The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 791-800

Nonparametric Regression with Left-Truncated and Right-Censored Data[1]

Jinho Park[2]

Abstract

Gross and Lai (1996) proposed a new approach for ordinary regression with left-truncated and right-censored (l.t.r.c) data. This paper shows how to apply nonparametric algorithms such as multivariate adaptive regression splines to l.t.r.c. data.

1. Introduction

In biostatistical applications we are often interested in survival times of individuals suffering from a disease. Typically, individuals suffering from a disease are followed and times of death of those of deceased are recorded. However, in many cases, their survival times are not completely observed. The most common kind of incomplete information is due to right censoring in survival data analysis. Recent interest has focused on left-truncated and right-censored data which arise in prospective studies of a disease. In prospective studies which begin at chronological time t_b, patients suffering from the disease who are alive at time t_b are recruited for the study. However, patients who have died before the beginning of the study cannot be subjects of this study. Suppose t_1 and t_2 denote the time of initial diagnosis and the time of death, respectively. Let $Y = t_2 - t_1$ and $T = t_b - t_1$. So Y is a survival time from the disease. In the above prospective studies, we can observe (Y, T) only when $Y \geq T$ (left truncation). Now suppose the study ends at time t_e, and let $C = t_e - t_1$. Because of the termination of the study, we can only observe $\min(Y, C)$ (right censoring). In this example, C and T are called a right-censoring variable and a left-truncation variable, respectively. An ordinary right-censored data without left truncation corresponds to the case $T = -\infty$.

In many applications, there are explanatory variables upon which survival time may depend. Let $X = (X_1, X_2, \ldots, X_M)^T$ and Y denote an M-dimensional covariate vector and a survival

[1] This research was supported by the Grant of Inha University, 1997-1998.

[2] Assistant Professor, Department of Statistics, Inha University, Inchon 402-751, Korea.

time, respectively. One of the main interests in survival data analysis is the effect of covariates on the hazard rate (hazard function) $\lambda(y|x) = f(y|x)/S(y|x)$, where $f(y|x)$ is the conditional density function of Y given $X=x$, and $S(y|x) = \Pr(Y \geq y \mid X=x)$. For right-censored data, Cox (1972) suggested the proportional hazard regression model, $\lambda(y|x) = \lambda_0(y)\exp(\beta^T x)$, and the main interest is in the estimation of the regression coefficient β based on the partial likelihood. The proportional hazard model was extended to nonparametric models, $\lambda(y|x) = \lambda_0(y)\exp(\eta(x))$, where non-linear covariate effects $\eta(x)$ was estimated by smoothing splines (O' Sullivan, 1988) and by regression splines (Sleeper and Harrington, 1990). The model was further generalized by Kooperberg et al. (1995).

Another approach for the censored data analysis is to use the linear regression model, $Y_i = \beta^T X_i + \varepsilon_i$, where the ε_i are independent and identically distributed with mean 0. This model has been studied by Miller (1976); Buckley and James (1979); Koul et al. (1981); Miller and Halpern (1982); Zhou (1992), among others. This approach was extended to left-truncated and right-censored data by Lai and Ying (1994), and Gross and Lai (1996). In this article we develop nonparametric regression methods for l.t.r.c. data by using the idea of weighted estimators with weights determined by jumps in the product-limit estimator, and we illustrate nonparametric methods with a simulated data set and the Stanford heart transplant data.

2. Nonparametric Regression

Let (X_1, Y_1), $(X_2, Y_2), \ldots$ be independent identically distributed random variables. Let C_i and T_i denote a right censoring variable and a left truncation variable, respectively. Suppose that (C_1, T_1), $(C_2, T_2), \ldots$ are independent of the (X_i, Y_i). When the Y_i are subject to right censoring, we observe $\min(Y_i, C_i)$ and the censoring indicator $I(Y_i \leq C_i)$, which is 1 if we observe uncensored data and 0 otherwise. If the Y_i are subject to left truncation in addition to right censoring, we observe $(\min(Y_i, C_i), I(Y_i \leq C_i), T_i)$ only when $\min(Y_i, C_i) \geq T_i$. Let $\tilde{Y}_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \leq C_i)$. Let

$$(X_i^o,\ \tilde{Y}_i^o,\ \delta_i^o,\ T_i^o) \quad i=1,2,\ldots n \quad \text{with} \quad \tilde{Y}_i^o \geq T_i^o$$

denote the observed data. We can regard the observed sample as being generated by a larger sample of (X_l, Y_l, C_l, T_l), $l=1,2,\ldots m(n)$, where

$$m(n) = \min\{m : \sum_{l=1}^{m} I(\tilde{Y}_l \geq T_l) = n\}.$$

In applications one is often interested in estimating $E[h(Y)]$ for some function h. However, $E[h(Y)]$ may not be estimable because of incomplete information about the distribution of Y due to left truncation and right censoring. For example, suppose that truncation and censoring occur at fixed points so that $T_i = t_0$ and $C_i = c_0$ for some constants t_0 and c_0, where $t_0 < c_0$. Then we cannot observe the Y_i which are outside the interval $[t_0, c_0]$. Hence $E[h(Y)]$ is not estimable. The conditional trimmed mean $E[h(Y) \mid a \le Y \le b]$, however, is estimable from the data for some constants a and b such that $a > t_0$ and $b < c_0$, as explained below.

Let $S(t)$ denote the survival function defined by $S(t) = \Pr(Y \ge t)$, and let

$$G_m(t) = \frac{1}{m} \sum_{l=1}^{m} \Pr\{T_l \le t \le C_l\}.$$

Define

$$\underline{\tau} = \inf\{ t : \liminf_{m \to \infty} G_m(t) > 0\},$$
$$\overline{\tau} = \inf\{t > \underline{\tau} : S(t) = 0 \text{ or } \liminf_{m \to \infty} G_m(t) = 0\}.$$

Then $\underline{\tau}$ and $\overline{\tau}$ are the left and right boundaries of the interval within which we can observe the data under left truncation and right censoring. Lai and Ying (1991) showed that the conditional distribution

$$F_{\underline{\tau}}(y) = \Pr(Y \le y \mid Y \ge \underline{\tau})$$

can be nonparametrically estimated for $y < \overline{\tau}$ from left-truncated and right-censored data.

Suppose a and b are some constants such that $a > \underline{\tau}$ and $b < \overline{\tau}$. Let $\hat{F}_a(y)$ be the product-limit estimator of $F_a(y) = \Pr(Y \le y \mid Y \ge a)$ given by

$$\hat{F}_a(y) = \begin{cases} 0 & \text{if } y < a \\ 1 - \prod_{i:\, a \le y_{(i)} \le y} \left[1 - \frac{d_{(i)}}{n_{(i)}}\right] & \text{otherwise,} \end{cases}$$

and let $\hat{S}_a(y)$ be an estimator of the conditional survival function $S_a(y) = \Pr(Y \ge y \mid Y \ge a)$ given by

$$\hat{S}_a(y) = \prod_{i:\, a \le y_{(i)} < y} \left[1 - \frac{d_{(i)}}{n_{(i)}}\right],$$

where $y_{(1)} < y_{(2)} < \cdots < y_{(n_1)}$ are the distinct uncensored observations; $d_{(i)}$ is the multiplicity of uncensored observations at $y_{(i)}$; $n_{(i)}$ is the size of the risk set at $y_{(i)}$, i.e., $n_{(i)} = \#\{j : T_j^o \le y_{(i)} \le \tilde{Y}_j^o\}$; and $\#(A)$ denotes the number of elements in set A. While $E[h(Y)]$ may not be estimable because of incomplete information about the distribution of Y, Gross and Lai (1996) showed that $E[h(Y) \mid a \le Y \le b]$, for a continuous function $h(\cdot)$, can be consistently estimated by

$$\frac{1}{\hat{F}_a(b)} \int_a^b h(y)\, d\,\hat{F}_a(y) = \frac{1}{\hat{F}_a(b)} \sum_{i=1}^{n} \delta_i^o\, I(a \le \tilde{Y}_i^o \le b)\, h(\tilde{Y}_i^o) \frac{\hat{S}_a(\tilde{Y}_i^o)}{\#(\tilde{Y}_i^o)}$$

under the condition that

$$\liminf_{m\to\infty}\frac{1}{m}\sum_{l=1}^{m}\Pr\{T_l\le\min(Y_l,C_l)\}>0.$$

When the (C_i, T_i) are i.i.d., note that the above condition is reduced to the assumption $\Pr\{T\le\min(Y,C)\}>0$, that is, a random sample is observable with a positive probability.

Consider a nonparametric regression model $E(Y\mid X)=\varphi(X)$ for some function $\varphi(\cdot)$ in L^2. As mentioned before, $\varphi(\cdot)$ may not be estimable because of incomplete information. The best approximation to the regression function $\varphi(\cdot)$ in L^2 at the presence of left truncation and right censoring can be defined as

$$\varphi^* = \operatorname{argmin}_{h\in L^2}E[(Y-h(X))^2\mid a\le Y\le b].$$

Suppose that an estimated regression function is chosen from a function space $\mathcal{F}_n$ based on a random sample of size n. The best approximation to the regression function $\varphi(\cdot)$ in $\mathcal{F}_n$ can be defined as

$$\tilde{\varphi}_n = \operatorname{argmin}_{g\in\mathcal{F}_n}E[(Y-g(X))^2\mid a\le Y\le b].$$

Using the argument in Gross and Lai (1996), it can be shown that

$$\frac{1}{\hat{F}_a(b)}\sum_{i=1}^{n}\delta_i^o\, I(a\le \tilde{Y}_i^o\le b)(\tilde{Y}_i^o-g(X_i^o))^2\frac{\hat{S}_{a}(\tilde{Y}_i^o)}{\#(\tilde{Y}_i^o)} \tag{1}$$

is a consistent estimator of $E[(Y-g(X))^2\mid a\le Y\le b]$. Then, a nonparametric regression estimator $\hat{\varphi}_n$ in $\mathcal{F}_n$ can be defined as

$$\hat{\varphi}_n = \operatorname{argmin}_{g\in\mathcal{F}_n}\frac{1}{\hat{F}_a(b)}\sum_{i=1}^{n}\delta_i^o\, I(a\le \tilde{Y}_i^o\le b)(\tilde{Y}_i^o-g(X_i^o))^2\frac{\hat{S}_{a}(\tilde{Y}_i^o)}{\#(\tilde{Y}_i^o)}.$$

Under certain conditions, Park (1995) showed that $\hat{\varphi}_n$ converges to φ^* with the optimal rate by using the space of tensor products of B-splines as $\mathcal{F}_n$.

For the case of complete data (without censoring and truncation), several nonparametric algorithms including MARS (multivariate adaptive regression splines) (Friedman, 1991) have been developed during the last two decades (see Hardle, 1990). For left-truncated and right-censored data, we can use the nonparametric regression algorithms using the weight

$$\frac{1}{\hat{F}_a(b)}\delta_i^o\, I(a\le \tilde{Y}_i^o\le b)\frac{\hat{S}_a(\tilde{Y}_i^o)}{\#(\tilde{Y}_i^o)}\quad\text{on }(X_i^o,\ \tilde{Y}_i^o)$$

using the consistent estimating equation (1).

In this article we assume that the Y_i are independent and identically distributed. In regression models, it is a reasonable assumption if the covariates X_i are random variables which are independent and identically distributed. However, the assumption that the Y_i are identically distributed does not hold in general, as in the case of nonrandom covariates.

Another assumption is that the (C_i, T_i) are independent of (X_i, Y_i). This assumption may not hold in some cases. Gross and Lai (1996) suggested the following procedure for the case that the assumptions do not hold. First, partition the range of X into several subregions $U_1, U_2, \ldots, U_K$, thereby stratifying the data into K subgroups corresponding to X values. When the range of X in each subregion is sufficiently small so that X values do not change much within each group, the Y_i can be regarded approximately as having the identical distribution given $X_i \in U_k$ and the (C_i, T_i) are conditionally independent of (X_i, Y_i) given $X_i \in U_k$. Since the necessary assumptions to prove that (1) is a consistent estimator of $E\,[\,(Y-g(X))^2 \mid a \le Y \le b\,]$ are approximately satisfied within each group, we can use (1) but replacing $\hat{S}_a(\tilde{Y}_i^o)\,/\,[\,\hat{F}_a(b)\cdot \#(\tilde{Y}_i^o)]$ by

$$\frac{n^{(k)}}{n}\,\frac{\hat{S}_a^{(k)}(\tilde{Y}_i^o)}{\hat{F}_a^{(k)}(b)\cdot \#^{(k)}(\tilde{Y}_i^o)}\,,$$

where $n^{(k)}$ is the number of observations such that $X_i^o \in U_k$, and $\hat{S}_a^{(k)}(\,\cdot\,)$, $\#^{(k)}(\,\cdot\,)$, and $\hat{F}_a^{(k)}(b)$ are the same as $\hat{S}_a(\,\cdot\,)$, $\#(\,\cdot\,)$, and $\hat{F}_a(b)$ in (1) but calculated using the observed data within each group. The motivation of the above weight is following. We may assume that the Y_i are approximately independent and identically distributed and the (C_i, T_i) are independent of (X_i, Y_i) within each group. Therefore, by the same argument which prove that (1) is a consistent estimator of $E\,[\,(Y-g(X))^2 \mid a \le Y \le b\,]$, it follows that

$$\frac{1}{\hat{F}_a^{(k)}(b)}\sum_{i=1}^{n}\delta_i^o\, I(\tilde{X}_i^o \in U_k)\, I(a \le \tilde{Y}_i^o \le b)\,(\tilde{Y}_i^o - g(X_i^o))^2\,\frac{\hat{S}_a^{(k)}(\tilde{Y}_i^o)}{\#^{(k)}(\tilde{Y}_i^o)}$$

is a consistent estimator of $E\,[\,(Y-g(X))^2 \mid a \le Y \le b,\ X \in U_k]$. Since there are $n^{(k)}$ observations in each group, the estimator of $E\,[\,(Y-g(X))^2 \mid a \le Y \le b\,]$ is given by

$$\sum_{k=1}^{K}\frac{n^{(k)}}{n\,\hat{F}_a^{(k)}(b)}\sum_{i=1}^{n}\delta_i^o\, I(\tilde{X}_i^o \in U_k)\, I(a \le \tilde{Y}_i^o \le b)\,(\tilde{Y}_i^o - g(X_i^o))^2\,\frac{\hat{S}_a^{(k)}(\tilde{Y}_i^o)}{\#^{(k)}(\tilde{Y}_i^o)}$$

by using weight $n^{(k)}/n$ on the estimator of $E\,[\,(Y-g(X))^2 \mid a \le Y \le b,\ X \in U_k]$. This idea of stratification was also used in Leurgans (1987) and Fygenson and Zhou (1994) for censored data.

3. Example

In this section, we present the results of applying MARS to a simulated data set and the Stanford heart transplant data. We fit the MARS model by using the Fortran program developed by Friedman.

Example 1 The simulated data are generated by

$$Y_i = \varphi\ (X_{1i},\ X_{2i}) + \varepsilon_i,$$

where

$$\varphi(x_1, x_2) = 1.3356(1.5(1 - x_1) + e^{2x_1 - 1}\sin(3\pi(x_1 - 0.6)^2) + e^{3(x_2 - 0.5)}\sin(4\pi(x_2 - 0.9)^2)),$$

$$X_{1i} \text{ and } X_{2i} \sim \text{Uniform}[0,1],\ \varepsilon_i \sim N(0, 0.75^2),\ C_i \sim \chi^4_2,\ T_i \sim \chi^2_1.$$

First we generate a sample of 200 vectors (X_i, Y_i, C_i, T_i). From 200 vectors, the observable sample $(X_i^o, \tilde{Y}_i^o, C_i^o, T_i^o)$ is then produced. In the sample generated, $n = 143$ so that 28.5% are left truncated. Among the observed data, 33% are right censored. For the observed data, we apply the MARS program to estimate the regression function $\varphi(x_1, x_2)$. In this example, a and b are chosen so that 95% of the observed data are included in the interval $[a, b]$. Figure 1 shows (a) the true regression function, (b) the observed data, and (c) and (d) the estimated regression functions from the observed data when the smoothing parameter (See Friedman, 1991) $d = 3$ and when $d = 1$. Compared with the regression function (a), the smoothing parameter $d = 3$ in (c) seems to have caused oversmoothing in the estimated regression function, and the choice of $d = 1$ in (d) seems to give a better estimate of (a). Note that the fitted surface gets smoother as the smoothing parameter becomes larger.

Example 2 Stanford heart transplant data.

In the Stanford heart transplant program 184 patients had received heart transplants from October 1967 to February 1980. The data consist of their survival times (days), age, and mismatch scores that measure the degree of tissue compatibility between the initial donor and recipient hearts. For 27 of the 184 patients, the mismatch scores are missing. The analysis of the data is based on the remaining 157 patients, and 55 are censored of 157 cases. As pointed out in Leurgans (1987), the simple random censorship model in which the censoring times are assumed i.i.d. may not be appropriate for this data. One way to avoid the difficulty is to use stratification. Following Leurgans (1987), the data are stratified into 4 groups using age as a criterion of stratification; less than 30, 30-39, 40-49, and 50 or older. We apply the MARS algorithm to the data. Parts (a) and (b) of Figure 2 show nonparametric regression on age and on mismatch score, separately. They show that age has a relatively constant effect up to a certain point and survival time decreases with age after that point. Figure 2 (c) shows the estimated regression function using age and mismatch score together, and Figure 2 (d) is the same estimated function but seen from a different point. Figure 2 (c) and (d) suggest the following interpretation: when the mismatch score is close to 0, that is, when tissues of the initial donor and the recipient are well matched, age has almost no effect on survival time. As the mismatch score becomes larger, age has a more (negative) effect on survival time. As a patient gets older, the mismatch score has a more (negative) effect on survival time.

The Stanford heart transplant data have been studied by several authors and the results of analyses are well summarized in Leurgans (1987) and Zhou (1992). When age and the mismatch score were used as covariates in linear model without interaction terms, the results were inconsistent among several methods as shown in Leurgans (1987). In the model with age and the square of age as covariates, the fitted line showed that survival time increases up to approximately age 30 but it decreases after that point, while Figure 2 shows that survival time remains constant up to age 30. It seems difficult to argue which result is more reasonable since there are not many patients before age 30 (23 observations out of 157 patients). In fact we had similar results if we used the smoothing parameter $d=0.01$ in MARS.

References

[1] Buckley, J. and James, I. (1979). Linear regression with censored data, *Biometrika,* Vol. 66, 429-464.

[2] Cox, D.R. (1972). Regression models and life-tables (with discussion), *J. Roy. Statist. Soc., Ser.* B, Vol. 34, 187-220.

[3] Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion), *Ann. Statist.,* Vol. 19, 1-141.

[4] Fygenson, M. and Zhou, M. (1994). On using stratification in the analysis of linear regression models with right censoring, *Ann. Statist.*, Vol. 22, 747-762.

[5] Gross, S.T. and Lai, T.L. (1996). Nonparametric estimation and regression analysis with left-truncated and right-censored data, *J. Amer. Statist. Assoc.,* Vol. 91, 1166-1180.

[6] Hardle, W. (1990). *Applied nonparametric regression,* Cambridge University Press, Cambridge.

[7] Kooperberg, C., Stone, C.J. and Truong, Y.K. (1995). Hazard regression, *J. Amer. Statist. Assoc.*, Vol. 90, 78-94.

[8] Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right censored data, *Ann. Statist.*, Vol. 9, 1276-1288.

[9] Lai, T.L. and Ying, Z. (1991). Estimating a distribution function with truncated and censored data, *Ann. Statist.*, Vol. 19, 417-442.

[10] Lai, T.L. and Ying, Z. (1994). A missing information principle and M-estimators in regression analysis with censored and truncated data, *Ann. Statist.*, Vol. 22, 1222-1255.

[11] Leurgans, S. (1987). Linear models, random censoring and synthetic data, *Biometrika,* Vol. 74, 301-309.

[12] Miller, R.G. (1976). Least squares regression with censored data, *Biometrika,* Vol. 63,

449-464.

[13] Miller, R.G. and Halpern, J. (1982). Regression with censored data, *Biometrika,* Vol. 69, 521-531.

[14] O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation, *SIAM J. on Scient. and Statist. Computing,* Vol. 9, 531-542.

[15] Park, J. (1995). Nonparametric function estimation with left-truncated and right-censored data, Ph.D. thesis, Stanford University.

[16] Sleeper, L.A. and Harrington, D.P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease, *J. Amer. Statist. Assoc.*, Vol. 85, 941-949.

[17] Zhou, M. (1992). M-estimation in censored linear models, *Biometrika,* Vol. 79, 837-841.

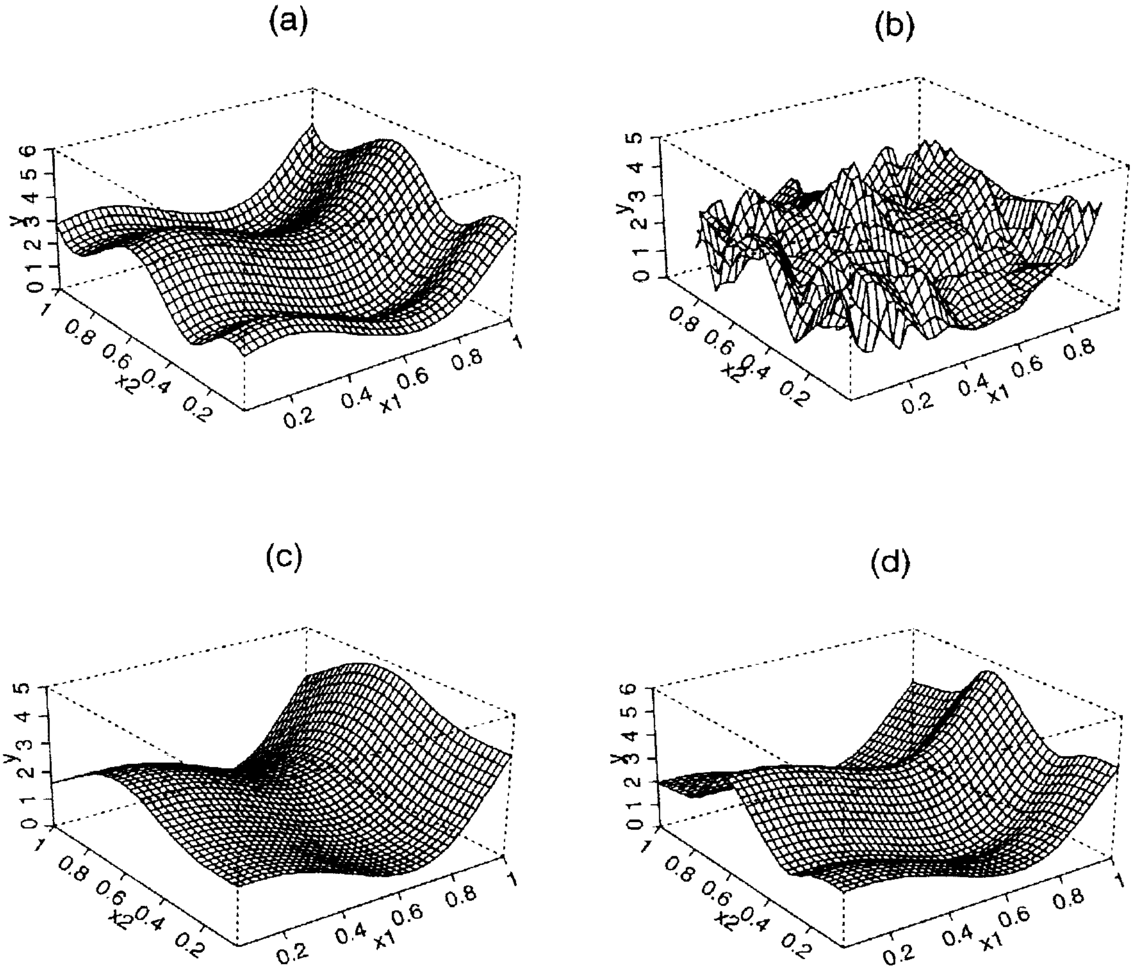

Figure 1. Nonparametric Regression using MARS (a) original model; (b) l.t.r.c. data; (c) reconstructed model ($d=3$); and (d) reconstructed model ($d=1$).

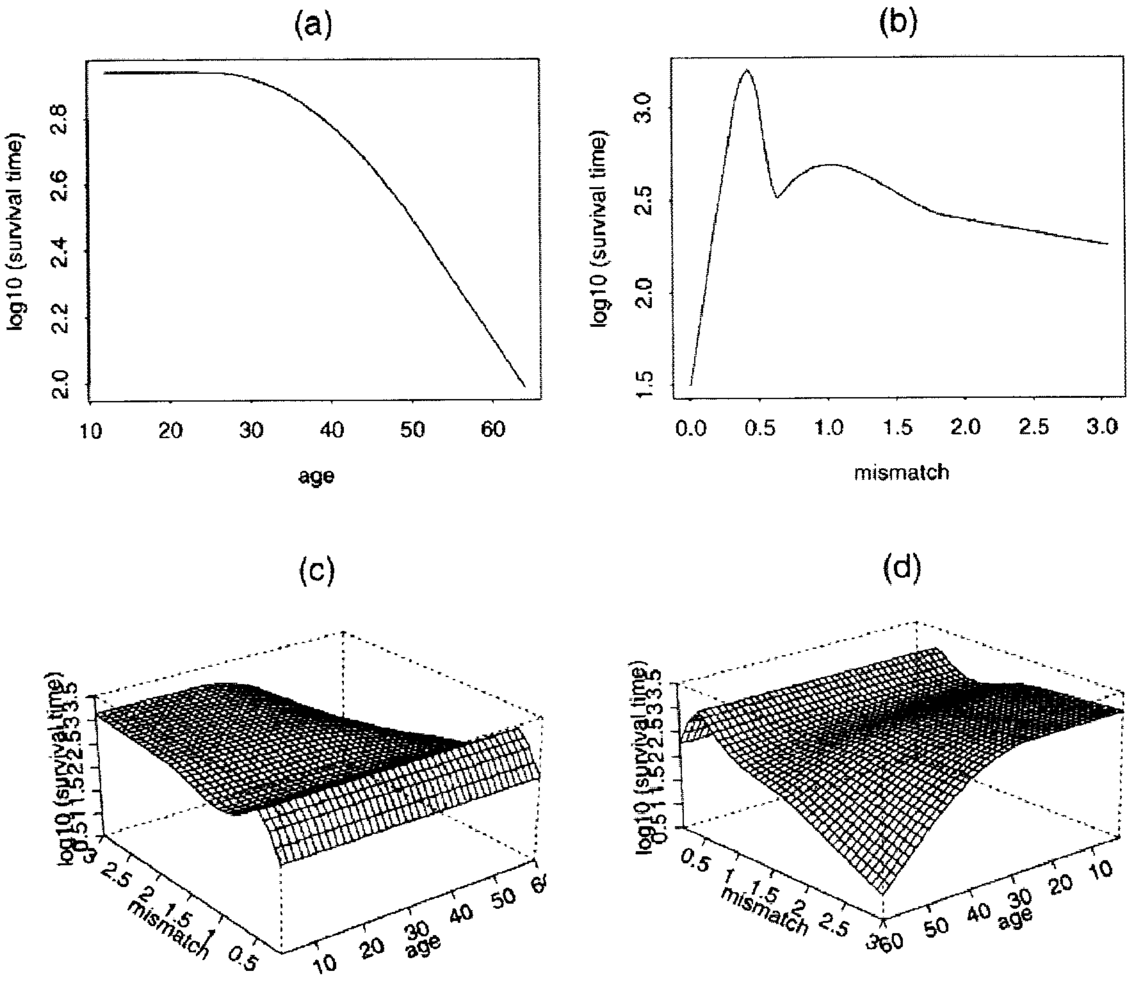

Figure 2. Stanford heart transplant data - Nonparametric regression using MAR ($d=0.1$) (a) on age; (b) on mismatch score; (c) on age and mismatch score; and (d) another view of (c)