AUTOMATED REQUEST TO THE BRITISH LIBRARY DOCUMENT SUPPLY CENTRE
1776661/KUL-M FXBK99
EMAIL REQUEST

B COPY

ARIEL

90288034

R.D. 03-Dec-04145

P.D. 03-Dec-04150 | Customer Code No. 05-0370 | Shelfmark 3287.223046

33

CME JOURNAL OF GYN...ONCOLOGY (HUNGARY)
ISSN 1219-9087
1999 VOL.4 NR.X P.262-270
DE BRABANDER, J,VERGOTE, I ET AL
UNIVARIATE AND MUL...RESSION ANALYSES..

OS FLOOR 6-2
CME JOURNAL OF GYNECOLOGIC ONCOL..
SEE ALSO CARD INDEX.
D1 S0 1999

05-0370

90288034

for RMS use only

* **********
* ATTN: LUC COOREMAN
* EMAIL: VIRLIB@LIB.AU.AC.BE T
* ---
* **********064

ARIEL

International Loan, Return Airmail within 4 weeks of date of receipt unless recalled earlier.

Request Ref. No. 1776661/KUL-M FXBK99 T

If no other library indicated please return loan to:- 064
The British Library Document Supply Centre, Boston Spa, Wetherby, West Yorkshire, United Kingdom, LS23 7BQ

BRITISH LIBRARY

DELIVERING THE WORLD'S KNOWLEDGE

This document has been supplied by the British Library
www.bl.uk

The contents of the attached document are copyright works. Unless you have the permission of the copyright owner, the Copyright Licensing Agency Ltd or another authorised licensing body, you may not copy, store in any electronic medium or otherwise reproduce or resell any of the content, even for internal purposes, except as may be allowed by law.

If the request form above has the word 'Copyright' written in large text, it was supplied under our Copyright Fee Paid service. You are therefore agreeing to the terms of supply for our Copyright Fee Paid service, available at **www.bl.uk/services/document/edd.html**

If the request form above does not have the word 'Copyright' written in large text, it was supplied through our Library Privilege service. You are therefore agreeing to the terms of supply for our Library Privilege service, available at **www.bl.uk/services/document/LPS.html**

Supplied by the British Library - "The world's knowledge" www.bl.uk

Univariate and multivariate regression analysis: some basic statistical principles

JOS DE BRABANTER[1], IGNACE VERGOTE, M.D.[2], HERMAN VERRELST[1], DIRK TIMMERMAN, M.D.[2], SABINE VAN HUFFEL[1], JOOS VANDEWALLE[1]

Department of Electrical Engineering (ESAT-SISTA/COSIC)[1], Catholic Universityt Leuven, Department of Gynecologic Oncology[2], University Hospitals, Leuven.

ABSTRACT In this review we have summerized the basic statistical principles for univariate and multivariate analysis. First, the different types of relations between variables, data structure, the types of variables and regression techniques are presented. Thereafter, four essential steps in the model building process are described: data collection, reduction of the number of independent variables, model refinement and selection, and finally model validation. Important issues such as the number of patients needed per variable in the pool are also discussed. The rules for using linear or logistic regression are extensively reviewed in the text. In the case of survival analysis, two possibly complicating features of the data set are censoring and time-dependent covariates. When analyzing a database it is important to know whether left or right censoring and random or interval censoring have been used. Survival can be analyzed as a survivor function or a hazard function. Some basic rules on the use of the Kaplan-Meier method, life-table method, parametric regression model, or the Cox regression model for survival analysis are explained .

Key words statistical principles, univariate regression, multivariate analysis

INTRODUCTION When evaluating the role of new prognostic variables some basic knowlegde of univariate and multivariate analyses is essential. In this chapter we shall present some basic rules and steps for performing statistical analyses in general, and more specific in the field of gynecologic oncology.

Address correspondence to:

Ignace Vergote, M.D.
Department of Gynecologic Oncology
University Hospitals Leuven
Gasthuisberg, B-3000, Leuven, Belgium
Phone (32 16) 344635 Fax (32 16) 344629
E-mail Ignace.Vergote@uz.kuleuven.ac.be

PRELIMINARIES

DATA STRUCTURE The raw input is usually a $n \times p$ table (n rows by p columns) of real numbers called a data matrix. The matrix summarizes observations made on n patients. Each of the n patients are characterized with respect to p variables. The values of the variables may represent the measurement of a quantity or a numerical code for a classification scheme. The following data matrix example illustrates the variety of data types that may be encountered.

Example: A hospital surgical unit is interested in predicting the survival of patients that have undergone a particular type of liver operation. A random selection of n patients is available for analysis. For each patient, the following information is extracted from the preoperative examination.

X_1 : sex (male = 0, female = 1)
X_2 : blood-clotting score
X_3 : age of the patient
X_4 : enzyme function test score
X_5 : liver function test score
X_6 : alcohol intake (no = 1, normal = 2, alcoholic = 3)

These constitute the potential independent variables (predictor variables) for a predictive regression model. The dependent variable (response variable) is the survival time, obtained through a follow-up study. The data are the values given to the potential independent variables and the dependent variable, as illustrated in *Table 1.*

Table 1. Potential independent variables and dependent variable - surgical unit example

Patient number	X1	X2	X3	X4	X5	X6	Survival time (weeks)
1	0	6.7	62	81	2.59	1	200
2	0	5.1	59	66	1.70	2	101
3	1	7.4	57	83	2.16	2	204
n	1	8.8	48	72	2.20	3	313

TYPES OF VARIABLES Variables can be classified as being quantitative (e.g. age, weight, etc.) or qualitative (e.g. marital status, sex, alcohol intake, hair color, nationality, etc.). Obtaining values for a quantitative variable involves measurement along a scale and a unit of measure. When the units of measure are infinitely divisible (e.g. meters, millimeter, liver function test score...) the variable is said to be continuous. In the case of an indivisible unit of measure (e.g. family size), the variable is

Supplied by the British Library - "The world's knowledge" www.bl.uk

considered to be discrete. Scales of measurement can also be classified on the basis of the relationship between among the elements composing the scale. A *ratio scale* has the following properties: (a) any two values along the scale may be expressed in a meaningful way as a ratio, (b) the distance between items on the scale is meaningful, and (c) the elements along the scale can be ordered from low to high; e.g. weight is usually measured on a ratio scale. An *interval scale*, unlike a ratio scale, does not have a fixed origin: e.g. elevation and temperature are measured relative to a fixed point (sea level or freezing point of water). An interval scale only satisfies the properties of (b) and (c), but not (a).

An *ordinal scale* is one in which only property (c) is present; e.g. the grades A, B, C, D, can be ordered from highest to lowest, but we cannot state that the difference between A and B is equivalent to the difference between B and C. The fourth type of scale, namely the *nominal scale*, corresponds to qualitative data. An example would be the variable 'alcohol intake' which has the categories 'no', 'normal', 'alcoholic'. The three categories can be indicated by a coded value such as 1, 2 or 3, but since these coded values are numerical, they should not be considered as quantitative. None of the three properties listed above can be applied to the coded data.

CLASSIFYING REGRESSION TECHNIQUES Once the data have been collected (random sample from a population), one must go through the following sequence of steps to identify the appropriate regression technique *(Figure 1)*. From the problem context, to identify the independent variables of interest and the dependent variable, to specify the aim of the analysis; to identify the type of the dependent variable, if the aim is to model and investigate the relationship between two or more variables. This type determines the kind of regression model: simple or multiple regression ($y \in \Re$), or logistic regression ($y \in (0,1)$). If the aim is to study the occurrence and timing of events, survival analysis is in place.

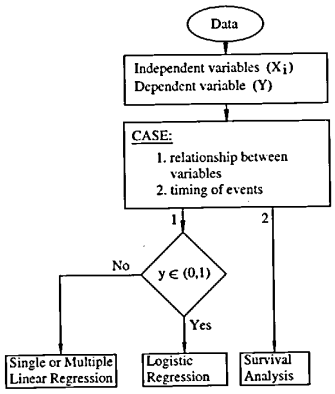


Figure 1. Classifying of regression techniques.

MODEL BUILDING In this section, we describe each of the following major steps in the model-building procedure *(Figure 2)*: 1. data collection and preparation, 2. reduction of the number of independent variables, 3. model refinement and selection, and 4. model validation.

DATA COLLECTION AND PREPARATION Having decided on which variables to measure, an experimental design must be set up which outlines how the data are to be obtained. This design must conform to the theory and practice of survey sampling and experimental design (1). A major characteristic of the quality of a research design is the consistency of the data. First of all, the consistency between different samples from the same population, and second, between measurements on the same sample but at different times. The majority of analy-

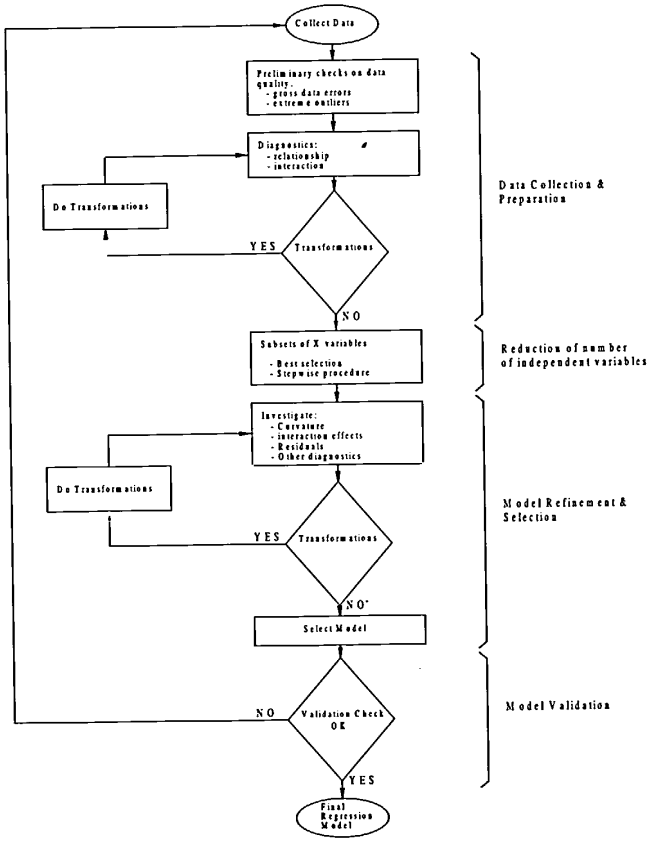


Figure 2. Strategy for building a regression model

Supplied by the British Library - "The world's knowledge" www.bl.uk

sis techniques assume that the selected objects (patients) represent a random sample from some well-defined population of objects (patients). Alternative probability sampling methods exist (e.g. stratified sampling), but adjustments have to be made to the standard inference procedures.

NUMBER OF PATIENTS The number of patients to be used for a regression study depends on the size of the pool of independent variables at the stage. When the pool of potentially useful independent variables is large, more cases are required than when the pool is small. A general rule of thumb states that there should be at least 6 to 10 patients for every variable in the pool. The requirement to have more cases for larger pools of independent variables also affects the model-building strategy. For example, if there is a large pool of potential independent variables, it may be difficult to investigate all two-variable interactions, whereas this might be feasible if the pool is small.

Once the data have been collected, edit checks and plots should be made to identify gross data errors, as well as extreme outliers (2-3). Data errors should be corrected or resolved before the model building begins. Once the data have been properly edited, the formal modeling process can begin.

IDENTIFYING INDEPENDENT VARIABLES A variety of diagnostics should be applied to identify the important independent variables, the functional forms in which these independent variables enter the regression model, and the important interactions between them. Scatter plots are useful for pointing out relationships and their strengths. Selected independent variables can be fitted in regression functions to explore relationships, possible strong interactions and transformations (4-6). Residual plots and partial regression plots should also play a major role here (7-9). Whenever possible, one should also rely on the investigator's prior knowledge and expertise to select the appropriate transformations and interactions to be investigated.

CODING OF QUALITATIVE INDEPENDENT VARIABLES It is inappropriate to model a nominal scaled variable as if it were numerical. To cope with this kinds of variables, indicator (or dummy) variables have to be introduced (10). For example, we may have variables that denote the hospital in which the patient was treated or the race of the patient. Each of these variables has a fixed number of discrete outcomes and the scale of measurement is nominal. Therefore, we must form a set of design variables (dummy variables) to represent the categories of the variable. Let's assume that in a study, the variable 'race' is coded at four levels (white = 1, black = 2, Hispanic = 3 and others = 4). *Table 2* gives the specification of the design variables for 'race' using white as the reference group.

Table 2. Specification of the design variables. White is the reference group.

	Design variables		
Race (code)	D_1	D_2	D_3
White (1)	0	0	0
Black (2)	1	0	0
Hispanic (3)	0	1	0
Others (4)	0	0	1

In general, if a nominal scale variable has k possible values, then k-1 design variables will be needed. The method of coding design variables which use a reference group is the most commonly employed method appearing in the literature.

REDUCTION OF THE NUMBER OF INDEPENDENT VARIABLES Once the investigator has decided on the functional form of the regression relation (whether given variables are to appear in linear form, quadratic form, etc. and whether any interaction terms are to be included), the next step is to select a few "good" subsets of X variables.

The reason for focusing on subsets of the pool of independent variables is that typically, the number of independent variables remaining after the initial screening is still large. Furthermore, very often these variables will be highly intercorrelated. The presence of many highly intercorrelated independent variables may add little to the predictive power of the model while it substantially increases the sampling variations of the regression coefficients, thereby reducing the model's descriptive abilities and increasing the problem of round-off errors.

A variety of computerized approaches have been developed to assist the investigator in reducing the number of independent variables. We describe two of these approaches in this section.

The first type of approach utilizes an automatic search procedure to arrive at a single subset of independent variables. This approach is recommended primarily for reductions involving large pools of independent variables. Forward-backward (or stepwise) selection for example, is based on a statistical algorithm that checks for the importance of variables, defined in terms of measuring the statistical significance of the variable (11). The algorithm includes or excludes variables on the basis of a fixed decision rule, e.g. at any step in the procedure, the most important variable will be the one that produces the largest change in log-likelihood relative to a model not containing the variable.

The second approach is applicable only to pools of independent variables that are small or moderate in size. It considers

Supplied by the British Library - "The world's knowledge" www.bl.uk

all possible regression models that can be developed from the pool of potential independent variables and identifies subsets of the independent variables that are "good" according to a criterion specified by the investigator.

While computerized approaches can be very helpful in identifying appropriate subsets, the process of developing a useful regression model must be pragmatic and needs to utilize large doses of subjective judgement (12). Independent variables that are considered to be essential should be included in the regression model before any computerized assistance is sought.

MODEL REFINEMENT AND SELECTION After reducing the number of independent variables by a computerized selection process, the investigator is left with a small number of potentially "good" regression models. At this stage, more detailed checks of curvature and interaction effects are desirable. Residual plots and partial regression plots are helpful in deciding whether one particular model is to be preferred. In addition, diagnostic checks to identify influential outlying observations (13-14), multicollinearity (8) and unequal error variances (6) are useful. Eventually, after checking and various remedial actions such as transformations, the investigator selects one regression model as the best. At this point, it is good statistical practice to validate the model.

MODEL VALIDATION The final step in the model-building procedure is the validation of the selected regression model. Model validation usually involves checking the model against independent data. Ideally, this model validation is prospective to examine whether the regression model properly generalizes the unseen data. This would strengthen the assumption that the model truly captures the properties of the underlying population and not only those of the retrospective data set.

A variety of methods exist to examine the validity of the regression model against the new data. One validation method is to re-estimate the model form chosen earlier, using the new data. The estimated regression coefficient and various characteristics of the fitted model are then compared for consistency with those of the regression model based on the earlier data. If the results are consistent, the chosen regression model is applicable under broader circumstances than those related to the original data. Another validation method is to re-estimate from the new data, all of the "good" models that had been considered in order to see whether the selected regression model is still the preferred model according to the new data.

Often it is neither practical nor feasible to collect new data. A reasonable alternative if the data set is large enough, is to then split the data into two sets. The first, called the model-building set, is used to develop the model. The second data set, called the validation or prediction set, is used to evaluate the generalization and predictive power of the selected model. This validation procedure is often called cross-validation.

MULTIPLE LINEAR REGRESSION We consider a general regression model with p independent variables and a linear regression function. This model can be described as follows:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \qquad (I)$$

where y is the response or dependent variable, $\beta_0, \beta_1, \ldots, \beta_p$ are unknown parameters, $x_0, x_1, \ldots, x_p$ are the regression or independent variables and is a random error term. *Figure 3* gives an example of the so-called response-plane in the case of two independent variables.

Regression model (I) is said to be "multiple," linear in the parameters, and linear in the independent variables. It is "multiple" because there are more than one independent variables, "linear in the parameters" because no parameters appear as an exponent or as part of a multiplication with other parameters, and "linear in the independent variables" because the variables appear only in the first power. A model that is linear in the parameters and in the independent variables is also called a first-order model.

Least squares is the most commonly used technique to estimate the parameters in the linear regression model (15-16). The goal is to find estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_p$ that minimize the sum of the squared differences between the actual y value and the value of y that is predicted by the equation. These estimates are called the least-squares estimates and the quantity minimized is called the error sum of squares. Essentially, the least-squares technique assumes that: 1. the expected values of the errors are zero, 2. that the variances of the errors (and thus the response variable) are constant across observations, 3. that the errors are uncorrelated, and 4. the errors are normally distributed (for hypothesis testing).

THE MEANING OF THE REGRESSION COEFFICIENTS The parameter β_0 is the *y*-intercept of the regression plane. The parameter β_1 indicates the change in the mean response per unit increase in X_1 when the other covariates are held constant. Therefore, the parameters $\beta_0, \beta_1, \ldots, \beta_p$ are called partial regression coefficients because they reflect the partial effect of one independent variable when the other independent variables included in the model are kept constant.

LOGISTIC REGRESSION

STATISTICAL BACKGROUND Logistic regression analysis, models the relationship between a binary or ordinal response variable

Supplied by the British Library - "The world's knowledge" www.bl.uk

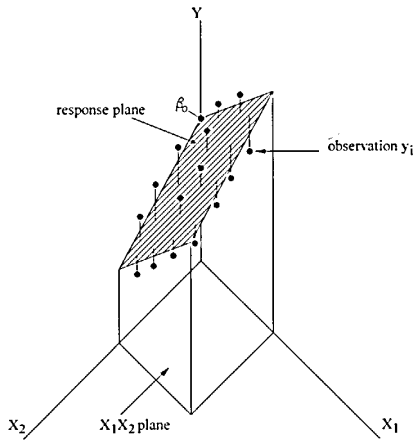


Figure 3. Example of response surface - a response plane with observations scattered about it

(dependent variable) and one or more explanatory variables (independent variables). The difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. The first difference concerns the nature of the relationship between the outcome and independent variables. The specific form of the logistic regression model is as follows:

$$y = \frac{\exp[\beta_0 + \beta_1 X_1 + ... + \beta_p X_p]}{1 + \exp[\beta_0 + \beta_1 X_1 + ... + \beta_p X_p]} + \varepsilon \qquad \text{(II)}$$

The logit transformation of the expected value of response y, $E\{y\}$, is defined as follows:

$$g(x) = \ln\left[\frac{E\{y\}}{1-E\{y\}}\right] = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \qquad \text{(III)}$$

The importance of this transformation is that the logit has many of the desirable properties of a linear regression model. The logit is linear in its parameters, may be continuous and may range from - ∞ to + ∞, depending on the range of x.

The second important difference between the linear and logistic regression models concerns the conditional distribution of the outcome variable. The quantity ε has a distribution with mean zero and some variance proportional to the first term of equation (II), and the conditional distribution of the outcome variable follows a binomial distribution.
For binary response models, the response y of an individual can take on one of two possible values, denoted by 0 and 1 (for example, $y = 1$ if a disease is present, otherwise $y = 0$) (17-20).

For ordinal response models, y is restricted to one of a small number of k possible outcomes. For example, the response has three possible outcomes: normal = 1, early disease state = 2, and advanced disease state = 3. Many researchers have considered this model (19, 21).

INTERPRETATION OF THE COEFFICIENTS OF THE LOGISTIC REGRESSION MODEL

UNIVARIATE LOGISTIC REGRESSION MODEL

DICHOTOMOUS (BINARY) INDEPENDENT VARIABLE We assume that x is coded as either zero or 1. The odds ratio (is a measure of association) denoted by Ψ, is defined as the ratio of the odds for $x = 1$ to the odds for $x = 0$, and is given by the equation $\Psi = \exp(\beta_1)$. For example, if y denotes the presence or absence of lung cancer and if x denotes whether or not the person is a smoker, then $\Psi = 2$ (odds ratio = 2) indicates that lung cancer occurs twice as often among smokers than among non-smokers in the study population.

POLYTOMOUS INDEPENDENT VARIABLE The notion of control group stands central in the method of coding the indicator variables used in data collection and preparaton. With this coding method, the odds ratios (calculated in the same way as above) estimate the risk of an exposed group relative to that control group.

CONTINUOUS INDEPENDENT VARIABLE Under the assumption that the logit is linear in the continuous variable x, it follows that the slope coefficient β_1 gives the change in the odds for an increase of one unit in x. Most often, this unit-value will not be of much interest for interpretation. For example, an increase of 1 year in age or of 1 mm Hg in systolic blood pressure may be too small to be considered important. A change of 10 years or 10 mm Hg might be considered more useful. To provide a useful interpretation for continuous scaled variables, one needs to develop a method for point and interval estimation for an arbitrary change of 'c' units in the variable. The odds ratio for a change of 'c' units in x is obtained by $\Psi(c) = \exp(c\beta_1)$. As an example, consider the univariate logistic regression model of age on coronary heart disease. The estimated odds ratio for an increase of 10 years in age is 3. This indicates that for every increase of 10 years in age, the risk of coronary heart disease increases 3 times.

MULTIPLE LOGISTIC REGRESSION In the previous section, we discussed the interpretation of an estimated logistic regression coefficient in case of a single variable. When multiple covariates are present, it is bad practice to analyze their risk factors individually by making a series of separate univariate regression models. Joining the covariates in one overall model allows for statistical adjustment of the effects of each of the covariates. The estimated coefficients in the multiple model will provide an estimate of the odds (calculated in the

Supplied by the British Library - "The world's knowledge" www.bl.uk

same way as in the univariate case) adjusted for all other variables included in the model.

SURVIVAL ANALYSIS

INTRODUCTION Survival analysis is a class of statistical methods to study the occurrence and timing of events. We define an event as a qualitative change that can be situated in time. A disease for example, is a transition from the state of non-disease to the state of disease.

To apply survival analysis, one needs to know 'when' the change occurred. Ideally, the transition occurs virtually instantaneously, and you know the exact time it occurred. Some transitions may take a little time and the exact time of onset may be unknown or ambiguous. If the event of interest is a political revolution for example, one might only know the year in which it began. Large transition periods are no problem as long as the interval during which the event occurs is short, relative to the overall duration of the observation.

For survival analysis, the best observation plan is prospective. It begins observing a set of individuals at some well-defined point in time and follows them for a substantial period, recording the times at which the events of interest occur. It is not necessary for all individual patients to experience the event. For some applications, you might also want to distinguish different kinds of events, but specialized methods of survival analysis are needed to handle these data appropriately.

A common aim of survival analysis is to estimate the risk of an event depending on covariates. If this is the goal, the data set must contain measurements of the covariates. Some of these covariates, such as race and sex, may be constant over time. Others like blood pressure or weight may vary with time. For time-varying covariates, the data set should include as many details as possible on their temporal variation.

Survival data have two common features that are difficult to handle with conventional statistical methods: censoring and time-dependent covariates. All methods of survival analysis allow for censoring and many also allow for time-dependent covariates. One of the confusing things about survival analysis is that there are so many different methods: life table, Kaplan-Meier estimators, exponential regression, log-normal regression, proportional hazards regression, competing risks models and discrete-time methods, to name only a few. More confusing is that sometimes these methods are complementary in some sense and can be used next to each other.

BASIC CONCEPTS OF SURVIVAL ANALYSIS In this section, we discuss several topics that are common to many different methods of survival analysis: censoring, the survivor function and the hazard.

CENSORING Censoring comes in many forms and occurs for many different reasons. The most basic distinction is between left censoring and right censoring. Cases are right censored when the observation is terminated before the event occurs. Left censoring is most likely to occur when the observation starts at a sample at a time when some of the individuals may already have experienced the event. Most computer programs for survival analysis do not allow for left censored data. Interval censoring combines both right and left censoring. For example, a sample of people are tested for HIV infection. If a person who was not infected by the end of year 1 is found to be infected by the end of year 2, the time of infection is interval censored between 1 and 2. When all observations are interval censored and the intervals are equally spaced, it is often convenient to treat such data as discrete-time data.

The distinction is important because some types of censoring require potentially dubious assumptions in order to justify the use of specific models, while other types do not impose these restrictions.

The simplest and most common situation is depicted in *Figure 4.* This figure depicts some of the data from a study in which all patients undergo surgery at time t = 0 and are followed for 60 months thereafter. The horizontal axis represents time. Each of the horizontal lines labeled A through D represent a single person. An "x" indicates that a disease occurred at that point in time. The vertical line at 60 is the point at which we stop following the patients. Any disease occurring at time 60 or earlier is observed, and those disease times are uncensored. Any diseases occurring after 60 are not observed, and those disease times are censored at time 60.

Therefore, persons B and C have uncensored disease times, while persons A and D have right-censored disease times. Observations that are censored in this way are referred to as

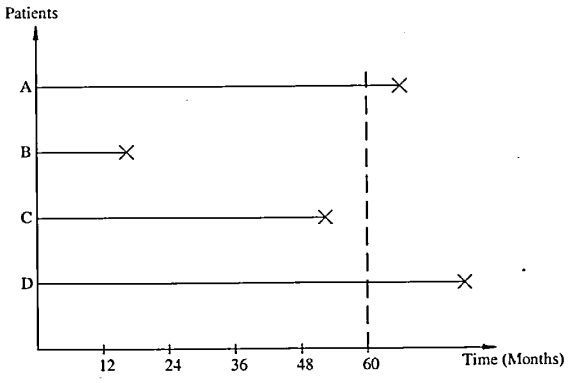


Figure 4. Singly right-censored data

Supplied by the British Library - "The world's knowledge" www.bl.uk

singly type I censored. "Type I" means that the censoring time is fixed (that is, under the control of the investigator), and "singly" refers to the fact that all observations had the same censoring time. Type II censoring occurs when the observation is terminated after the occurrence of a pre-specified number of events.

Random censoring occurs when observations are terminated for reasons that are not under the control of the investigator. These forms of censoring are depicted in *Figure 5*, where the "0" for patients E and F indicates that the observation is censored at that point of time.

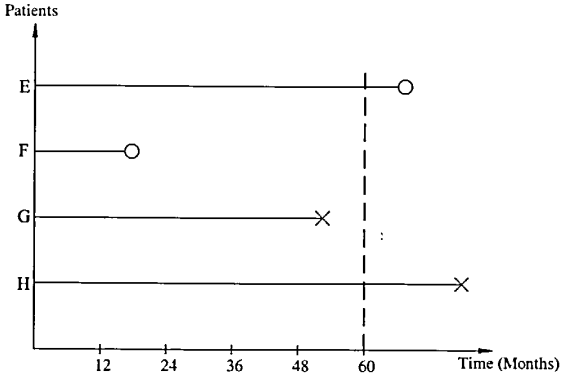


Figure 5. Randomly censored data.

Random censoring can also arise when there is a single termination time, but entry times vary randomly among individuals. Consider again the example in which patients are followed from surgery until disease. A more likely scenario is one in which patients undergo surgery at various points in time, but the study has to be terminated on a single date. All patients without disease on that date are considered censored, but their survival times starting from the moment of surgery will vary. This censoring is considered random because the entry times typically are not under the control of the investigator.

Standard methods of survival analysis do not make a distinction between type I, type II, and random censoring. They are all treated as generic right-censored observations. If only type I or type II censoring occur, everything is in order. The maximum likelihood and partial likelihood methods handle these types of censoring with negligible bias. Things are not so simple with random censoring. Standard methods require that random censoring be non-informative (22). Informative censoring can, at least in principle, lead to severe bias, but it is difficult in most situations to gauge the magnitude or direction of this bias.

An easy solution to random censoring caused by random entry times, is to include entry time as a covariate in a regression model. This solution should work well in most situations, but can lead to computational difficulties if a large part of the observations are censored.

No statistical test exists for checking whether censoring is informative or non-informative. The best solution is some kind of sensitivity analysis, the basic idea of which is to redo the analysis under two extreme assumptions on censored cases. One assumption is that censored observations experience events immediately after they are censored. This corresponds to the hypothesis that censored cases are those that tend to be at high risk of an event. The opposite assumption is that censored cases have longer times to events than anyone else in the sample has. This corresponds to the hypothesis that censored cases are those that tend to be at low risk of an event.

The results of these two "extreme" models then have to be compared to the original model. If, for each of the covariates, the two extreme estimates bracket the original estimate, we can be reasonably confident that treating observations terminated for reasons that are not under control of the investigator as non-informative censoring, has no appreciable effect on the conclusions.

In conclusion, when designing and conducting studies, the amount of random censoring should be reduced as much as possible. Covariates that are likely to affect the rate of censoring, should be measured and included in the model whenever possible. In studies with high levels of random censoring, less confidence should be given to the results than estimated by the calculated confidence intervals.

SURVIVAL DISTRIBUTIONS All standard approaches to survival analysis are probabilistic or stochastic. The times at which events occur are assumed to be realizations of some random process. This implies that the event time T for some particular individual is a random variable having a probability distribution (23).

SURVIVOR FUNCTION (S(t)) The cumulative distribution function (c.d.f.) of a variable T, denoted by $F(t)$, is a function describing the probability that the variable will be less than or equal to any chosen value ($F(t) = \text{Prob}\{T \leq t\}$). In survival analysis, we work with a closely related function called the survival function, defined as $S(t) = \text{Prob}\{T > t\} = 1-F(t)$. If the event of interest is a disease, the survival function gives the probability of surviving beyond t. *Figure 6* plots the survivor function for a fictitious data set. For example, at 24 months the probability of surviving is 0.6. This means that a patient has a 60% chance of surviving at least 24 months.

Supplied by the British Library - "The world's knowledge" www.bl.uk

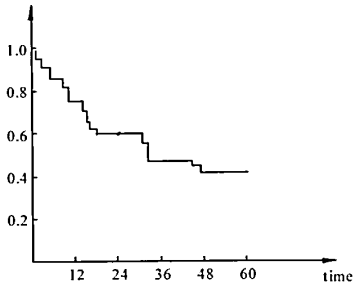


Figure 6. Plot of a survivor

HAZARD FUNCTION (H(t)) The hazard function is much more than a convenient way of describing a probability distribution. Roughly speaking, at any time instance t points out the risk of event occurrence at that time. More precisely, the hazard is a quantity expressing "the number of events per interval of time." The interpretation of the value of the hazard depends on the time scale. If for example, my hazard for death is 0.018 per year, then I can expect to live another 1/0.018 = 55.5 years. *Figure 7* plots hazard estimates for a fictitious data set of surgery. The most striking fact about this example is the rapid decline in the hazard of death from the origin to about 200 days after surgery after which the hazard fairly stabilizes. For the interval (0-50 days), the hazard is 0.0055 (evaluated at the midpoint of the interval) and for the interval (50-100 days) the hazard is 0.0040, etc.

HOW TO CHOOSE A METHOD This section discusses survival analysis with the SAS software package. It is not intended to explain the procedures of the SAS system thoroughly. Before using them, the official documentation in the *SAS/STAT User's Guide, Version 6, Fourth edition, Volume 1 and Volume 2* or in the appropriate technical report should be consulted.

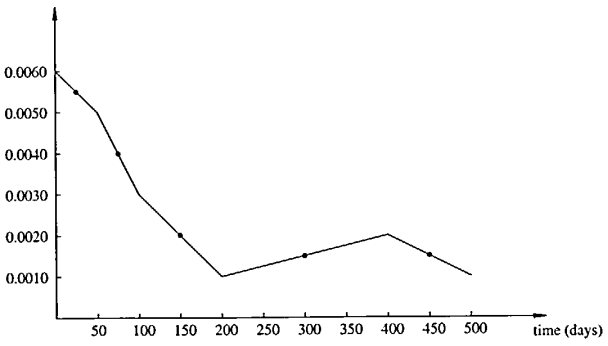


Figure 7. Hazard plot for a fictive dataset

Figure 8 presents the strategy for choosing a method of survival analysis that is SAS. In biomedicine, the Kaplan-Meier (KM) estimator is the most widely used method for estimating survivor functions (24). This method is most suitable for smaller data sets with precisely measured event times. The life-table or actuarial method may be better for large data sets or when the measurement of event times is crude, because the KM method then produces tables that may be too long for presentation and interpretation.

For estimating the effects of variables, *Cox* regression (25) has become the standard, but the only SAS procedure that allows left censoring is the parametric regression model. With the parametric model, you can test certain hypothesis about the shape of the Hazard function while Cox regression model only gives non-parametric estimates of the survivor function, which can be difficult to interpret.

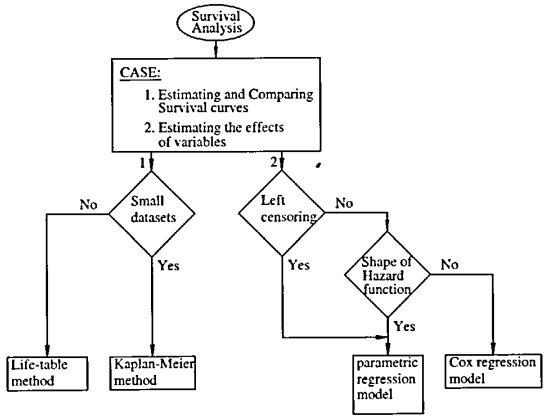


Figure 8. Strategy for choosing a method of survival analysis

REFERENCES

1. Box GEP, Hunter WG, Hunter JS. Statistics for experimenters New York, John Wiley and Sons. 1978.

2. Barnett V, LewisT. Outliers in statistical data. 2nd ed. New York, John Wiley and Sons. 1984.

3. McGill R, Tukey JW, Larsen W. Variations of box plots. Amer Stat 1978; 32:12-16.

4. Box GEP, Cox DR. An analysis of transformations. J Royal Stat Society 1964; B26:211-243.

5. Emerson JD, Stoto MA. Exploratory methods for choosing power transformations. J Amer Stat Ass 1982; 77:103-108.

6. Leinhardt S, Wasserman SS. Exploratory data analysis: An introduction to selected methods Sociological Methodology, ed. K.F. Schuessler, San Francisco, Jossey-Bass 1979:311-365.

7. Larsen WA, McCleary SJ. The use of partial residual plots in regression analysis. Technometrics 1972; 14,781-790.

Supplied by the British Library - "The world's knowledge" www.bl.uk

8. Belsley DA, Kuh E, Welsch RE. Regression diagnostics: Identifying influence data and sources of collinearity. New York, John Wiley and Sons, Inc. 1980.

9. Chambers JM, Cleveland WS, Kleiner B. et al. Graphical methods for data analysis Belmont, CA, Wadsworth Publishing Company. 1983.

10. Jobson JD. Applied multivariate data analysis. Vol I and II, Springer-Verlag, New York, Berlin, Heidelberg. 1992.

11. Mantel N. Why step-down procedures in variable selection. Technometrics 1970; 12:621-625.

12. Feedman DA. A note on screening regression equations. Amer Stat 1983; 37:152-155.

13. Cook RD, Weisberg S. Residuals and influence in regression. New York, Chapman and Hall. 1982.

14. Fox J. Regression diagnostics. Beverly Hills, CA, Sage Publications. 1991.

15. Daniel C, Wood F. Fitting equations to data. 2nd ed., New York, John Wiley and Sons, Inc. 1980

16. Draper N, Smith H. Applied regression analysis, 2nd ed. New-York, John Wiley and Sons, Inc. 1981.

17. Collett D. Modelling binary data. London, Chapman and Hall. 1991.

18. Agresti, A. Categorical data analysis. New-York, John Wiley and Sons, Inc. 1990.

19. Cox DR, Snell EJ. The analysis of binary data. 2nd edn. London, Chapman and Hall. 1989.

20. Hosmer DW, Lemeshow S. Applied logistic regression. New York, John Wiley and Sons, Inc. 1989.

21. Walker SH, Duncan DB. Estimating of the probability of an event as a function of several independent variables. Biometrika, 1967; 54:167-179.

22. Cox DR, Oakes D. Analysis of survival data. London, Chapman and Hall. 1984.

23. Gross AJ, Clark VA. Survival distributions: Reliability applications in the Biomedical Sciences. New York, John Wiley and Sons, Inc. 1975.

24. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Amer Stat Assoc 1958; 53:457-481.

25. Cox DR. Regression models and life tables. J Royal Stat Soc 1972; B34:187-220.

Supplied by the British Library - "The world's knowledge" www.bl.uk