The Korean Communications
in Statistics Vol. 10 No. 1, 2003
pp. 167-175

Simplicial Regression Depth with Censored and Truncated Data[1]

Jinho Park[2]

Abstract

In this paper we develop a robust procedure to estimate regression coefficients for a linear model with censored and truncated data based on simplicial regression depth. Simplicial depth of a point is defined as the proportion of data simplices containing it. This simplicial depth can be extended to regression problem with censored and truncated data. Any line can be given a depth and the deepest regression line is the line with the maximum simplicial regression depth. We show how the proposed regression performs through analyzing AIDS incubation data.

Keywords : censoring, simplicial regression depth, truncation.

1. Introduction

To investigate the relationship between covariates and censored response, one can use a linear regression model. An estimation of regression coefficients has been investigated by Miller (1976), Buckley and James (1979), Koul et al (1981), Miller and Halpern (1982), and Zhou (1992), among others. The linear model is further generalized to the censored and truncated data by Gross and Lai (1996). Estimation of the regression coefficients is usually based on the least squares type estimations or M-estimations, and they require that the conditional distribution of the response given the covariates (the distribution of the error term) has certain properties such as non-skewness or homoskedasticity. And these estimation procedures are not robust in the sense that a few observations can have a serious influence on the analysis of the model.

As a robust estimation procedure for regression coefficients, Rousseeuw and Hubert (1999) proposed a method to define regression depth of a line. For multi-dimensional location problem, there are several methods to define depth of a point (see Hwang et al, 2002). Rousseeuw and Hubert (1999) extended the concept of halfspace location depth to a line in regression problem. And any line can be given a rank using regression depth and the deepest

[1] This research was supported by the Grant of Inha University, 2001-2002

[2] Assistant Professor, Department of Statistics, Inha University, Incheon 402-751, Korea.
E-mail : jhpark@stat.inha.ac.kr

regression line is the line with the maximum regression depth. Rousseeuw and Hubert (1999) showed that the deepest regression line has a break-down value of $1/3$, while the least squares regression line has 0. That is, at least one third of the data need to be replaced to change arbitrarily the deepest regression line. Another definition of location depth is based in simplicial depth (Liu, 1990). Simplicial depth of a point is the proportion of the data simplices containing it. Simplicial depth was extended to regression problem by Rousseeuw and Hubert (1999).

In this paper, we develop a procedure to find the deepest regression line based on simplicial depth. In section 2, we define a simplicial regression depth of any line for censored and truncated data and find the deepest regression line with the maximum depth. Section 3 illustrates the suggested procedure for a real data set and shows how it performs.

2. Simplicial Regression Depth with Censored and Truncated Data

Liu (1990) introduced a notion of simplicial depth. Simplicial depth of a point θ in R^p with respect to a p-dimensional distribution F is defined as

$$sdepth(\theta; F) = P_F(\theta \in S(X_{1,} X_{2,} \ldots, X_{p+1})),$$

where $(X_1, X_2, \ldots, X_{p+1})$ is a random sample from F, and $S(X_{1,} X_{2,} \ldots, X_{p+1})$ is the simplex determined by $X_1, X_2, \ldots, X_{p+1}$For p-dimensional random sample $(X_1, X_2, \ldots X_n)$ the sample version of simplicial depth $sdepth(\theta; X_1, X_2, \ldots, X_n)$ of a location θ is defined as

$$sdepth(\theta; X_1, X_2, \ldots, X_n) = \binom{n}{p+1}^{-1} \sum_{i_1 < \cdots < i_{p+1}} I(\theta \in S(X_{i_1}, X_{i_2}, \ldots, X_{i_{p+1}})),$$

where i_j takes values in $1, 2, \ldots, n$ Note that $\binom{n}{p+1}$ is the number of possible simplices and simplicial depth is the proportion of the data simplices containing θ. The center of the distribution can be estimated as the point with the maximum simplicial depth. An underlying idea od simplicial depth is that the center should be inside of simplices constructed by data as often as possible.

The notion of simplicial depth was extended to regression problem by Rousseeuw and Hubert (1999). Suppose that X is a $(p-1)-$ dimensional covariate and Y is the response. To define depth of a line (plane), they used concept of dual plot. In dual plot, a line $y = \theta_1 x_1 + \cdots + \theta_{p-1} x_{p-1} + \theta_p$ is transformed to a point $\theta = (\theta_1, \ldots, \theta_{p-1}, \theta_p)$ in R^p and a point $(x_1, \ldots, x_{p-1}, y)$is transformed to a line $\theta_p = -x_1\theta_1 - \cdots - x_{p-1}\theta_{p-1} + y_p$ in $\theta_1, \ldots, \theta_{p-1}, \theta_p$ axes. The dual plot preserves the ordering of the lines and points in the sense that a point lying below (on or above) corresponds to a line below (through or above) after transformation in the dual plot. This is important since simplicial depth is determined by the order of lines and points.

For a complete data set (X_1, Y_1), $(X_2, Y_2), \ldots, (X_{n,} Y_n)$ without censoring and truncation, Rousseeuw and Hubert (1999) defined the simplicial regression depth of a line $y = \theta_1 x_1 + \cdots + \theta_{p-1} x_{p-1} + \theta_p$ as

$$sdepth(\theta_1, \ldots, \theta_p) = \begin{pmatrix} n \\ p+1 \end{pmatrix}^{-1} \sum_{i_1 < \cdots < i_{p+1}} I(\theta \in S(H_{i_1}, H_{i_2}, \ldots, H_{i_{p+1}})),$$

where H_i is the hyperplane in dual space corresponding to the i-the observation (X_i, Y_i) and $S(H_{i_1}, H_{i_2}, \ldots, H_{i_{p+1}})$ is the simplex determined by $(p+1)$ hyperplane. The estimated regression line is defined as the deepest regression line with the maximum simplicial regression depth. So the estimate of the regression coefficient is defined as

$$(\hat{\theta}_1, \ldots, \hat{\theta}_p) = \arg\max_{\theta_1, \ldots, \theta_p} sdepth(\theta_1, \ldots, \theta_p)$$

When the response is censored and truncated, iterative procedures or weighted procedures are usually used to estimate regression coefficient. In this article we use a weighted procedure to define simplicial regression depth for censored and truncated data. The above simplicial depth is defined as a sample proportion of simplices containing θ. For censored and truncated data, the definition of simplicial depth can be modified as a weighted proportion of simplices containing θ and the weights are determined by the Kaplan-Meier estimates.

Let C_i and T_i denote a right censoring variable and a left truncation variable, respectively. Suppose that the (C_i, T_i) 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 $I(Y_i \le C_i)$, 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 $(\min(Y_i, C_i), I(Y_i \le C_i), T_i)$ only when $\min(Y_i, C_i) \ge T_i$ Let $\tilde{Y}_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \le C_i)$ Let

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

denote the observed data. An ordinary right censored data without left truncation corresponds to the case $T = -\infty$

Let $S(t)$ denote the survival function defined by $S(t) = \Pr(Y \ge t)$ and let $G(t) = \Pr(T \le t \le C)$ Define

$$\underline{\tau} = \inf\{t : G(t) > 0\},$$

$$\bar{\tau} = \inf\{t > \underline{\tau} : S(t) = 0 \text{ or } G(t) = 0\}$$

Then $\underline{\tau}$ and $\bar{\tau}$ 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 $y < \bar{\tau}$ from left-truncated and right-censored data.

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

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

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 $y_{(1)} < y_{(2)} < \cdots$ 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., $n_{(i)} = \sum_{j=1}^{n} I(T_j \le y_{(i)} \le \tilde{Y}_j)$ While $E[h(X, Y)]$ may not be estimable because of incomplete information about the distribution of Y, Gross and Lai (1996) showed that $E[h(X, Y) \mid a \le Y \le b]$ for a $h(\cdot)$, can be consistently estimated by

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

where $\#(\tilde{Y}_i) = \sum_{j=1}^{n} I(T_j \le \tilde{Y}_i \le \tilde{Y}_j)$

If $F(\underline{\tau}) = 0$ and $F(\bar{\tau}) = 1$, the survival function is estimable without the condition $a \le Y \le b$. And $E[h(X, Y)]$can be consistently estimated by

$$\sum_{i=1}^{n} \delta_i\, h(X, \tilde{Y}_i) \frac{\hat{S}(\tilde{Y}_i)}{\#(\tilde{Y}_i)},$$

where $\hat{S}(t) = \prod_{i:\, y_{(i)} < t} \left[1 - \frac{d_{(i)}}{n_{(i)}}.\right]$This implies that $E[h(X, Y)]$can be estimated using the weights

$$W_i = \delta_i \frac{\hat{S}(\tilde{Y}_i)}{\#(\tilde{Y}_i)}$$

instead of the equal weight to each observation. Hence it seems natural to define the simplicial regression depth using the above weights. Note that the above weights are the jump sizes of the product-limit estimator $\hat{F}(t)$, and that the weights are the same as in Zhou (1992) if the response is only right-censored. Without the condition $F(\underline{\tau}) = 0$ and $F(\bar{\tau}) = 1$, we can use the weight

$$W_i = \frac{\delta_i\, I(a \le \tilde{Y}_i \le b)}{\hat{F}_a(b)} \frac{\hat{S}_a(\tilde{Y}_i)}{\#(\tilde{Y}_i)} \tag{2}$$

since $E[h(X, Y) \mid a \leq Y \leq b]$ can be consistently estimated by (1).

Following the above arguments, we can define simplicial regression depth for censored and truncated data as

$$sdepth(\theta_1, \dots, \theta_p) = \alpha^{-1} \sum_{i_1 < \cdots < i_{p+1}} I(\theta \in S(H_{i_1}, H_{i_2}, \dots, H_{i_{p+1}}))\ (W_{i_1} + \cdots + W_{i_{p+1}}),$$

where the summation is over all possible combinations to construct a simplex with uncensored observations, H_i and W_i are the hyperplane and the weight corresponding to the i-the observation (X_i, Y_i) and $\alpha = \sum_{i_1 < \cdots < i_{p+1}} (W_{i_1} + \cdots + W_{i_{p+1}})$ So the above simplicial regression depth is a weighted proportion of simplices containing θ, and the weight of a simplex is determined by the observations corresponding to the simplex.

The simplicial regression depth has value between 0 and 1. When all the uncensored data lie on a line, the simplicial regression depth of the line is 1. The maximum regression depth represents the degree of linearity in the data. And the estimated regression line is defined as the deepest regression line with the maximum simplicial regression depth.

3. Examples and Simulation

In this section we show how the deepest regression line performs for simulated data and the AIDS incubation data. To see performance of the deepest regression line, we compare the estimates of regression coefficients based on simplicial regression depth and those based on least squares. First we generate a data set of sample size 25. The first graph of Figure 1 shows the observations with the deepest regression line and the least squares regression line with the weights given in (2). We can see that there is not much difference between two lines. In fact the data are generated from $y = \theta_1 x + \theta_2 + \varepsilon$ with $\theta_1 = 1$ and $\theta_2 = 0$ The estimated deepest regression line is $\hat{y} = 0.830x + 0.118$ and the weighted least squares line is $\hat{y} = 0.880x - 0.043$. To investigate the effect of an outlier, we move the largest observation to another location as indicated in the second graph of Figure 1. The second graph shows two regression lines after the observation is moved. Then the weighted least squares line becomes $\hat{y} = 0.640x - 0.163$ but the deepest regression line does not change. From this example we can see the deepest regression line is more robust than the least squares regression line.

To compare two estimation methods, we have done some simulation study. The data are generated from the simple linear model

$$Y_i = \theta_1 X_i + \theta_2 +, \varepsilon_i$$

where $\theta_1 = 1$ $\theta_2 = 0$ The covariate X_i are generated from $N(0, 1)$ and the error ε_i are generated from $N(0, 0.5^2)$ and $(\mathrm{Gamma}(4, 4) - 1)$ so that the errors have the mean zero and variance 0.25. And censoring variables and truncation variables are generated from a uniform distribution. Table 1 shows estimates and their standard errors for sample sizes $n=30$ $n=50$ and $n=100$ when the errors are from the normal distribution. The estimates and standard errors are based on 1000 repetitions. For one case of simulation study, 30% of the data are censored and 25% are truncated. We also investigate the estimates when 50% of the dara are censored and 35% are truncated. The estimates and their standard errors are given at Table 2 when the errors are from a Gamma distribution. Throughout the simulation study, we can see that the biases of the estimates are not significant considering the standard errors. But the standard errors of estimates based on simplicial regression depth are about twice of those based on the least squares. The larger variances of estimates of the deepest regression line seems show the loss of efficiency due to using depth instead of the least squares estimation. So there is a trade off between efficiency and robustness.

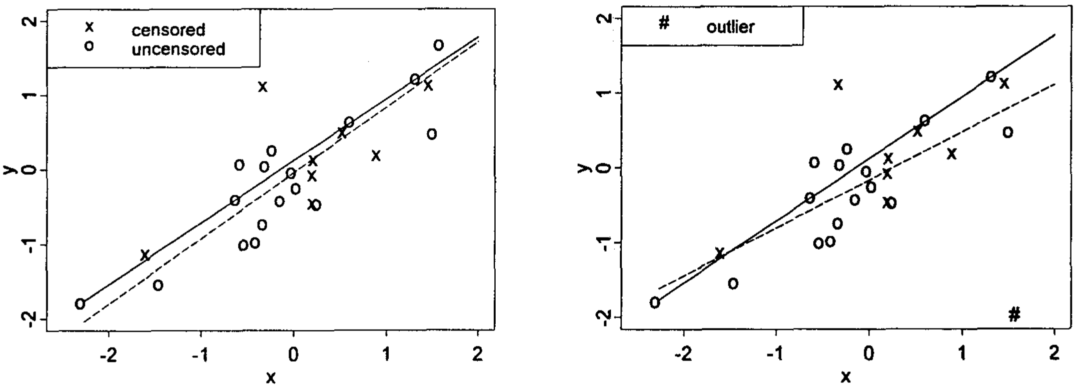

Figure 1. The solid line is the deepest regression line and the dashed line is the weighted least squares regression line

Table 1. Estimates and their standard errors with normal noise

		30% censoring, 25% truncation				50% censoring, 35% truncation			
		$\hat{\theta}_1$	$se(\hat{\theta}_1)$	$\hat{\theta}_2$	$se(\hat{\theta}_2)$	$\hat{\theta}_1$	$se(\hat{\theta}_1)$	$\hat{\theta}_2$	$se(\hat{\theta}_2)$
n=30	deepest line	0.984	0.260	-0.046	0.214	1.008	0.375	-0.099	0.303
	leaset squares	1.025	0.149	-0.088	0.148	1.031	0.194	-0.121	0.177
n=50	deepest line	0.998	0.234	-0.038	0.173	1.010	0.261	-0.113	0.201
	leaset squares	1.041	0.121	-0.094	0.116	1.046	0.143	-0.133	0.141
n=100	deepest line	1.013	0.203	-0.040	0.131	1.011	0.206	-0.099	0.152
	least squares	1.057	0.094	-0.099	0.082	1.060	0.098	-0.134	0.100

Table 2. Estimates and their standard errors with gamma noise

		30% censoring, 25% truncation				50% censoring, 35% truncation			
		$\hat{\theta}_1$	$se(\hat{\theta}_1)$	$\hat{\theta}_2$	$se(\hat{\theta}_2)$	$\hat{\theta}_1$	$se(\hat{\theta}_1)$	$\hat{\theta}_2$	$se(\hat{\theta}_2)$
n=30	deepest line	0.993	0.281	-0.108	0.206	0.991	0.337	-0.153	0.241
	leaset squares	1.029	0.146	-0.099	0.136	1.030	0.163	-0.133	0.155
n=50	deepest line	0.991	0.217	-0.105	0.162	1.005	0.230	-0.142	0.177
	leaset squares	1.036	0.102	-0.103	0.098	1.038	0.116	-0.135	0.113
n=100	deepest line	0.982	0.172	-0.094	0.120	0.983	0.170	-0.148	0.126
	least squares	1.044	0.075	-0.103	0.066	1.045	0.080	-0.135	0.077

The AIDS incubation data include 295 cases of HIV infection by blood or blood-product transfusion reported to the Center for Disease Control prior to January 1, 1987, and diagnosed prior to July 1, 1986. The data consists of three variables; INF is the month of infection with 1=January of 1987 and 101=June of 1986; DIAG is the duration of the induction period in months; and AGE is the age+1 (in years) at the time of infection. Following Kalbfleisch and Lawless (1989), the response variable Y is the incubation period defined as DIAG-0.5. Since only the patients diagnosed prior to July 1, 1986 are recruited into the study, the data are right truncated. The right truncation variable T is 101.5-INF. We observe (Y, T) when $Y \leq T$. The relationship between the age of patients (AGE=age + 1) and the incubation time (DIAG) was investigated by Gross and Lai (1996)) using a linear model, $-\log(DIAG) = \theta_1(AGE) + \theta_2 + \varepsilon$ Then response $Y' = -\log(DIAG)$ is left truncated by $T' = -\log(101.5 - INF)$. Because of the different characteristics between children (age≤4), adults (5≤age≤59), elderly (age≥60) patients, the data set is divided into three groups. Table 3 shows the estimates of regression coefficients and their bootstrap standard errors. The standard errors are estimated by 1000 bootstrap repetitions. The least squares estimates are taken from Gross and Lai (1996). We use the same values for a and b as in Gross and Lai (1996) and the weights given in (2). We can see the differences between estimates are not significant considering the bootstrap standard errors. However, note that the standard errors of the estimates by depth regression are greater than the least squares estimates as in the previous simulation study.

Table 3. Estimates of regression coefficients and their bootstrap standard errors

	least squares estimates				simplicial regression depth estimates			
AGE	$\hat{\theta}_1$	$se(\hat{\theta}_1)$	$\hat{\theta}_2$	$se(\hat{\theta}_2)$	$\hat{\theta}_1$	$se(\hat{\theta}_1)$	$\hat{\theta}_2$	$se(\hat{\theta}_2)$
	$a=-4.38, b=-1.84$							
age $\leq$ 4	-0.412	0.036	-1.915	0.086	-0.531	0.085	-1.772	0.229
5$\leq$ age $\leq$59	-0.006	0.003	-4.136	0.162	0.014	0.008	-4.302	0.308
age $\geq$ 60	0.024	0.016	-2.355	1.063	-0.065	0.051	0.771	3.595
	$a=-3.5, b=-1.84$							
age $\leq$ 4	-0.376	0.086	-1.974	0.163	-0.531	0.115	-1.772	0.253
5$\leq$ age $\leq$59	-0.004	0.004	-3.266	0.195	0.022	0.007	-4.354	0.290
age $\geq$ 60	-0.002	0.009	-2.980	0.590	-0.013	0.060	-2.396	4.179

In this article we have proposed a method to define the simplicial regression depth for censored and truncated data. To define regression depth, we have used the weights obtained from the product-limit estimator instead of equal weights.

From the simulation study, we can see that influence of a few observations on the deepest regression line is not serious compared to the least squares line. However there are some drawbacks with the regression depth. As we have seen in the simulation study, there is some loss of efficiency. Since the regression depth is not derived from the criterion to minimize the mean square error, the estimates by the regression depth may lose some efficiency compared to the least square estimates.

References

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

[2] 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.

[3] Hwang, J., Jorn, H. and Kim, J. (2002). On the performance of bivariate robust location

estimators under contamination. To appear in *Computational Statistics & Data Analysis.*

[4] Kalbfleisch, J.D. and Lawless, J.F. (1989). Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS, *J. Amer. Statist. Assoc.,* Vol. 84, 360-372.

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

[6] Lai, T.L. and Ying, Z. (1991). Estimating a distribution function with truncated and censored data, *Annals of Statistics*, Vol. 19, 417-442.

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

[8] Liu, R.Y. (1990). On a notion of data depth based on random simplices, *Annals of Statistics*, Vol. 18, 405-414.

[9] Miller, R.G. (1976). Least squares regression with censored data, *Biometrika,* Vol. 63, 449-464.

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

[11] Rousseeuw, P.J. and Hubert, M. (1999). Regression depth, *J. Amer. Statist. Assoc.,* Vol. 94, 388-402.

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

[Received October 2002, Accepted February 2003]