응용통계연구 (2008)
21(6), 981–996

로버스트 회귀모형을 이용한 자료결합방법

전명식[1] · 정지송[2] · 박혜진[3]

[1]고려대학교 통계학과, [2]미래에셋증권 마케팅지원본부, [3]고려대학교 통계학과

(2008년 8월 접수, 2008년 9월 채택)

요 약

서로 다른 출처로부터 얻어진 데이터 파일들을 하나의 데이터 파일로 만드는 통계적 자료결합방법은 공통변수와 서로 다른 고유변수를 포함하여 변수들 간에 존재하는 관련성에 대해 살펴볼 수 있다. Rubin (1986)이 제안한 일반회귀모형의 예측값을 이용한 통계적 결합방법은 자료에 대한 다변량 정규성을 가정하기 때문에 이 가정을 위반하는 자료를 이용하는 것은 많은 문제를 수반한다. 본 연구는 제공파일의 고유변수에 모분포를 반영하지 못하는 특이점이 존재하는 경우, 일반회귀모형을 이용한 통계적 결합방법의 대안으로 로버스트 회귀추정방법을 이용한 자료결합방법을 제안하였다. 나아가 로버스트 회귀모형을 이용한 결합방법과 일반회귀모형을 이용한 결합방법에서의 상관관계 및 결정계수 보존에 관한 성능을 비교하기 위하여 모의실험을 수행하였다.

주요용어: 통계적결합, 로버스트 회귀모형, 상관관계, 결정계수.

1. 서론

통계적 자료결합방법은 서로 다른 출처로부터 얻어진 데이터 파일들을 하나의 데이터 파일로 만드는데 목적을 두고 있다. 즉, 동일한 모집단에서 서로 다른 표본들로부터 관찰된 자료로 생성된 두 개 이상의 파일이 몇 개의 공통변수들을 가지며, 각 파일은 각각 독자적인 고유변수를 가진 경우에 이들 파일들의 자료를 결합하여 각 파일이 갖고 있는 변수들 간의 관계를 고찰하려고 할 때 시도하는 결합이다 (Rässler, 2002). 이렇게 결합된 파일은 공통변수와 서로 다른 고유변수를 포함하여 변수들 간에 존재하는 관련성에 대해 살펴볼 수 있을 것이다.

우선 자료 결합을 위해서 파일 A는 (Z_A, X)로 구성되어 있고, 파일 B는 (Z_B, Y)로 구성되어 있다고 하자. 여기서 파일 A와 파일 B에서 모두 관찰되는 변수 Z를 공통변수라고 하고 각각의 파일에서만 관찰되는 변수 X, Y를 고유변수라고 한다. 이때 파일 A를 수용파일이라고 하고 파일 B를 제공파일이라 하여 통계적 결합을 수행하면 수용파일에 없는 변수 Y가 제공파일로부터 추가로 얻어져 $(Z, X, \hat{Y})$로 구성된 새로운 데이터 파일을 만들게 되어 변수 Z, X, Y에 대하여 분석을 할 수 있게 될 것이다.

자료결합에서 얻어진 결합파일의 변수들 사이의 상관관계 보존에 관한 연구는 Kadane (1978), Rubin (1986), Moriarity와 Scheuren (2001) 그리고 Rässler (2004)에 의해 연구되었다. 그러나 이러한 연구는 다변량 정규성을 가정하기 때문에 특이점들이 존재하는 경우에 자료결합에 직접적으로 활용되기가

고려대학교 특별연구비에 의하여 수행되었음.
[1]교신저자: (136-701) 서울특별시 성북구 안암동 5-1, 고려대학교 통계학과, 교수. E-mail: jhun@korea.ac.kr
[2](150-878) 서울특별시 영등포구 여의도동 25-12 신송센터빌딩 7층 미래에셋증권 마케팅지원본부, 사원.
E-mail: wjdwlthd@miraeasset.com
[3](136-701) 서울특별시 성북구 안암동 5-1, 고려대학교 통계학과, 석사과정. E-mail: dabin220@korea.ac.kr

어려울 수 있다. 만약 서로 다른 두 파일이 모분포를 잘 반영하지 못하는 특이점을 포함한 경우, 자료결합과정 중 선형모형단계에서 최소제곱추정량을 사용한 예측값들이 특이점들로 인해 민감할 것이다. 따라서 특이점이 포함된 데이터의 통계적 결합상황에 최소제곱추정량에 근거한 예측값을 사용하는 것은 실제 자료의 상관관계를 왜곡할 수 있는 문제점을 가지고 있다. 본 연구에서는 최소제곱추정량의 강건하지 못한 성격을 보완하기 위해 로버스트 회귀추정방법을 이용한 자료결합방법을 제안하고 모의실험을 통해 그의 성질을 비교 설명하고자 한다.

2. 로버스트 회귀모형을 이용한 자료결합

파일 A는 $(Z_A,\ X)$로 파일 B는 $(Z_B,\ Y)$로 구성되어 있는 자료를 고려한다. 여기서 Z_A와 Z_B는 각각 n_A와 n_B개의 개체들에 대한 $n_A \times (k+1)$, $n_B \times (k+1)$ 자료행렬로서 첫 번째 열은 1로 그리고 나머지 열들은 공통변수 Z들로 이루어져 있다. 또한 고유변수에 대한 자료행렬 X와 Y는 각각 p개와 q개의 고유변수로 이루어진 $n_A \times p$, $n_B \times q$ 행렬로서 다변량정규분포를 따른다고 가정한다. 이때 각 파일의 고유변수들을 종속변수로 하고 공통변수 Z를 독립변수로 하는 선형모형은 (2.1)과 같이 나타낼 수 있다.

$$
\begin{aligned}
X &= Z_A\beta_{XZ_A} + U_A, \quad U_A \sim N(0,\ \Sigma_A^2 \otimes I_{n_A}),\\
Y &= Z_B\beta_{YZ_B} + U_B, \quad U_B \sim N(0,\ \Sigma_B^2 \otimes I_{n_B}),\\
\beta_{XZ_A} &= (\beta_{X0},\ \beta_{X1},\ldots,\beta_{Xp})', \qquad \beta_{YZ_B} = (\beta_{Y0},\ \beta_{Y1},\ldots,\beta_{Yq})',
\end{aligned} \tag{2.1}
$$

여기서 공통변수 $Z = z$라고 주어졌을 때 $(X,\ Y \mid Z = z) = (X_1,\ X_2,\ldots,X_p,\ Y_1,\ Y_2,\ldots,Y_q \mid Z = z)$의 X, Y의 조건부결합분포도 다변량정규성을 만족한다고 가정한다. 따라서 $E((X,\ Y) \mid Z = z) = \mu_{XY|Z}$이고 $\mathrm{Cov}((X,\ Y) \mid Z = z) = \Sigma_{XY|Z}$로 표기하면

$$
\boldsymbol{\Sigma_{XY|Z}} = \left(\begin{array}{ccc|ccc}
\sigma_{X_1X_1|Z} & \cdots & \sigma_{X_1X_p|Z} & \sigma_{X_1Y_1|Z} & \cdots & \sigma_{X_1Y_q|Z}\\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\
\sigma_{X_pX_1|Z} & \cdots & \sigma_{X_pX_p|Z} & \sigma_{X_pY_1|Z} & \cdots & \sigma_{X_pY_q|Z}\\
\hline
\sigma_{Y_1X_1|Z} & \cdots & \sigma_{Y_1X_p|Z} & \sigma_{Y_1Y_1|Z} & \cdots & \sigma_{Y_1Y_q|Z}\\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\
\sigma_{Y_qX_1|Z} & \cdots & \sigma_{Y_qX_p|Z} & \sigma_{Y_qY_1|Z} & \cdots & \sigma_{Y_qY_q|Z}
\end{array}\right) = \left(\begin{array}{c|c}
\Sigma_{11} & \Sigma_{12}\\
\hline
\Sigma_{21} & \Sigma_{22}
\end{array}\right) \tag{2.2}
$$

와 같이 표현된다.

이때 공통변수 Z_A와 고유변수 X를 사용한 β_{XZ_A}의 추정량 $\hat{\beta}_{XZ_A}$과 공통변수 Z_B와 고유변수 Y를 사용한 β_{YZ_B}의 추정량 $\hat{\beta}_{YZ_B}$를 통하여, 식 (2.3)과 같이 모수들을 추정한다 (Rässler, 2002).

$$
\begin{aligned}
&\text{(파일 A)} \quad \hat{\beta}_{YZ.X} = \hat{\beta}_{YZ_B} - \hat{\beta}_{XZ_A}\hat{\beta}_{YX.Z}, \qquad \text{단 } \hat{\beta}_{YX.Z} = \hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12}\\
&\text{(파일 B)} \quad \hat{\beta}_{XZ.Y} = \hat{\beta}_{XZ_A} - \hat{\beta}_{YZ_B}\hat{\beta}_{XY.Z}, \qquad \text{단 } \hat{\beta}_{XY.Z} = \hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{21}.
\end{aligned} \tag{2.3}
$$

이렇게 구해진 회귀모수들을 가지고, 결합단계에서 사용될 파일 A와 파일 B의 모든 변수 X와 Y들의 예측값을 식 (2.4)와 같이 구할 수 있다.

$$
\begin{aligned}
\text{(파일 A)} \quad \hat{Y} &= \hat{Y}_A = Z_A\hat{\beta}_{YZ.X} + X\hat{\beta}_{YX.Z}\\
\hat{X} &= \hat{X}_A = Z_A\hat{\beta}_{XZ.Y} + \hat{Y}_A\hat{\beta}_{XY.Z}\\
\text{(파일 B)} \quad \hat{X} &= \hat{X}_B = Z_B\hat{\beta}_{XZ.Y} + Y\hat{\beta}_{XY.Z}\\
\hat{Y} &= \hat{Y}_B = Z_B\hat{\beta}_{YZ.X} + \hat{X}_B\hat{\beta}_{YX.Z}.
\end{aligned} \tag{2.4}
$$

식 (2.4)에서 구한 파일 A와 파일 B의 예측값을 이용하여 개체간의 근사성 측정을 통해 파일 A에 존재하지 않는 변수 Y와 파일 B에 존재하지 않는 변수 X를 각각 파일 B와 파일 A로부터 제공받아 파일 A와 파일 B는 모든 변수 값을 갖는 '완전한' 파일이 된다. 개체 사이의 근사성 측정은 절대차이를 사용하여 식 (2.5), (2.6)과 같이 한다.

$*$ 파일 A의 $\hat{X}_{Ai}$와 파일 B의 $\hat{X}_{Bj}$의 근사성 측정

$$d_{ij}^{X} = |\hat{X}_{Bj} - \hat{X}_{Ai}|, \quad i = 1, 2, \ldots, n_A,\ j = 1, 2, \ldots, n_B. \tag{2.5}$$

$*$ 파일 A의 $\hat{Y}_{Ai}$와 파일 B의 $\hat{Y}_{Bj}$의 근사성 측정

$$d_{ij}^{Y} = |\hat{Y}_{Ai} - \hat{Y}_{Bj}|, \quad i = 1, 2, \ldots, n_A,\ j = 1, 2, \ldots, n_B. \tag{2.6}$$

d_{ij}^{X}는 파일 B의 j번째 개체의 예측값 $\hat{X}_{Bj}$와 파일 A의 i번째 개체의 예측값 $\hat{X}_{Ai}$의 차이를 구한 것으로서, 파일 B의 j번째 개체는 가장 작은 d_{ij}^{X}를 갖는 파일 A의 i번째 개체와 결합하게 된다. 또한 d_{ij}^{Y}는 파일 A의 i번째 개체의 예측값 $\hat{Y}_{Ai}$과 파일 B의 j번째 개체의 예측값 $\hat{Y}_{Bj}$의 차이를 구한 것으로서 파일 A의 i번째 개체는 가장 작은 d_{ij}^{Y}를 갖는 파일 B의 j번째 개체와 결합하게 된다. 이런 과정을 통해 파일 A와 B에 결합된 부분을 각각 $\tilde{Y}_{Bj}$, $\tilde{X}_{Ai}$로 표기하면 결합된 파일 A와 파일 B는 식 (2.7)과 같이 나타낼 수 있다.

$$\begin{aligned} &(\text{파일 A}) : (Z_A, X_{Ai}, \tilde{Y}_{Bj}), \quad i = 1, 2, \ldots, n_A,\ j = 1, 2, \ldots, n_B. \\ &(\text{파일 B}) : (Z_B, \tilde{X}_{Ai}, Y_{Bj}), \quad i = 1, 2, \ldots, n_A,\ j = 1, 2, \ldots, n_B. \end{aligned} \tag{2.7}$$

그런데 제공파일에 특이값이 존재하는 경우, 일반회귀모형에서 파일 A의 예측값 $\hat{Y}_A$와 파일 B의 예측값 $\hat{Y}_B$을 이용하여 자료결합을 수행하면, 특이점이 수용파일에 결합될 수 있기 때문에 실제자료의 특성을 왜곡하게 된다. 왜냐하면, 제공파일에 특이점이 존재하는 경우, 선형회귀모형의 적용은 공통변수의 정보를 왜곡하는 예측값을 산출하여 잘못된 결합을 초래할 수 있기 때문이다. 따라서, 본 연구에서는 일반회귀모형을 이용한 자료결합방법에서 β_{XZ_A}와 β_{YZ_B}의 최소제곱추정량을 사용하는 것과 달리 식 (2.8), (2.9)와 같이 잔차제곱의 절사합을 최소화하는 최소절사제곱추정량을 활용하고자 한다.

$$\hat{\beta}_{XZ_A} = \min_{\beta_{XZ_A}} \arg\, Q(\beta_{XZ_A}), \quad \text{단, } Q(\beta_{XZ_A}) = \sum_{i=1}^{h} \left(r^2\right)_{i:n},\ r = x - z_A\prime\hat{\beta}_{XZ_A}, \tag{2.8}$$

$$\hat{\beta}_{YZ_B} = \min_{\beta_{YZ_B}} \arg\, Q(\beta_{YZ_B}), \quad \text{단, } Q(\beta_{YZ_B}) = \sum_{i=1}^{h} \left(r^2\right)_{i:n},\ r = y - z_B\prime\hat{\beta}_{YZ_B}, \tag{2.9}$$

$\left((r^2)_{1:n} \le (r^2)_{2:n} \le \cdots \le \left(r^2\right)_{n:n}\right)$: 제곱된 잔차의 순서 통계량,

단, $i = 1, 2, \ldots, h$, $h = (n + p + 1)/2$ (p: 설명변수의 개수).

3. 통계적 결합에서의 상관관계 및 결정계수 보존

특이점을 포함한 사례에 대한 결합을 고려하여 일반회귀모형을 이용한 자료결합방법을 적용해보고 결합파일에서 나타나는 문제점을 본 논문에서 제안하는 로버스트 회귀모형을 이용한 자료결합방법과 비교하고자 한다.

3.1. 일반회귀모형을 이용한 자료결합

결합하고자하는 두 파일이 다변량 정규분포를 따르는 경우, 일반회귀모형을 이용한 자료결합방법으로 얻어진 결합자료는 결합 전 자료에서 변수들의 상관관계를 잘 유지하고 모형에 대한 결정계수를 잘 보

표 3.1. 모분포 (3.1)로 부터 생성된 파일 A와 파일 B

〈파일 A〉			〈파일 B〉		
Z_A	X	Y	Z_B	X	Y
3.05	2.97	2.91	2.26	3.11	2.55
3.44	4.09	3.85	3.18	3.19	2.98
2.67	2.97	3.09	2.70	2.68	3.00
5.71	5.42	5.40	3.47	3.69	3.48
3.36	3.13	3.49	3.58	2.94	3.58
5.47	5.00	5.78	3.92	4.21	4.17
3.20	3.60	3.31	4.79	4.26	4.21
4.21	4.36	4.31	3.76	3.83	3.67
4.02	4.56	3.73	4.24	3.79	4.36
5.23	4.91	4.46	3.66	3.73	3.80
4.37	4.64	4.28	4.03	4.00	4.34
3.91	4.76	3.43	5.21	4.67	5.22
5.02	4.06	4.96	3.63	3.44	4.12
4.89	4.71	4.41	6.16	6.02	6.10
3.38	3.68	3.39	2.49	2.21	2.00

존하였다. 그렇다면 두 파일의 자료결합에 있어서 제공파일 B의 고유변수 Y에 특이점이 존재하는 경우에 대하여 일반회귀모형을 이용한 통계적 결합을 고려해보자. 결합방법의 과정을 간단히 설명하기 위해 (Z, X, Y)는 (3.1)과 같은 모분포를 따른다고 가정한다.

$$(Z, X, Y) \sim N_3(\underset{\sim}{\mu}, \Sigma_{ZXY}), \quad \text{여기서, } \underset{\sim}{\mu} = \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix}, \quad \Sigma_{ZXY} = \begin{pmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.8 \\ 0.9 & 0.8 & 1 \end{pmatrix}. \tag{3.1}$$

표 3.1은 자료결합의 예를 설명하기 위해 인공적으로 만든 자료이다. 통계적 결합을 수행하기 위해 파일 A에 변수 Y를 파일 B에는 변수 X를 제거하여 표 3.2와 같이 구성한다. 이때 제공파일 B의 고유변수 Y에 모분포를 반영하지 못하는 특이점으로 고려하기 위해 Z_B=3.63, 6.16, 2.49의 개체에 해당되는 $Y_{13} = 4.12$, $Y_{14} = 6.10$, $Y_{15} = 2.00$을 제거하고 그 개체 값에 각각 13.89, 16.09, 12.07을 부여한다. 또한 파일 A를 수용파일이라 하고, 파일 B를 제공파일이라 하여 수용파일의 존재하지 않는 변수 Y를 제공파일로부터 제공받는 상황을 고려한다. 그리고 결합방법에 있어서는 제약이 없는 결합을 실시한다. 제약이 있는 결합은 상관관계를 보존하지 못할 뿐만 아니라 결합하고자 하는 자료에 특이점이 존재하는 경우에 개체들이 한번 씩 모두 결합해야 한다는 제약이 있으므로 특이점이 결합 자료에 항상 포함되는 문제점이 있다. 따라서 특이점이 존재하는 자료의 결합에서 제약이 있는 결합방식은 좋은 방법이 될 수 없다.

표 3.2와 같이 제공파일 B에 특이점이 존재하는 경우, 일반회귀모형을 이용한 자료결합 이후 얻어지는 결합자료가 어떠한 문제점을 갖는지를 살펴보도록 하자. 제공파일로 고려한 파일 B에 존재하는 Z_B와 Y를 이용하여 일반회귀모형을 적용하고 추정된 $\hat{\beta}_{YZ_B} = (-0.025, 1.538)$에 의해 각 파일의 예측값을 다음과 같은 식 (3.2)로 구한다. 또한 2절에서 소개한 자료결합 알고리즘에서 이용할 수 있는 사전정보가 없는 경우에는 $R_{XY|Z}= 0_{p\times q}$을 선택하여 구한다. 따라서 식 (2.4)는 다음과 같은 식 (3.2)가 됨을 보

표 3.2. 통계적 결합에 이용될 수용파일 A와 제공파일 B

〈수용파일 A〉

Z_A	X	Y
3.05	2.97	
3.44	4.09	
2.67	2.97	
5.71	5.42	
3.36	3.13	
5.47	5.00	
3.20	3.60	
4.21	4.36	
4.02	4.56	
5.23	4.91	
4.37	4.64	
3.91	4.76	
5.02	4.06	
4.89	4.71	
3.38	3.68	

〈제공파일 B〉

Z_B	X	Y
2.26		2.55
3.18		2.98
2.70		3.00
3.47		3.48
3.58		3.58
3.92		4.17
4.79		4.21
3.76		3.67
4.24		4.36
3.66		3.80
4.03		4.34
5.21		5.22
3.63		**13.89**
6.16		**16.09**
2.49		**12.07**

표 3.3. $\hat{\beta}_{YZ_B}$에 의해 구해진 각 파일의 예측값

〈수용파일 A〉

Z_A	X	$\hat{Y}_A$
3.05	2.97	4.66
3.44	4.09	5.26
2.67	2.97	4.08
5.71	5.42	8.76
3.36	3.13	5.14
5.47	5.00	8.39
3.20	3.60	4.90
4.21	4.36	6.45
4.02	4.56	6.16
5.23	4.91	8.02
4.37	4.64	6.69
3.91	4.76	5.99
5.02	4.06	7.69
4.89	4.71	7.49
3.38	3.68	5.17

〈제공파일 B〉

Z_B	Y	$\hat{Y}_B$
2.26	2.55	3.45
3.18	2.98	4.86
2.70	3.00	4.13
3.47	3.48	5.31
3.58	3.58	5.48
3.92	4.17	6.00
4.79	4.21	7.34
3.76	3.67	5.76
4.24	4.36	6.50
3.66	3.80	5.60
4.03	4.34	6.17
5.21	5.22	7.99
3.63	13.89	5.56
6.16	16.09	9.45
2.49	12.07	3.80

일 수 있다.

$$\begin{aligned} &\text{(제공파일 B)의 예측값 : } \hat{Y}_B = Z_B\hat{\beta}_{YZ_B}. \\ &\text{(수용파일 A)의 예측값 : } \hat{Y}_A = Z_A\hat{\beta}_{YZ_B}. \end{aligned} \tag{3.2}$$

이때 구해진 예측값은 표 3.3과 같이 정리 할 수 있다.

표 3.3의 수용파일 A와 제공파일 B에서 $\hat{Y}_A$와 $\hat{Y}_B$의 근사성 측정을 통해 수용파일 A의 개체와 가장 가까운 제공파일 B의 개체값 Y_j, $j = 1, 2, \ldots, 15$를 수용파일 A에 결합한다. 수용파일 A의 i번째 개체와

표 3.4. $\hat{Y}_{Ai}$와 $\hat{Y}_{Bj}$의 절대차이

$\hat{Y}_{Ai}$ \ $\hat{Y}_{Bj}$	3.45 (1)	4.86 (2)	4.13 (3)	5.31 (4)	5.48 (5)	6.00 (6)	7.34 (7)	5.76 (8)	6.50 (9)	5.60 (10)	6.17 (11)	7.99 (12)	5.56 (13)	9.45 (14)	3.80 (15)
4.66	1.21	**0.20**	0.53	0.65	0.82	1.34	2.68	1.10	1.84	0.94	1.51	3.33	0.9	4.79	0.86
5.26	1.81	0.40	1.13	**0.05**	0.22	0.74	2.08	0.50	1.24	0.34	0.91	2.73	0.30	4.19	1.46
4.08	0.63	0.78	**0.05**	1.23	1.40	1.92	3.26	1.68	2.41	1.52	2.09	3.91	1.48	5.37	0.28
8.76	5.31	3.89	4.63	3.45	3.28	2.76	1.42	3.00	2.26	3.16	2.59	0.77	3.20	**0.69**	4.96
5.14	1.69	0.28	1.01	**0.17**	0.34	0.86	2.20	0.61	1.36	0.46	1.03	2.85	0.42	4.31	1.34
8.39	4.94	3.53	4.26	3.08	2.91	2.39	1.05	2.63	1.89	2.79	2.22	**0.40**	2.83	1.06	4.59
4.90	1.45	**0.04**	0.77	0.41	0.58	1.10	2.44	0.86	1.60	0.70	1.27	3.09	0.66	4.55	1.10
6.45	3.00	1.59	2.32	1.14	0.97	0.45	0.89	0.69	**0.05**	0.85	0.28	1.54	0.89	3.00	2.65
6.16	2.71	1.30	2.03	0.85	0.68	0.16	1.18	0.40	0.34	0.56	**0.01**	1.83	0.60	3.29	2.35
8.02	4.57	3.16	3.89	2.71	2.54	2.02	0.68	2.26	1.52	2.41	1.85	**0.03**	2.46	1.43	4.22
6.69	3.24	1.83	2.56	1.38	1.21	0.69	0.65	0.93	**0.19**	1.09	0.52	1.30	1.13	2.76	2.89
5.99	2.54	1.13	1.86	0.68	0.51	**0.01**	1.35	0.23	0.51	0.39	0.18	2.00	0.43	3.46	2.19
7.69	4.24	2.83	3.56	2.38	2.21	1.69	0.35	1.93	1.19	2.09	1.52	**0.30**	2.13	1.76	3.89
7.49	4.04	2.63	3.36	2.18	2.01	1.49	**0.15**	1.73	0.99	1.89	1.32	0.50	1.93	1.96	3.69
5.17	1.72	0.31	1.04	**0.14**	0.31	0.83	2.17	0.59	1.33	0.43	1.00	2.82	0.39	4.28	1.37

표 3.5. 결합결과

Z_A	X	$\tilde{Y}$
3.05	2.97	2.98
3.44	4.09	3.48
2.67	2.97	3.00
5.71	5.42	**16.09**
3.36	3.13	3.48
5.47	5.00	5.22
3.20	3.60	2.98
4.21	4.36	4.36
4.02	4.56	4.34
5.23	4.91	5.22
4.37	4.64	4.36
3.91	4.76	4.17
5.02	4.06	5.22
4.89	4.71	4.21
3.38	3.68	3.48

제공파일 B의 j번째 개체의 근사성을 측정하기 위한 절대차이 $d_{ij}^{Y} = |\hat{Y}_{Ai} - \hat{Y}_{Bj}|$, $i, j = 1, 2 \ldots, 15$를 정리하면 표 3.4와 같다.

표 3.4를 보면 수용파일 A의 예측값 $\hat{Y}_{Ai}$, $i = 1, 2 \ldots, 15$와 제공파일 B의 예측값 $\hat{Y}_{Bj}$, $j = 1, 2, \ldots, 15$ 간의 절대차이를 구하여 차이가 가장 작은 개체를 선택함을 알 수 있다. 예를 들어 수용파일 A의 첫 번째 개체의 예측값 $\hat{Y}_{A1} = 4.66$은 제공파일 B의 예측값 $\hat{Y}_{Bj}$, $j = 1, 2, \ldots, 15$와의 절대차이를 구한 결과 제공파일 B의 두 번째 개체의 예측값 $\hat{Y}_{B2} = 4.86$이 $d_{12}^{Y} = |\hat{Y}_{A1} - \hat{Y}_{B2}| = 0.2$로 차이가 가장 작음을 알 수 있다. 따라서 수용파일 A의 첫 번째 개체는 제공파일 B의 두 번째 개체의 변수 Y값 2.98를 제공받게 된다. 이렇게 결합된 결과는 표 3.5와 같다.

표 3.6. 결합 전과 후의 자료로부터 모형을 적합한 결정계수

		모형에 포함되는 변수	모형	결정계수
결합 전		(Z_A, X, Y)	$Z_A = \beta_0 + \beta_1 X + \beta_2 Y$	0.93
		(Z_A, X)	$Z_A = \beta_0 + \beta_1 X$	0.75
결합 후	일반회귀	$(Z_A, X, \tilde{Y})$	$Z_A = \beta_0 + \beta_1 X + \beta_2 \tilde{Y}$	0.78

표 3.5의 결합된 변수 $\tilde{Y}$의 값을 보면 수용파일 A의 각 개체에 대하여 제공파일 B의 가장 근사한 개체가 결합되었다는 것을 볼 수 있으나, 제공파일 B에 존재하는 특이점이 결합되었음을 알 수 있다. 따라서 표 3.5의 결합결과는 결합 전 자료의 구조가 특이점으로 인해 보존되지 못할 것으로 판단된다. 실제로, 결합자료의 Z_A, X, $\tilde{Y}$의 상관계수의 추정값을 구해보면, $\rho_{Z\tilde{Y}} = 0.652$이고 $\rho_{X\tilde{Y}} = 0.605$이다. 따라서 모상관계수 $\rho_{ZY} = 0.9$, $\rho_{XY} = 0.8$과 비교해보더라도 결합자료가 결합 전 자료의 상관관계를 충분히 보존하지 못한다고 할 수 있다. 특히, 결합자료가 특이점을 포함하기 때문에 고유변수 X와 Y의 상관관계에 있어서 왜곡이 일어났음을 알 수 있다. 또한 결합자료에서의 모형에 대한 결정계수는 결합 전 자료에서의 모형에 대한 결정계수를 잘 보존하지 못할 것이라 판단된다. 이때 공통변수 Z_A를 반응변수로 하는 선형회귀모형에 대한 설명변수의 설명력을 결정계수로 표현할 수 있다. 제공파일 B의 고유변수 Y에 특이점이 존재하는 경우, 결합 전과 후의 자료로부터 일반회귀모형을 적합한 결정계수 결과 비교는 표 3.6과 같다.

표 3.6을 보면 통계적 결합을 수행한 후에 얻어진 결합파일로부터 모형 $Z_A = \beta_0 + \beta_1 X + \beta_2 \tilde{Y}$에 대한 결정계수(0.78)이 결합 전 자료에서 모형 $Z_A = \beta_0 + \beta_1 X + \beta_2 Y$에 대한 결정계수(0.93)을 충분히 보존 하지 못한다. 또한 결합 전 통계적 결합을 하기 위해 변수 Y를 제거한 자료 (Z_A, X)에서의 결정계수(0.75)가 변수 Y의 정보를 추가적으로 얻은 결합파일 $(Z_A, X, \tilde{Y})$에서의 결정계수(0.78)과 별 차이가 없다는 것을 알 수 있다. 따라서 결합자료에 특이점이 포함되기 때문에 변수 Z의 예측에 대한 변수 Y의 결합정보는 왜곡될 수 있다는 것을 알 수 있다.

3.2. 로버스트 회귀모형을 이용한 자료결합

앞서 제안된 로버스트 회귀모형을 이용한 자료결합방법을 적용하기 위해 표 3.2의 자료를 그대로 사용한다. 또한 앞서 살펴본 일반회귀모형을 이용한 자료결합방법과 마찬가지로 제약이 없는 경우에 대하여 실시한다.

최소절사제곱방법으로 추정된 $\hat{\hat{\beta}}_{YZ_B} = (0.534, 0.870)$에 의해 수용파일 A의 예측값을 식 (3.3)과 같이 구한다. 이때 제공파일 B의 고유변수 Y에 3개$(n_B \times p = 15 \times 0.2)$의 특이점이 포함되어 있으므로 최소절사제곱추정량은 절사량 $h = 12$을 선택하여 구한다. 또한 앞서 본 바와 같이 자료결합 알고리즘에서 이용할 수 있는 사전정보가 없는 경우에는 $R_{XY|Z} = 0_{p\times q}$을 선택하여 구하므로 식 (2.4)는 다음과 같은 식 (3.3)이 됨을 보일 수 있다.

$$\begin{aligned} &\text{(제공파일 B)의 예측값 : } \hat{\hat{Y}}_B = Z_B \hat{\hat{\beta}}_{YZ_B}, \\ &\text{(수용파일 A)의 예측값 : } \hat{\hat{Y}}_A = Z_A \hat{\hat{\beta}}_{YZ_B}. \end{aligned} \tag{3.3}$$

이때 구해진 예측값은 표 3.7과 같이 정리 할 수 있다. 표 3.7에서 구해진 수용파일 A의 예측값 $\hat{\hat{Y}}_A$과 제공파일 B의 예측값 $\hat{\hat{Y}}_B$으로 근사성을 측정할 때 선형모형단계에서 로버스트 회귀추정방법으로 사용한 최소절사제곱의 절사량 h를 활용해보도록 한다. 즉, 잔차제곱의 절단 합을 최소화하는 최소절사제곱방

표 **3.7.** $\hat{\beta}_{YZ_B}$에 의해 구해진 각 파일의 예측값

⟨수용파일 A⟩

Z_A	X	$\hat{Y}_A$
3.05	2.97	3.19
3.44	4.09	3.52
2.67	2.97	2.86
5.71	5.42	5.50
3.36	3.13	3.46
5.47	5.00	5.29
3.20	3.60	3.32
4.21	4.36	4.19
4.02	4.56	4.03
5.23	4.91	5.08
4.37	4.64	4.33
3.91	4.76	3.93
5.02	4.06	4.90
4.89	4.71	4.79
3.38	3.68	3.47

⟨제공파일 B⟩

Z_B	Y	$\hat{Y}_B$
2.26	2.55	2.50
3.18	2.98	3.30
2.70	3.00	2.88
3.47	3.48	3.55
3.58	3.58	3.65
3.92	4.17	3.94
4.79	4.21	4.70
3.76	3.67	3.80
4.24	4.36	4.22
3.66	3.80	3.72
4.03	4.34	4.04
5.21	5.22	5.06
3.63	13.89	3.69
6.16	16.09	5.89
2.49	12.07	2.70

표 **3.8.** 잔차 $Y - \hat{Y}_B$를 순서화된 잔차 $Y - \hat{Y}_B$

개체 i	Y	$\hat{Y}_B$	$Y - \hat{Y}_B$
1	2.55	2.50	0.05
2	2.98	3.30	−0.32
3	3.00	2.88	0.12
4	3.48	3.55	−0.07
5	3.58	3.65	−0.07
6	4.17	3.94	0.23
7	4.21	4.70	−0.49
8	3.67	3.80	−0.13
9	4.36	4.22	0.14
10	3.80	3.72	0.08
11	4.34	4.04	0.30
12	5.22	5.06	0.16
13	13.89	3.69	10.20
14	16.09	5.89	10.20
15	12.07	2.70	9.37

개체 i	Y	$\hat{Y}_B$	$Y - \hat{Y}_B$
7	4.21	4.70	−0.49
2	2.98	3.30	−0.32
8	3.67	3.80	−0.13
4	3.48	3.55	−0.07
5	3.58	3.65	−0.07
1	2.55	2.50	0.05
10	3.80	3.72	0.08
3	3.00	2.88	0.12
9	4.36	4.22	0.14
12	5.22	5.06	0.16
6	4.17	3.94	0.23
11	4.34	4.04	0.30
15	**12.07**	**2.70**	**9.37**
14	**16.09**	**5.89**	**10.20**
13	**13.89**	**3.69**	**10.20**

법을 고려하여 잔차 $Y - \hat{Y}_B$를 순서대로 정렬한 후 절사량(h) 12번째까지 해당 개체의 $\hat{Y}_B$을 근사성 측정에 이용하고 나머지 개체는 절사한다. 잔차 $Y - \hat{Y}_B$를 순서화한 후 잔차 $Y - \hat{Y}_B$는 표 3.8과 같다.

표 3.8를 보면 13, 14, 15번째 개체가 절사되었다는 것을 알 수 있다. 따라서 해당 개체의 변수 Y의 실제값 $Y_{13} = 13.89$, $Y_{14} = 16.09$, $Y_{15} = 12.07$과 예측값 $\hat{Y}_{13} = 3.69$, $\hat{Y}_{14} = 5.89$, $\hat{Y}_{15} = 2.70$이 절사되어 근사성 측정에 이용되지 않을 것이다. 수용파일 A의 예측값 $\hat{Y}_A$과 절사 후 제공파일 B의 예측값 $\hat{Y}_B$으로 근사성측정하기 위한 절대차이 $d_{ij}^Y = |\hat{Y}_{Ai} - \hat{Y}_{Bj}|$, $i = 1, 2, \ldots, 15$, $j = 1, 2, \ldots, 12$를 정리하면 표 3.9와 같다.

표 3.9. $\hat{Y}_{Ai}$와 $\hat{Y}_{Bj}$의 절대차이

$\hat{Y}_{Ai}$ \ $\hat{Y}_{Bj}$	4.70 (1)	3.30 (2)	3.80 (3)	3.55 (4)	3.65 (5)	2.50 (6)	3.72 (7)	2.88 (8)	4.22 (9)	5.06 (10)	3.94 (11)	4.04 (12)
3.19	1.51	**0.11**	0.62	0.37	0.46	0.69	0.53	0.30	1.03	1.88	0.76	0.85
3.52	1.17	0.23	0.28	**0.03**	0.12	1.03	0.19	0.64	0.70	1.54	0.42	0.51
2.86	1.84	0.44	0.95	0.70	0.79	0.36	0.86	**0.03**	1.37	2.21	1.09	1.18
5.50	0.80	2.20	1.70	1.95	1.85	3.00	1.78	2.62	1.28	**0.43**	1.56	1.46
3.46	1.24	0.16	0.35	**0.10**	0.19	0.96	0.26	0.57	0.77	1.61	0.49	0.58
5.29	0.59	1.99	1.49	1.74	1.64	2.79	1.57	2.41	1.07	**0.23**	1.35	1.25
3.32	1.38	**0.02**	0.49	0.23	0.33	0.82	0.40	0.43	0.90	1.75	0.63	0.72
4.19	0.50	0.90	0.39	0.64	0.55	1.70	0.48	1.31	**0.03**	0.87	0.25	0.16
4.03	0.67	0.73	0.23	0.48	0.38	1.53	0.31	1.15	0.19	1.03	0.09	**0.01**
5.08	0.38	1.78	1.28	1.53	1.43	2.58	1.37	2.20	0.86	**0.02**	1.14	1.04
4.33	0.37	1.03	0.53	0.78	0.69	1.83	0.62	1.45	**0.11**	0.73	0.39	0.30
3.93	0.77	0.63	0.13	0.38	0.29	1.43	0.22	1.05	0.29	1.13	**0.01**	0.10
4.90	0.20	1.60	1.10	1.35	1.25	2.40	1.18	2.02	0.68	**0.17**	0.96	0.86
4.79	**0.09**	1.49	0.98	1.23	1.14	2.29	1.07	1.90	0.57	0.28	0.84	0.75
3.47	1.23	0.17	0.33	**0.08**	0.17	0.97	0.24	0.59	0.75	1.59	0.47	0.57

표 3.10. 결합결과

Z_A	X	$\tilde{Y}$
3.05	2.97	2.98
3.44	4.09	3.48
2.67	2.97	3.00
5.71	5.42	5.22
3.36	3.13	3.48
5.47	5.00	5.22
3.20	3.60	2.98
4.21	4.36	4.36
4.02	4.56	4.34
5.23	4.91	5.22
4.37	4.64	4.36
3.91	4.76	4.17
5.02	4.06	5.22
4.89	4.71	4.21
3.38	3.68	3.48

표 3.9를 보면 수용파일 A의 예측값 $\hat{\hat{Y}}_{Ai}$, $i = 1, 2, \ldots, 15$와 제공파일 B의 예측값 $\hat{\hat{Y}}_{Bj}$, $j = 1, 2, \ldots, 12$ 간의 절대차이를 구하여 차이가 가장 작은 개체를 선택함을 알 수 있다. 예를 들어 수용파일 A의 첫 번째 개체의 예측값 $\hat{\hat{Y}}_{A1} = 3.19$는 제공파일 B의 예측값 $\hat{\hat{Y}}_{Bj}$, $j = 1, 2, \ldots, 12$와의 절대차이를 구한 결과 제공파일 B의 두 번째 개체의 예측값 $\hat{\hat{Y}}_{B2} = 3.30$이 $d_{12}^{Y} = |\hat{\hat{Y}}_{A1} - \hat{\hat{Y}}_{B2}| = 0.11$로 차이가 가장 작음을 알 수 있다. 따라서 수용파일 A의 첫 번째 개체는 제공파일 B의 두 번째 개체 변수 Y값 2.98를 제공받게 된다. 이렇게 결합된 결과는 표 3.10과 같다.

표 3.10의 결합된 변수 $\tilde{Y}$의 값을 보면 수용파일 A의 각 개체에 대하여 제공파일 B의 가장 근사한 개체

표 **3.11.** 결합자료의 변수 Z, X, Y의 상관관계

	Corr(Z, X)	Corr(Z, Y)	Corr(X, Y)
모분포	0.900	0.900	0.800
일반회귀모형	0.866	0.652	0.605
로버스트 회귀모형	0.866	0.954	0.836

표 **3.12.** 결합 전과 후의 자료로부터 모형을 적합한 결정계수

		모형에 포함되는 변수	모형	결정계수
결합 전		(Z_A, X, Y)	$Z_A = \beta_0 + \beta_1 X + \beta_2 Y$	0.93
		(Z_A, X)	$Z_A = \beta_0 + \beta_1 X$	0.75
결합 후	일반회귀	(Z_A, X, $\tilde{Y}$)	$Z_A = \beta_0 + \beta_1 X + \beta_2 \tilde{Y}$	0.78
	로버스트 회귀	(Z_A, X, $\tilde{Y}$)	$Z_A = \beta_0 + \beta_1 X + \beta_2 \tilde{Y}$	0.93

가 결합했다는 것을 볼 수 있다. 그러나 일반회귀모형을 이용한 자료결합에서와 달리 결합된 변수 $\tilde{Y}$에 특이점이 포함되지 않았다는 것을 알 수 있다. 그러므로 결합파일은 결합 전 자료의 구조를 충분히 잘 보존할 것으로 판단된다. 결합자료의 Z, X, $\tilde{Y}$의 상관계수의 추정값을 구해보면, $\rho_{Z\tilde{Y}} = 0.954$이고 $\rho_{X\tilde{Y}} = 0.836$이다. 따라서 모상관계수 $\rho_{ZY} = 0.9$, $\rho_{XY} = 0.8$과 비교해보더라도 결합자료에서 변수들의 상관관계는 결합 전 자료에서 변수들의 상관관계를 잘 유지한다고 할 수 있다. 제공파일 B의 고유변수 Y에 특이점이 존재하는 경우, 일반회귀모형과 로버스트 회귀모형을 이용해 만들어진 결합자료의 변수 Z, X, Y의 상관관계는 표 3.11과 같다.

표 3.11의 결과를 보면 제공파일 B의 고유변수 Y에 특이점이 존재하는 경우, 본 연구에서 제안하는 로버스트 회귀모형을 이용한 자료결합방법이 일반회귀모형을 이용한 자료결합방법에 비해 결합 전 자료에서 변수들의 상관관계를 보다 잘 유지한다는 것을 알 수 있었다. 다음으로, 공통변수 Z를 반응변수로 하는 선형회귀모형에서 제공파일 B로부터 변수 Y가 결합되었을 때 모형의 설명력을 나타내는 결정계수의 변화에 대해 살펴보자. 제공파일 B의 고유변수 Y에 특이점이 존재하는 경우, 결합 전과 후의 자료로부터 일반회귀모형을 적합 시켰을 때 결정계수 결과 비교는 표 3.12와 같다.

표 3.12를 보면 로버스트 회귀모형을 이용한 자료결합을 수행한 후에 얻어진 결합파일로부터 모형 $Z_A = \beta_0 + \beta_1 X + \beta_2 \tilde{Y}$에 대한 결정계수(0.93)이 결합 전 자료에서 모형 $Z_A = \beta_0 + \beta_1 X + \beta_2 Y$에 대한 결정계수(0.93)을 잘 보존한다. 또한 결합 전 통계적 결합을 하기 위해 변수 Y를 제거한 자료 (Z_A, X)에서의 결정계수(0.75)보다 변수 Y의 정보를 추가적으로 얻은 결합파일(Z_A, X, $\tilde{Y}$)에서의 결정계수(0.93)이 증가했으므로 변수 Z_A에 대한 설명력이 향상 되었음을 알 수 있다. 따라서 로버스트 회귀모형을 이용해 통계적 결합을 한 후에 얻어진 결합자료가 일반회귀모형을 이용해 얻어진 결합자료보다 결합 전의 자료를 잘 유지하고 모형에 대한 설명력을 더 향상 시킬 수 있다는 것을 알 수 있다.

4. 모의실험

본 연구에서 제안한 로버스트 회귀모형을 이용한 결합방법과 일반회귀모형을 이용한 결합방법에서의 상관관계 및 결정계수 보존에 관한 성능을 쉽게 비교하기 위하여 모의실험을 수행하였다.

4.1. 두 자료결합방법에서의 고유변수간의 상관계수 보존

파일 A와 파일 B가 다변량 정규분포를 따르는 경우와 특이점을 가진 오염된 자료일 경우에 대하여 일

표 4.1. $D_{XY} = |\rho_{XY} - \rho_{X\tilde{Y}}|$의 평균

	$n_A = n_B = 20$	$n_A = n_B = 50$	$n_A = n_B = 100$
일반회귀모형	0.075	0.040	0.027
로버스트 회귀모형	0.065	0.037	0.038

반회귀모형을 이용한 자료결합방법과 로버스트 회귀모형을 이용한 자료결합을 각각 적용하고, 각 방법에 의해 생성된 결합자료에서 변수 $Z = (Z_1, Z_2, Z_3)$, X, Y사이의 상관관계의 보존정도를 비교실험 하였다. 상관관계 보존의 비교를 위해, 식 (4.1)를 사용하여 구한 원자료와 결합자료의 상관계수 차이의 평균을 100회 독립반복시행에 근거하여 구했다.

$$D_{XY} = \left|\mathrm{Corr}(X,Y) - \widehat{\mathrm{Corr}}(X,\tilde{Y})\right|. \tag{4.1}$$

• **다변량 정규분포를 따르는 자료의 경우** 파일 A와 파일 B에 공통적으로 존재하는 공통변수 Z_i, $i = 1, 2, 3$은 식 (4.2)와 같은 분포로부터 임의 생성한다.

$$(Z_1, Z_2, Z_3) \sim N_3(\underline{0}, \Sigma_{Z_1Z_2Z_3}), \quad \text{단 , } \underline{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \Sigma_{Z_1Z_2Z_3} = \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{pmatrix}. \tag{4.2}$$

다음으로 각각의 파일 A와 파일 B에 존재하는 변수 X, Y를 생성하기 위해 선형회귀모형에서의 오차 $\epsilon_1 \sim N(0,1)$과 $\epsilon_2 \sim N(0,1)$를 고려한다. 파일의 구성은 (4.3)과 같으며 표본크기 n_A, n_B가 각각 20, 50, 100인 경우를 고려하였다.

$$\begin{aligned} &\text{파일 A : } \{Z_{A1}, Z_{A2}, Z_{A3}, X_A, Y_A\}, \\ &\quad \text{단 , } X_A = Z_{A1} + Z_{A2} + Z_{A3} + \epsilon_1, \quad Y_A = Z_{A1} + Z_{A2} + Z_{A3} + \epsilon_2. \\ &\text{파일 B : } \{Z_{B1}, Z_{B2}, Z_{B3}, X_B, Y_B\}, \\ &\quad \text{단 , } X_B = Z_{B1} + Z_{B2} + Z_{B3} + \epsilon_1, \quad Y_B = Z_{B1} + Z_{B2} + Z_{B3} + \epsilon_2. \end{aligned} \tag{4.3}$$

다음으로는 파일 A에 존재하는 변수 Y_A를 제거하고, 파일 B에 존재하는 변수 X_B를 제거하여 파일 A를 수용파일로 파일 B를 제공파일로 사용하여 제약이 없는 통계적 결합을 수행한다.

표 4.1의 결과는 공통변수 Z_i, $i = 1, 2, 3$이 다변량 자료일 경우에 대하여 일반회귀모형을 이용한 자료결합과 로버스트 회귀모형을 이용한 자료결합을 100회 독립반복 수행한 결과이다.

표 4.1에서 보는 바와 같이 일반회귀모형을 이용한 자료결합방법과 로버스트 회귀모형을 이용한 자료결합방법 모두 모분포의 고유변수 X, Y에 대한 상관관계를 결합자료에서 잘 유지하게 한다는 것을 알 수 있다.

• **제공파일 B의 고유변수에 특이점이 존재하는 경우** 일반적으로 공통변수가 일변량일 경우 특이점 검색은 비교적 쉬울 수 있다. 그러나 공통변수가 다변량 자료일 경우 차원이 높아질수록 특이점 검색은 점점 더 어려워진다. 이제, 고유변수에 특이점이 존재하는 경우, 로버스트 회귀모형을 이용한 자료결합

표 4.2. $D_{XY} = |\rho_{XY} - \rho_{X\hat{Y}}|$의 평균($p = 0.1$)

	$n_A = n_B = 20$	$n_A = n_B = 50$	$n_A = n_B = 100$
일반회귀모형	0.483	0.477	0.577
로버스트 회귀모형	0.076	0.037	0.034

표 4.3. $D_{XY} = |\rho_{XY} - \rho_{X\hat{Y}}|$의 평균($p = 0.2$)

	$n_A = n_B = 20$	$n_A = n_B = 50$	$n_A = n_B = 100$
일반회귀모형	0.601	0.608	0.581
로버스트 회귀모형	0.077	0.044	0.027

방법과 일반회귀모형을 이용한 자료결합방법을 비교하고자 한다. 실험조건은 (4.4)와 같다.

$$
\begin{aligned}
&\text{수용파일 A} : (Z_{A1}, Z_{A2}, Z_{A3}) \sim N_3(\underline{0}, \Sigma_{Z_{A1}Z_{A2}Z_{A3}}),\\
&\quad \text{여기서 , } \underline{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \ \Sigma_{Z_{A1}Z_{A2}Z_{A3}} = \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{pmatrix}\\
&X_A = Z_{A1} + Z_{A2} + Z_{A3} + \epsilon_1, \quad \epsilon_1 \sim N(0,1)\\
&\text{제공파일 B} : (Z_{B1}, Z_{B2}, Z_{B3}) \sim N_3(\underline{0}, \Sigma_{Z_{B1}Z_{B2}Z_{B3}}),\\
&\quad \text{여기서 , } \underline{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \ \Sigma_{Z_{B1}Z_{B2}Z_{B3}} = \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{pmatrix}\\
&Y_B = Z_{B1} + Z_{B2} + Z_{B3} + \epsilon_2,\\
&\quad \text{여기서 , } \epsilon_2 \sim N(0,1)\text{일 확률은 } 1-p, \ \epsilon_2 \sim N(15,1)\text{일 확률은 } p.
\end{aligned}
\tag{4.4}
$$

오염률 $p = 0.1$, $p = 0.2$에 대하여 표본크기 n_A, n_B가 각각 20, 50, 100인 경우, 통계적 결합을 100번 독립반복 수행하였다. 결합자료의 각 변수 사이의 상관계수와 (4.3)으로부터 구해진 각 변수에 대한 표본상관계수의 평균차이를 계산하여 두 방법의 상관관계 보존의 성능을 비교한 결과는 표 4.2, 4.3과 같다.

표 4.2, 4.3의 결과를 보면 모든 경우에서 로버스트 회귀모형을 이용한 방법이 일반회귀모형을 이용한 방법에 비해 모분포의 각 변수에 대한 상관관계를 훨씬 잘 보존한다는 것을 알 수 있다. 앞서 두 경우에 대한 모의실험을 통해 알 수 있듯이 본 연구에서 제안하는 로버스트 회귀모형을 이용한 자료결합방법은 다변량정규분포를 따르는 자료에서 뿐만 아니라 특이점이 존재하는 자료에 대해서도 모분포의 각 변수에 대한 상관관계를 결합자료에서 잘 유지하게 한다는 것을 알 수 있다. 따라서 자료결합의 목적이 각 파일에 존재하는 고유변수의 상관관계를 파악하는데 있다면 로버스트 회귀모형을 이용한 자료결합방법을 적용하는 것이 일반회귀모형을 이용한 자료결합방법을 적용하는 것 보다 적절한 방법이라고 할 수 있다.

4.2. 두 자료결합방법에서의 모형에 대한 결정계수 보존 및 향상

고유변수가 다변량 자료인 경우를 고려하여 일반회귀모형을 이용한 자료결합방법과 로버스트 회귀모형을 이용한 자료결합방법으로 얻어진 결합파일에서의 모형에 대한 결정계수 보존 및 향상에 관한 성능을 비교실험 하였다.

표 4.4. 모형을 적합한 결정계수의 평균

		모형에 포함되는 변수	$n_A = n_B = 20$	$n_A = n_B = 50$	$n_A = n_B = 100$
결합 전		$(Z_A = X_1, X_2, X_3, Y)$	0.62	0.61	0.60
		$(Z_A = X_1, X_2, X_3)$	0.23	0.19	0.15
결합 후	일반회귀	$(Z_A = X_1, X_2, X_3, \tilde{Y})$	0.59	0.57	0.54
	로버스트 회귀	$(Z_A = X_1, X_2, X_3, \tilde{Y})$	0.64	0.63	0.63

파일 A와 파일 B에 존재하는 변수 (Z, X_1, X_2, X_3, Y)는 (4.5)와 같은 모분포로부터 임의로 생성한다.

$$(Z, X_1, X_2, X_3, Y) \sim N_5(\underline{\mu}, \Sigma), \quad \text{여기서, } \underline{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.2 & 0.2 & 0.3 & 0.7 \\ 0.2 & 1 & 0.1 & 0.2 & 0.5 \\ 0.2 & 0.1 & 1 & 0.2 & 0.5 \\ 0.3 & 0.2 & 0.2 & 1 & 0.5 \\ 0.7 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}, \tag{4.5}$$

수용파일 A : (Z_A, X_1, X_2, X_3),

제공파일 B : (Z_B, Y).

이때 위와 같이 파일 A에 변수 Y를 제거하여 수용파일이라 하고, 파일 B에서 변수 X_1, X_2, X_3를 제거하여 제공파일이라 하여 제약이 없는 결합을 실시한다. 그리고 표본크기 n_A, n_B가 각각 20, 50, 100인 경우, 통계적 결합을 100회 독립반복 수행한다.

• **다변량 정규분포를 따르는 자료의 경우** 다변량 정규분포를 따르는 자료인 경우, 일반회귀모형을 이용한 자료결합을 수행하여 얻어진 결합파일에서 모형에 대한 설명력이 결합 전의 수용파일에서 모형에 대한 설명력을 잘 유지하고 향상시킬 수 있는지 실험하였다. 결과는 표 4.4와 같다.

표 4.4의 결과를 보면 두 자료결합방법에 의해서 얻어진 결합파일 $(Z_A, X_1, X_2, X_3, \tilde{Y})$에서의 모형에 대한 설명력이 결합 전 자료 (Z_A, X_1, X_2, X_3, Y)에서의 모형에 대한 설명력을 대체로 잘 유지할 수 있다는 것을 알 수 있다. 또한 통계적 결합을 하기 위해 변수 Y를 제거한 결합 전 자료 (Z_A, X_1, X_2, X_3)에서 변수 (X_1, X_2, X_3)가 변수 Z_A를 설명하는 것보다 변수 Y의 정보를 추가적으로 얻은 결합파일 $(Z_A, X_1, X_2, X_3, \tilde{Y})$에서 변수 $(X_1, X_2, X_3, \tilde{Y})$가 변수 Z_A에 대한 설명력을 향상시킨다는 것을 알 수 있다.

• **제공파일 B의 고유변수에 특이점이 존재하는 경우** 제공파일 B의 고유변수 Y값에 특이점을 부여한 방법으로 다음과 같은 실험조건을 고려한다. 우선 제공파일 B의 공통변수 Z_B를 순서대로 정렬한 후 순위에 따라 작은 값들과 큰 값들에 해당되는 개체들을 각각 선택한다. 다음으로 그 선택된 개체의 고유변수 Y값에 $N(9, 1)$분포로 부터 임의 생성된 오염된 자료의 값으로 대체한다. 따라서 제공파일 B의 고유변수 Y값은 $n_B - pn_B$개의 실제 Y값과 pn_B개의 오염된 특이점을 포함한다.

오염률 $p = 0.1$, $p = 0.2$에 따라 두 자료결합방법에 따른 모형에 대한 결정계수 보존 및 향상에 관한 성능을 비교한 결과는 표 4.5, 4.6과 같다. 단, 로버스트 회귀모형을 이용한 자료결합과정 중 선형모형단계에서 최소절사제곱추정량은 절사량 $h = (1 - p) \times n_B$으로 선택하여 구한다. 표 4.5, 4.6의 결과를 보면 로버스트 회귀모형을 이용한 자료결합방법이 일반회귀모형을 이용한 자료결합방법에 비해 결합 전의 자료에서의 모형에 대한 결정계수를 대체로 잘 보존한다는 것을 알 수 있다. 또한 통계적 결합을 하

표 **4.5.** 모형을 적합한 결정계수의 평균($p = 0.1$)

		모형에 포함되는 변수	$n_A = n_B = 20$	$n_A = n_B = 50$	$n_A = n_B = 100$
결합 전		$(Z_A = X_1, X_2, X_3, Y)$	0.66	0.58	0.58
		$(Z_A = X_1, X_2, X_3)$	0.26	0.18	0.15
결합 후	일반회귀	$(Z_A = X_1, X_2, X_3, \tilde{Y})$	0.44	0.33	0.23
	로버스트 회귀	$(Z_A = X_1, X_2, X_3, \tilde{Y})$	0.57	0.51	0.51

표 **4.6.** 모형을 적합한 결정계수의 평균($p = 0.2$)

		모형에 포함되는 변수	$n_A = n_B = 20$	$n_A = n_B = 50$	$n_A = n_B = 100$
결합 전		$(Z_A = X_1, X_2, X_3, Y)$	0.66	0.60	0.58
		$(Z_A = X_1, X_2, X_3)$	0.28	0.19	0.14
결합 후	일반회귀	$(Z_A = X_1, X_2, X_3, \tilde{Y})$	0.43	0.29	0.19
	로버스트 회귀	$(Z_A = X_1, X_2, X_3, \tilde{Y})$	0.58	0.49	0.47

기 위해 변수 Y를 제거한 자료 (Z_A, X_1, X_2, X_3)에서 변수 (X_1, X_2, X_3)이 변수 Z_A에 대한 설명력보다 로버스트 회귀모형을 이용한 자료결합방법으로 생성된 결합파일 $(Z_A, X_1, X_2, X_3, \tilde{Y})$에서 변수 $(X_1, X_2, X_3, \tilde{Y})$가 변수 Z_A에 대한 설명력을 더욱 향상시킨다는 것을 알 수 있다.

5. 결론

통계적 결합방법에 대한 기존연구에서 일반회귀모형을 이용한 자료결합방법은 다변량 정규성을 만족하는 경우에 고유변수 간 상관관계를 결합자료가 강건하게 보존함을 보여주었다. 그러나 자료에 특이점이 존재하는 경우 기존 방법은 많은 문제점을 가지고 있다. 따라서 본 논문에서는 이러한 문제점을 해결하기 위해 로버스트 회귀모형을 이용한 자료결합방법을 제안하였고, 자료결합에 있어서 일반회귀모형과 로버스트 회귀모형의 성능을 비교하였다. 이때 방법의 정확성을 판단하기 위해 다변량 정규분포를 따르는 자료의 결합과 특이점이 존재하는 자료의 결합을 각각 실시하였다. 결합 이후에 생성된 자료의 상관관계가 원 자료의 상관관계를 얼마나 잘 유지하는지를 판단하기 위하여 상관계수의 절대차이를 비교하였으며, 나아가 결합 후와 결합 전의 자료로부터 회귀모형에 대한 결정계수의 변화를 비교 검토하였다. 다변량 정규분포를 따르는 자료에 대한 결합은 일반회귀모형을 이용한 자료결합방법과 로버스트 회귀모형을 이용한 자료결합방법 모두 상관관계가 비교적 충분히 보존됨을 알 수 있었다. 그러나 특이점이 존재하는 자료에 대한 각 방법의 모의실험 결과 다변량 정규분포를 따르는 자료에서 강건한 상관관계를 보존했던 일반회귀모형을 이용한 자료결합방법이 특이점의 결합으로 인해 상관관계를 과대 혹은 과소 추정함을 보였다. 그에 반해 본 연구에서 제안한 로버스트 회귀모형을 이용한 자료결합방법은 다변량 정규분포를 따르는 자료에 대한 결합결과와 마찬가지로 특이점이 존재하는 자료에 대해서도 결정계수의 보존 및 향상과 더불어 상관관계 추정에 대한 통계적 강건성을 보였다.

참고문헌

Kadane, J. B. (1978). Some statistical problems in merging data files, In *1978 Compendium of Tax Research*, Washington, DC:U.S. Department of the Treasury, 159–171. (Reprinted in Journal of Official Statistics, **17**, 423–433.)

Moriarity, C. and Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure, *Journal of official Statistics*, **17**, 407–422.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer-Verlag, New York.

Rässler, S. (2004). Data fusion: Identification problems, validity and multiple imputation, *Austrian Journal of Statistics*, **33**, 153–171.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business & Economic Statistics*, **4**, 87–94.

Statistical Matching Techniques Using the Robust Regression Model

Myoungshic Jhun[1] · Ji Song Jung[2] · Hye Jin Park[3]

[1]Dept. of Statistics, Korea University; [2]MIRAEASSET Securities Co.;
[3]Dept. of Statistics, Korea University

(Received August 2008; accepted September 2008)

Abstract

Statistical matching techniques whose aim is to achieve a complete data file from different sources. Since the statistical matching method proposed by Rubin (1986) assumes the multivariate normality for data, using this method to data which violates the assumption would involve some problems. This research proposed the statistical matching method using robust regression as an alternative to the linear regression. Furthermore, we carried out a simulation study to compare the performance of the robust regression model and the linear regression model for the statistical matching.

Keywords: Statistical matching method, robust regression model, correlation, coefficient of determination.

This research was supported by a Korea University Grant.

[1]Corresponding author: Professor, Dept. of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: jhun@korea.ac.kr

[2]MIRAEASSET Securities Co., Ltd. 7F, Shinsong Building, 25-12, Seoul 150-711, Korea.
E-mail: wjdwlthd@miraeasset.com

[3]Graduate Student, Dept. of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: dabin220@korea.ac.kr