응용통계연구 제 **20**권 *2*호, *2007*년, *pp. 313–322*

# 뎁스를 이용한 생존회귀모형들의 비교연구*

김지연[1] 황진수[2]

요 약

오염이 있는 생존자료에서 여러 가지 회귀뎁스(regression depth)를 비교 연구하였다. 중도절단 자료에서 회귀뎁스에 대한 정의는 Park과 Hwang(2003)의 반공간회귀뎁스(halfspace regression depth)와 Park(2003)의 심플리셜 회귀뎁스(simplicial regression depth)가 있다. 본 논문은 Hubert 등(2001)이 제안한 사영회귀뎁스(projection regression depth)를 생존자료에서 사용하는 방법을 제시하고 이 방법과 기존의 뎁스기반 회귀모형과의 비교를 다양한 오염 상황에서 실시하였다.

주요용어: 중도절단, 회귀뎁스.

## 1. 서론

중도절단 자료를 포함하는 생존자료에서의 회귀모형은 Cox의 비례위험모형과 가속수명모형 등이 대표적으로 사용되고 있다. 그러나 생존자료에 오염된 자료가 포함되는 경우, 예를 들면 생존시간을 포함하는 마이크로어레이 자료에서의 전처리 과정 등에서 흔하게 발생할 수 있는 오염 상황에 필요한 로버스트한 생존 회귀모형에 대한 연구는 활발하지 않다. 본 연구에서는 여러 형태의 오염상황에서 로버스트한 데이터 뎁스의 개념을 이용한 여러 회귀뎁스 방법들을 생존자료에서 각각의 성능을 비교 분석하고자 한다.

뎁스를 이용한 로버스트한 회귀추정량들에 관한 비교 연구는 김지연 등(2004)에서 수행을 하였다. 그 연구에서는 일반 자료에서 순위에 기반하며 높은 붕괴점을 가지는 Chang 등(1999)의 HBR(High Breakdown Rank)방법과 사영뎁스에 기반을 둔 rcent(regression centrality) 방법 그리고 Rousseeuw와 Hubert(1999)의 반공간회귀뎁스 방법들간의 비교를 하였다. 모의실험의 결과는 사영뎁스에 기반을 둔 rcent 방법이 오염의 형태에 관계없이 비교적 좋은 결과를 보임을 알 수 있었다.

중도절단이 있는 생존자료에서의 회귀뎁스는 새로운 정의가 필요하다. Park과 Hwang (2003)은 반공간 회귀뎁스에서 중도절단 자료를 고려한 가중치를 제안하였으며 Park(2003)은 심플리셜 회귀뎁스에 이 가중치를 사용하는 방법을 제안하였다. 본 논문에서는 위치모

* 이 논문은 한국학술진흥재단 기초과학연구지원사업에서 지원되었음(KRF-2004-015-C00075).

1) (151-747) 서울특별시 관악구 신림동 산 56-1, 서울대학교 복잡계통계연구센터, 연구원
E-mail: jeeyun@inha.ac.kr
2) (402-751) 인천광역시 남구 용현동 253, 인하대학교 통계학과, 교수
E-mail: jshwang@inha.ac.kr

수추정에서 붕괴점이 1/2로 알려진 사영회귀뎁스를 생존자료에서 사용하는 방법을 제시하고 기존의 회귀뎁스 기반 추정량(반공간, 심플리셜)에 대한 비교연구를 수행하고자 한다.

제 2절에서는 여러 뎁스 기반 회귀추정량에 대한 소개를 하며 제 3절에서는 다양한 오염 상황에서의 모의실험 결과를 소개하며 끝으로 제 4절에서는 결론과 향후과제에 대한 토의를 진행하고자 한다.

## 2. 중도절단된 자료를 포함한 로버스트 회귀추정량

### 2.1. 생존회귀모형의 소개

중도절단 자료는 여러 분야에서 발생한다. 암이나 기타 여러가지 병에 걸린 환자가 얼마나 생존 할것인지가 관심인 의학분야나 기계나 부품의 수명이 얼마나 지속될 것인지가 중요한 공학분야에서 주로 많이 발생한다. 여기서 병에 걸린 환자의 생존시간이나 기계의 수명과 같은 자료가 반응변수에 해당된다. $T_1, \ldots, T_n$은 생존시간, $C_1, \ldots, C_n$은 중도절단 될 때까지의 시간이라고 하자. 생존시간과 같은 반응값은 모든 환자에 대하여 정확하게 관찰되지 않을 수 있다. 따라서 얻을 수 있는 자료의 형태는 $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$이 된다. 즉, 중도절단이 발생하면 그 환자는 중도절단이 발생한 시각 $(C_i)$ 까지만 관측이 된다. 이때, $Y_i$와 $\delta_i$는 각각 $Y_i = T_i \wedge C_i$, $\delta_i = I(T_i \leq C_i)$이 된다.

생존자료에 가중치를 주는 것은 여러 방법이 있을 수 있으나 여기에서는 Kaplan-Meier의 추정량을 이용한 $W_i = \delta_i \frac{\hat{S}(Y_i)}{\#(Y_i)}$ 를 사용하였다. $\hat{S}(y)$는 $P(Y > y)$의 Kaplan-Meier 추정량를 나타내며, $\#(Y_i)$는 관측치 중에서 $Y_i$보다 큰 값들의 갯수, 즉 위험집합의 크기를 나타낸다. 이 가중치는 실제로 Kaplan-Meier 추정량의 점프 크기를 나타낸다. 즉, 중도절단이 발생한 관측치는 가중치가 0이 되고 중도절단이 발생하지 않은 관측치의 경우에는 $W_i$의 가중치를 갖게 된다. 이 가중치를 사용한 추정량은 보다 일반적인 경우에 근사적으로 일치 추정량이 됨을 Gross와 Lai(1996)에서 밝혔다.

주어진 관측자료에서 반응변수에 해당하는 $Y_i$의 중앙값은 $Y_i$만을 이용하는 일반적인 중앙값이 아닌 $Y_i$ 와 $W_i$를 동시에 고려하는 방법을 사용한다. 일반적으로 중도절단 자료를 포함하는 경우의 중앙값은 추정생존함수 $\hat{S}$의 50 백분위수를 나타내는 $\hat{S}^{-1}(1/2)$인데, 이것은 Miller(1981)에서 언급한 바와 같이 원래값 보다 과대추정하게되므로 각 점프마다 보간법을 사용하여 보정을 한다.

로그생존시간이 공변량과 선형관계를 따른다고 가정하면 로그생존시간 $\log Y_i$는

$$\beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \beta_p = (\mathbf{x}_i, 1)\beta'$$

즉, $R^p$ 에서의 아핀초공간인 $(\mathbf{x}_i, 1)\beta'$으로 적합하게 된다. 여기서 $\mathbf{x}_i = (x_{i1}, \ldots, x_{i,p-1})$는 $R^{p-1}$에 속한다. 다음 절의 회귀뎁스의 설명에서는 독립변수와 종속변수를 합하여 자료들의 집합을

$$Z_n = \{(x_{i1}, \ldots, x_{i,p-1}, y_i); i = 1, \ldots, n\} \subset R^p$$

으로 표현하기로한다.

### 2.2. 반공간 회귀뎁스

Rousseeuw와 Hubert(1999)에 의하여 제안된 회귀뎁스는 Tukey(1975)의 반공간 위치뎁스를 기반으로 만들어졌으며 붕괴점도 일반적으로 최소 33%가 되며 양호한 로버스트 성질을 가지고 있다고 알려져 있다.

독립변수들로 이루어진 $X$ 공간의 임의의 아핀초공간(affine hyperplane) $V$는 관측치를 $L(V)$, $R(V)$ 둘로 나누며 $L^+(V)$는 $L(V)$에 속하는 관측치 중에서 잔차가 양인 집합이고 $L^-(V)$는 잔차가 음인 집합을 나타낸다. 마찬가지로 $R^+(V)$, $R^-(V)$도 정의된다. 반공간 회귀뎁스의 정의는

$$rdepth^{(H)}(\beta, Z_n) = \min_V(\min\{L^+(V) + R^-(V), R^+(V) + L^-(V)\})$$

로 주어진다. 여기서 $L^+(V)$, $L^-(V)$, $R^+(V)$, $R^-(V)$를 식으로 표현하면 다음과 같이 주어지며

$$L^+(V) = \#\{i : X_i \in L(V), Y_i - \beta' X_i \geq 0\},$$
$$L^-(V) = \#\{i : X_i \in L(V), Y_i - \beta' X_i \leq 0\},$$
$$R^+(V) = \#\{i : X_i \in R(V), Y_i - \beta' X_i \geq 0\},$$
$$R^-(V) = \#\{i : X_i \in R(V), Y_i - \beta' X_i \leq 0\}.$$

회귀뎁스 추정량은 $rdepth^{(H)}(\beta, Z_n)$를 최대로 하는 것이다. 즉, $\hat{\beta} = \text{argmax}_\beta rdepth^{(H)}(\beta, Z_n)$ 이 된다.

중도절단 자료가 포함된 경우의 가중치 $W_i$를 고려한 회귀뎁스 추정량 $wrdepth(\beta, Z_n)$은 Park과 Hwang(2003)에서 제안되었다.

$$wrdepth(\beta, Z_n) = \min_V(\min\{WL^+(V) + WR^-(V), WR^+(V) + WL^-(V)\})$$

이 되며 여기서

$$WL^+(V) = \sum I\{i : X_i \in L(V), Y_i - \beta' X_i \geq 0\} W_i$$

가 되면 나머지 $WL^-(V)$, $WR^+(V)$, $WR^-(V)$도 같은 방법으로 확장 정의된다. 따라서 중도절단 자료가 포함된 경우의 회귀뎁스 추정량은 $\hat{\beta_w} = \text{argmax}_\beta wrdepth(\beta, Z_n)$ 이 된다.

### 2.3. 심플리셜 회귀뎁스

Rousseeuw와 Hubert(1999)에 의해 소개된 심플리셜 회귀뎁스는 다음과 같다.

$$rdepth^{(S)}(\beta, Z_n) = \binom{n}{p+1}^{-1} \sum_{i_1 < \cdots < i_{p+1}} I(\beta \in S(H_{i_1}, \ldots, H_{i_{p+1}})).$$

여기서 $H_{i_1}$은 관측치 $Z_{i_1}$에 해당하는 초평면이고 $S$는 $(p+1)$개의 초평면으로 정의되는 심플렉스이다. 즉, 서로 다른 $(p+1)$개의 심플렉스 중에서 주어진 모수 $\beta$를 포함하는 비율

로써 심플리셜 회귀뎁스로 정의한다. 따라서 심플리셜 회귀뎁스의 추정량은 $T_r^{(s)}(Z_n) = \text{argmax}_\beta rdepth^{(s)}(\beta, Z_n)$이 되며 이 추정량의 붕괴점은 $1/2^p$로 알려져 있어서 자료의 차원($p$)이 높아지면 성능이 안좋아진다.

중도절단 자료가 포함된 경우 추정량은 Park(2003)에 의하여 다음과 같이 제안되었다.

$$sdepth(\beta_1, \ldots, \beta_p) = \alpha^{-1} \sum_{i_1 < \cdots < i_{p+1}} I(\beta \in S(H_{i_1}, H_{i_2}, \ldots, H_{i_{p+1}}))(W_{i_1} + \cdots + W_{i_{p+1}})$$

이 되고, 여기서 $\alpha$는 $\alpha = \sum_{i_1 < \cdots < i_{p+1}} (W_{i_1} + \cdots + W_{i_{p+1}})$ 이다. 즉, 서로 다른 $(p+1)$개의 심플렉스 중에서 주어진 모수 $\beta$를 포함하는 가중치의 비율로써 정의한다. 결국, 중도절단 자료가 포함된 자료의 심플리셜 회귀뎁스의 추정량은 일반적인 경우와 같이 심플리셜 회귀뎁스가 가장 큰 추정량이 된다.

### 2.4. 사영 회귀뎁스

Hubert 등(2001)에 의하여 제안된 사영 회귀뎁스는 반공간 회귀뎁스 추정량처럼 잔차의 부호에 의해서만 결정되는 것이 아니라 잔차의 크기까지 고려한 추정량이다. 사영뎁스는 붕괴점이 1/2로서 가장 좋은 붕괴점을 가지고 있다. 즉, 자료의 거의 반정도를 교체하여도 추정량이 영향을 받지 않는다.

$$rcent(\beta, Z_n) = \inf_{\substack{\|u\|=1 \\ \nu \in R}} \frac{M_r}{\left(M_r + \left|\text{med}_i \frac{r_i(\beta)}{u^t x_i - \nu}\right|\right)},$$

단

$$M_r = \frac{\text{med}_i |y_i - \text{med}_j y_j|}{\text{med}_i |u^t x_i - \nu|}.$$

여기서 $u$는 방향을 나타내는 단위 벡터이고 $\nu$는 일차원 공간상의 경계를 표현하는 값이다. 사영최대뎁스 회귀추정량은 $T_r^c(Z_n) = \text{argmax} rcent(\beta, Z_n)$이 된다. 여기서 $\text{med}_j y_j$ 는 $y$ 들의 중앙값에 해당된다. 그러나 중도절단 자료가 포함된 경우는 위 식 $M_r$에서 $\text{med}_j y_j$를 단순히 $y$들의 중앙값이 아닌 $\hat{S}^{-1}(1/2)$의 선형보간을 이용한 $\hat{\hat{S}}^{-1}(1/2)$를 사용한다.

## 3. 모의실험

모의실험에서는 앞절에서 언급한 세가지 추정량 즉, 반공간 회귀뎁스(WR(H)), 심플리셜 회귀뎁스(WR(S)), 사영 회귀뎁스(WRcent)의 성능을 여러가지 오염 상황에서 비교하여 보았다. 비교의 기준은 추정량의 안정성인 분산과 편의$^2$ 그리고 편의$^2$비율(편의$^2$/MSE)를 계산하여 보았다. 모의실험은 크게 두가지 형태로 나누어 실시하였다. 즉, 오염의 방향이 $x$축인 경우와 $y$축인 경우로 나누어 오염의 방향과 오염의 비율, 그리고 중도절단 비율에 따라 세가지 추정량의 결과가 어떻게 변화하는지를 실험하였다. 특히 오염의 방향이 $y$축인 경우는 오염이 원자료보다 큰쪽으로 위치한 경우와 원자료보다 작은쪽으로 위치한 경우로 나누어 비교해 보았다. 오염의 비율과 중도절단 비율은 각각 10%, 20%, 30%, 40%로 고려하였다.

원자료의 모형은 $\log Y_i = \beta_0 + \beta_1 x_i + e_i$ , $e_i \sim N(0, \sigma^2)$ 이고 독립변수 $x_1, \ldots, x_n$ 는 $U(-1, 1)$ 에서 생성한다. ($\beta_0, \beta_1$ 은 0과 1이다.). 오염자료는 평균이 $\mu_1$ 과 $\mu_2$ 이고 분산 공분산 행렬이 $\tau I_2$인 이변량정규분포에서 생성한다.($\tau = 0.1$). 따라서 전체 생성자료는 $(1-\alpha)$ 만큼의 원자료와 $\alpha$ 만큼의 오염자료로 이루어졌다. 여기서 $\alpha$는 오염비율을 나타낸다. 중도절단 자료는 원자료와 오염자료에서 랜덤하게 정해진 비율만큼 생성하여 중도절단 자료로 만들었다. 원자료와 오염자료에서는 각각 $\gamma$ 만큼 중도절단 자료를 포함하고 있으므로 원자료에서는 $(1-\alpha)(1-\gamma)$만큼의 정상 자료(중도절단 자료가 아닌 자료)가 있고 오염자료에서는 $\alpha(1-\gamma)$만큼 정상 자료가 있다. $\gamma$는 중도절단 비율을 나타낸다. 모의실험 표본의 크기는 50이며 반복은 200번 행하였다.

### 3.1. $y$축 오염

원자료에서 $y$축 방향으로 오염이 있는 경우, 원자료 보다 오염이 $y$축으로 큰 경우와 작은 경우로 나누었고 그리고 원자료의 분산이 큰 경우($\sigma^2 = 1.0$)와 작은 경우($\sigma^2 = 0.01$)로 나누어 비교하였다. 분산이 작은 경우는 분산이 큰 경우에 비해 상대적으로 오염자료와 원자료간의 구별이 좀더 명확한 경우라고 할 수 있다. 표 3.1과 표 3.2는 오염자료의 $y$값이 원자료보다 큰 경우에 해당하고 표 3.3과 표 3.4는 그 반대로 오염자료의 $y$값이 원자료보다 작은 경우이다.

표 3.1은 원자료의 분산이 1.0 인 경우에 해당된다. 표를 살펴보면 WR(S)의 편의$^2$비율이 WRcent와 WR(H)보다 오염의 비율 10%를 제외한 모든 오염비율과 모든 중도절단 비율에서 작음을 보여주고 있다. 분산은 다른 방법들과 비교해 볼때 작은 값을 보여 주는 것

표 3.1: $y$축 오염에서의 성능 비교 ($\sigma^2 = 1.0$ , $(\mu_1, \mu_2)^t = (1, 5)^t$)

| $\alpha$ | $\gamma$ | Method | Slope | | | $\alpha$ | $\gamma$ | Method | Slope | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 분산 | 편의$^2$ | 편의$^2$비율 | | | | 분산 | 편의$^2$ | 편의$^2$비율 |
| 0.1 | 0.1 | WRcent | 0.2037 | 0.1552 | 0.4324 | 0.3 | 0.1 | WRcent | 0.3979 | 2.3128 | 0.8532 |
| | | WR(H) | 0.1578 | 0.0264 | 0.1433 | | | WR(H) | 0.3833 | 1.7852 | 0.8232 |
| | | WR(S) | 0.1816 | 0.0162 | 0.0819 | | | WR(S) | 0.5480 | 1.4045 | 0.7193 |
| | 0.2 | WRcent | 0.1962 | 0.1203 | 0.3801 | | 0.2 | WRcent | 0.4031 | 2.2902 | 0.8503 |
| | | WR(H) | 0.2161 | 0.0014 | 0.0064 | | | WR(H) | 0.3516 | 0.7717 | 0.6870 |
| | | WR(S) | 0.2258 | 0.0319 | 0.1238 | | | WR(S) | 0.4207 | 0.4225 | 0.5011 |
| | 0.3 | WRcent | 0.1585 | 0.1628 | 0.5067 | | 0.3 | WRcent | 0.4401 | 2.3347 | 0.8414 |
| | | WR(H) | 0.1956 | 0.0025 | 0.0126 | | | WR(H) | 0.3675 | 0.4247 | 0.5361 |
| | | WR(S) | 0.2571 | 0.0567 | 0.1807 | | | WR(S) | 0.3240 | 0.0600 | 0.1563 |
| | 0.4 | WRcent | 0.2719 | 0.1231 | 0.3116 | | 0.4 | WRcent | 0.4771 | 2.4452 | 0.8367 |
| | | WR(H) | 0.6952 | 0.0763 | 0.0989 | | | WR(H) | 1.0711 | 1.0325 | 0.4908 |
| | | WR(S) | 0.2725 | 0.0860 | 0.2399 | | | WR(S) | 0.4788 | 0.0052 | 0.0107 |
| 0.2 | 0.1 | WRcent | 0.2079 | 0.6178 | 0.7482 | 0.4 | 0.1 | WRcent | 0.1610 | 4.9378 | 0.9684 |
| | | WR(H) | 0.1819 | 0.2676 | 0.5953 | | | WR(H) | 0.1866 | 5.0348 | 0.9643 |
| | | WR(S) | 0.2716 | 0.1503 | 0.3562 | | | WR(S) | 0.4197 | 3.8947 | 0.9027 |
| | 0.2 | WRcent | 0.3154 | 0.6545 | 0.6748 | | 0.2 | WRcent | 0.1597 | 4.7818 | 0.9677 |
| | | WR(H) | 0.1995 | 0.0766 | 0.2774 | | | WR(H) | 0.3950 | 3.7413 | 0.9045 |
| | | WR(S) | 0.2022 | 0.0133 | 0.0617 | | | WR(S) | 0.6239 | 2.2097 | 0.7798 |
| | 0.3 | WRcent | 0.2557 | 0.5957 | 0.6997 | | 0.3 | WRcent | 0.2065 | 4.8152 | 0.9589 |
| | | WR(H) | 0.2892 | 0.0675 | 0.1892 | | | WR(H) | 0.6468 | 3.0014 | 0.8227 |
| | | WR(S) | 0.2446 | 0.0079 | 0.0313 | | | WR(S) | 0.6547 | 0.8737 | 0.5716 |
| | 0.4 | WRcent | 0.3581 | 0.6987 | 0.6611 | | 0.4 | WRcent | 0.2542 | 4.9310 | 0.9510 |
| | | WR(H) | 1.1140 | 0.4418 | 0.2840 | | | WR(H) | 1.0957 | 2.2465 | 0.6722 |
| | | WR(S) | 0.3717 | 0.0392 | 0.0954 | | | WR(S) | 0.5615 | 0.1073 | 0.1604 |

표 3.2: $y$축 오염에서의 성능 비교 ($\sigma^2 = 0.01$ , $(\mu_1, \mu_2)^t = (1,5)^t$)

| $\alpha$ | $\gamma$ | Method | Slope 분산 | Slope 편의$^2$ | Slope 편의$^2$비율 | $\alpha$ | $\gamma$ | Method | Slope 분산 | Slope 편의$^2$ | Slope 편의$^2$비율 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | WRcent | 0.0011 | 0.0009 | 0.4500 | 0.3 | 0.1 | WRcent | 0.0037 | 0.0163 | 0.8150 |
| | | WR(H) | 0.0013 | 0.0000 | 0.0000 | | | WR(H) | 0.0051 | 0.0196 | 0.7935 |
| | | WR(S) | 0.0017 | 0.0000 | 0.0000 | | | WR(S) | 0.0101 | 0.0184 | 0.6456 |
| | 0.2 | WRcent | 0.0014 | 0.0015 | 0.5172 | | 0.2 | WRcent | 0.0046 | 0.0126 | 0.7326 |
| | | WR(H) | 0.0021 | 0.0000 | 0.0000 | | | WR(H) | 0.0041 | 0.0077 | 0.6525 |
| | | WR(S) | 0.0035 | 0.0005 | 0.1250 | | | WR(S) | 0.0054 | 0.0035 | 0.3933 |
| | 0.3 | WRcent | 0.0020 | 0.0016 | 0.4444 | | 0.3 | WRcent | 0.0070 | 0.0152 | 0.6847 |
| | | WR(H) | 0.0041 | 0.0001 | 0.0238 | | | WR(H) | 0.0055 | 0.0047 | 0.4608 |
| | | WR(S) | 0.0061 | 0.0001 | 0.0161 | | | WR(S) | 0.0038 | 0.0004 | 0.0952 |
| | 0.4 | WRcent | 0.0018 | 0.0018 | 0.5000 | | 0.4 | WRcent | 0.0065 | 0.0160 | 0.7111 |
| | | WR(H) | 0.0598 | 0.0065 | 0.0980 | | | WR(H) | 0.7316 | 0.2023 | 0.2166 |
| | | WR(S) | 0.0087 | 0.0001 | 0.0114 | | | WR(S) | 0.0088 | 0.0002 | 0.0222 |
| 0.2 | 0.1 | WRcent | 0.0025 | 0.0058 | 0.6988 | 0.4 | 0.1 | WRcent | 0.0431 | 0.0235 | 0.3529 |
| | | WR(H) | 0.0024 | 0.0026 | 0.5200 | | | WR(H) | 0.8548 | 1.7906 | 0.6769 |
| | | WR(S) | 0.0029 | 0.0020 | 0.4082 | | | WR(S) | 0.6459 | 0.5081 | 0.4403 |
| | 0.2 | WRcent | 0.0027 | 0.0060 | 0.6897 | | 0.2 | WRcent | 0.0882 | 0.0300 | 0.2538 |
| | | WR(H) | 0.0027 | 0.0004 | 0.1290 | | | WR(H) | 0.5483 | 0.2925 | 0.3479 |
| | | WR(S) | 0.0030 | 0.0000 | 0.0000 | | | WR(S) | 0.2189 | 0.0649 | 0.2287 |
| | 0.3 | WRcent | 0.0032 | 0.0057 | 0.6404 | | 0.3 | WRcent | 0.2218 | 0.0607 | 0.2149 |
| | | WR(H) | 0.0055 | 0.0015 | 0.2143 | | | WR(H) | 0.5312 | 0.2210 | 0.2938 |
| | | WR(S) | 0.0047 | 0.0001 | 0.0208 | | | WR(S) | 0.0097 | 0.0073 | 0.4294 |
| | 0.4 | WRcent | 0.0033 | 0.0066 | 0.6667 | | 0.4 | WRcent | 0.3124 | 0.0827 | 0.2093 |
| | | WR(H) | 0.2115 | 0.0314 | 0.1293 | | | WR(H) | 1.1359 | 0.5747 | 0.3360 |
| | | WR(S) | 0.0106 | 0.0001 | 0.0093 | | | WR(S) | 0.0079 | 0.0018 | 0.1856 |

도 있지만 큰 값을 보여주는 것들도 있으나 편의$^2$ 값은 다른 방법에 비해 훨씬 작음을 보여준다.

표 3.2는 원자료의 분산이 0.01 인 경우에 해당된다. 표 3.1과 표 3.2와는 원자료의 분산만 다를 뿐인데 결과는 많이 다른 경향을 보이고 있다. 오염의 비율이 10%, 20%일때는 대체적으로 분산은 WRcent의 값이, 편의$^2$과 편의$^2$비율은 WR(S)의 값이 작음을 알 수 있다. 그러나 오염의 비율이 30%, 40%일때에는 대체적으로 중도절단 비율이 클때 WR(S)의 편의$^2$이 다른 방법들에 비해 작음을 알 수 있다.

표 3.3은 원자료의 분산이 1.0 인 경우에 해당된다. 표를 살펴보면 오염의 비율이 20%이고 중도절단의 비율이 10%, 40%를 제외하고는 모든 경우에 WRcent의 편의$^2$비율이 다른 방법들 보다 훨씬 작음을 알 수 있다. 또한 오염의 비율 40%를 제외하면 분산, 편의$^2$, 편의$^2$비율 모두가 WRcent에서 좋음을 볼 수 있다. 오염의 비율이 40%이고 중도절단 비율이 30%, 40%일 경우에는 WR(S)의 방법이 분산, 편의$^2$에서 좋은 결과를 보여준다.

표 3.4는 원자료의 분산이 0.01 인 경우에 해당된다. 이 경우는 모든 오염비율과 모든 중도절단 비율에서 WRcent가 분산, 편의$^2$에서 다른 방법들에 비해 훨씬 좋은 결과를 보여준다. 오염의 비율이 증가할수록 또한 중도절단 비율이 증가할수록 WRcent의 결과가 다른 방법들보다 더 좋음을 알 수 있다. 오염의 비율이 20%, 30%에서 편의$^2$비율의 값이 WR(H)와 WR(S)에서 적음을 볼 수 있는데 그것은 상대적으로 분산이 크기 때문이다.

### 3.2. $x$축 오염

오염 자료가 $x$축 방향에서 발생한 경우이다. 이 오염의 자료는 원자료의 값들보다 크지 않은 $y$값들을 갖는 경우를 고려하였다. 다른 경우의 결과도 이 결과와 다르지 않다.

표 3.3: $y$축 오염에서의 성능 비교 ($\sigma^2 = 1.0$ , $(\mu_1, \mu_2)^t = (1, -5)^t$)

| $\alpha$ | $\gamma$ | Method | Slope 분산 | Slope 편의$^2$ | Slope 편의$^2$비율 | $\alpha$ | $\gamma$ | Method | Slope 분산 | Slope 편의$^2$ | Slope 편의$^2$비율 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | WRcent | 0.1380 | 0.1327 | 0.4902 | 0.3 | 0.1 | WRcent | 0.6656 | 1.9506 | 0.7456 |
| | | WR(H) | 0.1424 | 0.2257 | 0.6131 | | | WR(H) | 0.9237 | 4.6278 | 0.8336 |
| | | WR(S) | 0.2381 | 0.3203 | 0.5736 | | | WR(S) | 1.2673 | 6.7152 | 0.8412 |
| | 0.2 | WRcent | 0.1938 | 0.1000 | 0.3404 | | 0.2 | WRcent | 0.6834 | 1.7232 | 0.7160 |
| | | WR(H) | 0.2018 | 0.3439 | 0.6302 | | | WR(H) | 1.2781 | 6.6031 | 0.8378 |
| | | WR(S) | 0.2789 | 0.5599 | 0.6675 | | | WR(S) | 1.0840 | 9.1715 | 0.8943 |
| | 0.3 | WRcent | 0.2090 | 0.1335 | 0.3898 | | 0.3 | WRcent | 0.7732 | 1.8496 | 0.7052 |
| | | WR(H) | 0.2405 | 0.4240 | 0.6381 | | | WR(H) | 1.2444 | 7.0810 | 0.8505 |
| | | WR(S) | 0.4601 | 0.8198 | 0.6405 | | | WR(S) | 0.6997 | 10.8754 | 0.9396 |
| | 0.4 | WRcent | 0.2125 | 0.0813 | 0.2767 | | 0.4 | WRcent | 1.0714 | 1.6515 | 0.6065 |
| | | WR(H) | 0.5832 | 0.1936 | 0.2492 | | | WR(H) | 2.3597 | 6.6378 | 0.7377 |
| | | WR(S) | 0.3794 | 0.9867 | 0.7223 | | | WR(S) | 0.3939 | 11.0975 | 0.9657 |
| 0.2 | 0.1 | WRcent | 0.2080 | 0.5685 | 0.7321 | 0.4 | 0.1 | WRcent | 0.6094 | 11.4273 | 0.9494 |
| | | WR(H) | 0.2215 | 0.9773 | 0.8152 | | | WR(H) | 0.2498 | 14.5599 | 0.9831 |
| | | WR(S) | 0.5493 | 1.3797 | 0.7152 | | | WR(S) | 0.4098 | 13.3550 | 0.9702 |
| | 0.2 | WRcent | 0.2678 | 0.4478 | 0.6258 | | 0.2 | WRcent | 0.9433 | 11.1386 | 0.9219 |
| | | WR(H) | 0.3606 | 1.3108 | 0.7843 | | | WR(H) | 0.4653 | 14.5998 | 0.9691 |
| | | WR(S) | 0.7044 | 1.9499 | 0.7346 | | | WR(S) | 0.4110 | 13.0303 | 0.9694 |
| | 0.3 | WRcent | 0.2585 | 0.4807 | 0.6503 | | 0.3 | WRcent | 1.1174 | 12.4884 | 0.9179 |
| | | WR(H) | 0.3983 | 1.2591 | 0.7597 | | | WR(H) | 0.5547 | 15.0724 | 0.9645 |
| | | WR(S) | 1.1325 | 3.2846 | 0.7436 | | | WR(S) | 0.3258 | 11.2750 | 0.9719 |
| | 0.4 | WRcent | 0.3656 | 0.5293 | 0.5915 | | 0.4 | WRcent | 0.9618 | 10.5436 | 0.9164 |
| | | WR(H) | 0.8904 | 1.1513 | 0.5639 | | | WR(H) | 0.6936 | 13.3339 | 0.9506 |
| | | WR(S) | 1.1203 | 4.4523 | 0.7990 | | | WR(S) | 0.4015 | 9.1708 | 0.9581 |

표 3.4: $y$축 오염에서의 성능 비교 ($\sigma^2 = 0.01$ , $(\mu_1, \mu_2)^t = (1, -5)^t$)

| $\alpha$ | $\gamma$ | Method | Slope 분산 | Slope 편의$^2$ | Slope 편의$^2$비율 | $\alpha$ | $\gamma$ | Method | Slope 분산 | Slope 편의$^2$ | Slope 편의$^2$비율 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | WRcent | 0.0016 | 0.0014 | 0.4667 | 0.3 | 0.1 | WRcent | 0.0044 | 0.0169 | 0.7934 |
| | | WR(H) | 0.0017 | 0.0025 | 0.5952 | | | WR(H) | 0.0525 | 0.0798 | 0.6032 |
| | | WR(S) | 0.0028 | 0.0035 | 0.5556 | | | WR(S) | 1.3406 | 0.4532 | 0.2526 |
| | 0.2 | WRcent | 0.0016 | 0.0009 | 0.3600 | | 0.2 | WRcent | 0.0057 | 0.0153 | 0.7286 |
| | | WR(H) | 0.0032 | 0.0026 | 0.4483 | | | WR(H) | 1.7676 | 0.8039 | 0.3126 |
| | | WR(S) | 0.0057 | 0.0043 | 0.4300 | | | WR(S) | 2.5270 | 3.3453 | 0.5697 |
| | 0.3 | WRcent | 0.0020 | 0.0013 | 0.3939 | | 0.3 | WRcent | 0.0057 | 0.0140 | 0.7107 |
| | | WR(H) | 0.0037 | 0.0039 | 0.5132 | | | WR(H) | 2.1090 | 1.3796 | 0.3955 |
| | | WR(S) | 0.0089 | 0.0116 | 0.5659 | | | WR(S) | 2.1004 | 6.1426 | 0.7452 |
| | 0.4 | WRcent | 0.0019 | 0.0007 | 0.2692 | | 0.4 | WRcent | 0.0061 | 0.0139 | 0.6950 |
| | | WR(H) | 0.0059 | 0.0009 | 0.1324 | | | WR(H) | 2.1322 | 1.1507 | 0.3505 |
| | | WR(S) | 0.0422 | 0.0217 | 0.3396 | | | WR(S) | 1.0879 | 9.4033 | 0.8963 |
| 0.2 | 0.1 | WRcent | 0.0025 | 0.0053 | 0.6795 | 0.4 | 0.1 | WRcent | 0.0157 | 0.0229 | 0.5933 |
| | | WR(H) | 0.0042 | 0.0130 | 0.7558 | | | WR(H) | 0.4837 | 15.8041 | 0.9703 |
| | | WR(S) | 0.0055 | 0.0165 | 0.7500 | | | WR(S) | 1.8616 | 11.3841 | 0.8595 |
| | 0.2 | WRcent | 0.0023 | 0.0040 | 0.6349 | | 0.2 | WRcent | 0.0148 | 0.0187 | 0.5582 |
| | | WR(H) | 0.0068 | 0.0166 | 0.7094 | | | WR(H) | 0.3799 | 15.6194 | 0.9763 |
| | | WR(S) | 0.0204 | 0.0383 | 0.6525 | | | WR(S) | 0.5464 | 13.5508 | 0.9612 |
| | 0.3 | WRcent | 0.0028 | 0.0049 | 0.6364 | | 0.3 | WRcent | 0.0162 | 0.0140 | 0.4636 |
| | | WR(H) | 0.0195 | 0.0238 | 0.5497 | | | WR(H) | 0.2816 | 14.1157 | 0.9804 |
| | | WR(S) | 0.2399 | 0.1281 | 0.3481 | | | WR(S) | 0.2407 | 12.1682 | 0.9806 |
| | 0.4 | WRcent | 0.0032 | 0.0056 | 0.6364 | | 0.4 | WRcent | 0.0319 | 0.0182 | 0.3633 |
| | | WR(H) | 0.0734 | 0.0212 | 0.2241 | | | WR(H) | 1.2044 | 11.2144 | 0.9030 |
| | | WR(S) | 0.8263 | 0.5138 | 0.3834 | | | WR(S) | 0.3381 | 9.8197 | 0.9667 |

표 3.5: $x$축 오염에서의 성능 비교 ($\sigma^2 = 0.01$ , $(\mu_1, \mu_2)^t = (4, 0)^t$)

| $\alpha$ | $\gamma$ | Method | Slope 분산 | Slope 편의$^2$ | Slope 편의$^2$비율 | $\alpha$ | $\gamma$ | Method | Slope 분산 | Slope 편의$^2$ | Slope 편의$^2$비율 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | WRcent | 0.0014 | 0.0016 | 0.5333 | 0.3 | 0.1 | WRcent | 0.0035 | 0.0027 | 0.4355 |
| | | WR(H) | 0.0019 | 0.0038 | 0.6667 | | | WR(H) | 0.1173 | 0.3461 | 0.7469 |
| | | WR(S) | 0.0023 | 0.0045 | 0.6618 | | | WR(S) | 0.1005 | 0.1960 | 0.6610 |
| | 0.2 | WRcent | 0.0017 | 0.0011 | 0.3929 | | 0.2 | WRcent | 0.0034 | 0.0029 | 0.4603 |
| | | WR(H) | 0.0033 | 0.0045 | 0.5769 | | | WR(H) | 0.1069 | 0.4417 | 0.8051 |
| | | WR(S) | 0.0058 | 0.0069 | 0.5433 | | | WR(S) | 0.1041 | 0.3081 | 0.7475 |
| | 0.3 | WRcent | 0.0018 | 0.0009 | 0.3333 | | 0.3 | WRcent | 0.0028 | 0.0031 | 0.5254 |
| | | WR(H) | 0.0050 | 0.0040 | 0.4444 | | | WR(H) | 0.1142 | 0.4429 | 0.7950 |
| | | WR(S) | 0.0119 | 0.0127 | 0.5163 | | | WR(S) | 0.0945 | 0.4444 | 0.8246 |
| | 0.4 | WRcent | 0.0018 | 0.0009 | 0.3333 | | 0.4 | WRcent | 0.0084 | 0.0042 | 0.3333 |
| | | WR(H) | 0.0066 | 0.0015 | 0.1852 | | | WR(H) | 0.1545 | 0.2601 | 0.6274 |
| | | WR(S) | 0.0176 | 0.0117 | 0.3993 | | | WR(S) | 0.1030 | 0.3533 | 0.7743 |
| 0.2 | 0.1 | WRcent | 0.0025 | 0.0035 | 0.5833 | 0.4 | 0.1 | WRcent | 0.1974 | 0.1814 | 0.4789 |
| | | WR(H) | 0.0084 | 0.0187 | 0.6900 | | | WR(H) | 0.0057 | 0.8616 | 0.9934 |
| | | WR(S) | 0.0131 | 0.0237 | 0.6440 | | | WR(S) | 0.0095 | 0.8239 | 0.9886 |
| | 0.2 | WRcent | 0.0029 | 0.0040 | 0.5797 | | 0.2 | WRcent | 0.2048 | 0.2530 | 0.5526 |
| | | WR(H) | 0.0230 | 0.0259 | 0.5297 | | | WR(H) | 0.0025 | 0.8515 | 0.9971 |
| | | WR(S) | 0.0317 | 0.0405 | 0.5609 | | | WR(S) | 0.0142 | 0.7954 | 0.9825 |
| | 0.3 | WRcent | 0.0032 | 0.0055 | 0.6322 | | 0.3 | WRcent | 0.2040 | 0.2316 | 0.5317 |
| | | WR(H) | 0.0259 | 0.0262 | 0.5029 | | | WR(H) | 0.0209 | 0.8273 | 0.9754 |
| | | WR(S) | 0.0303 | 0.0465 | 0.6055 | | | WR(S) | 0.0262 | 0.7518 | 0.9663 |
| | 0.4 | WRcent | 0.0035 | 0.0038 | 0.5205 | | 0.4 | WRcent | 0.1982 | 0.3175 | 0.6157 |
| | | WR(H) | 0.0273 | 0.0147 | 0.3500 | | | WR(H) | 0.0714 | 0.7286 | 0.9108 |
| | | WR(S) | 0.0459 | 0.0479 | 0.5107 | | | WR(S) | 0.0376 | 0.7402 | 0.9517 |

표 3.5는 원자료의 분산이 0.01인 경우에 해당된다. 모든 오염비율에서 WRcent의 편의$^2$이 가장 작음을 알 수 있다. 특히 오염의 비율이 10%, 20%, 30%에서는 WRcent가 분산, 편의$^2$에서 가장 작은 결과를 얻었다. 오염의 비율이 20%를 제외하면 편의$^2$비율값도 WRcent가 다른 방법들 보다 작은 결과를 보여주고 있다. 오염의 비율이 40%에서는 WRcent가 분산은 다른 방법들에 비해 큰 값을 보이지만 편의$^2$은 상대적으로 작은 값을 보여 편의$^2$비율의 값이 작게됨을 알 수 있다.

## 4. 결론 및 토의

중도절단 자료를 포함하면서 또한 오염된 자료가 있는 경우 WRcent, WR(H), WR(S) 추정량들을 비교하였다. 오염은 $y$축과 $x$축의 방향을 고려하였고 특별히 $y$축 방향은 다시 원자료보다 큰 경우와 작은 경우로 나누었다. 오염이 $y$축의 방향에 있으면서 원자료보다 큰 $y$값을 갖는 경우에는 모든 경우는 아니지만 대체적으로 WR(S)값이 좋은 결과를 보여주는 반면에 오염이 $y$축의 방향에 있으면서 원자료보다 작은 $y$값을 갖는 경우에는 WRcent의 값이 좋은 결과를 보여준다. 그리고 오염이 $x$축의 방향에 있으면서 원자료보다 크지 않은 $y$값을 갖는 경우에도 WRcent의 값이 다른 방법들 보다 좋은 결과를 보여준다. 현재까지 진행되어온 모의 실험결과를 분석해 보면 오염 자료가 원자료 보다 큰 $y$값을 갖는 경우를 제외하면, 즉 오염의 자료가 원자료 보다 크지 않은 $y$값을 갖을때 중도절단 자료를 포함한는 경우에는 WRcent의 방법이 제일 좋다고 할 수 있다. 이것은 WRcent에서 $y$값들의 중앙값을 $\hat{\hat{S}}^{-1}(1/2)$로 대치하여 사용했기 때문이다. 즉, 오염의 자료가 원자료 보다 큰 $y$값을 갖는 경우는 $\hat{\hat{S}}^{-1}(1/2)$의 값이 오염부분에서 나타날 가능성이 커지게 되므로 WRcent는 좋지 않은

결과를 보여준다. 그러나 오염의 자료가 원자료 보다 작거나 또는 크지않은 $y$ 값을 갖는 경우에는 $\hat{\hat{S}}^{-1}(1/2)$의 값이 원자료에서 나타나게 되므로 좋은 결과를 보여주게 된다. 모의실험의 횟수와 반복의 수는 현 프로그램의 한계로 인하여 통상적으로 실시하는 것보다 적게 하였는데 이는 프로그래밍의 기법을 보완하거나 근사적인 프로그램 기법을 이용하여 추후 보완해야할 부분으로 생각한다. 또한 사영회귀뎁스에서 중도절단 자료를 보정하는 방법의 한계로 인하여 나타나는 사영회귀뎁스의 로버스트성의 문제는 추후에 해결해야할 과제로 생각된다.

## 참고문헌

김지연, 황진수, 김진경. (2004). 다양한 오염 상황에서의 여러 로버스트 회귀추정량의 비교연구, <응용통계연구>, **17**, 475–488.

Chang, W. H., Mckean,J. W., Naranjo, J. D. and Sheather, S. J. (1999). High-breakdown rank regression, *Journal of the American Statistical Association*, **94**, 205–219.

Gross, S. T. and Lai, T. L. (1996). Nonparametric estimation and regression analysis with left-truncated and right-censored data, *Journal of the American Statistical Association*, **91**, 1166–1180.

Hubert, M., Rousseeuw, P. J. and Van Aelst, S. (2001). Similarities between location depth and regression depth, *Trends in Mathematics*, 159–172.

Miller, R. (1981). *Survival Analysis*, Wiley, New York.

Park, J. (2003). Simplicial regression depth with censored and truncated data, *The Korean Communications in Statistics*, **10**, 167–175.

Park, J. and Hwang, J. (2003). Regression depth with censored and truncated data, *Communications in Statistics. Theory and Methods*, **32**, 997–1008.

Rousseeuw, P. J. and Hubert, M. (1999). Regression depth, *Journal of the American Statistical Association*, **94**, 388–402.

Tukey, J. W. (1975). Mathematics and the picturing of data, In *Proceedings of the International Congress of Mathematicians, Vancouver*, **2**, 523–531.

[ 2006년 11월 접수, 2007년 3월 채택 ]

# A Comparison Study of Survival Regression Models Based on Data Depths*

Jeeyun Kim[1] Jinsoo Hwang[2]

ABSTRACT

Several robust censored depth regression methods are compared under contamination. Park and Hwang(2003) suggested a way to circumvent the censoring issue by incorporating Kaplan-Meier type weight in halfspace regression depth and Park(2003) used a similar technique to simplicial regression depth. Hubert *et al.*(2001) suggested a high breakdown point regression depth based on projection called *rcent*. A new method to implement censoring in *rcent* is suggested and compared with two precedents under various contamination and censoring schemes.

*Keywords:* Censored, depth regression.

* This work was supported by a grant from KRF(2004-015-C00075).

1) Statistical Research Center for Complex Systems, Seoul National University, Seoul 151-742, Korea
E-mail: jeeyun@inha.ac.kr

2) Professor, Department of Statistics Inha University, 253 Yonghyun-Dong, Nam-Gu, 402-751, Incheon, Korea
E-mail: jshwang@inha.ac.kr