Estimation of continuous odds ratio function with censored data*

중도절단된 자료를 포함한 승산비 연속함수의 추정

Jung-suk Kim** Chang-Hee Kwon***

목 차

Key Words : Case-control study, Odds rato function, Censored data

Abstract

The odds ratio is used for assessing the disease-exposure association, because epidemiological data for case-control or cohort studies are often summarized into 2x2 tables.

In this paper we define the odds ratio function(ORF) that extends odds ratio used on discrete survival event data to continuous survival time data, and propose estimation procedures with censored data. The first one is a nonparametric estimator based on the Nelson-Aalen estimator of comulative hazard function, and the others are obtained using the concept of empirical odds ratio. Asymptotic properties such as consistency and weak convergence results are also provided. The ORF provides a simple interpretation and is comparable to survival function or comulative hazard function in comparing two groups.

The mean square errors are investigated via Monte Carlo simulation. The result are finally illustrated using the Melanoma data.

* This paper was the Proceedings paper of the 52th Session of the International Statistical Institute, in helsinki. 1999.8.12

** Dean of researcher of The Research Association for National Regional Competitiveness. E-mail: rose0220@paran.com

*** Professor, Dept. of Computer Engineering in Hansei University

Ⅰ. Introduction

Consider two independent random variables X and Y with continuous distribution F1 and F2 respectively. In Epidemiologic setting, we may view X and Y sa a survival time from the individual expose and mom-exposed to a possible risk factor, repectively.

For a given time t>0, we define the odds ratio function $\psi(t)$ as follows:

$$\Psi(t) = \frac{S_1(t)/(1-F_1(t))}{S_2(t)/(1-F_2(t))},$$

where $S_1(t) = 1 - F_1(t) = P\{X < t\}$ is the corresponding survival function. Here the number of $\psi(t)$ is the odds of an exposed individual being survival and denominator the odds of an unexposed individual being survival. Then the odds ratio function $\psi(t)$ is can be represented as $\psi(t) = \dfrac{e^{\Lambda_2(t)}-1}{e^{\Lambda_1(t)}-1}$,

where $\Lambda_i(t) = -\ln S_i(t)$, $i=1,2$ is the corresponding comulative hazard function.

Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ be independent copies' from X and Y, and $C_1, \cdots, C_n$ and $D_1, \cdots, D_m$ be independent censoring times with survival functions G_1 and G_2 repectively.

We assume that the survival and censoring times are independent. Under the random censoring model, we may only observe $X_i = T_i \wedge C_i$, $i = 1, \cdots, n$, $Y_j = U_j \wedge D_j$, $j = 1, \cdots, m$, $\delta_i = I(T_i \leq C_i)$, $\epsilon_j = I(U_j \leq D_j)$.

Hence the survival functions corresponding to the observed T_i and U_i are $H_i(t) = S_i(t) G_i(t)$, $i = 1,2$, i.e. $(1-H) = (1-F)(1-G)$

In this case of the unexposed group, we can observe $(U_1, \epsilon_1)(U_2, \epsilon_2) \cdots (U_m, \epsilon_m)$

Ⅱ. Estimation of the odds ratio function

In the first case, let's call the estimator $\widehat{\Psi}_1(t)$ which is obtained from Nelson-Aalen estimator $\widehat{\Lambda}_1(t)$ and $\widehat{\Lambda}_2(t)$ instead of comulative risk function $\Lambda_1(t)$ and $\Lambda_2(t)$ in the definition of odds ratio

$$\widehat{\Psi}_1(t) = \frac{e^{\widehat{\Lambda}_2(t)}-1}{e^{\widehat{\Lambda}_1(t)}-1} \qquad \widehat{\Lambda}_i(t): \text{Nelson-Aalen estimator of } \Lambda_i(t).$$

Here the cummulative hazard function $\Lambda_1(t)$ and $\Lambda_2(t)$ corresponding to the exposed and unexposed group are usually estimated by the Nelson-Aalen estimator

$$\widehat{\Lambda}_1(t) = \sum_{i: T(i) \leq t} \frac{a_i}{R_{1i}} \quad \text{and} \quad \widehat{\Lambda}_2(t) = \sum_{j: U(j) \leq t} \frac{b_j}{R_{2j}}$$

where $T_{(1)} < \cdot \cdot < T_{(k)}$ and $U_{(1)} < \cdot \cdot < U_{(k)}$ denote the ordered distinct exposed-group and unexposed-group in observed survival times, and $a_i(b_j)$ the number of tied exposed-group (unexposed-group) survival times tied at $T_{(i)}(U_{(j)})$.

Hence $R_{1i}(R_{2j})$ represents the numbers of exposed-group (unexposed-group) survival times at risk just before times $T_{(i)}(U_{(j)})$.

The second case, in 2×2 table of cohort study, if A is the number of event in exposed group and C is the number of event in unexposed group, then the likelihood function of A and C is

$$\binom{n}{a}\binom{m}{c} p_1^a(1-p_1)^b p_2^c(1-p_2)^d,$$

that is shown

$$\binom{n}{a}\binom{m}{c}(1-p_1)^n(1-p_2)^m\left(\frac{p_2}{1-p_2}\right)^{a+c}\Psi^a$$

So A and A+C are sufficient statistic, the estimator of Ψ is based on conditional probability of A when given A+C=k.(Cox 1970, Cox & Hinkley:1974)

i.e.

$$P(a|n, m, k, \Psi) = \frac{\binom{n}{a}\binom{m}{c}\Psi^a}{\sum_{k=0}^{\infty}\binom{n}{i}\binom{m}{k-i}\Psi^a}$$

The conditional MLE of odds ratio function Ψ is the maximaum value of Ψ which maximize the above conditional distribution and so which satisfies $a = E(A|n, m, k, \Psi)$ (Breslow &Day:1980) On the other hand , (Woolf:1955) from normal asymptotic of Ψ

the mean of $\ln\hat{\Psi}$ is asymptotically $\ln\Psi$ and the variance is

$$Var(\ln\Psi) = \frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}$$

Let for given t>0

$$a(t) = \sum_{i=1}^{n} I(X_i > t) + \sum_{i=1}^{n} I(X_i = t, \delta_i = 0),$$

$$b(t) = n - \sum_{i=1}^{n} I(X_i < t, \ \delta_i = 0) - a(t),$$

$$c(t) = \sum_{j=1}^{m} I(Y_j > t) + \sum_{j=1}^{m} I(Y_j = t, \varepsilon_j = 0), \quad d(t) = m - \sum_{j=1}^{m} I(Y_j < t, \ \varepsilon_j = 0) - c(t),$$

where a(t) is the number of the survival times which is exactly larger then t and b(t) is the number of the survival times which is exactly shorter then t in exposed group. So, when we don't know whether the survival time T is larger or smaller than t, the number of $\sum_{i=1}^{n} I(X_i < t, \ \delta_i = 0)$ is excepted among n-number exposed group.

The same way, c(t)(or d(t)) is the number of survival time T which is larger(or shorter) time than t and also the number of $\sum_{j=1}^{m} I(Y_j < t, \ \varepsilon_j = 0)$ is excepted in m-number unexposed group. So

$$\widehat{\Psi}_2(t)=\frac{a(t)/b(t)}{c(t)/d(t)}=\frac{a(t)d(t)}{b(t)c(t)}$$

is the repaired $\widehat{\Psi}_2$. When we don't know the survival time is larger or shorter than t by incomplete obsevation in censoring time t, we can repair as follows:

3rd estimator of odds ratio function

Let
$$a'(t)=a(t)+\sum_{i:X_i<t,\ \delta_i=0}\frac{\widehat{S}_1(t)}{\widehat{S}_1(X_i)}$$

$$b'(t)=b(t)+\sum_{i:X_i<t,\ \delta_i=0}\left(1-\frac{\widehat{S}_1(t)}{\widehat{S}_1(X_i)}\right)$$

$$c'(t)=c(t)+\sum_{j:Y_j<t,\ \varepsilon_j=0}\frac{\widehat{S}_2(t)}{\widehat{S}_2(Y_j)}$$

$$d'(t)=d(t)+\sum_{j:Y_j<t,\ \varepsilon_j=0}\left(1-\frac{\widehat{S}_2(t)}{\widehat{S}_2(Y_j)}\right)$$

$\widehat{S}_1(t)$ and $\widehat{S}_2(t)$ are the estimator of K-M(Kaplan & Maier : 1958) and are shown like as

$$\widehat{S}_1(t)=\sum_{i:X_i\le t}\left(\frac{n-R_{1i}}{R_{1i}}\right)^{\delta_i},\qquad \widehat{S}_2(t)=\sum_{j:Y_j\le t}\left(\frac{m-R_{2i}}{R_{2j}}\right)^{\varepsilon_i}$$

Here, R_{1i} and R_{2i} are the risk set of X_i and Y_j in exposed group and unexposed group. So, the 3rd estimator of odds ratio function is

$$\widehat{\Psi}_3(t)=\frac{a'(t)/b'(t)}{c'(t)/d'(t)}=\frac{a'(t)d'(t)}{b'(t)c'(t)}.$$

III. Asymptotic properties

For a survival function S, define $t_s=\max\{t: S(t)>0\}$. Then we have the following lemma

.

Auxilary 3.1 The process $\sqrt{n}(\widehat{\Lambda}_1(t|z)-\Lambda_1(t|z))$ and $\sqrt{m}(\widehat{\Lambda}_2(t|z)-\Lambda_2(t|z))$ converge weakly to the process $W_1(t)$ and $W_2(t)$ in the spaces $D[0,t_{s_1}]$ and $D[0,t_{s_2}]$, where the process $W_1(t)$ and $W_2(t)$ are independent mean zero Gaussian process with covariance functions $V_1(s\wedge t)$ and $V_2(s\wedge t)$ respectively.

Here

$$V_1(t)=\int_0^t H_1^{-1}(s)\,d\Lambda_1(s) \text{ and } V_2(t)=\int_0^t H_2^{-1}(s)\,d\Lambda_2(s)$$

The asymptotic variances can be estimated consistantly by

$$\widehat{V}_1(t) = n\sum_{i:T(1)_i\le t}\frac{1}{R_{1i}^2},\qquad \widehat{V}_2(t) = n\sum_{j:U(j)_i\le t}\frac{1}{R_{2i}^2}$$

Thm 3.1. (uniform consistency of $\widehat{\Psi}_1(t|z)$)

For arbitrary $0<k<t_{H_1}\wedge t_{H_2}$ when $n, m\to\infty$

$\sup |\widehat{\Psi}_1(t|z)-\Psi(t|z)|\to 0\ a.e.$, i=1,2

proof) By definition of $\widehat{\Psi}_1(t|z)$

Since

$$\widehat{\Psi}_1(t|z)-\Psi_1(t|z) = \frac{1}{(e^{\widehat{\Lambda}_1(t|z)}-1)(e^{\Lambda_1(t|z)}-1)}\times\{(e^{\Lambda_1(t|z)}-1)(e^{\widehat{\Lambda}_2(t|z)}-1) - (e^{\widehat{\Lambda}_1(t|z)}-1)(e^{\Lambda_2(t|z)}-1)\}$$

$$= \frac{1}{(e^{\widehat{\Lambda}_1(t|z)}-1)(e^{\Lambda_1(t|z)}-1)}\times\{e^{\Lambda_1(t|z)}(e^{\widehat{\Lambda}_2(t|z)}-e^{\Lambda_2(t|z)}) - e^{\Lambda_2(t|z)}(e^{\widehat{\Lambda}_1(t|z)}-e^{\Lambda_1(t|z)}) + (e^{\widehat{\Lambda}_1(t|z)}-e^{\Lambda_1(t|z)})-(e^{\widehat{\Lambda}_2(t|z)}-e^{\Lambda_2(t|z)})\}$$

So we can get the result by auxilary2.1 and Taylor Thm.

Let $M=\frac{mn}{m+n}$ and $\widehat{\eta}_1=\frac{n}{m+n}$

Then we can get the following result.

Thm 3.2 (weak convergency of $\widehat{\Psi}_1(t|z)$)

When $n+m\to\infty$, suppose $\widehat{\eta}_1\to\eta_1\ (0<\eta_1<1)$.

Then the process $\sqrt{M}(e^{\widehat{\Lambda}_1(t|z)}-1)(\widehat{\Psi}_1(t|z)-\Psi_1(t|z))$ weakly converges to the Gaussian process Z(t|z) with mean 0 and variance $\gamma(t|z)$ in the space $D[0, t_{H_1}\wedge t_{H_2}]$ when M converges to ∞.

where $\gamma(t|z) = \{\eta_1 e^{2\Lambda_2(t|z)}V_2(t|z) - (1-\eta_1)e^{2\Lambda_1(t|z)}\Psi^2(t|z)V_1(t|z))\}$.

Reference 3.1)

(1) For given t, $\widehat{\Psi}_1(t|z)$has asymptotically mean $\Psi(t|z)$ variance

$$\mathrm{Var}(\widehat{\Psi}_1(t|z)) = \frac{1}{\sqrt{M}(e^{\Lambda_1(t)}-1)^2}(\eta_1 e^{2\Lambda_2(t|z)}V_2(t|z) - (1-\eta_1)e^{2\Lambda_1(t|z)}\Psi^2(t|z)V_1(t|z))).$$

(2) The consistency estimator of asymptotic variance Var($\widehat{\Psi}_1(t|z)$)

$$Var(\widehat{\Psi}_1(t)) = \frac{1}{\sqrt{M}(e^{\widehat{\Lambda}_1(t)}-1)^2}(\eta_1 e^{2\widehat{\Lambda}_2(t|z)}V_2(t|z) + (1-\widehat{\eta}_1)e^{2\widehat{\Lambda}_1(t|z)}\widehat{\Psi}^2(t|z)V(t|z)))$$

$$= \frac{\widehat{V}_2(t|z)}{m}\left(\frac{e^{\widehat{\Lambda}_2(t|z)}}{e^{\widehat{\Lambda}_2(t|z)}-1}\right)^2\widehat{\Psi}_1^2(t) + \frac{\widehat{V}_1(t|z)}{n}\left(\frac{e^{\widehat{\Lambda}_1(t|z)}}{e^{\widehat{\Lambda}_1(t|z)}-1}\right)^2\widehat{\Psi}_1^2(t).$$

IV. Monte Carlo Simulation

In this section, we compare MSE and Bias via Monte Carlo Simulation to compare the efficieny of estimator of the proposed three type function.

The censoring rate change with 10%, 30%, 50%, 70%, and uses n and m random number (n=m=30, 50, and 100) via IMSL system for each group and tries 500 replication in all simulation.

The Monte Carlo Simulation is show in(table4.1), we used increasing Weib(0.5, 2), decreasing Weib(1.0, 0.5) and constant Exp(0.2) in the distribution of exposed group.

The range of time t is given in

$$\max(F_1-1(0.1),F_2-1(0.1)),\min(F_1-1(0.9),F_2-1(0.9))$$

(1) In all 3 casem as time t goes the morre out point, the larger MSE and Bias are.

(2) Over all case, $\widehat{\Psi}_1(t)$ is good. Especially time t comes to the center point, the more good results can be seen.

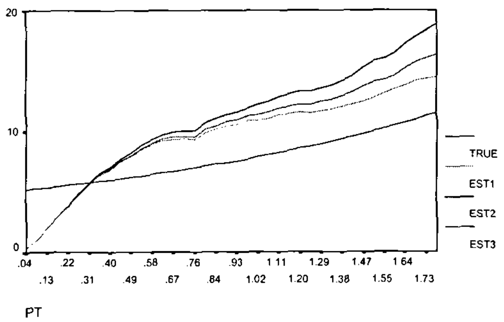
(3) $\widehat{\Psi}_3(t)$ is better than $\widehat{\Psi}_2(t)$.

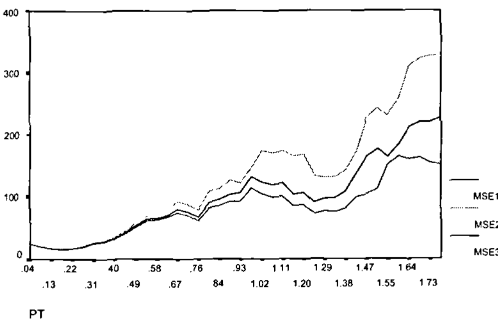
(4) As the sample size is larger, the MSE is smaller.

(5) When censoring rates become larger, in all case MSE is larger as much as, so uncertainty is higher.

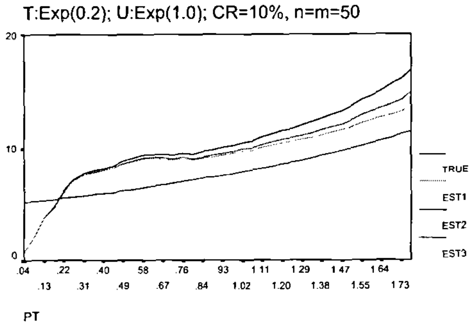
Case Group		Control Group		Censoring rate
F_1	G_1	F_2	G_2	
Weib(0.5, 2)	Exp(0.0853)	Exp(1)	Exp(0.1111)	10%
〃	Exp(0.2993)	〃	Exp(0.4286)	30%
〃	Exp(0.6120)	〃	Exp(1.0000)	50%
〃	Exp(1.1560)	〃	Exp(2.3333)	70%
Weib(1.0, 0.5)	Exp(0.0672)	Exp(1)	Exp(0.1111)	10%
〃	Exp(0.3757)	〃	Exp(0.4286)	30%
〃	Exp(1.3559)	〃	Exp(1.0000)	50%
〃	Exp(6.5260)	〃	Exp(2.3333)	70%
Weib(0.2, 1)	Exp(0.0222)	Exp(1)	Exp(0.1111)	10%
〃	Exp(0.0857)	〃	Exp(0.4286)	30%
〃	Exp(0.2000)	〃	Exp(1.0000)	50%
〃	Exp(0.4666)	〃	Exp(2.3333)	70 %

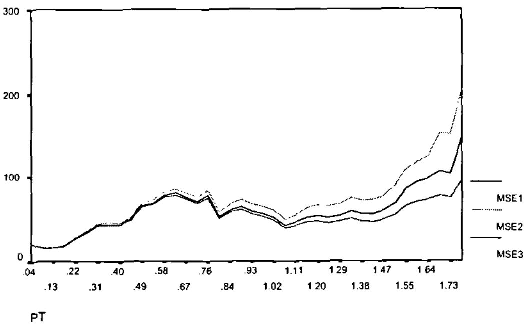
T:Exp(0.2); U:Exp(1.0); CR=10%, n=m=30



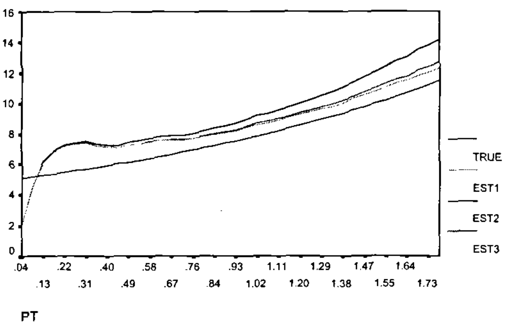


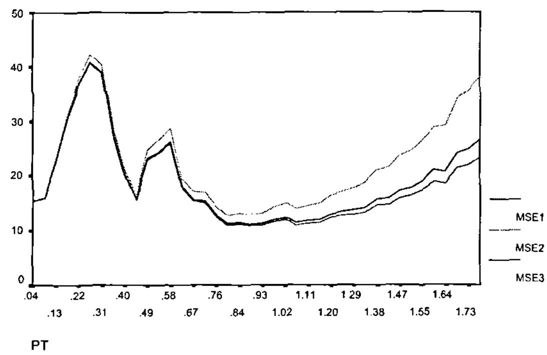
T:Exp(0.2); U:Exp(1.0); CR=10%, n=m=50



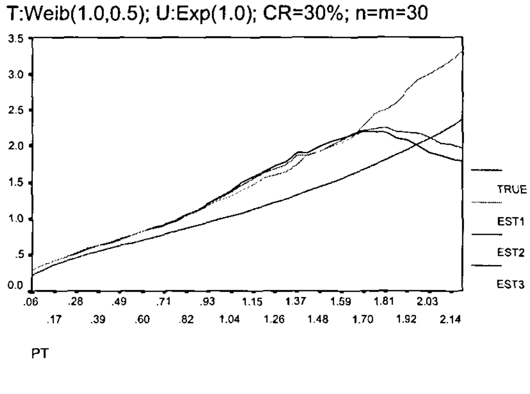


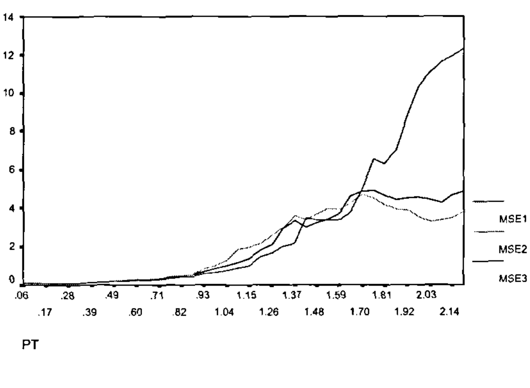
T:Exp(0.2); U:Exp(1.0); CR=10%, n=m=100



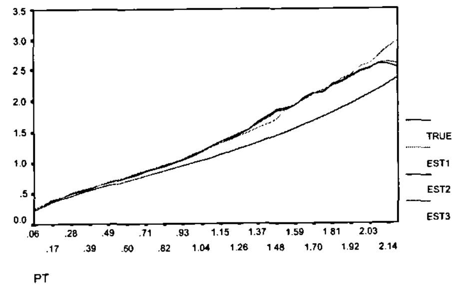


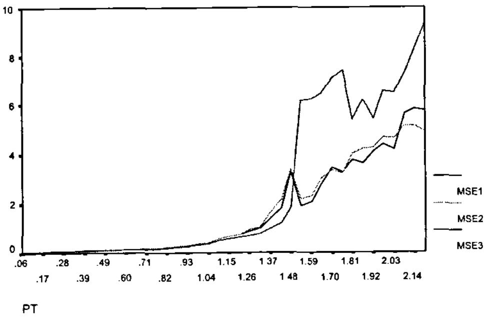
T:Weib(1.0,0.5); U:Exp(1.0); CR=30%; n=m=30



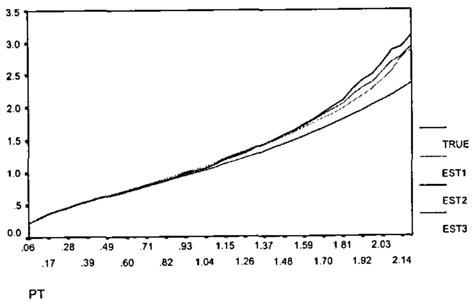


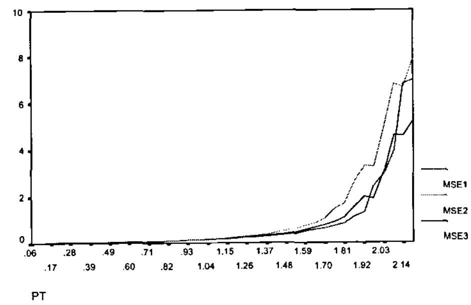
T:Weib(1.0,0.5); U:Exp(1.0); CR=30%; n=m=50



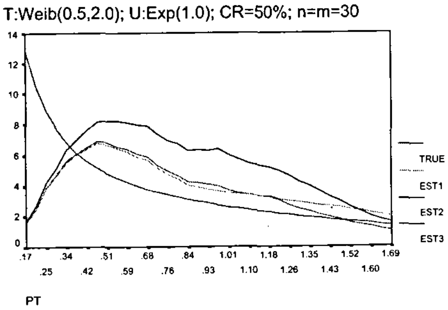


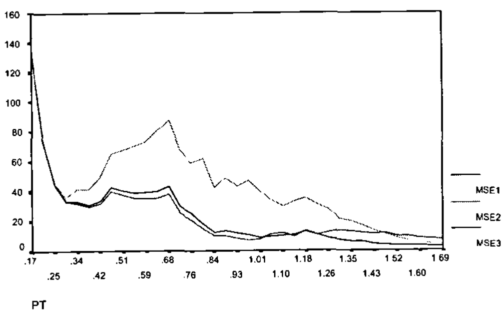
T:Weib(1.0,0.5); U:Exp(1.0); CR=30%; n=m=100



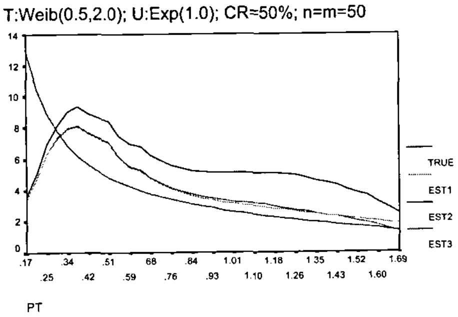


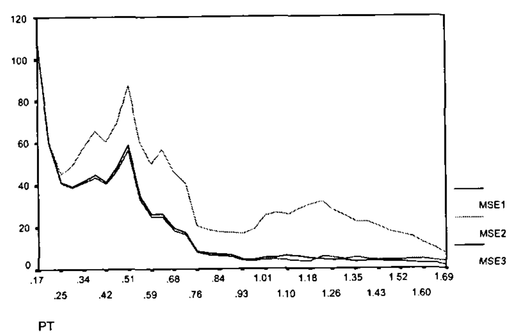
T:Weib(0.5,2.0); U:Exp(1.0); CR=50%; n=m=30

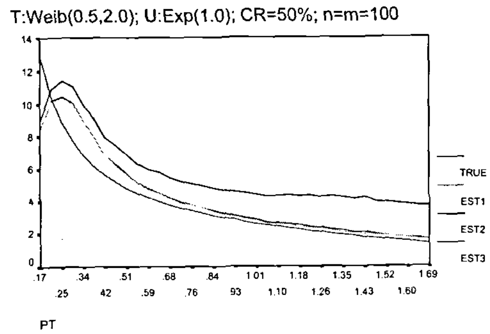


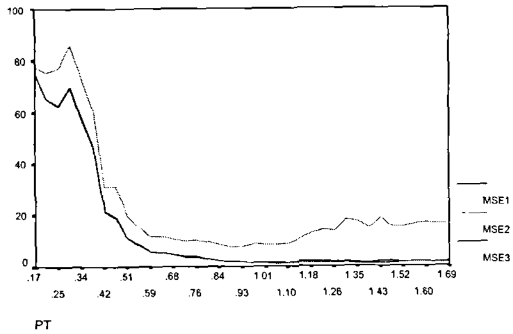


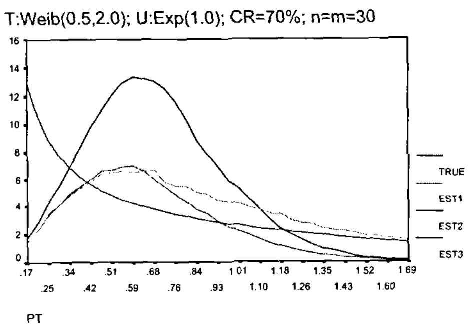
T:Weib(0.5,2.0); U:Exp(1.0); CR=50%; n=m=50

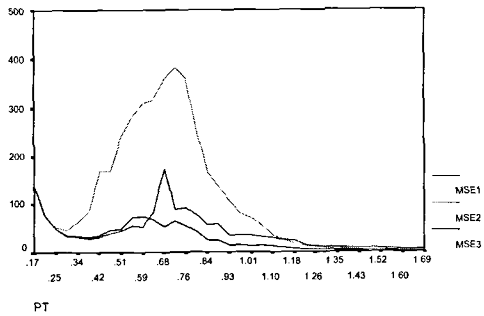


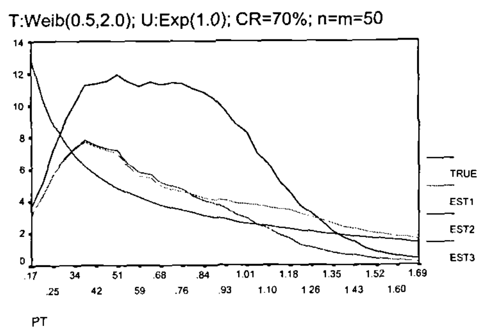


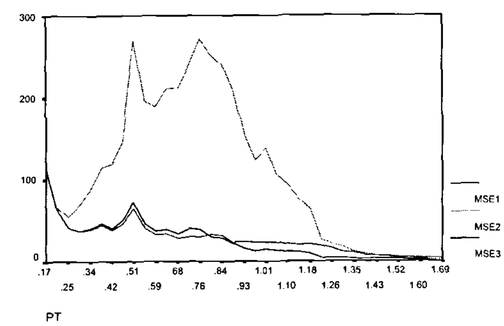












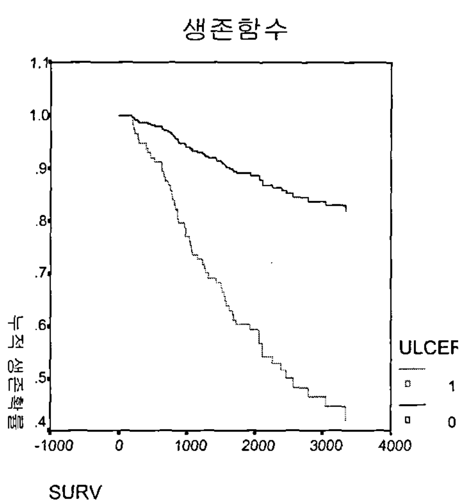
V. Malignant Melanoma Example

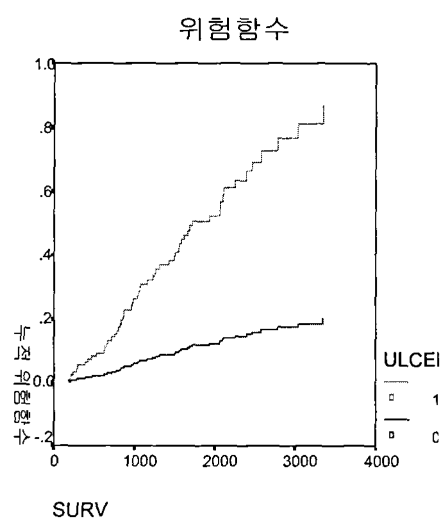
In the period 1962-77, 225 patients with malignant melanoma(cancer of the skin) had a radical operation performed at the Department of Plastic Surgery, University Hospital of Odense, Denmark. That is the tumor was completely removed together with the skin within a distance of about 2.5cm around it. All patients were followed until the end of 1977, that is, it was noted if and when any of the patients died.

This is a historically prospective clinical study with the object of assessing the effect of the risk factors on survival. The time variable viewed as most important is time since operation. Among the possible risk factors screened for significance were the sex and age at operation of the patient. Furthermore, clinical characteristics of the tumor such as tumor width and location on the body were considered as well as various histological classifications(that is, obtained by examination of the tissue), including tumor thickness, growth patterns, types of malignant cells, and ulceration. The latter factor is dichotomous and scored as "present" if the surface of the melanoma viewed in a microscope shows signs of ulcers and as "absent" otherwise. The material from 20 patients did not permit a histological evaluation and only the remaining 205 patients are considered here.

<u>Summary of the Number of Censored and Uncensored Values</u>

Ulcer	Total	Failed	Censored	%Censored
0	115	16	99	86.09
1	90	41	49	54.44
Total	205	57	148	72.20





References

1. Aalen, O. (1978). Nonparametric Inference for a Family of Counting Process, The Annals of Statistics, Vol. 6, 701-726.
2. Andersen, P.K. and Borgan, R.D. Gill, Keiding, N.(1992).Statistical Models Based on Counting Procosses, Springer-Verlag, New York.
3. Barmi,H.E.(1997). Testing For or Against a Trend in Odds Ratios, Communications in Statistics-Theory and Methods, Vol. 26, 1877-1891.
4. Breslow, N.E.(1996). Statistics in Epidemiology, Journal of the American Statistical Association, Vol. 91, No. 433, 14-28.
5. Cox, D.R. & Hinkley, D.V.(1974).Theoretical Statistics, Chapman and Hall, London.
6. Hanfelt, J.J. and Liang, K.Y(1998). Inference for Odds Ratio Models with Sparse Dependent Data, Biometrika, Vol. 85, 136-147
7. Kaplan and Meier.(1958). Nonparametric Estimation from Incomplete Observations, Journal of the American; Statistical Association, Vol. 53, 457-481.
8. Klein, J.P. and Moeschberger, M.L.(1997). Survival Analysis Techniques for Censored and Truncated Data, Springer-Verlag, New York.
9. Qaqish, B.F., Zhou, H. and Cai, J.(1997). On Case-Control Sampling of Clustered Data, Biometrika, Vol. 84, No. 4, 983-986.
10. Shen, X.(1998). Proportional Odds Regression and Sieve Maximum Likelihood Estimation, Biometrika, Vol. 85, 165-177.