Bayesian Cox Model with Categorical Predictors for Time to Event Breast Cancer Data

### Keywords
- Survival data
- Cox PH Model
- Bayesian approach
- Gibbs sampler

### Abstract
Survival analysis has become a standard tool for modeling cancer trial data when the event of interest is "time to event". Cox regression, which implements the proportional hazards model, is designed for analysis of time until an event or time between events, introduced by Cox (1972) in order to estimate the effects of different covariates influencing the time-to-event data. This model has been used extensively in time to event of cancer trial data for given categorical predictor variables. The Bayesian analysis has advantageous in dealing with small sample of censored data more than a frequentist method. The main objective of this article is to apply a Bayesian Cox model and is being compared with a frequentist method. Gibbs sampling technique is used to assess the posterior quantities of interest and to avoid the complexity in calculations. The posterior is arrived using SAS package.

### Introduction
In many cancer trials, the applications of proportional hazards model is often a more realistic model than the other survival models in the analysis of time to event data. The Cox proportional hazards model constructs an analytical model for time-to-event breast cancer data. The model produces a survival function that predicts the probability that the event of interest has occurred at a given time $t$ for given values of the predictor variables. The predictors are in nature of categorical type, shape of the survival function and the regression coefficients for the predictors are estimated from observed subjects.

To describe the distribution of survival time has assumed that the hazard function is completely specified given the baseline hazard function and the values of the covariates. In cancer studies, there may be factors other than the measured covariates that significantly affect the distribution of survival time. This condition is often referred to as heterogeneity of the subjects. Among the early papers of Vaupel, Manton and Stallard (1979) who used the concept of to describe the differences in survival time apparently among similar individuals. Hougaard (1995) presented an overview of the models, proposed for the use in time to event data. Aalen (1994) also provides a relatively non-technical summary with a focus on fully parametric models. Klein and Moeschberger (1997) presented methods based on incorporating in proportional hazards models and its technical details.

There are two approaches to a Bayesian analysis of the Cox Proportional Hazards Model (PHM). This is based on the partial likelihood $L(β, y)$ combined with a prior $π(β)$ which produces the posterior $π(β|y)$. The baseline hazard is left unspecified. Inference is made from samples drawn from $π(β|y)$. Thus the partial likelihood is treated as a likelihood function just as in the classical analysis. The basic idea of a model is to incorporate a comparison between Bayesian Cox PHM with non Bayesian approach of Cox PHM.

which may be re-expressed as

\[
h(t) = h_0(t) \exp(\beta x_i) \quad (1)
\]

\[
h(t) = h_0(t) \exp(\beta x_i + \eta_i) \quad (2)
\]

Using the relationship between the survival and hazard function, it has the conditional survival function as

\[
S(t) = \exp \left[ -\Lambda(t) \exp(\beta x_i) \right] \quad (3)
\]

and the conditional likelihood as

\[
L(γ, β) = \prod_{i=1}^{n} \left[ h(t_i) \exp(\beta x_i) S(t_i) \right] \quad (4)
\]

where there are $i$ clusters, $j^a$ one being of size $η_i$ and $g$ and $b$ represent baseline hazard and regression parameters, respectively. On substitution it gives

\[
L(γ, β) = \prod_{i=1}^{n} \left[ h(t_i) \exp(\beta x_i) \right]^{η_i} \exp \left[ -\Lambda(t_i) \exp(\beta x_i) \right] \quad (5)
\]

In the Bayesian point of view, the variance $\sigma^2$ would be expressed as a hyperparameter and prior knowledge concerning its value which will be summarized in a hyper prior distribution.

### Gibbs Sampler Algorithm
The Gibbs sampler technique is one of the best known MCMC sampling algorithms in the Bayesian computational methods. The Gibbs sampler by Grenander (1983),
the prescribed term is introduced by Geman and Geman (1984). Gibbs sampling is the landmark in problem of Bayesian inference (Gelfand and Smith, 1991). The Gibbs sampler tutorial is provided by Casella and George (1992).

Let \( \theta = (\theta_1, \theta_2, \ldots, \theta_p) \) be a p-dimensional vector of parameters and let \( \pi(\theta D) \) be its posterior distribution given the data D. Then, the fundamental format of the Gibbs sampler is given as

\[
\theta_0 = (\theta_{10}, \theta_{20}, \ldots, \theta_{p0}) \text{ and set } i = 0
\]

Step 1. Select an arbitrary starting point

\[
\theta_{i+1} = \left( \theta_{1,i+1}, \theta_{2,i+1}, \ldots, \theta_{p,i+1} \right)
\]

Step 2. Generate \( \theta_{1,i+1} = \left( \theta_{1,i+1}, \theta_{2,i+1}, \ldots, \theta_{p,i+1} \right) \)
- Generate \( \theta_{1,i+1} \sim \pi(\theta_1 | \theta_{2,i}, \ldots, \theta_{p,i}, D) \)
- Generate \( \theta_{2,i+1} \sim \pi(\theta_2 | \theta_{1,i+1}, \theta_{3,i}, \ldots, \theta_{p,i}, D) \)
- \ldots
- Generate \( \theta_{p,i+1} \sim \pi(\theta_p | \theta_{1,i+1}, \theta_{2,i+1}, \ldots, \theta_{p-1,i+1}, D) \)

Step 3. Set \( i = i + 1 \), and go to step 2

Each component of \( \theta \) is in the natural order and a cycle in this scheme requires generation of \( p \) random variates. Gelfand and Smith (1990) show that under certain regularity conditions, the vector sequence \( \{\theta_i, i=1,2,\ldots\} \) has a stationary distribution \( \pi(\theta | D) \). The performance of a Metropolis-Hastings algorithm depends on the choice of a proposal density \( q \). The Metropolis-Hastings algorithm can be used within the Gibbs sampler when direct sampling from the full conditional posterior is difficult.

**PRIOR**

Prior elicitation perhaps plays the most crucial role in Bayesian inference. Survival analysis with covariates, the most popular choice of informative prior for \( b \) is the normal prior, and the most common choice of non-informative prior for \( b \) is the uniform prior. The non-informative and improper priors may be useful and easier to specify for certain problems, but they cannot be used in all applications, such as model selection or model comparison, as it is well known that proper priors are required to compute Bayes factors and posterior model probabilities (Ibrahim, et al., 2004). Also non informative priors may cause instability in the posterior estimates and lead to convergence problems for the Gibbs sampler. Moreover, non informative prior do not make use of real prior information that one may use for a specific application.

**APPLICATION**

We consider the database consisting of 368 breast cancer women patients diagnosed at Cancer Institute (WIA), Chennai, India and follow-up period up to 180 months, represented by the variable Time. The event of interest was time to death in months. A censoring indicator variable stage has three categories and is represented by the variable Time. The event of interest is the hazard model is as follows

\[
\lambda(t) = \alpha + \beta_1AGE + \beta_2AGE + \beta_3STAGE_1 + \beta_4STAGE_2 + \beta_5STAGE_3 + \beta_6 \text{ Age group}
\]

where \( AGE_1 \) has two levels (Age <50yrs(=1 as reference) and Age>50yrs), and stage has 3-level of categorical predictor covariates (stage2B = 1(as reference), stage 3A = 2 and stage 3B = 3) \( STAGE_1 (k=1,2,3) \) are dummy variables representing the 3-level factor for underlying stage.

**Cox Proportion Hazards regression**

The Cox proportional hazards model to these data, variables age group and stages, which are categorical variables, are also take part as stratified aspects. By default, they categorized by using the reference coding with the last category as the reference category. However, it can explicitly specify the reference category of our own preference. Here, Age Group(less than 50yrs=0) is chosen as the reference category for age, stage (stage2B=1) is chosen as the reference category for stages of cancer. Coded variables are resulted (table below “Class Level Information”) with the reference coding has zero and the rest are having one for all variables. Since age group variable is binary, the variable has a value of 0 for the reference category. The variable stage has three categories and is represented by two dummy variables.

**Table 1: Classification of death according to Stages and Age group**

| Status   | Stages | Age groups |
|----------|--------|------------|
|          | Stages2B | Stage3A | Age <50 years | Age > 50 years |
| Alive    | 61 (55) | 72 (56) | 48 (37) | 115 (53) | 66 (44) |
| Dead     | 49 (45) | 56 (44) | 82 (63) | 103 (47) | 84 (56) |
| Total    | 110     | 128      | 130      | 218      | 150      |

From the table1, we see that death increases with the severity of stages and age. The event experienced cases among age group in more than 50 years is higher than the less than 50 years (Pari Dayal et al., 2013). The linear predictor is set equal to the intercept in the reference group (stage = 3); this defines the baseline hazard. The corresponding distribution of survival time is Gamma distribution (Cox and Oakes, 1984).

The Cox model with a parameter for each individual is using for identifying the risk variables for breast cancer patients. Here, the age and stages are considered as risk factors of categorical predictor variables. We analyzed the data assuming a Weibull distribution for the survivor function, and including random effect (b) for each patient. The hazard model is as follows

\[
t_i = \text{Weibull}(\alpha, \beta_i) \quad i = 1, 2, \ldots, n
\]

where \( \alpha \) has two levels(Age <50yrs(=1 as reference) and Age>50yrs), and stage has 3-level of categorical predictor covariates (stage2B = 1(as reference), stage 3A = 2 and stage 3B = 3) \( STAGE_1 (k=1,2,3) \) are dummy variables representing the 3-level factor for underlying stage.

**Class Level Information**

| Class | Value | Design Variables |
|-------|-------|------------------|
| Age group | 0 | 1 | Reference |
| stage  | 1 | 0 | Reference |
|       | 2 | 1 | 0 |
|       | 3 | 0 | 1 |

The test results of individual model effects are shown in table2. There is a strong prognostic effect of stages on patient’s survivivorship (p=0.0007), and the survival times for patients of different age groups play a role other way that it differ non significantly (0.4083). In the Cox proportional
hazards model, the effects of the covariates are to act multiplicatively on the hazard of the survival time, and therefore it is a little easier to interpret the corresponding hazard ratios than the regression parameters. For a categorical variable parameter, the hazard ratio is the ratio of the hazard rates between the given category and the reference category. The hazard rate of age group of less than 50 years is 1.137 times more that of Age Group more than 50 years, the hazard rate of stage of 3A is 0.951 times more that of stage2B, and the hazard rate of stage3B is 1.715 times more that of stage2B. Moreover there is evidence that the stage influences the event of interest.

### Table2: Cox PH Model: Parameters Estimates with Reference Coding

| Parameter   | levels | Parameter Estimate | Standard Error | Chi-Square | P     | Hazard Ratio | Wald ChiSq | p     |
|-------------|--------|--------------------|----------------|------------|-------|--------------|-----------|-------|
| age group   | 1      | 0.12805            | 0.15485        | 0.6838     | 0.4083 | 1.137        | 0.68      | 0.4083 |
| stage3A     | 2      | -0.05045           | 0.19651        | 0.0659     | 0.7974 | 0.951        | 14.45     | 0.0007 |
| stage3B     | 3      | 0.53946            | 0.18359        | 8.634      | 0.0033 | 1.715        |           |       |

This analysis generates a posterior chain of 10,000 iterations after 2,000 iterations of burn-in and it also displayed the names of the parameters and their corresponding effects and categories. Further it computes the maximum likelihood estimates of regression parameters are depicted in table3. These estimates are used as the starting values for the simulation of posterior samples.

### Table3: Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits |
|-----------|----|----------|----------------|-----------------------|
| a1        | 1  | 0.128    | 0.1549         | -0.1755 0.4315        |
| sta2      | 1  | -0.0503  | 0.1965         | -0.4354 0.3349        |
| sta3      | 1  | 0.5394   | 0.1836         | -0.1796 0.8993        |
| Theta     | 1  | 0.0001   |                 |                       |

Summary statistics, convergence diagnostics, and diagnostic plots are provided for each parameter of categorical predictors. Summary statistics of the posterior samples are shown in Table5. These results are quite comparable to the classical results based on maximizing the likelihood as shown in Table3, since the prior distribution for the regression coefficients is relatively flat. The table4 results the initial values of the chain on each parameters.

### Table4: Initial Values of the Chain on each Parameter

|                | age1 | sta2 | sta3 | id1 | id2 | id3 | id4 | id5 | id6 | id7 | id8 | id9 | id10 |
|----------------|------|------|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|
| Parameter      | 0.128| -0.027| 0.764| -0.617| -0.358| -0.258| 0.088| -0.807| 0.537| 0.21| -0.65| 0.513| 0.669| 0.646|
| df1            | 0.128| -0.612| 0.169| -0.807| -1.159| -0.752| -0.899| -0.673| 0.537| 0.361| 0.393| 0.524| 0.496|
| df2            |       |       |      |      |      |      |      |      |      |      |      |      |      |
| df3            |       |       |      |      |      |      |      |      |      |      |      |      |      |
| df4            |       |       |      |      |      |      |      |      |      |      |      |      |      |
| df5            |       |       |      |      |      |      |      |      |      |      |      |      |      |
| df6            |       |       |      |      |      |      |      |      |      |      |      |      |      |
| df7            |       |       |      |      |      |      |      |      |      |      |      |      |      |
| df8            |       |       |      |      |      |      |      |      |      |      |      |      |      |
| df9            |       |       |      |      |      |      |      |      |      |      |      |      |      |
| df10           |       |       |      |      |      |      |      |      |      |      |      |      |      |

The further illustrate the use of the backward elimination process to identify the effects that affect the survivorship of the breast cancer patients. It is specified to carry out the backward elimination which specifies the significant effects and categories. Further it computes the maximum likelihood estimates of regression parameters are depicted in table3. These estimates are used as the starting values for the simulation of posterior samples.

### Bayesian Cox Proportion Hazards Regression

Cox Proportion Hazards Regression uses the partial likelihood of the Cox model as the likelihood and generates a chain of posterior distribution samples by the Gibbs Sampler.

The Bayesian Cox PH model invokes the Bayesian analysis with generating samples by Gibbs sampler to maintain reproducibility to accumulate the posterior distribution samples using the data. By default, a uniform prior distribution is assumed on the regression coefficient Group. The uniform prior is a flat prior on the real line with a distribution that reflects ignorance of the location of the parameter, placing equal probability on all possible values the regression coefficient can take. Using the uniform prior, it would expect the Bayesian estimates to resemble the classical results of maximizing the likelihood (see: Table2). It should make sure that the posterior distribution samples have achieved convergence before using them for Bayesian inference.
With autocorrelations retreating quickly to 0 and large effective sample sizes (Table 5), both diagnostics indicate a reasonably good mixing of the Markov chain.

| Param    | N   | Mean | SD    | 95% HPD Interval | ESS   | Efficiency |
|----------|-----|------|-------|------------------|-------|------------|
| Age>50   | 10000 | 0.107 | 0.160 | -0.204 0.423 0.939 0.741 0.550 0.079 | 265 | 37.67 | 0.026 |
| Stage3A  | 10000 | -0.027 | 0.201 | -0.431 0.359 0.950 0.776 0.608 0.0924 | 248 | 40.27 | 0.024 |
| stage3B  | 10000 | 0.609 | 0.195 | 0.2383 1 0.943 0.75 0.568 0.0506 | 245 | 40.84 | 0.024 |
| id2      | 10000 | -0.081 | 0.33 | -0.791 0.5 0.907 0.62 0.395 0.0297 | 393 | 25.46 | 0.039 |
| id4      | 10000 | -0.040 | 0.280 | -0.545 0.574 0.955 0.800 0.646 0.2458 | 92 | 108.7 | 0.009 |
| Theta    | 10000 | 0.115 | 0.081 | 0.0417 0.31 0.984 0.979 0.976 0.9473 | 22 | 455.4 | 0.002 |

Trace, autocorrelation and density plots are produced for each of the four parameters in figure 1 also confirm the convergence of the Markov chain. It is imperative that these are examined before any conclusions are drawn from the simulated posterior samples. The results shown on the table 6 are almost perfect. The trace plot show excellent mixing, the autocorrelation decreases to near zero, and the density is bell-shaped. The trace plots are centered near

Table 5: Posterior Summaries and Intervals, Posterior Autocorrelations and Effective Sample Sizes
their respective posterior mean and later the posterior space with small fluctuations. For the theta which corresponds to the all group, the trace plot is centered near the posterior mean. Samples in both tails are covered. These results exhibit convergence of the Markov chain to its stationary distribution.

Figure 1: Confirmation of convergence by Trace, autocorrelation and density plots

The first hazard ratio table compares the hazards between the age group less than 50 years and more than 50 years. Summaries of the posterior distribution of the corresponding hazard ratio are shown in Table 6. There is a 95% chance that the hazard ratio of the age group less than 50 years and more than 50 years lies between 0.655 and 1.244. The second hazard ratio tables for three types of combinations with three different pairs of stages are being compared. Assesses the change of hazards for each pairs of stages and compares the stage 2A vs Stage 3A, stage 2A vs Stage 3B and stage 3A vs Stage 3B. Hazards ratio for Stage 2A vs Stage 3A lies between 0.694 and 1.538, hazards ratio for stage 2A vs Stage 3B lies between 0.369 and 0.795. Similarly the hazards ratio for stage 3A vs Stage 3B lies between 0.371 and 0.767. Also it reports the simple statistics, percentiles, credible intervals, and high probability density (HPD) intervals for each of the parameters based on the posterior sample of 10000. Because the priors used are non-informative, the mean, standard deviation and credible interval should be fairly close to the corresponding maximum likelihood estimates (estimate, standard error, 95% CI).

Table 6: Hazard Ratios for age group

| Description | N   | Mean | Percentiles /Quantiles | Posterior Intervals |
|-------------|-----|------|-------------------------|---------------------|
|             |     |      | SD 25% 50% 75% 95% Equal-Tail 95% HPD Interval |
| Age<50 vs >50 | 10000 | 0.9099 | 0.149 0.820 0.893 1.001 0.655 1.244 0.644 1.215 |
| Hazard Ratios for stage |
| sta 1 vs 2 | 10000 | 1.0485 | 0.2138 0.889 1.035 1.174 0.694 1.538 0.631 1.448 |
| sta 1 vs 3 | 10000 | 0.554 | 0.1085 0.477 0.547 0.623 0.369 0.795 0.359 0.771 |
| sta 2 vs 3 | 10000 | 0.5383 | 0.1011 0.464 0.529 0.594 0.371 0.767 0.349 0.729 |

Discussion
Bayesian Cox PH model proposed to fit flexible survival models for non-informative censored breast cancer data. Using SAS University Edition Virtual Application Software, we presented the comparable results as compared with the results of the seminal paper (Pari Dayal et al. 2013). Draw information based on different types of deviance criteria along with various additional supportive measures. The results which are presented in this paper followed the same trend and in fact it showed the reality. Results in all tables and all visual approximate estimates which are presented in this paper are consistent. The DIC is a suitable device to draw conclusions. Before drawing inferences from the posterior sample, we should examine the trace, autocorrelation and density plots for each parameter to be content that the underlying chain has converged. The plots for the two parameters shown the mixing in the chain is acceptable, although we notice long correlation times. The hazard ratio statement delivers the Bayes solution corresponding to the previous classical analysis in Table 2.

These results we can also use post sample to assess the posterior probability that the HR for age vs stage after the event of interest is <1. The probability is over 99%. However, an enormous statistical knowledge is required for it to be used correctly. This approach provides an alternative validation that could be used to confirm results of ‘frequentist’ approach. Bayesian inference has a number of advantages over the frequentist approaches, mostly in the flexibility of model-building for time to event breast cancer survival data. In addition, for many models, ‘frequentist’ inference can be obtained as a special case of Bayesian inference with the use of non-informative priors (Ibrahim et al., 2001). The Bayesian approach enables us to formulate accurate inference based on the posterior distribution for any sample size, whereas the ‘frequentist’ approach relies heavily on the large sample approximation. The most important concern is that there is a risk involved in the erroneous usage of the Bayesian methods which could lead to improper data analysis.

REFERENCE
1. Aalen O. O. (1994). Effects of frailty in Survival Analysis. Statistical Methods in Medical Research, 3, 227-43. | 2. Casella, G., and George, E.I. (1992). Explaining the Gibbs sampler. The American Statistician, 46, 167-74. | 3. Chib, S. and Greenberg, E. (1995), “Understanding the Metropolis-Hastings Algorithm,” American Statistician, 49, 327-335. | 4. Cox, D. R. (1972). Regression model and life tables (with discussion). Journal of the Royal Statistical Society(B), 34, 187-220. | 5. Cox, D. R and Oakes, D. (1984). Analysis of Survival Data. London Chapman and Hall. | 6. Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409. | 7. Gelfand, A. E., and Smith, A. F. M. (1991). Gibbs sampling for marginal posterior expectations, Communications in Statistics, A, 20, 1747-66. | 8. Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on pattern analysis and Machine Intelligence, 6, 721-41. | 9. Grenander, U. (1983). Tutorial in pattern theory. Technical Report. Providence, R.I: Division of Applied Mathematics, Brown University. | 10. Ibrahim, G. J., Chen, M.-H., and Sinha, D. (2001). Bayesian Survival Analysis. New York: Springer. | 11. Ibrahim, G. J., Chen, M.-H., and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. Statistica Sinica, 14, 863-83. | 12. Pari Dayal L, Leo Alexander T, Ponnuraja C, and Venkatesan P. Modelling of breast cancer survival data: A frailty model approach. Indian Journal of Applied Research, 2013, 3(10), 22-24 |