Regression modelling of event times in reliability assisted by the covariate order method

Jan Terje Kvaløy
Department of Mathematics and Natural Sciences
University of Stavanger
N-4036 Stavanger
Norway
jan.t.kvaloy@uis.no

Bo Henry Lindqvist
Department of Mathematical Sciences
Norwegian University of Science and Technology
N-7491 Trondheim
Norway

Abstract

In analyses of event times in reliability it is often relevant to fit some kind of regression model describing the relationship between the event times and various factors, or covariates, potentially having an influence on the event times. When fitting such models it is important to find a model which adequately relates the effect of the covariates to a quantity describing the event times, for instance a hazard rate. It turns out that an approach based on ordering the covariates can provide helpful tools in several phases of this process by providing omnibus tests for relationships between the covariates and the event times, suggesting functional forms for the effect of the covariates in the model and providing useful plots and tests for residuals.

1 INTRODUCTION

In analyses of event times in reliability (failures, degradation, etc), data on various factors potentially affecting the event times, often called covariates, have to be taken into account. To accomplish this, some type of regression model could be fitted, for instance a Weibull regression model, Cox regression model, accelerated failure time model or similar.

The covariate order method is in its basic form a kernel estimation method for nonparametric exponential regression, but the method easily extends to other applications and can for instance be used for nonparametric Cox-regression and as a basis for constructing omnibus tests for covariate effect. See for instance Kvaløy (2002) and Kvaløy and Lindqvist (2003, 2004). Besides fitting nonparametric models and testing for covariate effect, the covariate order method can be useful in regression modelling by providing useful plots and tests for residuals and for suggesting functional forms in parametric models. The latter applications are the main focus here. The covariate order method is described in Section 2 and applications to regression models in reliability are discussed in Section 3.

2 THE COVARIATE ORDER METHOD

The covariate order method is based on an idea of arranging data in increasing order of a covariate, and then defining a certain point process based on the corresponding event data. In the simplest case of exponentially distributed event time data, the hazard function can be directly estimated from this point process. In other cases the hazard function can be estimated by adding suitable time transformations.

2.1 The covariate order method for exponential regression

To explain the basic idea of the covariate order method we start by considering the simple situation where we have n independent observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$, where $T = \min(Z, C)$, Z is an exponentially distributed event time, C is a censoring time, $\delta = I(Z \leq C)$, X is a single continuous covariate and we want to estimate the hazard rate $\lambda(x)$ of the event time distribution. The method starts by arranging the observations $(T_1, \delta_1, X_1), \ldots, (T_n, \delta_n, X_n)$ such that $X_1 \leq X_2 \leq \cdots \leq X_n$. Next, for convenience, divide the observation times by the number of observations, n. Then let the scaled observation times $T_1/n, \ldots, T_n/n$, irrespectively if they are censored or not, be subsequent intervals of an artificial point process on a "time" axis s. For this process, let points which are endpoints of intervals corresponding to *uncensored* observations be considered as events, occurring at times denoted $S_1, \ldots, S_r$ where $r = \sum_{j=1}^{n} \delta_j$. This is visualised in Figure 1, for an example where the ordered observations are $(T_1, \delta_1 = 1), (T_2, \delta_2 = 0), (T_3, \delta_3 = 1), \ldots, (T_{n-1}, \delta_{n-1} = 0), (T_n, \delta_n = 1)$.

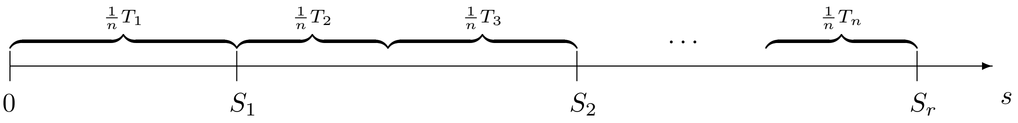

Figure 1: Construction of the process $S_1, \ldots, S_r$.

Now the conditional intensity of the process $S_1, \ldots, S_r$ at a point w is $n\lambda(X_I)$ where I is defined from $\sum_{i=1}^{I-1} T_i/n < w \le \sum_{i=1}^{I} T_i/n$. Thus from an estimate of the intensity of the process, say $\hat{\rho}_n(w)$, an estimate of $\lambda(X_I)$ is found as $\hat{\lambda}(X_I) = \hat{\rho}_n(w)/n$. The relationship between covariate values and corresponding points in the process $S_1, \ldots, S_r$ can generally be defined for instance by the simple function $\tilde{s}(x) = \frac{1}{n}\sum_{i=1}^{j} T_i$, where $X_j \le x < X_{j+1}$. The estimator can then be written $\hat{\lambda}(x) = \hat{\rho}(\tilde{s}(x))/n$. Various estimators of $\hat{\rho}(\cdot)$ can be used, one simple approach is to use a kernel estimator giving

$$\hat{\lambda}(x) = \frac{1}{nh_s}\sum_{i=1}^{r} K\left(\frac{\tilde{s}(x) - S_i}{h_s}\right) \tag{1}$$

Here $K(\cdot)$ is a positive kernel function which vanishes outside [-1,1] and has integral 1, and h_s is a smoothing parameter. Under certain mild regularity conditions it can be shown that this is a uniformly consistent estimator of $\lambda(x)$. See Kvaløy and Lindqvist (2004) for proofs and further details.

For higher dimensional covariates, $\mathbf{x} = (x_1, \ldots, x_m)$, we assume that the hazard rate is on the form of a generalised additive model

$$\lambda(\mathbf{x}) = \exp(\alpha + g_1(x_1) + \ldots + g_m(x_m)), \tag{2}$$

where $g_1(\cdot), \ldots, g_m(\cdot)$ are unspecified smooth functions. These functions can easily be estimated by the covariate order method using an iterative algorithm. The key point is that if Z is exponentially distributed with parameter $\exp(\alpha + g_1(x_1) + \ldots + g_m(x_m))$, then $Z\exp(\alpha + g_1(x_1) + \ldots + g_{j-1}(x_{j-1}) + g_{j+1}(x_{j+1}) + \ldots + g_m(x_m))$ will be exponentially distributed with parameter $\exp(g_j(x_j))$. Also note that it is possible to let some of the g-functions be parametric, for instance for discrete covariates.

2.2 Testing for covariate effect

Notice that if there is no covariate effect, that is $\lambda(x) \equiv \lambda$, then the process $S_1, \ldots, S_r$ is a homogeneous Poisson process (HPP). This observation suggests that in principle any statistical test for the null hypothesis of an HPP versus various non-HPP alternatives can be applied to test for covariate effect in exponential regression models. Moreover, such an approach can be extended to non-exponentially distributed event times by transforming the observation times to approximately exponentially distributed data. A detailed account of this approach for testing for covariate effects is given in Kvaløy (2002), where it is recommended to use an Anderson-Darling type test which turns out to have very good power properties against both monotonic and non-monotonic alternatives to constant $\lambda(x)$, and thus is a good omnibus test for covariate effect which can be used in any event time model.

2.3 Cox regression

In the classical Cox regression model the hazard rate is assumed to be on the form $\lambda_0(t)\exp(\beta_1 x_1 + \cdots + \beta_m x_m)$, and this can be generalised to a model on the form $\lambda_0(t)\exp(g_1(x_1) + \cdots + g_m(x_m))$ where $g_1(\cdot), \ldots, g_m(\cdot)$ are smooth unspecified functions of the covariates. The covariate order method can easily be applied to estimate the covariate functions $g_1(\cdot), \ldots, g_m(\cdot)$. Let Z be an event time with hazard rate as specified above, and let $\Lambda(t) = \int_0^t \lambda(u)du$. Then it is easy to show that $\Lambda_0(Z)\exp(g_2(x_2) + \cdots + g_m(x_m))$ is exponentially distributed with hazard rate $\exp(g_1(x_1))$. Thus if the other parts of the model are known it is easy to estimate $g_1(\cdot)$ using the covariate order method. The same is of course the case for the other covariate functions $g_2(\cdot), \ldots, g_m(\cdot)$, and it turns out that the entire model can be estimate by invoking an iterative algorithm where the covariate order method is used for estimating each of the covariate functions. See Kvaløy and Lindqvist (2003) for details.

3 APPLICATIONS IN RELIABILITY

The covariate order method has various applications to analyses of event data with covariates besides estimating nonparametric exponential regression and Cox regression models. Some key words are testing for covariate effect, suggesting functional form of the covariate effect and analysing residuals.

We now assume that we have n independent observations $(T_1, \delta_1, \mathbf{X}_1), \ldots, (T_n, \delta_n, \mathbf{X}_n)$, where $T = \min(Z, C)$. Based on some model we want to estimate the relationship between $\mathbf{X}$ and the hazard rate of Z. One example is a Weibull hazard regression model where the hazard rate for instance can be assumed to be on the form $\lambda(t; \mathbf{x}) = abt^{b-1} \exp(\beta_1 x_1 + \cdots + \beta_m x_m)$. The approaches presented below also extend to recurrent event situations, but the keep focus on the basic ideas we only consider the single event case.

3.1 Testing for covariate effect

One of the first steps in the analysis of data of the type considered should be to test whether each of the m covariates is individually having an effect on the event times. To test this we propose to use the Anderson-Darling test for covariate effect, Section 2.2, which is good at picking up any functional form of the effect of the covariate, and which does not make any assumptions about the event time distribution. In particular if any of the covariates is having a non-monotonic influence on the event time distribution this is useful. Many tests, for instance tests based on a specific model like the Weibull regression model suggested above, where the default is a monotonic effect of the covariate, will not be good at picking up such nonmonotonic relationships.

3.2 Suggesting functional form in parametric models

In parametric models we suggest as the next step, after having identified which covariates are having an individual effect, to use the covariate order method to suggest the functional form of the effect of the covariates. This is recommended instead of starting directly on the parametric fitting in order to indicate how the various covariates should be included in the parametric model.

A crude estimate of the relationship between the covariate and the hazard rate can be obtained either using the covariate order method for exponential regression, Section 2.1, or the extension to Cox regression models, Section 2.3. Using the method for exponential regression should be sufficient in most cases, as the point here is only to get a rough idea of the influence of each covariate on the hazard rate. In particular it is useful to pick up clearly non-linear and non-monotonic relationships so that covariates with such effects are properly modelled. For instance a second order term or some transformation of the covariate may be relevant in the parametric model in such cases.

3.3 Analysing residuals

An integral part of the modelling should be to check the adequacy of the model, e.g. based on some types of residuals. See for instance chapter 17 in Meeker and Escobar (1998) for an overview. To check the fitted model, various plots of the residuals could be made, for instance plots of the residuals against each of the covariates and against fitted values. However, in general the censoring and the skew distributions complicates the interpretation of such plots. The covariate order method then comes in as a useful alternative by providing both useful plots and formal tests. Recall that if $\Lambda(t; \mathbf{x}) = \int_0^t \lambda(u; \mathbf{x}) du$ then $\Lambda(Z; \mathbf{x})$ is exponentially distributed with parameter 1. Thus if the estimated model is adequate, $\hat{\Lambda}(T_1; \mathbf{X}_1), \ldots, \hat{\Lambda}(T_n; \mathbf{X}_n)$ should be approximately a censored sample from the exponential distribution with parameter 1. This is a commonly used type of Cox-Snell residuals (Cox and Snell, 1968). A probability plot of these residuals versus the exponential distribution can be made to check the overall adequacy of the model. How to use the residuals to check the fit of each covariate, however, is less clear. The censoring and the fact that the exponential distribution is skew makes a simple plot of the residuals versus the covariates difficult to interpret. An idea then is to use the covariate order method to estimate the hazard rate of the residuals as a function of each of the covariates. If the fitted model is adequate, all these estimated hazard rates should be approximately constant with value around 1.

Notice that plots of estimated hazard rates of the residuals versus each of the covariates should be made both for the covariates included in the model and for those not included in the model. Moreover, the covariate order test, Section 2.2, could be used to calculate a p-value for a test of relationship between each covariate and the residuals. If a significant relationship is found for any covariate, this is a strong indication that this covariate is not well modelled.

3.4 Example

In this short example we look at a dataset on lifetimes, measured in number of cycles, for 26 superalloys of a certain kind subject to different levels of psedostress in a straincontrolled tests. The data are previously analysed an can be found in for instance Meeker and Escobar (1998).

If we start by running the Anderson-Darling covariate order test we get a p-value of $1 \cdot 10^{-6}$. In other words, psedostress is clearly having a significant impact on the lifetime. Next to get an idea of the form of the relationship between psedustress and the hazard rate we fit a simple exponential regression model to the data using the covariate order method. In the left plot in Figure 2 a plot of the estimated log hazard rate, in other words a plot of $g(x) = \log(\lambda(x))$, is given. The plot indicates that the pseudostress is having a monotonically increasing effect on the log hazard rate, and possibly a non-linear effect.

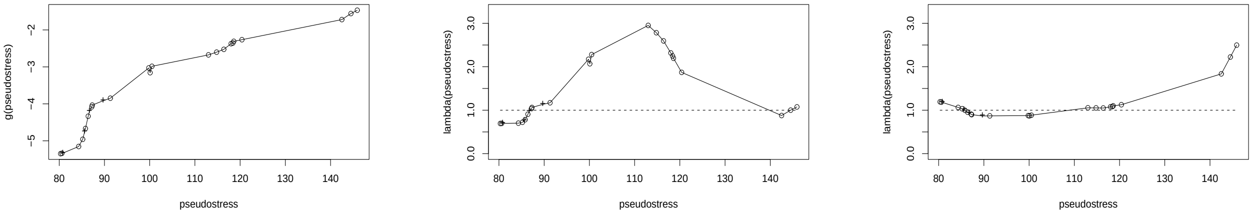

Figure 2: Left: Estimated log hazard rate for a simple exponential regression model fitted to the alloy data. Middle: Estimated hazard rate for the residuals of the simple Weibull regression model versus pseudostress. Right: Estimated hazard rate for the residuals of the Weibull regression model with second order term versus pseudostress. Points corresponding to uncensored observations are marked by a “o” in the plot and points corresponding to censored observations by a “+”.

Next we follow Meeker and Escobar (1998) and estimate a Weibull regression model. Based on the left plot in Figure 2 a natural starting point could then be a model on the form $\lambda(t; \mathbf{x}) = abt^{b-1} \exp(\beta x)$ where pseudostress, x, is modelled to have a linear effect on the log hazard rate. Since the plot indicates that the relationship is possibly non-linear, it would also be natural to for instance add a second order term or trying a transformation. But for illustration, let us first fit the suggested model, calculate the Cox-Snell type residuals discussed in Section 3.3 and make a plot of the estimated hazard rate of these residuals versus the covariate. The resulting plot is the middle plot in Figure 2. This plot shows a clear and nonmonotonic deviation from being constant around one. This is a clear indication that a better modelling of the effect of the covariate could be found, for instance by including a second order term.

Estimating the model $\lambda(t; \mathbf{x}) = abt^{b-1} \exp(\beta_1 x + \beta_2 x^2)$ we find that both the first and second order terms are clearly significant. If we calculate residuals for this model and make a plot of the estimated hazard rate for the residuals of this model we get the right plot in Figure 2. This plot indicates that this model gives a much better fit. There is still some deviation for the largest covariate values, but there are few observations in this region and thus a considerable uncertainty.

References

[1] Cox, D. R. and E. J. Snell (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B 30*, 248–275.

[2] Kvaløy, J. T. (2002). Covariate order tests for covariate effect. *Lifetime Data Analysis 8*, 35–52.

[3] Kvaløy, J. T. and B. H. Lindqvist (2003). Estimation and inference in nonparametric Cox-models: Time transformation methods. *Computational Statistics 18*, 205–221.

[4] Kvaløy, J. T. and B. H. Lindqvist (2004). The covariate order method for nonparametric exponential regression and some applications in other lifetime models. In M. S. Nikulin, N. Balakrishnan, M. Mesbah, and N. Limnios (Eds.), *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*, pp. 221–237. Birkhäuser.

[5] Meeker, W. Q. and L. A. Escobar (1998). *Statistical Methods for Reliability Data.* Wiley, New York.