Technical Report - 7/87

# AN APPLICATION OF COX REGRESSION MODEL TO THE ANALYSIS OF GROUPED PULMONARY TUBERCULOSIS SURVIVAL DATA

P. VENKATESAN*

K. VISWANATHAN+

R. PRABHAKAR*

* Tuberculosis Research Centre,
I.C.M.R., MADRAS, INDIA.

December 1987.

+ Department of Statistics,
University of Madras;
MADRAS, INDIA.

# AN APPLICATION OF COX REGRESSION MODEL TO THE ANALYSIS OF GROUPED PULMONARY TUBERCULOSIS SURVIVAL DATA

P. VENKATESAN
Department of Statistics
Tuberculosis Research Centre
I.C.M.R., MADRAS

and

K. VISWANATHAN
Department of Statistics
University of Madras
MADRAS.

R. PRABHAKAR
Tuberculosis Research Centre, (ICMR)
MADRAS

## 1. INTRODUCTION:

The recent statistical literature attests to a considerable current interest in specialised statistical methods for the analysis of time to occurrence or failure time data in medical research.

Frequently the primary objective of a failure time Study concerns with the association between certain co-variates $Z = (Z_1, \ldots, Z_p)$ and the time $T \geq 0$ to the occurrence of a certain event. For example a clinical study may be designed to compare several treatment programmes in respect to the time T to recurrence of a disease or response to treatment of a disease. The regression vector Z would include indicator components for treatment as well as other prognostic factors.

The distribution of failure time T can be represented in the usual manner in terms of density or distribution functions as well as in more specialised ways such as the hazard function.

If the underlying process is continuous, hazard function at time t among individuals with covariates Z is defined as

$$h(t, Z) = \lim_{\partial t \to 0} P[t \leq T \leq t + \partial t / T > t, Z] / \partial t \qquad (1.1)$$

The hazard function gives the risk of failure at any time given that the individual has not failed prior to t. Since the notion of failure rate is basic and conceptually simple it provides a starting point for modelling the association between Z and t.

## 2. COX REGRESSION MODEL:

One such model introduced by Cox (1972) provides the basis for most of the discussion of this paper. This model presumes that covariatos affect the hazard function in a multiplicative manner according to

$$h(t ; Z) = h_o(t)\, c^{Z \beta} \qquad (2.1)$$

where $h_o(t)$ is an unspecified function of time, $\beta$ is a column vector of the regression coefficients and the factor $e^{\beta}$ describes the risk of failure for an individual with regression variables Z relative to that of starting value Z = 0. Since the ratio of hazard functions corresponding to any two Z values is a constant it is often referred to as Proportional Hazard (PH) Model.

Cox model is a particular case of the general model that has hazard function

$$h(t; Z) = h_o\, g\,(\underline{Z}, \underline{\beta}) \qquad (2.3)$$

The Cox model is also equivalent to the assumption of Lehmann alternatives

$$1 - F(t; Z) = [1-F(t,0)]^{\exp Z\beta} \qquad (2.4)$$

The PH model provides a unified approach to inferences from a variety of typos of failure time data.

## 3. ESTIMATION OF $\beta$:

The approach taken here is a distribution free one where no specific form is assumed for $h_o(t)$. The main idea for this originated with Cox (1972) who proposed a method for estimating $\beta$ in the absence of the knowledge of $h_o(t)$ and also estimating for $h_o(t)$.

Consider a random sample of n individuals with k distinct observed lifetimes and (n-k) censored times, let the k observed times be

$$t(1) < t(2) < \ldots < t(k)$$

and $R_i = R(t(i))$ risk set at time t(i)

The likelihood function for estimating $\beta$ in the absence of knowledge of $h_o(t)$ is

$$L(\beta) = \prod_{i=1}^{k} \left[ e^{Z(i)\beta} \Big/ \sum_{l \in R_i} e^{Z_l \beta} \right] \qquad (3.1)$$

where Z(i) is the regressor vector associated with the individuals observed to fail at t(i)(Cox 1975). The motivation is the probability an individual being fail is

$$h(t, Zi) = \sum_{l \in R(t)} h(t, z_l)$$

$$= e^{z \beta} / \sum_{l \in R(t)} e^{Zl \beta} \qquad (3.2)$$

The likelihood is formed by the product of all such factors over the k uncensored lifetimes. The estimate $\hat{\beta}$ of $\beta$ is obtained by solving the set of equations

$$\frac{\partial L(\beta)}{\partial \beta} = 0 \qquad (3.3)$$

The another method used for estimating $\beta$ is

$$L(\beta) = \prod_{i=1}^{k} [e^{S_i \beta} / (\sum_{l \in R_i} e^{Zl \beta})^{di}] \qquad (3.4)$$

where di is the number of lifetimes equal to t(i) and $S_i$ is the sum of the regression vector Z for these di individuals.

4. STRATIFICATION:

When a factor does not affect the hazard multiplicatively, stratification may be useful in model building. Hazard function for an individual in stratum f with regressor variable Z is

$$h(t/Z) = h_{oj}(t) e^{Z \beta}$$

That is individuals in the same stratum have the same proportional hazards function but it is not necessarily the case for individuals in different strata. It is also assumed that the relative effect of the regressor variables is the same in each stratum. This condition sometimes needs to be relaxed with $\beta$ varying from stratum to stratum. A partial likelihood function $L_j(\beta)$ is obtained for each stratum and then the overall partial likelihood function for $\beta$ is

$$L(\beta) = L_i(\beta) \ldots L_s(\beta) \tag{4.1}$$

After estimating $\beta$ by maximising $L(\beta)$ the survivor functions can be estimated by

$$s_{oj}(t) = \exp\left[-\int_0^t h_{oj}(u)\, du\right] \quad j = 1,2, \ldots, s \tag{4.2}$$

## 5. PARAMETRIC MODELS:

If $h(t, Z)$ is specified upto a finite number of parameter $\theta$, ordinary parametric likelihood methods can be applied. However, the presence of censoring except in a few extremely special cases precludes the possibility of tractable exact distribution theory. One parametric model that has found extensive application is the Weibull Model is given by

$$h_o(t) = \alpha p\, (\alpha t)^{p-1} \tag{5.1}$$

This model is the intersection of PH and AFT models. Weibull is characterised by that model in which the regressor variables

act multiplicatively both upon hazard rate and upon failure time. The asymptotic, theory is readily applied to the Weibull model with $\theta = (\alpha, p, \beta)$. The Weibull model can be written as log linear form as

$$y + Z\beta = \delta + \sigma v \qquad (5.2)$$

where $y = \log t$, $\delta = -\log \alpha, \sigma = p^{-1}$, $\beta' = p^{-1}\beta$.
The error variable v has an extreme value density

$$f(v) = \exp(v - e^{v}). \qquad (5.3)$$

Other parametric cases include lognormal, log logistic and generalised gamma regression model corresponding to error quantities v that are normal, logistic and log gamma respectively. On fully parametric set up the precise conditions of censoring mechanism, regressor variables and the form of h(t, Z) that would be sufficient to ensure modelling is not yet known fully.

In studies involving human subjects no theoretical form for survival distributions can be assumed and in such situations techniques based on PH model appear to be superior to other models. The relative efficiency of Cox regression model compared with likelihood methods that assume full parametric form for the hazard function is generally good. (Breslow 1974, Kalfleisch 1974, Efron 1977, Oakes 1977).

## 6. A MODEL BASED ON GROUPING DATA FROM COX MODEL:

In many instances the life time T of individuals with regressor variable Z might be assumed to come from a continuous proportional hazard model. If the life times are grouped into k classes, then a group regression model is obtained for which survival probability for the i-th interval $[a_i, a_{i+1})$ is given by

$$P_i(Z) = S(a_i/Z) = S_o(a_i)^{\exp(Z\ \beta)} = P_i(O)^{\exp(Z\ \beta)} \qquad (6.1)$$

$$i = 1, 2, \ldots, k \text{ with } P_0(O) = 0.$$

This gives $p_i(z) = \dfrac{p_i(Z)}{p_{i-1}(Z)} = p_i(O)^{\exp(Z\beta)}$

where $p_i(O) = \dfrac{p_i(O)}{p_{i-1}(O)}$ $i = 1, 2, \ldots, k.$

Use of this model produces a likelihood

$$\pi_{i=1}^{k} \left[ \pi_{l \in D_i} (1 p_i(Z_l)) \ \ l \in \pi R_i - D_i \ p_i(Z_l) \right] \qquad (6.2)$$

which can be used to estimate $\beta$ and $p_1, p_2, \ldots, p_k$. This approach was first introduced by Kalbfleisch and Prentice (1973). The m.l.e. for (6.2) is discussed by Prentice and Gloeckler (1978). The log likelihood can be written as

$$\log L(\beta, p) = \pi_{i=1}^{k} \left[ \Sigma_{l \in D_i} \log \frac{1 - p_i^{\exp(Z_1 \beta)}}{\exp(Z_1 \beta)} + \Sigma_{l \in R_i} \log (p_i^{\exp(Z_l \beta)} \right. \qquad (6.3)$$

$$p_i$$

Prentice and Gloeckler suggest using the parameters

$$\gamma_i = \log(-\log p_i) \quad i = 1, 2, \ldots, k. \qquad (6.4)$$

The $\gamma_i$ are unrestricted and they found that convergence in a Newton-Raphson iteration procedure was improved when $\gamma = (\gamma_i \ldots \gamma_k)$ was used in the place of p.

## 7. APPLICATION TO PULMONARY TUBERCULOSIS DATA:

The studies at the Tuberculosis Research Centre, Madras have shown a negative relationship between time to response and the covariates Extent of Disease, Extent of Cavity and sputum culture positivity on admission (TRC 1966). Differential treatment methods between lightly and heavily diseased patients could distort the magnitude of relationship between two treatment groups. The estimation of joint relationship of Time to response with each of the covariates Extent of Disease, Cavity and culture Positivity not only gives the potential for making detailed and accurate assessment but also may provide valuable insight into the response mechanism. Of particular interest in terms of the response mechanism is the ability of Relative Risk (RR) models that are multiplicative or additive to describe data. In general terms a multiplicative RR model would correspond to a process where in an intial exposure induces a fractional increase in the age specific failure rate which then serves as a new base line rate that the RR function for the second exposure variable multiply.

## 8. METHODS AND MATERIALS:

Tuberculosis Research Centre (ICMR) Madras is conducting studies since 1956 in pulmonary and extrapulmonary tuberculosis patients to assess the therapautic effects when drugs are administered to the patients. Various tests have been carried out to assess the eligibility of patients to the study before admission and periodically to assess the response to treatments after admission. The assessment of response to treatment is based on clinical, radiological, bacteriological, biochemical, immunological and other investigations. The patients are followed up to a maximum period of 5 years.

Our study population consists of 261 patients taken from a study conducted at the Tuberculosis Research Centre, Madras (TRC 1983). The patients belong to 2 treatment groups of a controlled clinical study of 3 short course regimens. Out of the two group one contained rifampicin (R/7 series) and the other no rifampicin (Z/7 series).

R/7 Series: Daily chemotherapy for 2 months with rifampicin (R) 12 mg/kg body weight plus streptomycin sulphate (S) 0.75 g plus isoniazid (H) 400 mg (incorporating Pyridoxine 6 mg) plus pyrazinamide (Z) 40 mg/kg followed by twice weekly chemotherapy for 5 months with streptomycin 0.75 g plus isoniazid 15 mg/kg plus pyrazinamide 70 mg/kg the total duration being 7 months.

Z/7 Series: The same as the R/7 series but without rifampicin the total duration being 7 months.

Of the 261 patients (132 R/7, 129 Z/7) 67% were males, 24% of the patients were less than 25 years of age, 63% were between 25 and 44 and 13% were greater than 45 years. The patients in the two groups were almost similar with respect to age, sex and weight. Response to treatment times were available in monthly intervals. Over 30 disease characteristics wore available. A regression vector of 16 components is used to illustrate the above procedures. The components used are as follows:

1. Age - 3 components (1 for < 25 years, 2 for 25-45, 3 for > 45)
2. Sex - 2 components (1 male, 2 female)
3. Extent of Disease (EOD)- 3 components (BODI, EOD2, EOD3)
4. Extent of Cavity (BOC) - 3 components (EOC1, EOC2, EOC3)
5. Sputum Culture (CUL) - 3 components (CULl, CUL2, CUL3).

An additional indicator variable distinguishing two treatment groups taking value 0 for Z/7 series and 1 for R/7 series is considered. The final component is another indicator variable for patients response time is censored (0) or un-censored (1).

The radiographic and bacteriological findings on admission to treatment were given in Table-l.

TABLE-1

| Characteristic | Levels | Percentage of Patients | |
|---|---|---|---|
| | | R/7 | Z/7 |
| EOD | EOD1 | 29 | 27 |
| | EOD2 | 39 | 39 |
| | EOD3 | 32 | 34 |
| EOC | EOC1 | 43 | 39 |
| | EOC2 | 35 | 30 |
| | EOC3 | 22 | 31 |
| CUL | CUL1 | 42 | 44 |
| | CUL2 | 43 | 38 |
| | CUL3 | 15 | 18 |

The above variables are related to the hazard function to response time to treatment for tuberculosis using the PH model. A partial likelihood (Cox 1975) and Newton Raphson iteration is applied to the data which gives a maximum partial likelihood estimates for the corresponding regression coefficients as given in Table-2. Many refinement are necessary before claiming any association.

From Table-2 we see that the estimated regression coefficient of R/7 series $\beta = -0.344$ with an estimated standard error of $\beta$ from the observed information matrix is 0.117, giving a T-value of 2.94 which is significant at 0.02 level. This suggests that the response time is reduced by an

estimated multiplicated factor exp $(\beta) = 0.71$ in comparison to Z/7 series. The other estimates of $\beta$ for the covariates EOD, EOC and CUL are not significant. But we observe from the table that the $\beta$'s increases as we move from the lower levels to the higher levels of all the characteristics. For example the duration of response is prolonged by a factor exp (.083) = 1.09 for a patient with slight or limited disease (EOD1). The factors for moderate (EOD2) and extensive (EOD3) disease are 1.29 and 1.39 respectively. For a patient with 2 or more covariates the duration will be prolonged by exp (sum of the $\beta$'s). For example the duration of response to treatment for a patient having moderate disease (EOD2), extent of cavity (EOC2) and culture result 3 + (CUL3) will be prolonged by a factor exp (.173 + .238 + .112) = 1.67. If this patient is treated with R/7 series the duration will be reduced by a factor 0.71 in comparison to a patient in Z/7 series.

TABLE-2

M.L.E. FIT OF COX MODELS TO DATA ON 261 PULMONARY TUBERCULOSIS PATIENTS.

| Covariate | $\beta$ | SE ($\beta$) | Sig. |
|---|---|---|---|
| EOD1 | 0.083 | 0.101 | NS |
| EOD2 | 0.173 | 0.210 | NS |
| EOD3 | 0.255 | 0.314 | NS |
| EOC1 | 0.073 | 0.100 | NS |
| EOC2 | 0.238 | 0.348 | NS |
| EOC3 | 0.255 | 0.368 | NS |
| CUL1 | 0.026 | 0.151 | NS |
| CUL2 | 0.068 | 0.250 | NS |
| CUL3 | 0.112 | 0.254 | NS |
| R/7 | 0.344 | 0.117 | 0.01 |

NS = Non-Significant

## 9. DISCUSSION:

An important point particularly with strong regression effects concerns the appropriateness of PH assumption. One possibility of testing proportionality is to stratify the data and fit separate PH models for each stratum specific plots of log (-log p(t,Z) at specified Z values should then the separated by approximately a constant difference over time if the proportionality assumption is appropriate. More formal procedures arc available by introducing time dependent covariates into the model and testing for $\beta = 0$. The starting point is usually to obtain estimators for the value of $\beta$ which will indicate whether individual covariates influence prognosis, whether the effect is favourable or unfavourable and the magnitude of the effect. The standardised component of $\beta$ can be aged to determine whether the corresponding covariate has a statistically significant effect. A significant $\beta_i$ indicates that a highly significant prognostic factor is $Z_i$. The value of $\beta_i$ are used to estimate the effect of a particular covariate.

A number of parametric models have been proposed many of which arc unified in Farewell and Prentice (1977). Least Square Models arc proposed by Miller (1976). The later models are principly linear in logarithm of failure times. Log linear models are equally good.

## 10. REFERENCES :

1. BRESLOW, N.E. (1974): Biometrics 30, 89-100.
2. COX, D.R. (1972) : J.R.S.S., B-34, 187-200.
3. COX, D.R. (1975) : Biometrika 62, 269-276.
4. EFRON, B. (1977) : JASA, 72, 555-565.
5. FAREWELL, V.T. and PRENTICE, R.L. (1977): Technometrics 19, 69-75.
6. KALBFLEISCH, J.D. (1974): Biometrika 61, 31-38.
7. KALBFLEISCH, J.D. and PRENTICE, R.L. (1973): Biometrika 60, 267-278.
8. MILLER, R.G. (1976): Biometrika, 63, 449-464.
9. OAKES, S.D. (1977): Biometrika 64, 441-448.
10. PRENTICE, R.L. and GLOECKLER, L.A. (1978): Biometrics, 34, 57-67.
11. TUBERCULOSIS CHEMOTHERAPY CENTRE, MADRAS (1966): Bull. Wld. Hth. Org. 34, 483-515.
12. TUBERCULOSIS RESEARCH CENTRE, MADRAS (1983): TUBERCLE, 64, 73-91.

***

REPORT DOCUMENTATION PAGE:

| | |
|---|---|
| 1. Report No. 7/87 | 2. Library Catalogue No. |
| 3. Title (and Sub-title):<br>'An Application of Cox Regression Model to the Analysis of Grouped Pulmonary Tuberculosis Survival Data' | 4. Classification/Section:<br>i. Inference<br>ii. Survival Analysis |
| 5. Authors:<br>P. VENKATESAN<br>K. VISWANATHAN and<br>R.PRABHAKAR | 6. Contract/Sponsoring Agency<br>University of Madras |
| 7. Performing Organisation Name and Address:<br>Department of Statistics<br>University of Madras<br>Madras 600005 | 8. Type of Report:<br>Research Paper |
| 9. Report Date:<br>December 1987 | 10. Number of Pages:<br>14 |

11. Key-words:

Covariates - Hazard function - PH Model - AFT Model - Partial Likelihood - Survival Analysis - Clinical Trial - Prognostic factor - Pulmonary Tuberculosis - Newton Raphson Iteration Method.

12. Short Abstract:

The Technical Report is based on the paper presented at the 5th Conference of Indian Society for Medical Statistics hold at Srinagar in September 1987. The Cox model and its grouped version are applied to the tuberculosis data to explain the treatment differences by disease status and other demographic characteristics at diagnosis.

Place: MADRAS 600005
Date : DECEMBER 1987.

(K.N. PONNUSWAMY)
Professor and Head,
Dept. of Statistics,
University of Madras
MADRAS 600005.