Journal of Statistical Planning and Inference 140 (2010) 2904–2917

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Data-transformation approach to lifetimes data analysis: An overview

Govind S. Mudholkar [a], Kobby O. Asubonteng [b,*]

[a] Department of Statistics and Biostatistics, University of Rochester, Rochester, NY, USA
[b] Department of Biostatistics, University at Buffalo, Buffalo, NY, USA

ARTICLE INFO

Article history:
Received 30 August 2008
Accepted 5 July 2009
Available online 11 March 2010

MSC:
62N02
62N05
62P20

Keywords:
Weibull distribution
Transformations
Maximum likelihood estimation
Generalized Weibull
Exponentiated Weibull
Duration dependence
Reliability
Survival analysis

ABSTRACT

This article consists of a review and some remarks on the scope, common models, methods, their limitations and implications for the analysis of lifetime data. Also a new approach based upon data-transformations analogous to that of Box and Cox (1964) is introduced. The basic methods and theory of the subject are most familiarly and commonly encountered by the statistical community in the context of problems in reliability studies and survival analysis. However, they are also useful in areas of statistical applications such as goodness-of-fit and approximations for sampling distributions and are applicable in such diverse fields of applied research as economics, finance, sociology, meteorology and hydrology. The discussion includes examples from the mainstream statistical, social sciences and business literature.

1. Introduction

Most data types encountered in statistical applications in applied research are those that are modeled using linear models methodology and those occurring in lifetime data analysis. Assumptions such as normality, additivity, homoscedasticity and linearity commonly made in linear models and models with convenient distribution functions with appropriate hazard rates used for lifetimes, make the theories elegant and analytically tractable. But they are often not met in practice and corrective measures become essential. The most commonly employed corrective measures include seeking an appropriate model or modifying the customary model by introducing additional parameter. Also, alternatives based upon nonparametric reasoning and robustness requirements exists. All these approaches can often be demanding, valid in sufficiently large samples and very approximate. A simpler and more manageable omnibus approach is to transform the data to achieve model compatibility with a well understood and convenient customary model with possibly readily available software. The best known among such approaches is the one introduced by Box and Cox (1964) in the context of linear models. Analogous approach for the case of lifetime data analysis can be fruitful.

* Corresponding author.
E-mail address: asowusu@yahoo.com (K.O. Asubonteng).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.03.014

The beginning of modern lifetime data analysis is relatively recent, dating back to the second quarter of the 20th century, but has its roots in classical actuarial life-tables and risk-analysis mathematics. The current methodology was developed in the context of survival analysis and reliability engineering. However, it may be noted that the basic technology is applicable in such applied research as economics, finance, management science, psychology and sociology. A discussion of this broad scope together with comments on current models and methods and a new approach analogous to the Box and Cox transformations were the subject of our presentation entitled "Transformations for the Analysis of Lifetime Data in Natural, Social and Management Sciences" to the IMST2008-FIMXVI conference in Memphis, TN. Here we focus on reviewing the scope, the models and the methods with an emphasis on their limitations and briefly outline a new data-transformation based approach. Specifically, we present two transformations, one of which is appropriate when the data, on the basis of exploratory analysis, indicate unimodal or monotone hazard rate, whereas the other is suited to either bathtub or increasing hazard rate data. Because of its overview nature the paper may be considered to be "A Tale of Two Transformations for Lifetime Data Analysis". For a relatively recent overview of the subject see; "current status data analysis part XIII, chapter 35" by Jewell and van der Laan (2004).

The scope of the general methodology and applications are presented in Section 2. A review of current models appears in Section 3. The existing methods and a new approach based on data transformations are outlined in Section 4. The scope of the methods discussed in Sections 3 and 4 are illustrated in terms of examples involving real life data sets in Section 5. The final section is given to conclusions.

2. Applications in applied research

A general goal of research and deliberations in the area of lifetime data analysis, as in all areas of statistics, is to broaden the scope and advance data analysis and inference in diverse fields of applied research. In this section we outline applications in some of these fields where the methodology is applicable. Current statistical discussion of scope of lifetime data analysis appears in Jewell and van der Laan (2004).

Survival analysis: In survival analysis the outcome of interest is a random variable T (time), with unknown distribution, until some event such as death, the development of some disease, recurrence of a disease, conception, cessation of smoking, and so forth occurs. The interest is in characterizing the distribution of T especially in terms of the entities such as survival function, hazard function and mean residual life. When the interest is in comparisons of populations associated with multi-sample data or several dependent or independent random variables, notions such as proportional hazards, proportional odds, accelerated hazards models and more complex ideas come into play. Here we will focus on single-sample data. Lifetime data often involve various censoring of types such as random censoring, right or left censoring or both and interval censoring. In survival analysis, the most common type of censoring is either right or left or both and random censoring. Types I and II censoring normally occur in reliability studies. For a discussion of various types of censoring see Lawless (1982) and Klein and Moeschberger (2003). In lifetime data analysis it is customary to use survival function $S(t)=1-F(t)$ instead of $F(t)$. More importantly, the three concepts namely, the hazard function $h(t)$, cumulative hazard function $H(t)$ and mean residual lifetime MRL(t) are especially specific to lifetimes. They are now briefly described.

The hazard function h(t): The hazard function is foundational in survival analysis. It is also known as the conditional failure rate in reliability, the force of mortality in demography, the age-specific failure rate in epidemiology, the inverse of Mill's ratio in economics or is simply referred to as the hazard rate. The hazard rate is defined as $h(t)=\lim_{\Delta t\to 0}Pr[t\le T<t+\Delta t|T\ge]/\Delta t$, or more routinely as $h(t)=f(t)/(1-F(t))$, is a non-negative function. The quantity $h(t)\Delta t$ is the approximate probability that an individual who has survived to time t will experience the event in the interval $(t,t+\Delta t)$. Some common types of hazard rates encountered are increasing, decreasing, unimodal (bump-shaped) or bathtub-shaped.

Remark. Less common in literature is the discussion of *periodic hazard rate* relevant to such seasonal maladies as influenza. It is easily seen that such periodic hazard functions $h(t)$ can be expressed in terms of some sine-cosine series, i.e. $h(t)=g(\sin(t),\cos(t))$, for some $g(\cdot)$. The corresponding survival function may be obtained by solving the corresponding differential equation

$$dS(t)/dt=-g(\sin(t),\cos(t)). \tag{1}$$

A quantity related to the hazard function $h(t)$ is the cumulative hazard function $H(t)$, defined by

$$H(t)=\int_0^t h(u)\,du=-\ln[S(t)]. \tag{2}$$

The mean residual life function MRL(t): The mean residual life MRL, for individuals at time t measures their expected remaining lifetime. It is defined as

$$MRL(t)=E[T-t|T\ge t]. \tag{3}$$

It is the area under the survival curve to the right of t divided by $S(t)$. Note that the mean life, $\mu=MRL(0)$, is the total area under the survival curve.