The Korean Communications
in Statistics Vol. 7, No. 1, 2000
pp. 327-335

# Estimation on Hazard Rates Change-Point Model[1]

Kwang Mo Jeong[2]

## Abstract

We are mainly interested in hazard rate changes which are usually occur in survival times of manufactured products or patients. We may expect early failures with one hazard rate and next another hazard rate. For this type of data we apply a hazard rate change-point model and estimate the unknown time point to improve the model adequacy. We introduce change-point logistic model to the discrete time hazard rates. The MLEs are obtained routinely and we also explain the suggested model through a dataset of survival times.

## 1. Introduction

Statistical inference such as statistical hypothesis tests or confidence intervals on the unknown time point with respect to which parameters of interest change is called change-point problem. Many researchers have studied the change-point problem in various respects, for example, according to the objects of changes, parametric versus nonparametric methods. Change-point models on means, variances and regression coefficients are the well-known subjects. In classical change-point problem the main concern has been on the mean changes in a sequence of random variables. If the functional forms of distributions are known parametric methods such as the maximum-likelihood estimation(MLE) and the likelihood ratio test(LRT) are usually used. Hinkley(1970), Worsley(1986) and Siegmund(1988), among others, are the researches of this type. On the other hand Bhattacharyya and Johnson(1968), Darkhovskh(1976), Carlstein(1988), and Boukai(1993), Chang, Chen and Hsiung(1994) studied the change-point problem in a nonparametric set

When a data structure has changed after a certain point of time one regression model to study the data obviously leaves the data unfitted or poorly explained by the assumed model. By applying change-point hypothesis the switching regression models have been studied by, among others, Quandt(1958, 1960), Brown, Durbin, and Evans(1975), Kim(1994), Chen(1998). We are mainly interested in hazard rate changes which are usually occur in survival times of manufactured products or patients. Hazard rate is a

[1] This Research was supported in part by Pusan National University Research Fund, 1995.

[2] Professor, Department of Statistics, Pusan National University, Pusan, 609-735, Korea.

manufactured products or patients. Hazard rate is a very important concepts in reliability theory. We may expect early failures with one hazard rate and next another hazard rate. For survival time data there exists high initial risk but it settles down to lower long term risk. For this type of data we apply a hazard rate change-point model and estimate the unknown time point to improve the model adequacy for the given data. We introduce change-point logistic model to the discrete time hazard rates. The MLEs are obtained routinely and we also explain the suggested model through a real dataset of survival times.

## 2. Discrete Hazard Rates

Kaplan-Meier survival estimator is important in analyzing censored data and its survival curve can easily be obtained via usually used statistical softwares. But it sometimes is inefficient compared to parametric survival estimators. Furthermore we cannot directly compare them by eye even in the absence of statistical noise. The hazard rates plot is more efficient in comparing survival times of two or more treatments. In this section we introduce the discrete hazard rate for discretized survival data even if it originally is in continuous form.

**Example 1.** The data in Table 1 denotes survival times for 51 patients of head-and-neck cancer (Efron, 1988), which was originally conducted by the Northern California Oncology Group. We may discretize the data by one-month intervals as shown in Table2. The notation '+' denotes censored observation.

Table 1. Head-and-neck cancer survival times for 51 patients

7, 34, 42, 63, 64, 74+, 83, 84, 91, 108, 112, 129, 133, 133, 139, 140, 140, 146, 149, 154, 157,
160, 160, 165, 173, 176, 185+, 218, 225, 241, 248, 273, 277, 279+, 297, 319+, 405, 417, 420,
440, 523, 523+, 583, 594, 1101, 1116+, 1146, 1226+, 1349+, 1412+, 1417

We introduce notations to explain discrete hazard rates model. Let $n_i$ = number of patients at the beginning of month i, $y_i$ = number of patients who died during month i, $y_i'$ =number of patients lost to follow-up during month i. We may assume that the number of deaths $y_i$ is binomially distributed given $n_i$. Hence the binomial density is of the form

$$\binom{n_i}{y_i}\pi_i^{\,y_i}(1-\pi_i)^{\,n_i-y_i},\quad y_i=0,1,2,\cdots,n_i.$$

Table 2. Discretized survival times by one-month intervals

| month | $n_i$ | $y_i$ | $y_i$ | month | $n_i$ | $y_i$ | $y_i'$ |
|---|---|---|---|---|---|---|---|
| 1 | 51 | 1 | 0 | 25 | 7 | 0 | 0 |
| 2 | 50 | 2 | 0 | 26 | 7 | 0 | 0 |
| 3 | 48 | 5 | 1 | 27 | 7 | 0 | 0 |
| 4 | 42 | 2 | 0 | 28 | 7 | 0 | 0 |
| 5 | 40 | 8 | 0 | 29 | 7 | 0 | 0 |
| 6 | 32 | 7 | 0 | 30 | 7 | 0 | 0 |
| 7 | 25 | 0 | 1 | 31 | 7 | 0 | 0 |
| 8 | 24 | 3 | 0 | 32 | 7 | 0 | 0 |
| 9 | 21 | 2 | 0 | 33 | 7 | 0 | 0 |
| 10 | 19 | 2 | 1 | 34 | 7 | 0 | 0 |
| 11 | 16 | 0 | 1 | 35 | 7 | 0 | 0 |
| 12 | 15 | 0 | 0 | 36 | 7 | 0 | 0 |
| 13 | 15 | 0 | 0 | 37 | 7 | 1 | 1 |
| 14 | 15 | 3 | 0 | 38 | 5 | 1 | 0 |
| 15 | 12 | 1 | 0 | 39 | 4 | 0 | 0 |
| 16 | 11 | 0 | 0 | 40 | 4 | 0 | 0 |
| 17 | 11 | 0 | 0 | 41 | 4 | 0 | 1 |
| 18 | 11 | 1 | 1 | 42 | 3 | 0 | 0 |
| 19 | 9 | 0 | 0 | 43 | 3 | 0 | 0 |
| 20 | 9 | 2 | 0 | 44 | 3 | 0 | 0 |
| 21 | 7 | 0 | 0 | 45 | 3 | 0 | 1 |
| 22 | 7 | 0 | 0 | 46 | 2 | 0 | 0 |
| 23 | 7 | 0 | 0 | 47 | 2 | 1 | 1 |
| 24 | 7 | 0 | 0 |  |  |  |  |

In this density $\pi_i$ is interpreted as discrete hazard rate defined by $\pi_i = P$ (patients dies during $i$th interval | patients survives until beginning of $i$th interval).

The life-table survival estimates of $\pi_i$ is given by $\tilde{\pi}_i = y_i / n_i$. When $n_i > 0$ this estimate is always unbiased for $\pi_i$, but is usually too variable to be of direct use as shown in Figure 1. As was discussed by Efron(1988) we can do better with a parametric model if the parametric assumptions are correct. The plot of $\tilde{\pi}_i$ against survival time is shown in Figure 1. Because of noise in the observed data we cannot easily catch the functional form of the smoothing curve representing the plot. A logistic regression can be a very natural one in modelling the hazard rates plot. But one single logistic model does not seem to be appropriate to modelling the hazard rates plot and so in the next section we suggest a logistic model with one change-point at some unknown time point.

## 3. Change-Point Logistic Model

Let $x$ be a single covariate for the binomial response $Y$ among $n_i$ trials with success probability $\pi(x)$. Consider a logistic model defined as

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\alpha+\beta x.$$

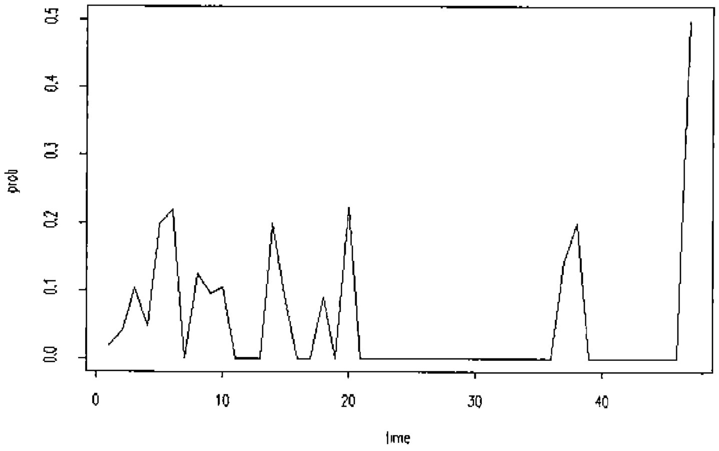

Figure 1. Plot of life-table hazard rates against time

To simplify the notation we denote $\log\{\pi(x)/(1-\pi(x))\}=\text{logit}(\pi)$. The observed covariates according to time sequences are denoted by $x_1,\cdots,x_n$. We are interested in the change-point logistic model with changing coefficients according to some time point $x_k$, which is the change-point. The change-point logistic model is defined by

$$\text{logit}(\pi_i)=\begin{cases}\alpha+\beta_1 x_i, & i=1,\cdots,k\\ \alpha+\beta_1 x_k+\beta_2(x_i-x_k), & i=k+1,\cdots,n\end{cases}\tag{3.1}$$

We note that the value of $\text{logit}(\pi_i)$ coincides at the change-point in the model. The main focus is to estimate the change-point and also to check for the model adequacy compared to a model with no change-point.

Let $\boldsymbol{\theta}=(\boldsymbol{\alpha},\beta_1,\beta_2)$ be a vector of unknown coefficients. Then the loglikelihood function $l(\boldsymbol{\theta})=\log \mathrm{L}(\boldsymbol{\theta})$ can be written as

$$\begin{aligned}\log L(\boldsymbol{\theta}) \propto & \sum_{i=1}^{k}\{y_i\log(\frac{\pi_i}{1-\pi_i})+n_i\log(1-\pi_i)\}\\ &+\sum_{i=k+1}^{n}\{y_i\log(\frac{\pi_i}{1-\pi_i})+n_i\log(1-\pi_i)\}\end{aligned}$$

If we substitute $\text{logit}(\pi_i)$ defined in (3.1) to the above equation then we can express it in terms of components of $\boldsymbol{\theta}$.

Next we discuss an estimation technique for the parameters. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is defined as an maximizer of loglikelihood function for given $x_k$. That is

$$\hat{\boldsymbol{\theta}} = \arg\max \log L(\boldsymbol{\theta}).$$

In order to obtain the MLEs of unknown parameters $\alpha$, $\beta_j$ and also unknown change-point $x_k$ we represent the likelihood function in the following form

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{k}[y_i(\alpha+\beta_1 x_i) - n_i \log(1+\exp(\alpha+\beta_1 x_i)] \\ &+ \sum_{i=k+1}^{N}[y_i\{\alpha+\beta_1 x_k + \beta_2(x_i - x_k)\} - n_i \log(1+\exp\{\alpha+\beta_1 x_k+\beta_2(x_i-x_k)\}] \end{aligned}$$

The MLEs of $\alpha$, $\beta_1$, $\beta_2$ can be obtained by a numerical methods such as Newton-Raphson iterative method. We briefly explain the Newton-Raphson procedure to obtain the MLEs of $\boldsymbol{\theta}$ given the change-point $x_k$. Let

$$\boldsymbol{q} = \left(\frac{\partial l(\boldsymbol{\theta})}{\partial \alpha}, \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_1}, \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_2}\right)'$$

be the vector of first order derivatives with respect to $\alpha$, $\beta_1$, $\beta_2$, respectively. The first order derivatives are calculated as

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \alpha} &= \sum_{i=1}^{N} y_i - \sum_{i=1}^{k} n_i \pi_{1i} - \sum_{i=k+1}^{N} n_i \pi_{2i} \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_1} &= \sum_{i=1}^{k} y_i x_i + x_k \sum_{i=k+1}^{N} y_i - \sum_{i=1}^{k} n_i x_i \pi_{1i} - x_k \sum_{i=k+1}^{N} n_i \pi_{2i} \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_2} &= \sum_{i=k+1}^{N} y_i x_i - x_k \sum_{i=k+1}^{N} y_i + x_k \sum_{i=k+1}^{N} n_i \pi_{2i} - \sum_{i=k+1}^{N} n_i x_i \pi_{2i} \end{aligned}$$

where

$$\pi_{1i} = \frac{\exp(\alpha+\beta_1 x_i)}{1+\exp(\alpha+\beta_1 x_i)}$$

and

$$\pi_{2i} = \frac{\exp\{\alpha+\beta_1 x_k+\beta_2(x_i-x_k)\}}{1+\exp\{\alpha+\beta_1 x_k+\beta_2(x_i-x_k)\}}.$$

Similarly we can obtain second order derivatives to find the Hessian matrix. The MLEs are obtained by Newton-Raphson algorithm in a routine method. As a byproduct we can also obtain the covariance matrix of MLEs from the Hessian matrix at the final step of iterations.

## 4. A Practical Example

In this section we explain the change-point hazard rate model via logistic regression for the data given in Table 2. We are interested in modelling the hazard rates using change-point logistic model defined in (3.1). One single logistic model does not seem to be appropriate to modelling the hazard rates plot and so we suggest a logistic model with one change-point at some unknown time point. By varying the assumed change-point $x_k$, $k=2,3,\cdots,n-1$, we first find the time point which maximizes likelihood function and next the estimates of regression coefficients for the given $x_k$. A Fortran program to find MLEs of logistic regression coefficients and change-point was performed on Unix Enterprise 3000. The MLEs of $\alpha$, $\beta_1$, $\beta_2$ are $\hat{\alpha}=-3.915$, $\hat{\beta_1}=0.400$, $\hat{\beta_2}=-0.065$ , respectively with change-point at $x_k=5$ . On the other hand if we assume a logistic model with no change-point the MLEs of $\alpha$ and $\beta$ are $\hat{\alpha}=-2.281$, $\hat{\beta}=-0.031$ , respectively. Hence the MLEs of discrete hazard rates are obtained from the relationship

$$\hat{\pi_i}=\begin{cases} \dfrac{\exp(-3.915+0.400x_i)}{1+\exp(-3.915+0.400x_i)}, & \text{for } i\leq 5 \\ \dfrac{\exp\{-3.915+0.400\,x_5-0.065\,(x_i-x_5)\}}{1+\exp\{-3.915+0.400\,x_5-0.065\,(x_i-x_5)\}}, & \text{for } i>5 \end{cases} \tag{4.1}$$

where $x_i$ is the survival time measured in one-month interval. The plot of $\hat{\pi_i}$ against $x_i$ is shown in Figure 2. We see a peak point at $x_k=5$ and the scheme of logistic regression changes with respect to this point. The covariance matrix of estimated parameters are given by

$$Cov(\hat{\theta})=\begin{pmatrix} 0.3580 & -0.0664 & 0.0024 \\ -0.0664 & 0.0135 & -0.0009 \\ 0.0024 & -0.0009 & 0.0004 \end{pmatrix}$$

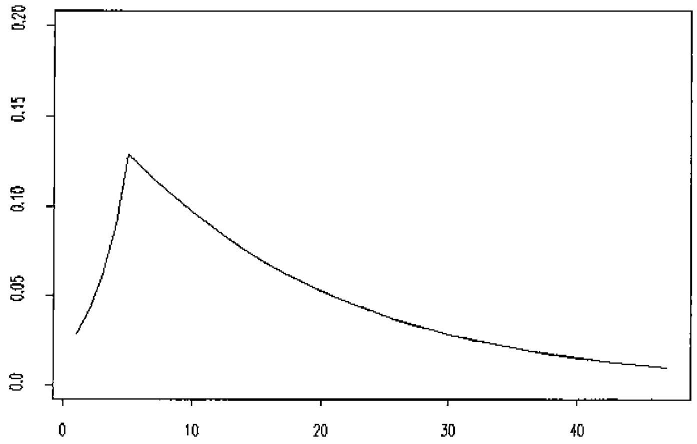

Figure 2. Plot of $\hat{\pi_i}$ against survival times under change-point model

The expected (or predicted) counts are obtained from the estimated probabilities multiplied by the number of patients surviving at the beginning of each interval. The expected counts and signed deviance residuals are given in Table 3. Here change-point corresponds to the change-point logistic model. On the other hand cubic-linear denotes the cubic-linear spline model with join point at 11 month. The cubic-linear logistic model was suggested by Efron(1988) to improve the model goodness-of-fit compared to linear or cubic logistic model.

Table 3. Expected counts and signed deviance residuals for two models

| month | $n_i$ | $y_i$ | expected counts | | signed deviance residuals | |
|---|---|---|---|---|---|---|
| | | | change-point | cubic-linear | change-point | cubic-linear |
| 1 | 51 | 1 | 1.47 | 0.76 | -0.42 | 0.27 |
| 2 | 50 | 2 | 2.12 | 2.18 | -0.09 | -0.13 |
| 3 | 48 | 5 | 2.98 | 4.16 | 1.11 | 0.42 |
| 4 | 42 | 2 | 3.78 | 5.31 | -1.04 | -1.73 |
| 5-6 | 72 | 15 | 9.02 | 10.40 | 1.97 | 1.46 |
| 7-8 | 49 | 3 | 5.46 | 5.41 | -1.21 | -1.19 |
| 9-11 | 56 | 4 | 5.42 | 3.54 | -0.67 | 0.25 |
| 12-14 | 45 | 3 | 3.63 | 2.18 | -0.35 | 0.54 |
| 15-18 | 45 | 2 | 2.95 | 2.05 | -0.60 | -0.03 |
| 19-24 | 46 | 2 | 2.25 | 1.91 | -0.17 | 0.06 |
| 25-31 | 49 | 0 | 1.58 | 1.80 | -1.79 | -1.91 |
| 32-38 | 47 | 2 | 0.98 | 1.52 | 0.91 | 0.38 |
| 39-47 | 28 | 1 | 0.36 | 0.78 | 0.88 | 0.24 |

The change-point logistic model seems to be well fitted in the respect of expected counts and signed deviance residuals. The sum of squares of signed deviance residuals is 13.64 with 9 degrees of freedom for the change-point logistic model. On the other hand for the cubic-linear logistic model it is 11.02 with 8 degrees of freedom. We note that there is no significant improvement for the more complex cubic-linear model.

## 5. Summary and Further Remarks

We considered a change-point logistic model on discrete hazard rates of survival time. The MLEs of logistic regression parameters and change-point were obtained. Newton-Raphson iterative algorithm was used to obtain MLEs. We can also obtain the covariance matrix of estimated parameters. The suggested model was explained through a real dataset of head-and-neck cancer survival times. We checked model goodness-of-fit in the respect of signed deviance residuals. The results were compared with that of cubic-linear logistic model with join point. The proposed change-point logistic model on hazard rates also performed quite well.

We didn't discuss the distribution of estimated change-point. But this topic is also an interesting problem in change-point model. The limiting distribution and other related problems will be remained as future researches. We also expect that other generalized linear models with different links may improve the model goodness-of-fit to the given data.

## References

[1] Bhattacharyya, G. K. and Johnson, R. A.(1968). Nonparametric Tests for Shift at Unknown Time Point, *Annals of Mathematical Statistics* 39, 1731-1743.

[2] Boukai, B.(1993). A Nonparametric Bootstrapped Estimate of the Change-point, *Nonparametric Statistics*, 3, 123-134.

[3] Brown, R. L., Durbin, J. and Evans, J. M.(1975). Techniques for Testing the Constancy of Regression Relationships over Time (with Discussion), *Journal of Royal Statistical Society B*, 149-192.

[4] Carlstein, E.(1988). Nonparametric Change-point Estimation, *The Annals of Statistics*, 16(1), 188-197.

[5] Chang, I. S., Chen, C. H. and Hsiung, C. A.(1994). Estimation in Change-Point Hazard Rate Models With Random Censorship, Change-Point Problems, *Institute of Mathematical Statistics, Lecture Notes* 23, 78-92.

[6] Chen, J.(1998). Testing for a Change-point in Linear Regression Models, *Communications in Statistics, Theory and Methods*, 27(10), 2481-2493.

[7] Chen, J. and Gupta, A. K. (1997). Testing and Locating Variance Change Points with Application to Stock Prices, *Journal of the American Statistical Association*, 92, 739-747.

[8] Darkhovskh, B. S.(1976). A Nonparametric Method for the Posterioi Detection of the "Disorder" Time of a Sequence of Independent Random Variables, *Theory of Probability and Application*, 21, 178-183.

[9] Hinkley, D. V.(1970). Inference About the Change-Point in a Sequence of Random Variables, *Biometrika* 57, 1, 1-17.

[10] Kim, H. (1994). Tests for a Change-point in Linear Regression, *Institute of Mathematical Statistics, Lecture Notes* 23, 170-176.

[11] Quandt, R. E.(1958). The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes, *Journal of the American Statistical Association*, 53, 873-880.

[12] Quandt, R. E.(1960). Tests of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes, *Journal of the American Statistical Association*, 55, 324-330.

[13] Efron, B.(1988). Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve, *Journal of the American Statistical Association* 83(402), 414-425.

[14] Siegmund, D.(1988). Confidence Sets in Change-Point Problems, *International Statistical Review*, 56, 1, 31-48.

[15] Worsley, K. J.(1986). Confidence Region and Tests for a Change-point in a Sequence of Exponential Family Random Variables, *Biometrika*, 73(1), 91-104.