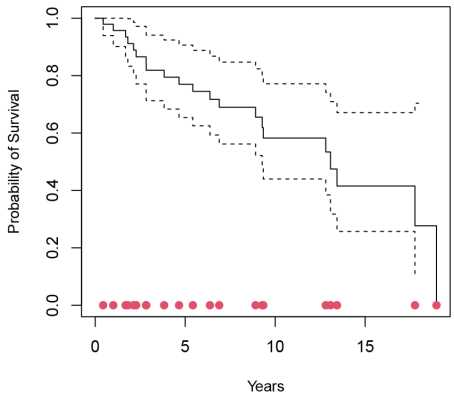

Figure 1: Kaplan–Meier survival curve (with 95% confidence bands) for a set of 50 patients. Red dots indicate the times at which a patient died.

For each engine part, or patient in a medical study, we suppose that there is a true survival time T, as well as a true censoring time C. (The survival time is also known as the failure time or the event time.) The survival time represents the time at which the event of interest occurs: for instance, the time at which the engine part fails or the patient dies. By contrast, the censoring time is the time at which censoring occurs: for example, the time at which a patient drops out of the study or the study ends.

For each part or patient, we observe either the survival time T or the censoring time C. Specifically, we observe the random variable

$$Y = \min(T, C)$$

In other words, if the event occurs before censoring (i.e., $T \leq C$) then we observe the true survival time T; however, if censoring occurs before the event ($T > C$) then we observe the censoring time. We also observe a status indicator δ, with $\delta = 1$ if $T \leq C$ and 0 if $T > C$. Thus, $\delta = 1$ if we observe the true survival time, and $\delta = 0$ if we instead observe the censoring time. Finally, our data consists of n (Y, δ) pairs for each part or patient in our study, which we denote by $(y_1, \delta_1), \ldots, (y_n, \delta_n)$.

The Kaplan–Meier survival curve

The first natural step in a survival analysis is to summarise the survival times y_i. We might start by computing the median survival time. But how do we do this: what do we do with the censored observations? The Kaplan–Meier estimator gives a nice solution. We line up the observations (say, patients) in time order, and focus on the times $t_1 < t_2 < \ldots < t_K$ at which an event occurred. At time t_1, we compute the number of patients n_1 still alive and the number who died at t_1, say d_1. Then the probability that a patient survives past t_1 is just $(n_1 - d_1)/n_1$.

Similarly, at the next event time, t_2, suppose there are n_2 patients alive and in the study just before that time and d_2 who died. Then the probability that a patient survives past t_2, given that they were alive just before time t_2, is $(n_2 - d_2)/n_2$. To be clear, the fact that someone has survived to t_2 is itself informative, which is why we ask "What is the probability you will survive past t_2, given you have survived to t_2?" and not just "What is the probability you will survive past t_2?" We continue in this way, computing a series of probabilities, each one analogous to the probability of a runner successfully jumping over a hurdle, given that they have not fallen down before that point. The Kaplan–Meier survival curve computes the product of these probabilities, reflecting the fact that to survive past time t, the patient must clear all hurdles up to and including that at time t.

Figure 1 shows a Kaplan–Meier survival curve (with 95% confidence bands) for a set of 50 patients. The median survival time – the time at which the survival probability equals 0.5 – is about 13 years in this example. The red dots indicate the times at which a patient died.

Cox's model

Moving beyond a simple summary, suppose

Mario Cortina Borja is chair of the *Significance* editorial board, and professor of biostatistics in the Population Policy and Practice Teaching and Research Department at the Great Ormond Street Institute of Child Health, University College London.

that we have predictor variables measured on each observation (e.g., age, health status, biomarkers). We would like to carry out a regression analysis to assess the effects of each predictor; if one of the predictors is one of two drugs, we would like to use regression to compare the two drugs while adjusting for confounder factors.

Cox's model is based on a quantity known as the *hazard function* $h(t \mid x)$. This is the probability that an individual with predictors x will die at time t, given that the individual is alive just before t. Cox's regression model starts with an assumption of *proportional hazards*: $h(t \mid x) = h_0(t)\exp(x\beta)$. This says that the hazard for an individual with predictors x is the product of a baseline hazard $h_0(t)$ (corresponding to $x = 0$) and a factor that depends on x and the regression parameters β.

To estimate β, Cox invented *a partial likelihood*: it uses the same "product in time" construction as in the Kaplan–Meier method, computing the probability under the model of surviving past each event time t_k as a product of conditional probabilities. This is a function of the unknown regression parameters β, and this function is maximised (with a numerical optimiser) to yield the partial maximum likelihood estimates $\hat{\beta}$. Furthermore, this elegant construction does not require specification of the baseline hazard $h_0(t)$, making the method flexible and robust.

With Cox's proportional hazards model we obtain all of the things that we enjoy in regression analysis, such as parameter estimates, standard errors and confidence intervals. But there is much more: it provides a natural way of modelling *time-dependent covariates*, where a patient characteristic (such as blood pressure) is measured not once at the beginning of a study but many times over the course of a study. And it can deal with other forms of censoring such as left and interval censoring. The method has also been generalised to handle high-dimensional data, incorporating sparsity penalties such as the lasso.[37,38] And there are even deep learning versions of Cox's method.[39]

The impact of this method on many fields, especially medical sciences, has been enormous; it surely will continue to grow in importance in the future. ■

Disclosure statement
The author declares no competing interests.

On Sir David Cox's publications

Mario Cortina Borja, **Julian Stander**, and **Giovanni Sebastiani**

analyse some of the tens of thousands of citations for the hundreds of papers produced by David Cox in a publishing career that spanned more than 70 years

Sir David Cox's publishing career was remarkable for the extent and influence of its contributions, as well as for its duration; his works have been published continually from 1947 to 2021. David's website at Nuffield College, Oxford, lists 385 publications (bit.ly/3M9eIB4) and we shall refer to this collection as N385. These numbers are impressive on their own, but we should also remember and celebrate David's unfailing kindness to generations of statisticians and other scientists, of which there are many accounts in this issue of *Significance*. Here we examine in more detail the extent and influence of David's work by focusing on trends in the number of references or mentions of his works in scientific publications. We will pay particular attention to his paper on "Regression models and life-tables",[14] which we will refer to as Cox (1972).

Data

The publications in N385 comprise 302 peer-reviewed articles, 23 books, and 60 entries classified as chapters in edited books, entries in encyclopaedias, proceedings, conference proceedings, reports, or lectures. (Publications are numbered from 1 to 384 on the website, but there are two entries with the number 191.) This publication list is incomplete as it does not include, for example, David's many contributions to the discussions of Royal Statistical Society read papers. His peer-reviewed output appeared in 104 different journals. There were 56 papers in the *Journal of the Royal Statistical Society* (*JRSS*; 35, 13, 7, and 1 in *Series B*, *A*, *C*, and *D*, respectively), 60 in *Biometrika*, and 24 in International Statistical Institute publications.

We also analysed data from two sources: Web of Science (WoS), which claims to be "the world's most trusted publisher-independent global citation database", and PubMed (PM), which "comprises more than 33 million citations for biomedical literature from MEDLINE, life science journals, and online books". These data were downloaded between 5 and 21 February 2022.

For each year after publication, we obtained information on *citations* from WoS, where the number of citations of one of David's outputs is the number of WoS papers listing the output as a reference. Similarly, PM yielded data on *mentions*, that is, the number of PM papers containing text referring to the output. A paper that contributes a mention does not necessarily contribute a citation. Below we consider mentions of Cox (1972), identified by searching on {"Cox regression" OR "Cox's regression" OR "Cox proportional hazard*" OR "Cox's proportional hazard*"}.

Trends in citations

WoS returned 74,406 citations for papers in N385. Figure 1 shows the cumulative number of citations for David's eight most cited papers. The two most cited papers, Cox (1972) and the Box–Cox (1964) paper on the analysis of transformations,[12] were published in *JRSS Series B*. These are followed by two papers with several co-authors in the *British Journal of Cancer* on the design and analysis of randomised clinical trials,[40,41] and David's (1975) *Biometrika* paper in which partial likelihood is defined.[15] This list concludes with three more *Series B* articles: a pivotal paper defining logistic regression for binary sequences in 1958;[5] a 1987 paper with Nancy Reid on parameter orthogonality and approximate conditional inference;[18] and a 1968 work with E. Joyce Snell proposing a general definition of residuals for linear models.[42] The sharp increase in the last five years of citations of David's (1958) analysis of binary sequences paper is driven by references made in areas including bioinformatics, machine learning, remote sensing, and neuroscience where logistic regression is mostly used as a classification method.