International Journal of Advance Research in Science and Engineering
Vol. No.4, Issue 11, November 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

SURVIVAL ANALYSIS OF DEGRADATION DATA

Manuel Reyes[1], Manuel Rodriguez[2], Viridiana Reyes[3]

[1,2,3]*Instituto Tecnológico de Cd. Juárez (México)*

ABSTRACT

Survival analysis examines and models the time it takes for events to occur, termed survival time. The Cox proportional-hazards regression model is the most common tool for studying the dependency of survival time on predictor variables. This paper describes the Cox regression model, and explains how to use the survival package in R to estimate Cox regressions. Survival estimates are an essential compliment to multivariable regression models for time-to-event data, both for prediction and illustration of covariate effects. They are easily obtained under the Cox proportional-hazards model. Finally the data from the accelerated life test are analyzed.

Keywords: Cox Proportional-Hazards Regression Model, R, Survival Estimation, Time-Dependent Covariates, Time-Varying Coefficients.

I. INTRODUCTION

Survival analysis examines the time it takes for events to occur. The prototypical such event is death, from which the name "survival analysis" and it will focus on analyzing the distribution of survival times. The survival modeling examines the relationship between survival and one or more predictors, usually termed covariates in the survival-analysis literature. The survival package in R contains the commonly employed tools of survival analysis [1], [2].

Let T represent survival time. We regard T as a random variable with cumulative distribution function $P(t) = \Pr(T \leq t)$. And probability density function $p(t) = dP(t)/dt$. The survival function $S(t)$ is the complement of the distribution function, $S(t) = \Pr(T > t) = 1 - P(t)$.A fourth representation of the distribution of survival times is the hazard function, which assesses the instantaneous risk of demise at time t, conditional on survival to that time:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)} \quad (1)$$

A feature of survival data is censoring, the most common form of which is right-censoring: Here, the period of observation expires, or an individual is removed from the study, before the event occurs. Censoring complicates the likelihood function, and hence the estimation, of survival models.

Survival analysis typically examines the relationship of the survival distribution to covariates. For example, a parametric model based on the exponential distribution may be written as$\log h_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$or, equivalently$h_i(t) = \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$.

The constant in this model represents a kind of log-baseline hazard, since$\log h_i(t) = \alpha$ or$h_i(t) = e^{\alpha}$ when all of the x's are 0. The Cox model leaves the baseline hazard function$\alpha(t) = \log h_0(t)$unspecified

$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$or,

equivalently$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$.This model is semi-parametric because while the baseline hazard can take any form, the covariates enter the model linearly.

1.1 The Proportional-Hazards Model

Consider, now, two observations i and i' that differ in their x-values, with the corresponding linear predictors$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ and$\eta_{i'} = \beta_1 x_{i'1} + \beta_2 x_{i'2} + \cdots + \beta_k x_{i'k}$. The hazard ratio for these two observations,

$$\frac{h_i(t)}{h_{i'}(t)} = \frac{h_0(t)e^{\eta_i}}{h_0(t)e^{\eta_{i'}}} = \frac{e^{\eta_i}}{e^{\eta_{i'}}} \quad (2)$$

is independent of time t. Consequently, the Cox model is a proportional-hazards model. Even though the baseline hazard is unspecified, the Cox model can still be estimated by the method of partial likelihood. In his 1972 paper [3], Cox introduced two key ideas: a simple model for the relationship between covariates and the hazard of experiencing an event, and a partial-likelihood approach to estimate the model parameters.

1.2 The Hazard of Failure

The Cox proportional-hazards regression is thoroughly described elsewhere ([4], [5], [6] and [7]). Following Fox (2011) [8], for subjects $i = 1, \ldots, n$ let T_i denote the failure time, C_i denote the censoring time, and $N_i(t)$ represent a counting process such that $N_i(t) = I((T_i \geq t)$ where $I(u)$ is the indicator function taking value 1 if event u occurs and 0 otherwise. A subject is at risk until they experience an event or are censored. $Y_i(t)$Indicate whether the ith subject is at risk at time$t$, i.e.$Y_i(t) = I\{\min(T_i, C_i) < t\}$. Let X_i denote a predictor of interest; and $\boldsymbol{Z}_i$ a $(p \times 1)$ vector of additional covariates, where T_i and $\boldsymbol{Z}_i$ are independent given X_i and$\boldsymbol{Z}_i$. The failure time T_i is not available for all subjects, but instead $\min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$ are observed. The hazard of failure $\lambda(t|X, \boldsymbol{Z})$ is related to the covariates by:

$$\lambda(t|X, Z) = \lim_{h \to 0^+}\{h^{-1} P(t \leq t + h | T \geq t, X, \boldsymbol{Z})\}$$

$$\lambda(t|X, Z) = \lambda_0(t)\exp(\beta X + \beta_Z^T \boldsymbol{Z}) \quad (3)$$

1.3 Accelerated Failure time Models

The accelerated failure time (AFT) model specifies that predictors act multiplicatively on the failure time (additively on the log of the failure time). The predictor alters the rate at which a subject proceeds along the time axis. The model is:

$$S(t|X) = \psi\big((\log(t) - X\beta)/\sigma\big) \quad (4)$$

where ψ is any standard survival distribution and σ is called the scale parameter.

We can also write this relationship as:

$$\log(T) = X\beta + \sigma\epsilon \quad (5)$$

where ϵ is a random variable from the ψ distribution.

Assumptions:

• The true form of ψ is correctly specified.

• Each X_j affects $\log T$ linearly (assuming no interactions).

• σ is a constant, independent of X.

The exponential and Weibull distributions are the only two distributions that can be used to describe both PH and AFT models. These models can be fit in R using the survreg() function.

1.4 Exponential Proportional Hazards Regression

The exponential survival regression model can be expressed as $h(t|\mathbf{X}) = \lambda \exp(\boldsymbol{X\beta})$

$$S(t|\mathbf{X}) = \exp[-\lambda t \exp(\mathbf{X\beta})] = \exp(-\lambda t)^{\exp(\mathbf{X\beta})} \quad (4)$$

The regression can also be written as $\log h(t|X) = \log(\lambda) + X\beta$ If we replace λ with $\lambda = \exp(\beta_0)$, then

$$h(t|X) = \exp(\beta_0 + X\beta) \quad (5)$$

Therefore, we can think of λ as a transformed intercept term.

1.5 Weibull and Extreme Value Distributions

The Weibull distribution is often used for product life [9]. It is also used to describe the life of electronic components in accelerated tests. According to extreme value theory, it may describe a "weakest link" product. Such a product consists of many parts from the same life distribution, and the product fails with the first part failure.

The population fraction failing by age t is $F(t) = 1 - \exp[-(t/\alpha)^{\beta}]$. The shape parameter β and the scale parameter α are positive. For a Weibull distribution, the population fraction surviving age t is $R(t) = \exp[-(t/\alpha)^{\beta}]$, the probability density is $f(t) = (\beta/\alpha^{\beta})t^{\beta-1}\exp[-(t/\alpha)^{\beta}]$, and for the hazard function, we have $h(t) = (\beta/\alpha)(t/\alpha)^{\beta-1}$

The extreme value distribution is an analytic methods for Weibull data. $\ln T$ for a Weibull distribution has an extreme value distribution. The population fraction below t is $F(t) = 1 - \exp\{-\exp[(t-\xi)/\delta]\}$. The location parameter is ξ and the scale parameter is δ. The extreme value reliability function is $R(t) = \exp\{-\exp[(t-\xi)/\delta]\}$, the probability density is $f(t) = (1/\delta)\exp[(t-\xi)/\delta]\exp\{-\exp[(t-\xi)/\delta]\}$ and the hazard function is $h(t) = \left(\frac{1}{\delta}\right)\exp[(t-\xi)/\delta]$.

Suppose a Weibull life distribution has shape and scale parameters β and α. The $\ln(t)$ has an extreme value distribution with $\xi = \ln\alpha$ and $\delta = 1/\beta$. The last equation shows that the spread in $\ln T$ is the reciprocal of β. The Weibull parameters can be expressed as $\alpha = \exp\xi$ and $\beta = 1/\delta$.

1.6 Weibull AFT Regression Functions in R

Weibull accelerated failure time (AFT) regression can be performed in R using the survreg function. The survreg fit a parametric survival regression model. These are location-scale models for an arbitrary transform of the time variable; the most common cases use a log transformation (in R, log computes logarithms, by default natural logarithms, and log10 computes common (base 10) logarithms), leading to accelerated failure time models.

The results are not, however, presented in a form in which the Weibull distribution is usually given. In Therneau (2014) [10], accelerated failure time models are usually given by:

$$\log T = Y = \mu + \boldsymbol{\alpha}^T \boldsymbol{z} + \sigma W \quad (6)$$

where $\boldsymbol{z}$ are set of covariates and W has the extreme value distribution. Given transformations $\gamma = 1/\sigma$, $\lambda = \exp(-\mu/\sigma)$, $\beta = -\alpha/\sigma$ we have a Weibull model with baseline hazard of

$$h(x|\boldsymbol{z}) = (\gamma \lambda t^{\gamma-1}) \exp(\boldsymbol{\beta}^T \boldsymbol{z}) \quad (7)$$

1.7 The Coxph Function

The Cox proportional-hazards regression model is fit in R with the coxph function (located in the survival package):

```
> library(survival)
> args(coxph)
function (formula, data, weights, subset, na.action, init, control,
   ties = c("efron", "breslow", "exact"), singular.ok = TRUE,
   robust = FALSE, model = FALSE, x = FALSE, y = TRUE, tt, method = ties,
   ...)
```

Most of the arguments to coxph, including data, weights, subset, na.action, singular. ok, model, x and y, are familiar from lm. The formula argument is a little different. The right-hand side of the formula for coxph is the same as for a linear model. The left-hand side is a survival object, created by the Surv function. In the simple case of right-censored data, the call to Surv takes the form Surv(time, event), where time is either the event time or the censoring time, and event is a dummy variable coded 1 if the event is observed or 0 if the observation is censored.

Among the remaining arguments to coxph: init (initial values) and control are technical arguments, method indicates how to handle observations that have tied (i.e., identical) survival times. The default "efron" method is generally preferred to the once-popular "breslow" method; the "exact" method is much more computationally intensive. If robust is TRUE, coxph calculates robust coefficient-variance estimates. The default is FALSE, unless the model includes non-independent observations, specified by the cluster function in the model formula.

II. METHOD

This study tested 5mm epoxy encapsulated AlGaInP LEDs, a type of red LED, which were operated in a specially designed heat chamber. The data are analyzed in three stages. First, the data are presented in a luminosity scattering study considering three suppliers; also the data degradation pathway is presented. Second, OpenBUGS regression coefficients are generated and failure time's 30% degradation is estimated. Finally, the data from the accelerated life test are analyzed

First, following the method of E. Hong (2004) [11], three groups of 6 LEDs 5mA 1.5 V were exposed to 80 ℃ for 1000 hours. Readings are made every 144 hours for 6 days to standardize test conditions. Brightness is measured in Lux. Second LED sets were exposed to 35, 45 and 55 ℃ and the data is presented next:

```
> d.1<- read.table("tfb2.csv", header=TRUE, sep=",")
> head(d.1)
```

	temp	volt	time	status	maker
1	29	110	4160	1	1
2	30	111	3984	1	2
3	30	118	4002	0	2
4	31	112	3931	0	2
5	32	106	4019	1	1
6	30	104	4002	0	1

> tail(d.1)

	temp	volt	time	status	maker
85	49	115	2782	0	1
86	52	117	2893	1	2
87	53	113	2776	1	1
88	52	103	2944	1	1
89	53	119	2860	1	2
90	53	109	2906	1	1

Where:temp: actual temperature in degrees Celsius, volt: actual operating voltage, time: recorded time in hours, status: failure = 1 and censored = 0, maker: diode manufacturer

```
> library(survival)
> c.1<-coxph(Surv(time, status) ~temp + volt + maker,data=d.1)
> summary(cox1)
Call:
coxph(formula = Surv(time, status) ~ temp + volt + maker, data = d.1)
n= 90, number of events= 69
coef exp(coef) se(coef)      z Pr(>|z|)
temp   0.32248   1.38055  0.04114  7.838 4.55e-15 ***
volt  -0.01028   0.98978  0.02110 -0.487    0.626
maker  0.29970   1.34946  0.25143  1.192    0.233
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
exp(coef) exp(-coef) lower .95 upper .95
temp     1.3805     0.7243    1.2736     1.496
volt     0.9898     1.0103    0.9497     1.032
maker    1.3495     0.7410    0.8244     2.209
Concordance= 0.857  (se = 0.041 )
Rsquare= 0.736   (max possible= 0.996 )
Likelihood ratio test= 120  on 3 df,   p=0
Wald test            = 61.56  on 3 df,   p=2.723e-13
Score (logrank) test = 106.5  on 3 df,   p=0
```

III. RESULTS

The column marked z in the output records the ratio of each regression coefficient to its standard error, a Wald statistic which is asymptotically standard normal under the hypothesis that the corresponding β is 0. The covariates temp and maker have highly statistically significant coefficients, while the coefficient for volt is marginally significant.

The exponentiated coefficients in the second column of the first panel (and in the first column of the second panel) of the output are interpretable as multiplicative effects on the hazard. Thus, for example, holding the other covariates constant, one degree Celsius additional increases the hazard of failure by a factor of $e^{\beta_1} = 1.3805$.

The likelihood-ratio, Wald, and score chi-square statistics at the bottom of the output are asymptotically equivalent tests of the omnibus null hypothesis that all of the βs are 0. In this instance, the test statistics are in close agreement.

The next R commander plot Fig. 1:

```
> plot(survfit(c.1), ylim=c(0.7, 1), xlab="Hours",ylab="Proportion failure")
```

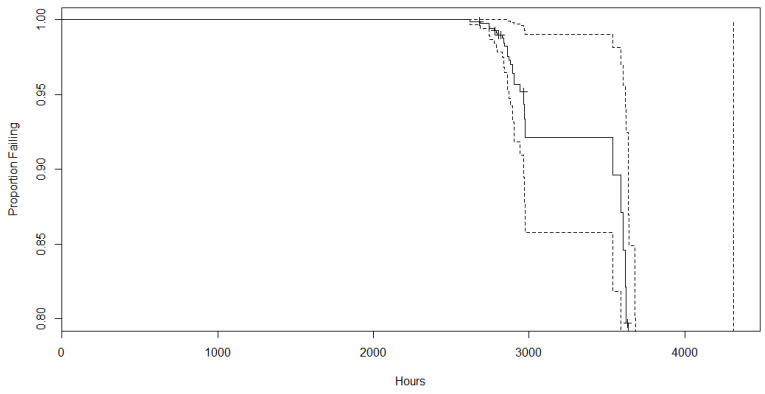

Figure 1: Estimated Survival Function $\hat{S}(t)$ for the Cox Regression

Next we will fit the model $h(t|rx) = \lambda \exp(\beta \text{maker})$ where $\text{maker} = 1, 2$ is a maker indicator:

```
> sr<-survreg(Surv(time, status)~maker, data=d.1, dist="exponential")
> summary(sr)
Call:
survreg(formula = Surv(time, status) ~ maker, data = d.1, dist = "exponential")
Value Std. Error      z       p
(Intercept)  8.44138      0.382 22.0756 5.43e-108
maker       -0.00787      0.241 -0.0327  9.74e-01
Scale fixed at 1
Exponential distribution
Loglik(model)= -650.6   Loglik(intercept only)= -650.6
            Chisq= 0 on 1 degrees of freedom, p= 0.97
Number of Newton-Raphson Iterations: 3
```

n= 90

We have to transform this output to interpret it in the proportional hazards setting $\lambda = \exp\left(-(\text{Intercept})\right) = \exp(-8.44138) = 0.00021$ and $\beta = \text{coefficient for } maker = -0.00787$.

Therefore:

$hr(maker = 2 : maker = 1) = \exp(\beta) = \exp(-0.00787) = 0.992$

$h(t|maker = 2) = \lambda \exp(2\beta) = 0.000206$

$h(t|maker = 1) = \lambda \exp(\beta) = 0.000208$

The PH regression model for a Weibull distribution is defined as $h(t|X) = \alpha\gamma t^{\gamma-1} \exp(X\beta)$. For our example, this becomes $h(t|X) = \alpha\gamma t^{\gamma-1} \exp(maker \times \beta)$:

```
> sw=survreg(Surv(time, status)~maker , data=d.1, dist="weibull")
> summary(sw)
Call:
survreg(formula = Surv(time, status) ~ maker, data = d.1, dist = "weibull")
Value Std. Error      z       p
(Intercept)  8.2684     0.0455 181.577 0.00e+00
maker       -0.0109     0.0287  -0.381 7.03e-01
Log(scale)  -2.1279     0.1015 -20.963 1.42e-97
Scale= 0.119
Weibull distribution
Loglik(model)= -546.7   Loglik(intercept only)= -546.7
        Chisq= 0.14 on 1 degrees of freedom, p= 0.7
Number of Newton-Raphson Iterations: 7 n= 90
```

Where: $\gamma = 1/\text{Scale} = 1/0.119 = 8.40$, $\alpha = \exp\left(-(\text{Intercept})/\gamma\right) = \exp(-8.2684/8.40) = 0.374$, $\beta = -\text{coefficient for } maker \times \gamma = 0.0109/0.119 = 0.092$ and $h(t|maker) = \alpha\gamma t^{\gamma-1} \exp(maker \times \beta) = 0.374 \times 8.40 t^{7.40} \exp(0.092 maker)$.

Using our data set, we fit the following Weibull regression model with volt and maker and predictors:

```
> sw2=survreg(Surv(time, status)~volt+maker , data=d.1, dist="weibull")
> summary(sw2)
Call:
survreg(formula = Surv(time, status) ~ volt + maker, data = d.1,
   dist = "weibull")
            Value Std. Error      z       p
(Intercept)  8.4664    0.24826  34.103 6.65e-255
volt        -0.0018    0.00222  -0.813  4.16e-01
maker       -0.0110    0.02859  -0.386  6.99e-01
Log(scale)  -2.1314    0.10127 -21.047  2.46e-98
Scale= 0.119
Weibull distribution
Loglik(model)= -546.3   Loglik(intercept only)= -546.7
        Chisq= 0.8 on 2 degrees of freedom, p= 0.67
Number of Newton-Raphson Iterations: 7 n= 90
```

The column labeled z is the Wald statistic $(\hat{\beta}_j/\hat{se}(\hat{\beta}_j))$ for testing$H_0:\beta_j = 0$. Parameter estimates are interpreted the same way as in parametric models, except no shape parameter is estimated because we are not making assumptions about the shape of the hazard, for example:

$$h(t|volt, maker) = h(t)\exp(\beta_1 \times volt + \beta_2 \times maker)$$

```
> cph1=coxph(Surv(time, status)~volt+maker , data=d.1)
> summary(cph1)
Call:
coxph(formula = Surv(time, status) ~ volt + maker, data = d.1)
  n= 90, number of events= 69
        coef exp(coef) se(coef)     z Pr(>|z|)
volt  0.01547   1.01559  0.01887 0.820    0.412
maker 0.09106   1.09533  0.24328 0.374    0.708
      exp(coef) exp(-coef) lower .95 upper .95
volt      1.016     0.9847    0.9787     1.054
maker     1.095     0.9130    0.6799     1.765
Concordance= 0.551  (se = 0.041 )
Rsquare= 0.009   (max possible= 0.996 )
Likelihood ratio test= 0.82  on 2 df,   p=0.665
Wald test            = 0.82  on 2 df,   p=0.6623
Score (logrank) test = 0.83  on 2 df,   p=0.6616
```

IV. DISCUSSION

As mentioned, tests for the proportional-hazards assumption are obtained from cox.zph, which computes a test for each covariate, along with a global test for the model as a whole:

```
> cox.zph(cph1)
         rho chisq     p
volt   -0.1346 1.034 0.309
maker  -0.0519 0.184 0.668
```

There is, therefore, strong evidence of proportional hazards for age, while the global test (on 3 degrees of freedom) is quite statistically significant. These tests are sensitive to linear trends in the hazard.Plotting the object returned by cox.zph produces graphs of the scaled Schoenfeld residuals against transformed time (see Fig.2):

```
> par(mfrow=c(1, 2))
> plot(cox.zph(cph1))
```

Interpretation of these graphs is greatly facilitated by smoothing, for which purpose cox.zph uses a smoothing spline, shown on each graph by a solid line; the broken lines represent ± 2-standard-error envelopes around the fit. Systematic departures from a horizontal line are indicative of non-proportional hazards. The assumption of proportional hazards appears to be supported for the covariates maker.

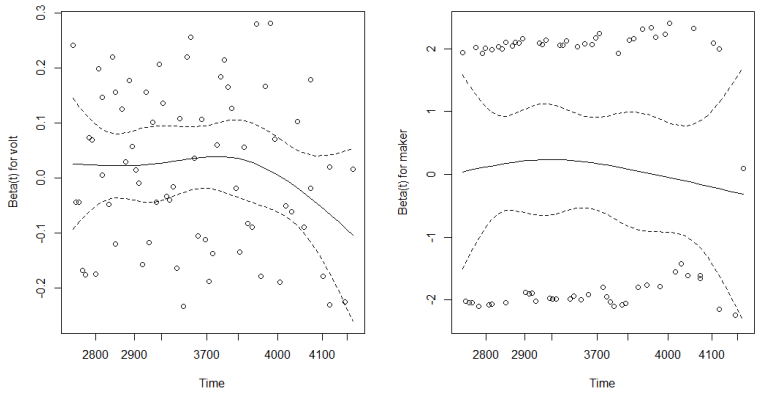

Figure 2: Plots of Scaled Schoenfeld Residuals Against Transformed Time for Each Covariate

REFERENCES

[1] T. M. Therneau, A package for survival analysis in S. Technical report, Mayo Foundation, Rochester MN., 1999.

[2] T. M. Therneau and P. M. Grambsch,Modeling Survival Data: Extending the Cox Model(Springer, New York, 2000)

[3] D. R. Cox, Regression Models and Life Tables, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2. (1972), pp. 187-220.

[4] J.D. Kalbeisch and R.L. Prentice,The Statistical Analysis of Failure Time Data(2nd edition. John Wiley & Sons, New York, 2002).

[5] J.P. Klein and M.L. Moeschberger,Survival Analysis: Techniques for Censored and Truncated Data(2nd edition. Springer-Verlag, New York, 2003).

[6] Harrell FE (2006). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag, New York.

[7] Allison PD (2010). Survival Analysis Using SAS: A Practical Guide. 2nd edition. SAS Publishing, Cary.

[8] J. Fox and S. Weisberg (2011). Appendix: An R Companion to Applied Regression. 2nd edition. Sage, Thousand Oaks.

[9] W.Nelson, Accelerated Testing: Statistical Models, Test Plans, and Data Analyses(John Wiley & Sons, Inc. 63-66, 2004)

[10] T. M. Therneau (2014). survival: A Package for Survival Analysis in S. R package version 2.37-7, URL http://CRAN.R-project.org/package=survival.

[11] E.Hong and N.Narendran, (2005). LED Life for General Lighting: Definition of Life. The Alliance for Solid-State Illumination Systems and Technologies (ASSIST). http://www.lrc.rpi.edu/programs/solidstate/assist/pdf/ASSIST-LEDLife-revised2007.pdf.