Analyzing Survival Data as Binary Outcomes with Logistic Regression

Johan Lim[a], Kyeong Eun Lee[1,b], Kyu S. Hahn[c], Kunwoo Park[a]

[a]Department of Statistics, Seoul National University,
[b] Department of Statistics, Kyungpook National University
[c]Underwood International College, Yonsei University

Abstract

Clinical researchers often analyze survival data as binary outcomes using the logistic regression method. This paper examines the information loss resulting from analyzing survival time as binary outcomes. We first demonstrate that, under the proportional hazard assumption, this binary discretization does result in a significant information loss. Second, when fitting a logistic model to survival time data, researchers inadvertently use the maximal statistic. We implement a numerical study to examine the properties of the reference distribution for this statistic. Finally, we show that the logistic regression method can still be a useful tool for analyzing survival data in particular when the proportional hazard assumption is questionable.

Keywords: Information loss, logistic regression, survival data.

1. Introduction

This study examines the efficiency of the logistic regression model when used as an alternative to the statistical methods specifically designed for analyzing time-to-event data. In biomedical studies, in assessing the effects of experimental treatment in survival outcome, many clinical researchers rely on a logistic regression model after discretizing continuous observations (Annesi et al., 1989; Cain et al., 1994; Moriguchi et al., 2006). For example, when assessing the effectiveness of a newly developed drug for treating myocardial infection, one may show that 65% of the participants in the treatment group had survived over 5 years, whereas only 27% in the control group did so.

In discretizing continuous or ordinal data into binary counts, the foremost important concern arises from the subjective nature of the choice concerning the cutoff point for aggregation. Sometimes the researcher chooses this cutoff point because of its substantive meaning. For example, cancer researchers may choose the five-year post surgery period as the cutoff point because experience shows that the likelihood of recurrence declines drastically thereafter. However, many researchers arbitrarily choose this cutoff point; or, even worse, researchers often choose the time point at which the resulting binary outcomes maximally differentiate the treatment and the control groups.

When subjectively choosing the cutoff point that maximally separates the survival times of the treatment and the control groups, we argue that the standard normal distribution should no longer be the reference distribution. Instead, one ought to consider the distribution of the maxima of dependent multivariate normal distributions as the reference distribution. Naturally, this reference distribution has heavier tails than the standard normal distribution.

[1] Assistant Professor, Department of Statistics, Kyungpook National University, 1370 Sankyuk-Dong, Buk-Gu, Daegu 702-701, Korea.
E-mail: artlee@knu.ac.kr

When discretizing continuous observations, another natural concern is the potential loss of information. As Abbott (1985) argued, conceptually it is evident that the inference based on the proportional hazard model is considerably more informative than that based on an analysis of the survival outcome at a fixed time point. However, for practical data analysis, if the statistical method designed for binary data (e.g. logistic regression or the methods for analyzing a 2×2 table) attains a high level of discriminant or predictive power, information loss can be mitigated. Further a set of logistic regressions could provide provide better understanding on the treatment effect than the proportional hazards model when it varies over time. Thus, it is important to keep track of the gain or the loss incurred by discretization.

The paper is organized as follows. In Section 2, we launch an in-depth analysis of the gain and loss incurred by discretization. We first study the danger from a subjectively chosen cut-off point for the discretization. Secondly, we discuss the information loss resulting from discretization when the true underlying model is the proportional hazards model. Finally, we discuss potential benefit of using a set of logistic regressions. In Section 3, we illustrate our arguments by analyzing the Veteran's lung cancer data reported in Kalbfleisch and Prentice (1980). Section 4 concludes our discussion.

2. Gain and loss by discretization

2.1. Choosing a cutoff point for discretization

We first assess the danger associated with arbitrarily choosing the cutoff point when discretizing data before fitting a logistic regression model to survival time data. Researchers often choose the cutoff point c to maximize the treatment effects. In other words, they choose the point where the survival probabilities between the control and the treatment group are most discrepant. Under this setting the testing statistic to be used is

$$\mathrm{T}_{\max} = \max_{c} \frac{\widehat{p}_1(c) - \widehat{p}_2(c)}{\sqrt{\widehat{p}_1(c)(1-\widehat{p}_1(c))/n + \widehat{p}_2(c)(1-\widehat{p}_2(c))/m}},$$

where n and m are the number of subjects from the control and the treatment group, and $p_1(c)$ and $p_2(c)$ are the probabilities that a subject from the control and the treatment group survives more than c.

It is clear that the asymptotic reference distribution of $\mathrm{T}_{\max}$ is no longer the standard normal distribution. To learn more about the reference distribution of the test statistic T, we implement a numerical study. In doing so, we generate 100 random samples for the treatment and the control groups, where each random sample is generated from the exponential distribution with mean 100. Thus, the total sample size is 200. We generate 1000 such data sets and conduct a permutation test. Here, we set $c = 50, 100, 150$, and 200 and obtain the test statistic while recording the maximum value from each sample. Figure 1 summarizes our results by comparing the cumulative distribution function of the test statistic with that under the standard normal distribution.

As shown in Figure 1, our results clearly show that the cumulative distribution of the test statistic T has much heavier tails than the standard normal distribution. Accordingly, using the standard normal distribution as the theoretical reference distribution will increase the likelihood of committing Type I error.

2.2. Information loss from discretization

Several studies have examined the information loss resulting from fitting a logistic regression model to survival data in various settings. Comparing the parameter estimates from the proportional hazard

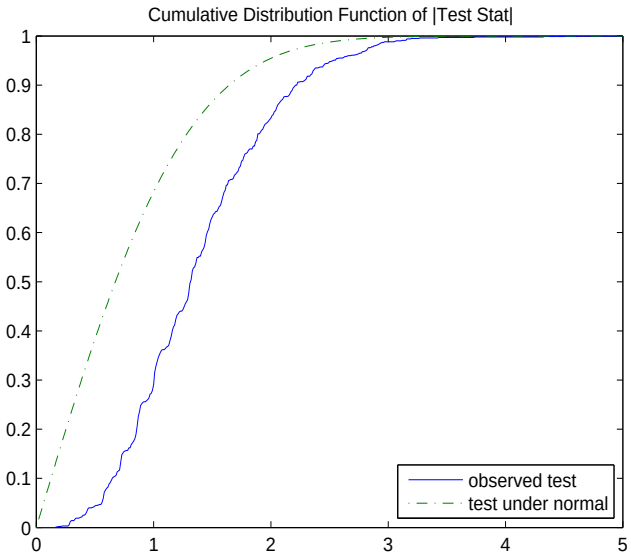


Figure 1: *The cumulative distribution functions of* T_{max} *and the standard normal distribution*

model (PHM) and a modified logistic regression model, Ingram and Kleinman (1989) showed that the two models yielded nearly identical results. Annesi et al.(1989) compared the asymptotic relative efficiency of the logistic regression approach and the proportional hazard model. The authors concluded that the latter is superior to the former when analyzing longitudinal data. Moriguchi et al. (2006) compared the PHM and the logistic regression models in a retrospective study concerning the prognosis of gastric cancer patients. Their results showed that both models identified the same set of risk factors, although the magnitudes of regression coefficients in the two sets of models were somewhat discrepant.

As in previous studies, we study the issue of information loss in a specific model setting that is very common in clinical studies. More specifically, we compare the efficiency of the logistic regression model with that of the survival regression model based on the partial likelihood in terms of their Fisher information (see Moriguchi et al. 2006). In order to do so, we assume the data are generated from two samples with differential hazards; a similar assumption is made in Efron (1977)'s study measuring the information loss resulting from using the partial likelihood. The author compared the partial and the logistic likelihoods in terms of the information on the parameters of the assumed survival model.

2.3. Information

Suppose that a clinical study is conducted to investigate a treatment of interest in a randomized design. n and m subjects are randomly assigned to control and treatment groups respectively. Without loss of generality, we assume that the true hazard rate of the underlying survival time in the placebo group is $\lambda(t) = 1$, whereas that for the treatment group is $\lambda(t) = \lambda$. Let $(U_{11}, \ldots, U_{1n})$ and $(U_{21}, \ldots, U_{2m})$ denote the survival times in the control and treatment groups respectively. U_{ij}s follow the exponential distribution with rate λ_i and independently and identically distributed for $i = 1, 2$ and $j = 1, 2, \ldots, n_i$. For simplicity, we let $\lambda_1 = 1$, $\lambda_2 = \lambda$, $n_1 = n$, and $n_2 = m$.

A simple approach to analyzing these data (without considering the censoring in survival time) is to define a binary outcome $W_i = I(U_{1i} < c)$ and $V_j = I(U_{2j} < c)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where c is the cutoff point for discretization. If one fits a logistic regression model to these transformed data, the treatment and the response variables can be written as

$$(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} 0 & \ldots & 0 & 1 & \ldots & 1 \\ w_1 & \ldots & w_n & v_1 & \ldots & v_m \end{pmatrix}^{\mathrm{T}}$$

where x is the indicator variable capturing the group membership of each patient (0 for the control group, 1 for the treatment group) and y is the binary survival outcome.

Alternatively, one can resort to the proportional hazard model. The most commonly assumed probability model is in the partial likelihood form, and it is constructed by comparing the risk of the failed subject to that of all the other subjects at each time point.

In this section, our interest is to compare the efficiency of the two methods in estimating treatment effects when the data is from a proportional hazards model. We compute the Fisher information of the parameter in the proportional hazard model in applying the logistic regression model and the partial likelihood for the proportional hazards model.

As a preliminary work, we derive the Fisher information on λ for the two methods. It would have been ideal if the two quantities have simple expressions with respect to the true λ; unfortunately, this is not the case with the partial likelihood approach.

2.3.1. Logistic regression

If one chooses to work with the binary survival outcome with the cutoff point c, the transformed variables $\{W_i,\ i = 1, 2, \ldots, n\}$ and $\{V_j,\ j = 1, 2, \ldots, m\}$ have the independent Bernoulli distribution

$$W_i \sim \text{Bernoulli}(p(0|c)), \quad V_j \sim \text{Bernoulli}(p(1|c)),$$

with corresponding success probabilities $p(0|c) = P(U_{1i} < c) = 1 - \exp(-c)$ and $p(1|c) = P(U_{2j} < c) = 1 - \exp(-\lambda c)$. For notational simplicity, we let $p_0 = p(0|c)$ and $p_1 = p(1|c)$ below. The logistic model is

$$\log\left(\frac{p(x|c)}{1 - p(x|c)}\right) = \mu + \alpha x \tag{2.1}$$

for $x = 0, 1$, where

$$\mu = \log\left(\frac{p_0}{1 - p_0}\right), \quad \text{and} \quad \alpha = \log\left(\frac{p_1}{1 - p_1}\right) - \log\left(\frac{p_0}{1 - p_0}\right).$$

The log-likelihood function for (2.1) (Jung, 2009) is

$$\sum_{i=1}^{n} w_i\mu - n\log\left(1 + \exp(\mu)\right) + \sum_{j=1}^{m} v_j(\mu + \alpha) - m\log\left(1 + \exp(\mu + \alpha)\right).$$

and the Fisher information matrix of α and μ is

$$\text{FI}(\alpha, \mu) = m\frac{\exp(\mu + \alpha)}{\left(1 + \exp(\mu + \alpha)\right)^2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + n\frac{\exp(\mu)}{\left(1 + \exp(\mu + \alpha)\right)^2}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Finally, the information on λ is

$$\begin{aligned} \mathrm{FI}^{\mathrm{logit}}(\lambda) &= \left\{\left(\partial\alpha/\partial\lambda, \partial\mu/\partial\lambda\right) \mathrm{FI}(\alpha,\mu)^{-1} \left(\partial\alpha/\partial\lambda, \partial\mu/\partial\lambda\right)^{\mathrm{T}}\right\}^{-1} \\ &= \left(\frac{e^{\lambda c}-1}{ce^{\lambda c}}\right)^2 \left\{\frac{e^{2\lambda c}}{m\left(e^{\lambda c}-1\right)^2} + \frac{e^{2c}}{n\left(e^{c}-1\right)^2}\right\}^{-1} \end{aligned} \tag{2.2}$$

by noting $p_0 = 1 - e^{-c}$ and $p_1 = 1 - e^{-\lambda c}$, and

$$\begin{aligned} \alpha &= \log\left(p_1/(1-p_1)\right) - \log\left(p_0/(1-p_0)\right) = \log\left(\frac{e^{\lambda c}-1}{e^{c}-1}\right) \\ \mu &= \log\left(\frac{p_0}{1-p_0}\right) = \log\left(e^{c}-1\right). \end{aligned}$$

2.3.2. Proportional hazard model with partial Likelihood

If survival times is modeled via the partial likelihood without discretization, the baseline hazard function for the control and the treatment groups can be written as

$$\lambda_0(x) = \exp\left(\log(\lambda)x\right) = \begin{cases} 1 & \text{if } x = 0 \\ \lambda & \text{if } x = 1, \end{cases}$$

where 0 and 1 indicates the control and the treatment group, respectively. Note that $\beta = \log\lambda$ in a survival regression form.

The log partial likelihood(PL) with respect to β is defined as

$$\log \mathrm{PL}(\beta) = \sum_{i=1}^{n+m} \beta x_i - \log\Big(\sum_{j\in R_i} \exp(\beta x_j)\Big),$$

where R_i is the risk set at the i-th observed survival time or the set of subjects' indices who survive beyond x_i. The the information on β can be calculated by attaining the expectation of the negative second derivative

$$\begin{aligned} \mathrm{FI}(\beta) &= -\mathrm{E}\Big(\frac{\partial^2 \log \mathrm{PL}(\beta)}{\partial\beta^2}\Big) \\ &= \sum_{i=1}^{n+m} \Big\{\frac{\sum_{j\in R_i} x_j^2 \exp(\beta x_j)}{\sum_{j\in R_i}\exp(\beta x_j)} - \Big(\frac{\sum_{j\in R_i} x_j \exp(\beta x_j)}{\sum_{j\in R_i}\exp(\beta x_j)}\Big)^2\Big\}. \end{aligned}$$

As earlier, our objective is to compute the Fisher information on λ, not β, and we denote it $\mathrm{FI}(\lambda)^{\mathrm{PL}}$. Since, $\beta = \log\lambda$,

$$\mathrm{FI}^{\mathrm{PL}}(\lambda) = \lambda^2 \mathrm{FI}(\log\lambda).$$

Note that it is not amenable to derive a short-hand expression for $I(\lambda)$ with respect to λ. For comparison, we apply a simple Monte Carlo approximation technique to obtain the value based on 100 data sets generated from a fixed λ in the ensuing section.

2.3.3. Comparison

Given the Fisher information of the two estimation methods, the asymptotic relative efficiency (ARE)

$$\mathrm{RE}(\lambda) = \frac{\mathrm{FI}^{\mathrm{PL}}(\lambda)}{\mathrm{FI}^{\mathrm{logit}}(\lambda)}$$

will be computed for various choices of λ and c. The degree to which the choice of c is inappropriate, or the severity of information loss, is closely related to the proximity of c to the mean survival time $1/\lambda$ as well as the difference in the mean survival time between the control and the treatment groups.

For the two sample models discussed in Section 1, we first calculate and graphically profile the ARE when $\lambda = 1, 2, 3, 5$ and $c \in [0, 10]$. As indicated earlier, for each λ, we generate 100 simulated data sets to calculate the approximate Fisher information for the partial likelihood approach. At this stage, for simplification, two additional assumptions are made. First, we assume the sample sizes in the control and the treatment groups are identical (i.e., $n = m$). Also, we do not simulate censoring, which it may cause additional complications for the comparison of the information contents.

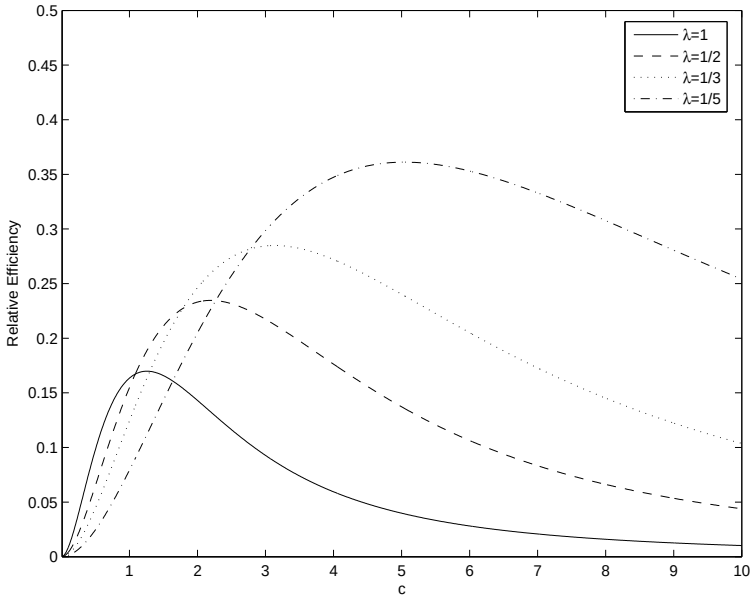


Figure 2: *The RE between the partial likelihood approach and the logistic regression for different choices of λ and c.*

Figure 2 plots the RE profiles with c's on the interval $[0, 10]$ for four choices of λ. Most notably, our results show that, for each λ, the optimal efficiency of logistic regression is achieved when the cutoff point c is approximately identical to the mean survival time $1/\lambda$.

It should also be noted that that the ARE increases as $1/\lambda$ increases, widening the gap in the mean survival times between the two groups. In other words, the amount of information loss incurred by using the logistic regression method decreases as the two groups' survival time become increasingly separable.

2.4. A set of logistic regressions

The previous section shows that the discretization of the data results in a severe information loss when the data is from the proportional hazard model. However, the underlying distribution is unknown, the dicretization and fitting a set of logistic regressions could provide a simple way to understand the treatment effect. In this section, we numerically illustrate how a set of logistic regressions can be used to understand the underlying model, particularly the non-proportional hazard model.

We consider the comparison of two survival curves, S_1 and S_2. The hazard rate of S_1 is

$$\lambda_1(t) = \begin{cases} 1/100, & 0 \le t < 50 \\ 1/10, & 50 \le t < \infty \end{cases}$$

and that of S_2 is

$$\lambda_2(t) = 1/50, \quad 0 \le t < \infty.$$

Figure 3 plots the survival curves S_1 and S_2.

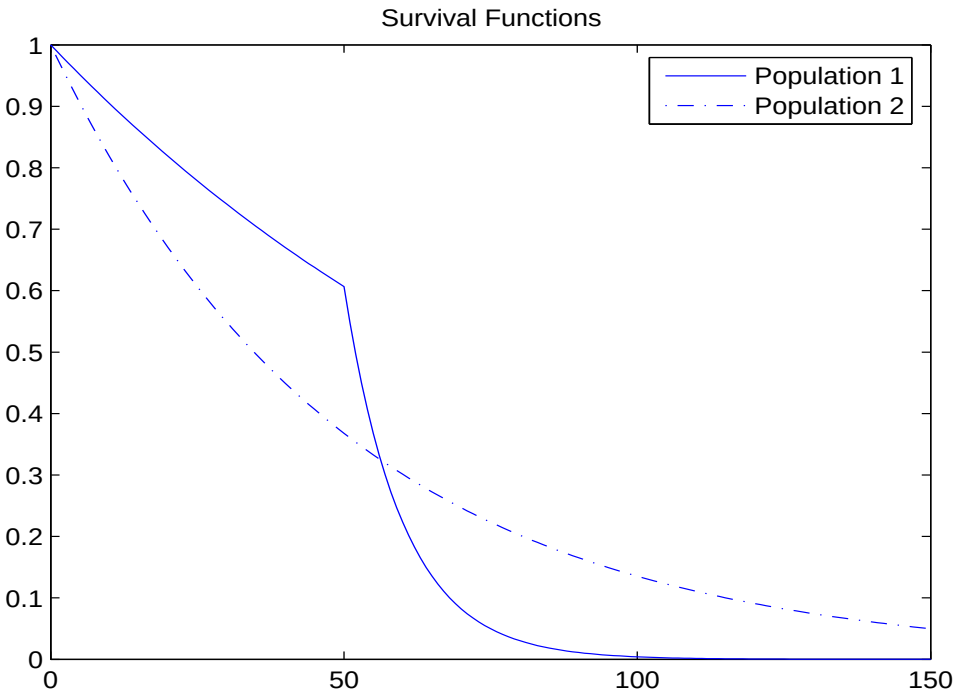


Figure 3: *True survival functions of two populations*

We generate 200 data sets with a size of $n = 30, 50, 100$ samples from each population. For each data set, we discretize the data using a given level c and fit the logistic regression to estimate the population effect. To be specific, for each t_i, the survival time of the $i-$th subject, we let

$$Y_i = \begin{cases} 0, & \text{if } t_i \le c, \\ 1, & \text{otherwise,} \end{cases}$$

and let

$$X_i = \begin{cases} 0, & \text{if the } i\text{−th subject is from Population 1} \\ 1, & \text{if the } i\text{−th subject is from Population 2.} \end{cases}$$

We then fit the logistic regression between Y_i and X_i:

$$\log \frac{\mathrm{P_c}(Y_i = 1 \big| X_i)}{1 - \mathrm{P_c}(Y_i = 1 \big| X_i)} = \beta_{0\mathrm{c}} + \beta_{1\mathrm{c}} X_i.$$

We choose $c = 30, 40, 50, 60, 70, 80, 90$, and 100. We estimate $\beta_{1\mathrm{c}}$ from the model and record their p-values. The 200 estimates and their p-values from 200 data sets are plotted in Figure 4.

The p-values and the t-values in Figure 4 show that $S_1(t)$ is larger than $S_2(t)$ when $t \leq 50$, whereas when it is smaller when $t \geq 70$. There is no difference between two functions around $t = 60$, the time point when $S_1(t)$ and $S_2(t)$ meet to each other. In summary, a series of regression analysis provides a good understanding of the underlying survival functions even when the hazards rates of two populations are not proportional to each other.

3. Example

In this section, we analyze the widely used Veteran's Administration lung cancer data (see Kalbfleisch and Prentice 1980). In the data sets analyzed in this section, males with inoperable lung cancer were randomly assigned to either the standard or the treatment chemotherapy conditions, and the end point for therapy was time to death. The data set also included the information concerning various covariates capturing the heterogeneity among patients such as disease extent and pathology, previous treatment of the disease, demographic background, and initial health conditions. Our analysis accounts for all the covariates included in the data set.

In our analysis, we discretize the survival time and fit a logistic regression model. In doing so, we choose the cutoff point c to be $20, 40, 60, 80$, or 100 (measured in weeks). Table 1 summarizes our results. To better understand the nature of treatment effect, we compare the marginal survival functions at each time point c for the patients in two experimental conditions.

factor	c=20	c=40	c=60	c=80	c=100	PHM
Intercept	0.2931	0.3096	0.4736	0.7949	0.8274	
Treatment	0.9685	0.3002	0.0928	0.0415	0.0071	0.085
Cell type	0.4773	0.0842	0.0077	0.0024	< 0.0001	0.0008
Squamous	0.6171	0.5273	0.1402	0.0590	0.0013	0.0003
Small	0.5073	0.1887	0.7623	0.6066	0.5521	0.29
Adeno 3	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Karno	0.3110	0.2437	0.4663	0.5497	0.3524	0.65
DD	0.0137	0.0934	0.2348	0.6094	0.3665	0.12
Pr. The.	0.0160	0.3996	0.4009	0.8502	0.2251	0.45

Table 1: The outputs of a set of chosen logistic regressions for the Veteran's Administration lung cancer data.

Our results show that, when fitting a logistic regression model, the statistical significance of treatment effects becomes inflated as c increases. When fitting the PHM, the p-value for the coefficient estimate concerning treatment effects was only .085. Against this baseline, for example, when $c = 80$ and 100, the p-values concerning treatment effects were .0415 and .0071 respectively. Thus, if we had discretized the data using $c = 80$ or 100, the logistic model would have seriously inflated the statistical significance of treatment effects. However, the results from a series of logistic regression analysis shows that the treatment effects become increasingly clear as c increases. This implies that, given all other covariates, the survival function of the treatment group is larger than that of the non-treatment group when t is large. In the current example, the proportional hazard assumption does not seem to be completely met. Under this circumstance, when used with caution, the logistic regression approach can provide a useful alternative to the proportional hazard model for analyzing survival time data.

4. Conclusion

In this article, we studied the information loss resulting from fitting a logistic regression model to survival time data after discretizing them into binary outcomes. Fitting the logistic model after discretizatizing the survival data results in a significant amount of information loss when compared with the correctly specified survival model. However, a careful use of a series of logistic regression analysis provides a good understanding of the underlying survival model. In particular, as shown in Section 3, it helps understand a non-proportional hazard model. We illustrate the gain and the loss resulting from discretization by analyzing the Veteran's Administration lung cancer data.

References

Abbott, R.D. (1985). Logistic regression in survival analysis. *American Journal of Epidemiology*, **121**, 465-471.

Annesi, I., Moreau, T., and Lellouch, J. (1989). Efficiency of the logistic regression and cox proportional hazards models in longitudinal studies. *Statistics in Medicine*, **8**, 1515-1521.

Cain, K.C., Martin, D.P., Holubkov, A.L., Raghunathan, T.E., Cole, W.G., Thompson, A., A logistic regression model of mortality following hospital admissions among medicare patients: comparison with HCFA's model. *AHSR FHSR Annual Meeting Abstract Book*, 1994, **11**, 81–82.

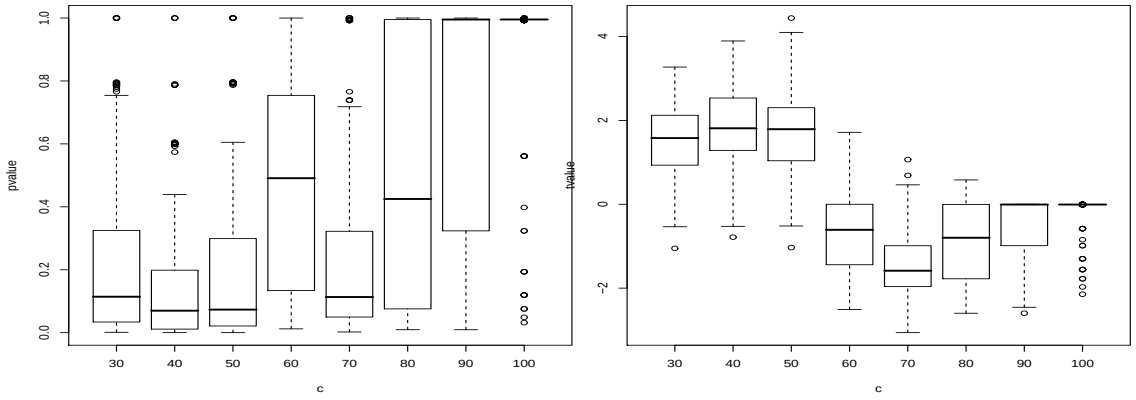
Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**, 557-565.

Ingram, D.D. and Kleinman, J.C. (1989). Empirical comparisons of proportional hazards amd logistic reression models. *Statistics in Medicine*, **8**, 525-538.

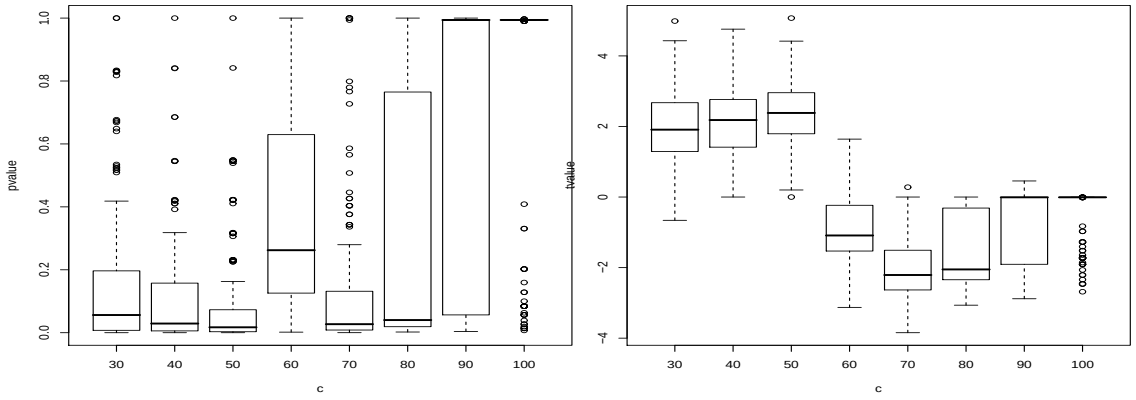
Jung, K-M. (2009). Multiple Deletions in Logistic Regression Models. *Communications of the Korean Statistical Society*, **16**, 309-315.

Kalbfleisch, J. D., Prentice, R. L.(1980) *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New York.

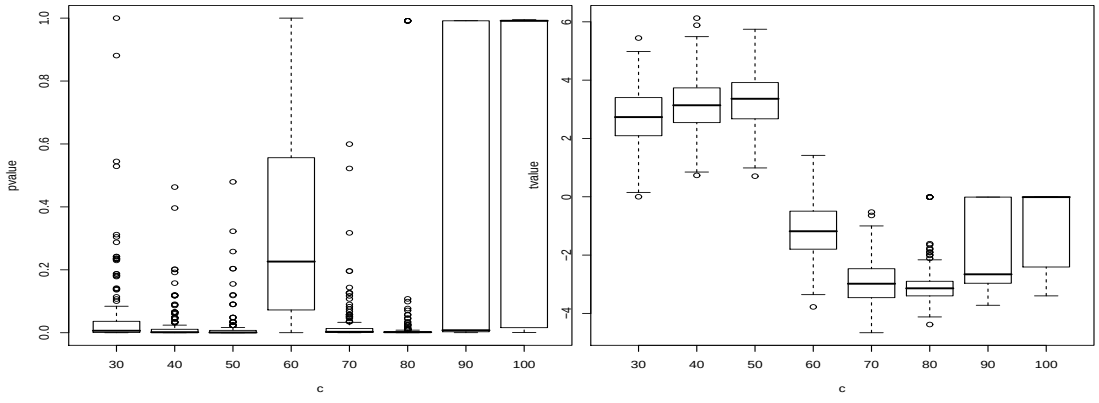
Moriguchi, S., Hayashi, Y., Nose, Y., Maehara, Y., Korenaga, D., Sugimachi, K. (2006). A comparison of the logistic regression and the cox proportional hazards models in the retrospective studies on the prognosis of patients with castric cancer. *Journal of Surgical Oncology.* **52**, 9–13.



(a) p-values for $n = 30$. (b) t-values for $n = 30$.



(c) p-values for $n = 50$. (d) t-values for $n = 50$.



(e) p-values for $n = 100$. (f) t-values for $n = 100$.

Figure 4: *p-values and t-values for testing β_1 in the logistic regressions.*