KYBERNETIKA — VOLUME 24 (1988), NUMBER 3

A NONPARAMETRIC METHOD OF REGRESSION ANALYSIS FROM CENSORED DATA

PETR VOLF

The linear regression model is analysed nonparametrically from censored sample. The simple kernel-estimator of regression function is presented. Its asymptotic normality is proved, using the results about mean lifetime estimation. The trimmed version of estimator is considered, too. The asymptotic distribution enables us to derive some methods for testing the form of regression curve.

1. INTRODUCTION AND MODEL

Recent development of statistical inference from noncomplete data leads to new modifications of traditional statistical methods. Censorning, as a special case of incompleteness, is often met in lifetime studies, both in biological and industrial research. Estimation of mean lifetime represents necessary step to more complicated tasks. It has been solved by several authors: Breslow and Crowley [1], Gill [2], Reid [5], Susarla and Van Ryzin [6] and others. Although the Cox's model is commonly used in order to describe the lifetime regression with censoring, some methods for analysis of parametrized linear model have been derived, cf. [3], [4]. Our interest is concentrated to nonparametric methods of regression function estimation, namely K-estimation. It is known that these methods at least give important information about the form of dependence. Moreover, the method is not too complicated.

Simple form of K-estimator with a constant kernel on finite support was studied in [7] and its P-consistency was proved. In the present paper the asymptotic normality is derived, using the mentioned results about mean lifetime estimation. The trimmed version of the estimator is examined, too. A method for testing the form of regression function is suggested. When dealing with lifetime data, linear model may be used for log of lifetime. That is why the negative values are also considered.

Let $Y_1, \ldots, Y_N$ be random variables satisfying the model

$$Y_i = r(x_i) + \varepsilon_i ,$$

where x_i's are known values of regressor, ε_i's are independent random variables identically distributed according to continuous distribution function F, with zero mean, $r(x)$ is a real function. By $P = 1 - F$ we denote the survival function. Under random right-censoring we observe $T_i = \min(Y_i, V_i)$ and $\delta_i = I[Y_i \leqq V_i]$, where $I[\cdot]$ denotes the indicator, V_i are random variables, they may also depend on x. We suppose that they fulfil the linear model

$$V_i = v(x_i) + e_i\,,$$

with e_i's independent mutually and of ε_j's. Let e_i's be distributed according to a continuous distribution function G, $Q = 1 - G$ (but identity of the distributions of e_i's is not substantial). Denote by $\mathscr{T}_F = \sup\{t: F(t) < 1\}$, by $\mathscr{T}_G$ the same quantity for the distribution of e_i.

2. ESTIMATOR AND ITS ASYMPTOTIC NORMALITY

The kernel estimator of $r(z)$ at fixed point z will be computed from the observations at the points x_i near to z. Denote $\mathscr{O}_{d_N}(z) = \{x: |z - x| \leqq d_N\}$ the neighbourhood of z, $M_N(z)$ the number of x_i's in it. Some supposition about the design of variable x is inevitable. Throughout the paper we shall assume that $N \to \infty$, $d_N \to 0$, $Nd_N \to \infty$ and $M_N(z)/2Nd_N \to h(z) > 0$, $h(z)$ finite. For instance, let us imagine that x_i's are the realizations of a random variable $\mathscr{X}$ in $(R_1, \mathscr{B}_1)$ possessing a continuous density function h, $h(z) > 0$. Then $d_N \to 0$ and $Nd_N \to \infty$ imply $M_N(z)/2Nd_N \to h(z)$ a.s., it follows from the theory of nonparametric density estimation.

Let us denote

$$T_{k1} \leqq T_{k2} \leqq \ldots \leqq T_{kM}\,, \quad \text{with} \quad M = M_N(z)\,, \tag{1}$$

the ordered results observed at points $x_i \in \mathscr{O}_{d_N}(z)$. From them we construct the product limit estimator of the distribution function of variable $\{Y \mid x_i \in \mathscr{O}_{d_N}(z)\}$:

$$F^y_{N,z}(t) = \begin{cases} 0 & \text{for} \quad t < T_{k1} \\ 1 - \prod_{j=1}^{M} \left(\dfrac{M - j}{M - j + 1}\right)^{\delta_{kj} I[T_{kj} \leqq t]} & \text{for} \quad t \in [T_{k1}, T_{kM}) \\ 1 & \text{for} \quad t \geqq T_{kM} \end{cases}$$

This function has jumps $\Delta F^y_{N,z}(T_{kj}) > 0$ if the corresponding $\delta_{kn} = 1$ and at T_{kM}. The estimator of the value of the regression function at point z is then defined by

$$r_N(z) = \frac{\sum_{i=1}^{N} T_i \, \Delta F^y_{N,z}(T_i) \, I[|T_i| \leqq A_N]}{\sum_{i=1}^{N} \Delta F^y_{N,z}(T_i) \, I[|T_i| \leqq A_N]}\,,$$

where A_N is a conveniently chosen real sequence tending to infinity. It helps us to solve the most problematic case, namely $\mathscr{T}_F = \infty$, cf. [6].

The asymptotic distribution of $r_N(z)$ is derived directly from the asymptotic

distribution of location parameter estimator. Define

$$R(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C_z(\min(s, t))\, P(s)\, P(t)\, \mathrm{d}s\, \mathrm{d}t$$

with

$$C_z(s) = \int_{-\infty}^{s} \frac{\mathrm{d}F(u)}{P^2(u)\, Q(u + r(z) - v(z))}.$$

representing the covariance function of the asymptotic normal distribution for the PL-estimator of F, as it was derived by Breslow and Crowley [1].

Theorem 1. Let the following assumptions hold:

A1. $d_N(Nd_N)^\gamma = O(1)$ for some $\gamma > 2$.

A2. $M_N(z)/Nd_N \to 2\, h(z) > 0$.

A3. The functions r and v are continuous at z.

A4. The distribution functions F, G are uniformly continuous.

A5. $\mathscr{T}_F + r(z) \leqq \mathscr{T}_G + v(z)$.

A6. $R(z) < \infty$.

A7. $\sqrt{(M_N(z))} \{ \int_{-\infty}^{\infty} t\, \mathrm{d}F(t) - \int_{-A_N - r(z)}^{A_N - r(z)} t\, \mathrm{d}F(t) \} \to 0$, as $N \to \infty$.

Then $\sqrt{(M_N(z))}\,(r_N(z) - r(z))$ is asymptotically distributed as $N(0, R(z))$.

Proof. Denote $\tau_i = \min(\varepsilon_i, e_i + v(z) - r(z))$, it represents the model of randomly censored variables ε_i. Denote the indicator of censoring by $\delta_i^* = I[\varepsilon_i \leqq \tau_i]$. When estimating the value of $r(z)$ at point z, only the part (1) from all sample is used. Then $T_{kj} - \tau_{kj} - r(z) = O(d_N)$ uniformly for the indices kj, $j = 1, 2, \ldots, M_N(z)$. It follows from A3. Let us imagine that the PL-estimator $F_{N,z}$ of F is constructed from the (unknown) sample $\tau_{kj}, \delta_{kj}^*, j = 1, \ldots, M_N(z)$. Let us denote by N_{kj} the number of members from this sample, which are greater than or equal to τ_{kj}. Then, with $M = M_N(z)$

$$F_{N,z}(t) = \begin{cases} 0 & \text{for } t < \min \tau_{kj} \\ 1 - \prod_{j=1}^{M} \{(N_{kj} - 1)/N_{kj}\}^{\delta^*_{kj} I[\tau_{kj} \leqq t]} & \min \tau_{kj} \leqq t < \max \tau_{kj} \\ 1 & t \geqq \max \tau_{kj}\,. \end{cases}$$

Then $\sum \tau_{kj}\, \Delta F_{N,z}(\tau_{kj}) / \sum \Delta F_{N,z}(\tau_{kj})$ could serve as an estimator of $\mathrm{E}\varepsilon_1 = 0$. Under A2, A5 – A7 this estimator has asymptotically $N(0, R(z))$ distribution (cf. also [2]). The sums are over $j = 1, \ldots, M_N(z)$ such that $|T_{kj}| \leqq A_N$.

In Volf [7], in order to prove the P-consistency of $r_N(z)$, A1 – A4 imply that for every real L

$$\mathsf{P}\left\{ \min_{i \neq j} \frac{|\tau_{ki} - \tau_{kj}|}{d_N} > L \right\} \to 1\,.$$

Therefore the ranks of τ_{kj} and T_{kj} tend to be the same in probability, moreover $\mathsf{P}\{\delta_{kj} = \delta_{kj}^*,\ j = 1, \ldots, M_N(z)\} \to 1$. Then also

$$\mathsf{P}\{\Delta F_{N,z}^{\nu}(T_{kj}) = \Delta F_{N,z}(\tau_{kj}) \text{ for every } j = 1, \ldots, M_N(z)\} \to 1\,.$$

Now, the first right-sided term in the expression

$$\sqrt{(M_N(z))}\,(r_N(z) - r(z)) = \frac{\sqrt{(M_N(z))}\sum \tau_{kj}\,\Delta F^{v}_{N,z}(T_{kj})}{\sum \Delta F^{v}_{N,z}(T_{kj})} + \sqrt{(M_N(z))}\,O(d_N)$$

tends in distribution to $N(0, R(z))$, too. The last term tends to zero as the consequence of A1 and A2. □

Assumption A1 seems to be rather strong restriction. When the form $d_N = CN^{-\beta}$ is chosen, then A1 implies $\beta \in (\frac{2}{3}, 1)$. Experience with simulated data indicates this restriction as unnecessary. But it is well known that, concerning to K-estimates, the influence of asymptotic properties is rather slow. In practice, sometimes we prefer to choose d_N such that prescribed number $M_N(z)$ is ensured – the method of M-nearest neighbour.

Sometimes the pattern of censoring is not purely random. Especially A5 can be violated by truncation. In case of mixed censoring the trimmed version of the estimator is available. For $\alpha \in (0, 1)$ let us denote

$$k^{v}_{N,z}(\alpha) = \min\{t: F^{v}_{N,z}(t) \geqq \alpha\}\,, \quad k_{N,z}(\alpha) = \min\{t: F_{N,z}(t) \geqq \alpha\}\,,$$
$$k(\alpha) = \min\{t: F(t) \geqq \alpha\}\,.$$

For $\alpha \in (0, \frac{1}{2})$ define

$$R^{\alpha}(z) = (1 - 2\alpha)^{-2} \int_{k(\alpha)}^{k(1-\alpha)} \int_{k(\alpha)}^{k(1-\alpha)} C_z(\min(s, t))\,P(s)\,P(t)\,\mathrm{d}s\,\mathrm{d}t$$

and finally the estimator

$$r^{\alpha}_{N}(z) = \frac{\sum_{i=1}^{N} T_i\,D_{i,N,z}(\alpha)}{\sum_{i=1}^{N} D_{i,N,z}(\alpha)}\,,$$

where $D_{i,N,z}(\alpha)$ stands for $\Delta F^{v}_{N,z}(T_i)\,I[k^{v}_{N,z}(\alpha) \leqq T_i \leqq k^{v}_{N,z}(1-\alpha)]$.

Theorem 2. Let assumptions A1 – A4 hold, moreover let us suppose

A8. F is symmetrical, i.e. $F(t) + F(-t) = 1$.

A9. $\alpha \in (0, \frac{1}{2})$ is such that $G(k(1-\alpha) + r(z) - v(z)) < 1$.

Then $r^{\alpha}_{N}(z)$ is a probability consistent estimator of $r(z)$ and the distribution of $\sqrt{(M_N(z))}\,(r^{\alpha}_{N}(z) - r(z))$ tends to $N(0, R^{\alpha}(z))$.

Proof. As it was outlined in the proof of previous Theorem 1, assumptions A1 – A4 lead to the conclusion that

$$\mathsf{P}\{F^{v}_{N,z}(T_{kj}) = F_{N,z}(\tau_{kj}) \quad \text{for whole sample (1), i.e.} \quad j = 1, 2, \ldots, M_N(z)\} \to 1\,.$$

Therefore

$$\mathsf{P}\{r^{\alpha}_{N}(z) = r(z) + O(d_N) + \textstyle\int \tau\,\mathrm{d}F_{N,z}(\tau) / \int \mathrm{d}F_{N,z}(\tau)\} \to 1\,.$$

The integration is in bounds $k_{N,z}(\alpha)$, $k_{N,z}(1-\alpha)$. The problem is then reduced to consistency and asymptotic normality of the estimation of $\mathsf{E}\varepsilon_1 = 0$. It is represented by the last term in the brackets.

Remind that $S_N = \sup_{t \leqq B} |F_{N,z}(t) - F(t)| = O\{(\log M/M)^{1/2}\}$ a.s., with $M = M_N(z)$, when $B < \mathscr{T}_F$, F, G are continuous. We have $t \leqq \max\{k(1-\alpha),\ k_{N,z}(1-\alpha)\}$, which asymptotically is less then $\mathscr{T}_F$ a.s. Then the maximal jump of $F_{N,z}(t)$ tends to be less than $2S_N$ almost surely. That is why $|\alpha - F_{N,z}(k_{N,z}(\alpha))| \leqq 2S_N$ in the limit and $(1-2\alpha)/\int \mathrm{d}F_{N,z}(t) \to 1$ a.s. The consistency will be proved, if we check that the following term tends to zero:

$$\left|\int_{k_{N,z}(\alpha)}^{k_{N,z}(1-\alpha)} t\,\mathrm{d}F_{N,z}(t) - \int_{k(\alpha)}^{k(1-\alpha)} t\,\mathrm{d}F(t)\right| \leqq$$

$$\leqq \{|k(1-\alpha)| + |k(\alpha)|\}\, S_N + \left|\int_{k_{N,z}(\alpha)}^{k(\alpha)} t\,\mathrm{d}F_{N,z}(t)\right| + \left|\int_{k_{N,z}(1-\alpha)}^{k(1-\alpha)} t\,\mathrm{d}F_{N,z}(t)\right|.$$

For some b_N which is in the limit a.s. less than $\mathscr{T}_F$, the second term is equal to

$$|b_N\{F_{N,z}(k(\alpha)) - F_{N,z}(k_{N,z}(\alpha))\}| \leqq |b_N|\,\{|F_{N,z}(k(\alpha)) - \alpha| +$$

$$+ |\alpha - F_{N,z}(k_{N,z}(\alpha))|\} \leqq |b_N|\,(S_N + 2S_N) \quad \text{a.s.}$$

The third term can be bounded in the same way.

The asymptotic distribution of the trimmed version of the mean estimator was derived by Reid [5]. Her results are the same as that of us. □

3. TESTS OF THE FORM OF REGRESSION FUNCTION

Let us estimate the values of regression function at K different points $z_1, \ldots, z_K$. The estimators $r_N(z_j)$ are asymptotically independent because the neighbourhoods $\mathscr{O}_{d_N}(z_j)$ are disjoint for N sufficiently large. This arises the possibility to test a correspondence between a hypothetical regression function $r_0(x)$ and the real one. When assumptions of Theorem 1 hold and hypothesis (r_0) is valid then the statistic

$$\sum_{j=1}^{K} M_N(z_j)\,(r_N(z_j) - r_0(z_j))^2/R(z_j)$$

has asymptotically χ^2_K distribution. Estimator of R is naturally constructed from the sample (1) as

$$R_N(z) = \sum\sum C^{\nu}_{N,z}(\min(T_{ki}, T_{kj}))\,(1 - F^{\nu}_{N,z}(T_{ki}))\,(1 - F^{\nu}_{N,z}(T_{kj}))\,\Delta T_{ki}\,\Delta T_{kj},$$

where the summation is over $i, j = 1, \ldots, M_N(z) - 1$ and $\Delta T_{ki} = T_{k(i+1)} - T_{ki}$. Put

$$C^{\nu}_{N,z}(t) = \sum M_N(z)\,\delta_{kj}\,I[T_{kj} \leqq t]/(M_N(z) - j + 1)^2$$

where the summation is over $j = 1, \ldots, M_N(z)$.

As it was mentioned before, the ranks of T_{kj} and τ_{kj} tend to be the same in probability if A1 – A5 hold. From this the convergence of $C^{\nu}_{N,z}(t)$ to $C_z(t - r(z))$ in probability follows. But the consistency of $R_N(z)$ would require some additional assumptions.

Let us suppose that F is strictly increasing at $k(\alpha)$ and $k(1-\alpha)$. Then the quantiles can be estimated consistently and the trimmed version $R^{\alpha}_N(z)$ would be a P-consistent

estimator of $R^a(z)$ as soon as the assumptions of Theorem 2 hold. Sometimes the hypothesis $\{H_0: r(x) \equiv \text{const.}\}$ is tested, i.e., the regression seems not to be significant. Put

$$c_{j,N} = M_N(z_j)/R(z_j), \quad \hat{r}_N = \sum_{j=1}^{K} c_{j,N}\, r_N(z_j) / \sum_{j=1}^{K} c_{j,N}$$

Theorem 3. Let A1 – A7 hold for every z_j, $j = 1, \ldots, K$. Moreover, assume that there exist $\lim_{N\to\infty} M_N(z_j)/M_N(z_i) < \infty$, $i, j = 1, \ldots, K$. Then the statistic

$$U_N = \sum_{j=1}^{K} c_{j,N}(r_N(z_j) - \hat{r}_N)^2 \tag{2}$$

is asymptotically distributed as χ^2_{K-1} when H_0 holds.

Proof. Let us construct a symmetric matrix $\boldsymbol{B}_N(K \times K)$ with the elements

$$(\boldsymbol{B}_N)_{ij} = I[i = j] - (c_{i,N}c_{j,N})^{1/2} / \sum_{l=1}^{K} c_{l,N}\,.$$

It is an idempotent matrix, $\boldsymbol{B}_N^2 = \boldsymbol{B}_N$. Therefore its rank may be computed as the sum of its diagonal elements, which gives $K - 1$. The statistic (2) can be written as $U_N = \boldsymbol{W}_N'\boldsymbol{B}_N\boldsymbol{W}_N$, where $\boldsymbol{W}_N$ is an $K \times 1$ vector with elements $\sqrt{(c_{j,N})}\,(r_N(z_j) - r)$, $r = r(z_j)$ is our hypothetical constant.

Theorem 1 ensures that $\boldsymbol{W}_N \xrightarrow{\mathcal{D}} \boldsymbol{W}$, where $\boldsymbol{W} = (w_1, \ldots, w_K)'$ are independent random variables with standard normal distribution. Let $b_{ij} = \lim_{N\to\infty} (\boldsymbol{B}_N)_{ij}$ be the elements of matrix $\boldsymbol{B}$. Such a matrix is also idempotent with the same rank $K - 1$. That is why the statistic $\boldsymbol{W}'\boldsymbol{B}\boldsymbol{W}$ has χ^2_{K-1} distribution. As $\boldsymbol{B}_N \to \boldsymbol{B}$, we may conclude that the asymptotic distribution of U_N is the same as the distribution of $\boldsymbol{W}'\boldsymbol{B}\boldsymbol{W}$. □

4. CONCLUDING REMARKS

First devote to the choice of the width of window. One-dimensional window around the point z is defined as the interval $[z - d_N, z + d_N]$, K-dimensional window as the product of such intervals. Denote by $\text{Ext}(x)$ "effective extent of variable x". It could mean $x_{\max} - x_{\min}$ or width of the interval, in which x is distributed uniformly (approximately). For the one-dimensional problem some examples indicate that $d_N = \text{Ext}(x)/N^{0.7}$ is quite sufficient. Maybe, it is rather narrow, the order of $N^{-0.5}$ would suit better, but it would be in contradiction with Theorem 1, assumption A1.

Then the number of points in such window is

$$M_N \sim 2d_N N/\text{Ext}(x) = 2N^{0.3}\,.$$

Denote

$$P_1 = M_N/N = 2N^{-0.7}\,.$$

For K-dimensional x we demand the same number M_N and we choose

$$d_{Nj} = D_K\, \text{Ext}(x_j)/N^{0.7}, \quad \text{for every dimension} \quad j = 1, 2, \ldots, K\,.$$

Then

$$\frac{M_N}{N} \sim (D_K P_1)^K$$

$$2N^{-0.7} = (D_K\, 2N^{-0.7})^K$$

$$D_K = \left(\frac{N^{0.7}}{2}\right)^{1-\frac{1}{K}}.$$

When testing the method by simulated examples, we approximately followed these formulas and compared our results with the results obtained for other values of window-width. The results did not contradict our way of choice of d_N.

As it was noted, sometimes the method of M-nearest failure neighbour is more practical. The attempts were made to establish some optimal width of widnow $2d_N$ by the cross-validation method. This procedure is clear intuitively, but theoretically and even practically disputable. Cross-validated estimator has the tendency to follow all departures of data, such estimator gives not smooth curve. Some authors add a penalty term based on the second derivative of obtained curve to the cross-validation criterion. It leads to more complicated computations. That is why we did not use this procedure.

The kernel estimator with a Gaussian kernel (instead of bounded "window") was checked, too. In such case we need not repeated computing of PL-estimates, instead we compute the values of Gaussian density function. The results are quite similar. It is known that such continuous kernels give the better estimates at the ends of the regressor domain. But only some extensive Monte-Carlo study could lead to definitive conclusions as to the advantages of various kernels.

(Received February 9, 1987.)

REFERENCES

[1] N. Breslow and J. Crowley: A large sample study of the life table and product limit estimates under random censorship. Ann. Statist. *2* (1974), 437—453.

[2] R. Gill: Large sample behaviour of the product-limit estimator on the whole line. Ann. Statist. *11* (1983), 49—58.

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

[4] R. G. Miller: Least squares regression with censored data. Biometrika *63* (1976), 449—464.

[5] N. Reid: Influence functions for censored data. Ann. Statist. *9* (1981), 78—92.

[6] V. Susarla and J. Van Ryzin: Large sample theory for an estimator of the mean survival time from censored samples. Ann. Statist. *8* (1980), 1002—1016.

[7] P. Volf: Estimation in linear model with censored data. In: Proc. of 5th Pannonian Symposium on Math. Statist., Visegrád 1985. Akadémiai Kiadó, Budapest 1987, 431—438.

Petr Volf, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia.