응용통계연구 제 12권 2호, 1999년, pp. 573-584

회귀모형에서 이변량 자료에 합산된 일변량 자료를 첨가시킬 때의 효과*

박래현[1] 이석훈[2] 김노만[3]

요 약

본 연구는 이변량 회귀모형을 이변량 자료에 적용할 때 이변량 자료(분리형 자료) 이외에 이변량 자료를 합산한 일변량 자료(통합형 자료)를 동시에 사용하는 문제를 고찰하였다. 특징을 파악하기 위하여 설명변수가 하나인 경우를 다루었는데 통합형 자료의 첨가효과를 회귀계수의 추정량의 평균제곱오차의 크기로서 측정하면서 효과와 이변량 모형과의 관계를 조사하였다. 최대우도 추정량의 특성으로부터 대표본의 성질을 추출하고 또한 모의실험을 통하여 소표본에서도 대표본의 성질이 만족하는지 조사하였고 끝으로 실제 자료에 적용하여 보았다.

1. 서론

컴퓨터와 정보통신의 발전과 함께 다량의 정보 또는 자료가 계속적으로 새로운 관점과 새로운 방식으로 생성, 수집, 결합되고 있다. 이러한 현상은 특히 분류의 영역에서 두드러지게 나타나고 있는데, 예를 들자면 고도의 첨단 의료기기들이 환자의 혈액내 성분인 프로테인(protein)을 보다 세분화하여 알부민(albumine)과 글로불린(globulin)으로 측정하도록 하게 한다거나 평범하게 우리의 일상 생활 면에서는 쓰레기의 총량을 가연성과 비가연성으로 나누어 측정한다거나 또는 사회현상을 분석하는데 있어서 보다 세밀한 분류 기준을 사용하는 추세에서 쉽게 발견할 수 있다. 이러한 현상을 통계적인 관점에서 조명하여 볼 때 과거의 분류방식에 의해서 생성된 많은 귀중한 자료 또는 정보가 보다 세밀하게 분류되지 못했다고 해서 그 자료가 아직 유용한 정보를 포함하고 있음에도 불구하고 효과적으로 사용되지 못하거나 또는 신중한 고찰 없이 사용되어 그 결과에 있어서 이론적 근거를 확보하지 못하는 경우가 나타나고 있다고 판단된다. 특히 세밀하게 분류하기가 어렵거나 비용이 많이 들 때는 세밀하게 분류가 안된 과거 자료의 사용이 더욱 중요하다고 생각된다. 이러한 판단에서 본 논문에서는 특별히 회귀 모형에서 나타나는 문제의 특성, 해결 방안 및 관련되는 이론적인 배경 등을 고찰한다.

과거의 연구로는 이러한 문제를 다룬 예가 별로 많지 않으나, 유사한 동기에서 출발되었던 연구로는 범주형 자료분석에서 Park(1996)의 연구가 있었다. 이것을 이석훈(1998 a, b)이 사회 현상에서 주로 제기되는 분류 문제로 발전시켜 두 집단으로 분류된 자료와 두 집

* 이 논문은 1997년 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음
1) (305-764)대전광역시 유성구 궁동 220번지, 충남대학교 통계학과, 교수
2) (305-764)대전광역시 유성구 궁동 220번지, 충남대학교 통계학과, 교수
3) (305-600)유성우체국 사서함 35, 국방과학 연구소, 선임 연구원

단 중 하나가 세분화됨으로서 생성된 세 집단으로 분류된 자료의 통합 과정을 연구한 바 있다. 한편, 인구통계를 사용하는 분야에서 나타나는 현상 중에서 하나의 특성변수로 나누어진 자료와 또 하나의 특성변수와 함께 두 개의 변수로 나누어진 자료를 통합하는 상황을 고찰하는 연구가 있다. 최근의 대표적인 응용 예로서는 경영학의 마케팅분야에서 중요한 주제로 생각하는 시장세분화(Segmentation)가 있다. 어느 지역의 주민이 일부 자료는 성별과 주택 소유 여부 각각의 주변분포만 나타내고 있다고 하자. 다른 자료는 성별과 주택 소유 두 변수의 결합분포까지 제공한다고 할 때 이들 두 자료의 결합문제를 Putler(1996)가 다룬 바 있다.

종속변수로 사용하려는 크기가 n인 현재의 세분화 자료를 $(y_{1i}, y_{2i})'$, $i = 1, 2, \cdots, n$이라 하고, 세분화되지 않은 크기가 k인 과거의 자료를 $z_j = y_{j1} + y_{j2}$, $j = 1, 2, \cdots, k$라고 하자. p개의 독립변수를 $x_1, x_2, \cdots, x_p$라 할 때 제2장에서는 회귀분석시 과거의 자료를 이용하는 방법을 제시하여 과거의 자료가 어느 경우에 얼마나 효과가 있는지를 논하는데, 이때 효과의 측도는 추정량의 분산으로 한다. 과거 자료의 사용효과를 y_1과 y_2의 상관계수 ρ와 두 표본의 크기 n, k에 따라 살펴보려고 하는데 계산상의 어려움 때문에 절편이 없고 $p = 1$인 경우와 n, k가 대표본일 때로 국한시켜 전개한다. 제3장에서는 2장의 내용이 대표본일 경우의 이론이므로 소표본에서의 효과를 시뮬레이션을 통하여 근사하고, 또한 실제자료를 제안한 모델에 적용하여 보고, 4장에서는 결론을 맺는다.

2. 모형 및 분석

2.1. 모형과 기본식

현재의 분리된 자료에 대해 적용한 회귀분석 모형은 다음과 같다.

$$\vec{y_i} = \begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} = \begin{pmatrix} \beta_{01} + \beta_{11}x_{i1} + \cdots + \beta_{p1}x_{ip} \\ \beta_{02} + \beta_{12}x_{i1} + \cdots + \beta_{p2}x_{ip} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix}, i = 1, 2, \cdots, n$$

여기서 $(\epsilon_{1i}, \epsilon_{2i})'$는 평균이 $(0, 0)'$이고 분산공분산행렬이 Σ인 이변량 정규분포를 따른다. 따라서 위 모형 하에서 $z_j = y_{1j} + y_{2j}, j = 1, 2, \cdots, k$는 평균이 $(\beta_{01}+\beta_{02})+(\beta_{11}+\beta_{12})w_{j1}+(\beta_{21}+\beta_{22})w_{j2}+\cdots+(\beta_{p1}+\beta_{p2})w_{jp}$이고, 분산이 $\vec{1}'\Sigma\vec{1}$의 정규분포를 따른다. 여기서, $\vec{1}' = (1, 1)$이고, w_{jk}는 x_{jk}와는 다른 독립변수 x_k의 관찰치이다.

위 모형에서 우리가 취급할 내용은 과거 자료 z_j를 현재 자료에 어떻게 도입할 것이며 또 도입시 회귀계수 β_{ij} 또는 $E(\vec{y_0})$(주어진 독립변수값 $\vec{x_0}$에서 $\vec{y_0} = (y_{10}, y_{20})'$의 기대값)의 추정에 어느 경우에, 얼마만큼 도움이 되 는가등 첨가 효과를 살펴보는 것이다. 구체적인 방법으로는 $\beta_{ij}, E(\vec{y_0})$의 최우추정량의 분산을 두 가지 경우 - 과거의 통합자료를 첨가했을 경우와 안 했을 경우 - 로 나누어 계산하여 비교하는 것을 택하려 한다.

현재 자료 $\vec{y_1}, \cdots, \vec{y_n}$만의 우도함수를 L_1, 현재 자료에 과거 자료 $z_1, \cdots, z_k$까지를 합친

자료(이를 결합 자료라 하자.)에서의 우도함수를 L_2라 하면 L_1, L_2는 다음과 같다.

$$L_1 = (2\pi)^{-n}|\Sigma|^{-\frac{1}{2}n} exp[-\frac{1}{2}tr\{(Y-XB)\Sigma^{-1}(Y-XB)'\}]$$

$$L_2 = L_1 \cdot (2\pi)^{-\frac{1}{2}k}(\vec{1}'\Sigma\vec{1})^{-\frac{1}{2}k} exp[-\frac{1}{2}tr\{(\vec{Z}-X^*B\vec{1})(\vec{1}'\Sigma\vec{1})^{-1}(\vec{Z}-X^*B\vec{1})'\}]$$

위에서 $Y = (\vec{y_1}, \vec{y_2}, \cdots, \vec{y_n})'$, $B = \begin{pmatrix} \beta_{11} & \cdots & \beta_{p1} \\ \beta_{12} & \cdots & \beta_{p2} \end{pmatrix}'$, $\vec{Z} = (z_1, \cdots, z_k)'$,

$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$, $X^* = \begin{pmatrix} w_{11} & \cdots & w_{1p} \\ \cdots & \cdots & \cdots \\ w_{k1} & \cdots & w_{kp} \end{pmatrix}$

이다. 계산의 편의를 위하여 $l_1 = lnL_1$, $l_2 = lnL_2$라 하고, l_1에서의 최대화하는 최우추정량은 알려진 것이고 l_2에서의 최우추정량을 구하기 위한 일차편도함수는 아래와 같다.

$$\frac{\partial l_2}{\partial B} = X'Y\Sigma^{-1} - X'XB\Sigma^{-1} + (\vec{1}'\Sigma\vec{1})^{-1}X^{*'}\vec{z}\vec{1}' - (\vec{1}'\Sigma\vec{1})^{-1}X^{*'}X^*B\vec{1}\vec{1}'$$

$$\frac{\partial l_2}{\partial \sigma_{11}} = -\frac{n}{2}C_2 + C_1 + \frac{1}{2}G_1,\ \frac{\partial l_2}{\partial \sigma_{22}} = -\frac{n}{2}C_3 + C_1 + \frac{1}{2}G_2,\ \frac{\partial l_2}{\partial \sigma_{12}} = nC_4 + 2C_1 + G_3$$

여기서, $C_1 = -\frac{k}{2}\frac{1}{VarZ} + \frac{1}{2}\frac{SSZ}{(VarZ)^2}$, $C_2 = \frac{\sigma_{22}}{DET}$, $C_3 = \frac{\sigma_{11}}{DET}$, $C_4 = \frac{\sigma_{12}}{DET}$,

$G_1 = \frac{\sigma_{22}^2 s_{11} - 2\sigma_{12}\sigma_{22}s_{12} + \sigma_{12}^2 s_{22}}{(DET)^2}$,

$G_2 = \frac{\sigma_{12}^2 s_{11} - 2\sigma_{12}\sigma_{11}s_{12} + \sigma_{11}^2 s_{22}}{(DET)^2}$,

$G_3 = \frac{(\sigma_{11}\sigma_{22} + \sigma_{12}^2)s_{12} - \sigma_{12}\sigma_{22}s_{11} - \sigma_{12}\sigma_{11}s_{22}}{(DET)^2}$,

$DET = \sigma_{11}\sigma_{22} - \sigma_{12}^2$, $VarZ = \sigma_{11} + 2\sigma_{12} + \sigma_{22}$, $SSZ = (\vec{z} - X^*B\vec{1})'(\vec{z} - X^*B\vec{1})$,

$\begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = (Y-XB)'(Y-XB)$ 이다.

2.2. 모수의 추정과 첨가 효과 분석

Marquardt 방법을 이용하여 앞에서 구한 우도방정식의 해를 최우추정량으로 구하고, 이 때 Fisher의 정보행렬의 역행렬로부터 근사 분산공분산을 얻으려고 하는데 첨가효과를 구체적으로 수식화하기 위하여 이후로는 절편이 없는 $p = 1$인 경우만을 고려한다.

Fisher 정보행열 I는 로그우도함수의 이차편도함수의 행렬 $G = (g_{ij})$ (부록참조)의 각 항의 기대값 $E(g_{ij})$으로 다음과 같은 항을 갖는다.

$$E(g_{ij}) = g_{ij},\ i = 1,2,\ j = 1,2,\ E(\frac{\partial^2 l_2}{\partial\sigma_{ij}\partial\beta_i}) = 0,\ i = 1,2, j = 1,2$$

$$E(g_{33}) = \frac{k}{2(VarZ)^2} - \frac{n}{2}{C_2}^2,\ E(g_{44}) = -\frac{n}{2}{C_3}^2 - \frac{k}{2(VarZ)^2},$$

$$E(g_{55}) = \frac{-2k}{(VarZ)^2} + \frac{n(\sigma_{11}\sigma_{22} + {\sigma_{12}}^2)}{(DET)^2},\ E(g_{34}) = -\frac{k}{2(VarZ)^2} - \frac{n}{2}{C_4}^2,$$

$$E(g_{35}) = -\frac{k}{(VarZ)^2} + nC_2C_4,\ E(g_{45}) = -\frac{k}{(VarZ)^2} + nC_3C_4$$

따라서, n, k가 모두 클 때 $\vec{\theta} = (\beta_1, \beta_2, \sigma_{11}, \sigma_{22}, \sigma_{12})'$의 최우추정량 $\tilde{\vec{\theta}}$의 공분산행렬 $cov(\tilde{\vec{\theta}}) \doteq I^{-1}$을 이용하면 $\tilde{\beta}_1$, $\tilde{\beta}_2$ 의 근사공분산 행렬(asymptotic covariance matrix)은

$$cov\left(\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix}\right) \doteq \begin{pmatrix} \frac{\sigma_{11}}{SSX+SSW} + Q & \frac{-\sigma_{12}}{SSX+SSW} + Q \\ \frac{-\sigma_{12}}{SSX+SSW} + Q & \frac{\sigma_{22}}{SSX+SSW} + Q \end{pmatrix}$$

이고,

$$Q = \frac{DET \cdot SSW}{VarZ \cdot SSX(SSX + SSW)}$$

이다.

한편 분리형 자료로부터 구한 $\vec{\beta} = (\beta_1, \beta_2)'$의 최우추정량 $\hat{\vec{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$의 공분산행렬은

$$cov\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = SSX^{-1}\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

이므로, 두 가지 경우 β_1, β_2의 최우추정량에 대한 점근분산의 비는 아래와 같이 두 종속변수 y_1과 y_2의 상관계수 ρ와 각각의 표준편차 σ_1과 σ_2의 비인 $R = \sigma_1/\sigma_2$ (단, $\sigma_1 = \sqrt{\sigma_{11}}, \sigma_2 = \sqrt{\sigma_{22}}$)의 함수로 표현된다.

$$f_1(\rho, R) = \frac{Var(\tilde{\beta}_1)}{Var(\hat{\beta}_1)} \doteq \frac{SSX}{SSX + SSW} + \frac{SSW(1-\rho^2)}{(1 + R^2 + 2\rho R)(SSX + SSW)}$$

$$f_2(\rho, R) = \frac{Var(\tilde{\beta}_2)}{Var(\hat{\beta}_2)} \doteq \frac{SSX}{SSX + SSW} + \frac{SSW \cdot R^2(1-\rho^2)}{(1 + R^2 + 2\rho R)(SSX + SSW)}$$

여기서, 일반성을 잃지 않고 $\sigma_1 \leq \sigma_2$ (즉, $0 < R \leq 1$)라고 가정하기로 한다.

위의 두식은 β_1, β_2와는 독립적인 식이므로 R과 λ에 따르는 변화만을 조사하면 다음과 같은 성질을 쉽게 발견할 수 있다.

성질 :

f_1, f_2는 $\rho = -R$일 때, 최대값을 각각 1, $\frac{SSW+R^2SSW}{SSX+SSW}$ (≤ 1) 갖고,

$\rho = -R$에서 멀수록 작아지는 단조 감소형이며,

$\rho = \pm 1$일 때, 모두 최소값이 $\frac{SSX}{SSX+SSW}$을 갖는다.

따라서, 분산이 작은 변수를 설명하는 회귀계수 β_1의 추정량의 분산의 크기에서 볼 때 $\rho = -R$이면 과거 자료의 첨가효과가 전혀 없고, 분산이 큰 변수를 설명하는 β_2의 추정량은 $\rho = -R$일 때 최소의 효과를 갖는다. 두 변수의 상관계수 ρ 값이 $-R$에서 멀어질수록 첨가효과는 커지며 $\rho = \pm 1$일 때, 효과가 최대가 된다. 한편, 추론 대상이 주어진 독립변수의 값 x_0에서 $E(y_{i0})$의 추정량 $\tilde{\beta}_i x_0$의 분산의 비는

$$\frac{Var(\tilde{\beta}_i x_0)}{Var(\hat{\beta}_i x_0)} = \frac{Var(\tilde{\beta}_i)}{Var(\hat{\beta}_i)},\ i = 1, 2$$

이므로 추론 대상이 β_i일 때와 같은 경우가 되어 따로 다룰 필요가 없다.

위 식 f_1, f_2로부터 얻은 첨가효과의 특성들 중 특히 두 변수 y_1과 y_2의 상관계수 ρ가 두 변수의 표준편차의 비의 음수값, 즉 $-R$일 때 최소효과가 나타나는 것을 이해하기 위하여 독립변수 x, w 값을 모두 1로 하고 분리형 자료에 대한 통합형 자료 표본크기의 비 k/n을 λ라 놓고 f_1과 f_2를 다시 표현하면 다음과 같다.

$$f_1(\lambda,\rho,R)=\frac{1+\lambda+R^2+2\rho R-\lambda\rho^2}{(1+\lambda)(1+R^2+2\rho R)},\ f_2(\lambda,\rho,R)=\frac{1+\lambda+R^{-2}+2\rho R^{-1}-\lambda\rho^2}{(1+\lambda)(1+R^{-2}+2\rho R^{-1})}$$

이 결과는 이경아(1997)가 한 개의 모집단에 대한 추론과정의 연구에서도 언급한 바가 있는데 $\rho=-R$일 때, $f_1=1$로서 첨가효과가 전혀 없음을 보여준다.

그러나, $\rho=-R$일 때 f_2의 값은 1보다 작은 $\frac{1+\lambda R^2}{1+\lambda}$로 통합된 자료의 첨가가 분산이 큰 변수의 모수의 추정에는 도움이 된다. $\rho=-R$일 때, $\hat{\beta}_1$의 추정시 첨가 효과가 없는 현상은 다음의 정리로 설명될 수 있다.

정리 :

$\vec{y}=(y_1,y_2)'\sim N_2(\vec{\beta},\sigma)$일 때, $y_1+y_2=z$가 주어졌을 때의 y_1의 조건부 확률분포는

$$N\left(\beta_1+\frac{\sigma_{11}+\sigma_{12}}{\sigma_{11}+2\sigma_{12}+\sigma_{22}}(z-\beta_1-\beta_2),\ \sigma_{11}-\frac{(\sigma_{11}+\sigma_{12})^2}{\sigma_{11}+2\sigma_{12}+\sigma_{22}}\right)$$

이고, $\rho=-R$일 때 $Var(y_1|y_1+y_2)=Var(y_1)$, $E(y_1|y_1+y_2=z)=E(y_1)$이다.

이 정리에서 드러난 $\rho=-R$일 때 통합 자료 y_1+y_2가 z로 주어졌을 때 y_1의 조건부 분산이 y_1의 주변 분산과 같다는 사실을 통합 자료의 추가가 y_1에 대한 정보에 전혀 기여하지 않는 것으로 해석한다면 y_1의 기대값의 추정량의 분산이 통합 자료를 추가하여도 전혀 감소하지 않는 특성을 이해할 수가 있다.

3. 모의실험 및 토의

2.3절에서 논의한 모든 결과는 표본의 크기 n, k가 모두 대표본일 경우에 근사적으로 성립되는 내용이므로 여기서는 n, k가 소표본 경우에는 어떠한 현상이 있는지를 모의실험을 통하여 살펴보고, 실제 자료로 1980-1997년까지 우리 나라의 15세 이상 경제활동가능인구를 가지고, 1980-1993년까지 경제활동가능인구를 통합된 자료로, 1994-1997년까지 경제활동인구와 비경제활동인구로 분리된 자료로 하여 당시의 국민 1인당 GNP에 대하여 회귀절편(β_0)이 있는 모형으로분석을 한다. 모의실험 결과 분석시 평균제곱오차의 제곱근을 $RMSE$, 분리형 자료의 평균제곱오차의 제곱근을 $RMSE1$, 결합형 자료의 평균제곱오차의 제곱근을 $RMSE2$, 그 비를 $MR=RMSE2/RMSE1$로 표시한다.

3.1. 설명변수 x와 w가 모두 1인 경우

$\beta_1=1$, $\beta_2=1$, $\sigma_{11}=1$, $\sigma_{22}=4$라 하고, σ_{12}를 $-1,0,1$로 즉, ρ를 $-R,0,R$로 바꾸면서 표본의 크기 n과 k를 변화시켰다. 다음의 표 3.1는 1,000번을 반복하면서 구한 각 모수의 추정치들의 평 균제곱오차의 제곱근을 n과 k에 따라서 정리한 표이다.

표 3.1: $\beta_1 = 1, \beta_2 = 1, \sigma_{11} = 1, \sigma_{22} = 4$일 때의 β_1, β_2의 추정치

				$\rho = -R$		$\rho = 0$		$\rho = R$	
n	k	구 분		β_1	β_2	β_1	β_2	β_1	β_2
5	5	분리형 자료	추정치	1.0042	1.0370	0.9896	1.0438	1.0033	1.0258
			RMSE1	0.2107	0.8199	0.1957	0.7954	0.1788	0.7119
		통합자료 추가	추정치	1.0035	1.2042	0.9902	1.0312	0.9922	1.0001
			RMSE2	0.2145	0.5602	0.2028	0.4879	0.1582	0.4428
		RMSE2/RMSE1		1.163	0.683	1.036	0.613	0.884	0.622
5	10	분리형 자료	추정치	1.0117	1.0201	0.9721	1.0117	1.0151	1.0139
			RMSE1	0.2074	0.8393	0.2022	0.8149	0.1894	0.7817
		통합자료 추가	추정치	1.0094	1.0158	0.9795	1.0355	1.0095	0.9827
			RMSE2	0.2522	0.4562	0.2022	0.4229	0.1457	0.3603
		RMSE2/RMSE1		1.215	0.543	1.000	0.519	0.769	0.461
5	20	분리형 자료	추정치	1.0159	0.9969	0.9881	1.0342	0.9960	1.0091
			RMSE1	0.2044	0.8012	0.1965	0.8582	0.1904	0.7945
		통합자료 추가	추정치	1.0395	0.9937	0.9775	1.0159	0.9883	1.0100
			RMSE2	0.2600	0.3597	0.2004	0.3194	0.1349	0.2640
		RMSE2/RMSE1		1.272	0.449	1.019	0.372	0.709	0.332
10	10	분리형 자료	추정치	0.9930	0.9933	0.9984	0.9884	0.9983	1.0001
			RMSE1	0.1011	0.3830	0.1045	0.3944	0.1002	0.3720
		통합자료 추가	추정치	0.9946	0.9996	1.0012	0.9921	0.9978	0.9977
			RMSE2	0.1062	0.2448	0.0967	0.2435	0.0781	0.2144
		RMSE2/RMSE1		1.050	0.639	0.926	0.617	0.779	0.576
10	20	분리형 자료	추정치	0.9942	1.0003	1.0013	0.9882	1.0031	0.9822
			RMSE1	0.1024	0.3793	0.1032	0.3817	0.0987	0.3869
		통합자료 추가	추정치	0.9958	1.0080	1.0002	0.9937	1.0049	0.9683
			RMSE2	0.1098	0.2006	0.0950	0.1853	0.0669	0.1649
		RMSE2/RMSE1		1.072	0.528	0.920	0.485	0.677	0.426
10	40	분리형 자료	추정치	1.0013	0.9865	0.9973	0.9848	1.0074	1.0103
			RMSE1	0.1013	0.3685	0.1042	0.4037	0.0946	0.3847
		통합자료 추가	추정치	1.0018	0.9930	0.9979	0.9949	1.0033	1.0038
			RMSE2	0.1088	0.1509	0.0932	0.1474	0.0587	0.1201
		RMSE2/RMSE1		1.073	0.409	0.895	0.365	0.621	0.312

모의실험 결과를 $\lambda = k/n$(표본의 비)의 변화에 따라 $RMSE$ (Root Mean Square Error)의 변동 추이를 보면 $\rho = -0.5$이면서, $n = 5$일 때, $\hat{\beta}_1$의 $RMSE2$는 기대와는 달리 증가 추세를 나타내었는데 이는 소표본의 현상으로 보아야 할 것 같다. 반면에 $\hat{\beta}_2$의 $RMSE2$는 0.5602에서 0.3597로 크게 감소했다. $\hat{\beta}_1$의 MR은 이론상 1이 나와야 하지만 1보다 약간 큰

값이 나타났고, $\hat{\beta}_2$의 MR은 0.683에서 0.449로 34% 감소한다. $n = 10$일 때에도 역시 $\hat{\beta}_1$의 $RMSE2$은 거의 변화가 없고, $\hat{\beta}_2$의 $RMSE2$는 크게 감소하며, $\hat{\beta}_1$의 MR은 1보다 약간 큰 값이지만, $\hat{\beta}_2$의 MR은 크게 감소한다. $\rho = 0$이면서, $n = 5$일 때, 표본의 비가 증가할수록 $\hat{\beta}_1$의 $RMSE2$는 작게 감소, $\hat{\beta}_2$의 $RMSE2$는 0.4879에서 0.3194로 크게 감소 한다. 또, $\hat{\beta}_1$의 MR은 작게 감소, $\hat{\beta}_2$의 MR은 크게 감소 한다. $n = 10$일 때, 역시 $n = 5$일 때와 거의 유사한 현상이 나타난다.

$\rho = 0.5$인 경우는, $n = 5, 10$일 때 모두, $\hat{\beta}_1$의 $RMSE2$는 0.1582에서 0.1349로 다른 경우와 달리 확실히 감소하고, $\hat{\beta}_2$의 $RMSE2$도 감소(40%)한다. 전체적으로는 상관계수 ρ가 $-0.5, 0, 0.5$로 변화할 때 추정치 $\hat{\beta}_1$과 $\hat{\beta}_2$에 대한 $RMSE$가 작아졌으며, 다른 말로 하면 통합 자료의 첨가 효과가 높아지는 것으로 나타났으며, 분산이 큰 변수의 기대값의 추정량에 대한 RMSE가 그렇지 않은 자료 보다 크게 나타났다.

3.2. 설명변수 x와 w의 각 값에서 관찰치가 얻어진 경우

설명변수 x와 w를 1, 2, 3, 4, 5로 하고 각 점에서 n_1, n_2, n_3, n_4, n_5개의 통합형 자료를 얻었을 때의 경우를 다음과 같이 고려하여 모의실험에서 1,000번을 반복 실험하여 나타나는 추정치 와 추정치들의 RMSE를 조사하였고, 그 결과는 표 3.2이다.

표 3.2: $\beta_1 = 1$, $\beta_2 = 1$, $\sigma_{11} = 1$, $\sigma_{22} = 4$일 때 β_1, β_2의 추정치

CASE	n_1	n_2	n_3	n_4	n_5	구분		$\rho = -R$		$\rho = 0$		$\rho = R$	
	k_1	k_2	k_3	k_4	k_5			β_1	β_2	β_1	β_2	β_1	β_2
CASE1	1	1	1	1	1	분리형 자료	추정치	0.9944	1.0071	0.9908	1.0088	1.0012	1.0088
							RMSE1	0.0180	0.0745	0.0185	0.0755	0.0183	0.0794
	10	10	10	10	10	통합자료 추가	추정치	0.9950	1.0036	0.9910	1.0094	1.0006	0.9992
							RMSE2	0.0237	0.0278	0.0186	0.0227	0.0119	0.0167
						RMSE2/RMSE1		1.321	0.373	1.007	0.300	0.649	0.211
CASE2	3	3	3	3	3	분리형 자료	추정치	0.9994	1.0055	1.0001	1.0001	0.9969	0.9960
							RMSE1	0.0064	0.0249	0.0064	0.0248	0.0060	0.0230
	10	10	10	10	10	통합자료 추가	추정치	0.9981	1.0039	0.9992	1.0023	1.0006	1.0032
							RMSE2	0.0066	0.0107	0.0056	0.0095	0.0039	0.0037
						RMSE2/RMSE1		1.026	0.431	0.882	0.383	0.652	0.337
CASE3	3	0	1	0	3	분리형 자료	추정치	0.9977	1.0013	0.9949	0.9941	0.9987	1.0026
							RMSE1	0.0111	0.0433	0.0117	0.0462	0.0115	0.0442
	10	10	10	10	10	통합자료 추가	추정치	0.9963	1.0021	0.9965	1.0029	0.9986	1.0014
							RMSE2	0.0132	0.0171	0.0104	0.0151	0.0072	0.0111
						RMSE2/RMSE1		1.179	0.394	0.891	0.327	0.631	0.252
CASE4	3	3	3	3	3	분리형 자료	추정치	1.0015	1.0011	1.0014	1.0021	0.9982	1.0046
							RMSE1	0.0065	0.0246	0.0063	0.0254	0.0067	0.0249
	20	20	20	20	20	통합자료 추가	추정치	1.0003	1.0012	0.9996	1.0012	0.9978	1.0033
							RMSE2	0.0067	0.0089	0.0055	0.0075	0.0036	0.0059
						RMSE2/RMSE1		1.034	0.362	0.869	0.296	0.541	0.237

모의실험결과를 각 CASE별로 구분하여 나타난 결과를 비교해 보면, CASE1과 CASE2에서는 통합형 자료의 크기를 각 설명변수에서 10개로 고정시키고 분리형 자료 표본의 크기를 각 설명변수에서 1개와 3개씩 변화시켰을 때로서 나타난 결과는 다음과 같다. 상관계수 ρ를 $-0.5, 0, 0.5$로 변화시킬 때 분리형 자료로부터 얻은 $\hat{\beta}_1$, $\hat{\beta}_2$의 $RMSE1$은 변화가 매

우 작으나, 통합형 자료를 첨가 시켰을 때 추정치 $\hat{\beta}_1$, $\hat{\beta}_2$의 $RMSE2$는 크게 감소하는데, 이는 3.1절의 모의실험에서 나타난 결과와 유사하다. 구체적으로 살펴보면, 추정치 $\hat{\beta}_1$의 $RMSE2$는 CASE1에서 0.0237 → 0.0119로 50% 감소하고, CASE2에서 0.0066 → 0.0039로 41% 감소함을 보였고, 추정치 $\hat{\beta}_2$의 $RMSE2$는 CASE1에서 0.0278 → 0.0167로 40% 감소하고, CASE2에서 0.0107 → 0.0037로 66%로 크게 감소함을 보였다. 이 결과는 분산이 큰 자료의 모수 추정치에 대한 RMSE2 가 작아지는 결과로 첨가 효과가 더 크다는 의미이며, 대표본에서의 결과와 일치하며, 3.1절에 나타나는 결과와 유사하다.

그리고 CASE2와 CASE4의 MR을 비교하면 상관계수가 $-0.5, 0, 0.5$ 각각에서 통합형자료가 10에서 20으로 늘어난 효과를 볼 수 있는데 역시 변수의 분산이 큰 쪽에서 더 큰 효과가 나타나면서 대표본의 성질을 함께 따르는 것으로 나타났다. CASE3는 CASE2와는 달리 분리형 자료에 해당되는 설명변수 값과 통합형 자료에서의 설명변수 값이 다른 경우이다. MR값을 기준으로 보면, CASE2보다는 전반적으로 작은 반면, CASE1보다는 차이가 크지는 않지만 약간 큰 값이 나타났다. 이 현상은 총 50개의 통합형 자료의 첨가 효과가 분리형 자료의 총 개수가 커짐에 따라 작아지는 것으로 설명할 수 있겠다. 이와 유사한 현상은 표 3.1의 $n = 5$, $k = 10$인 경우와 $n = 5$, $k = 20$인 경우에서도 나타난 바 있다.

3.3. 제안한 모델에 실제자료적용 예

이제는 실제자료에 적용하여 보기로 한다. 1980~1997년까지 우리 나라의 15세 이상 경제활동가능인구 자료를 표 3.3에서와 같이, 1980~1993년까지 경제활동가능인구를 통합 자료로, 1994~1997년까지 경제활동인구와 비경제활동인구로 분리한 분리된 자료로 구분하여 당시의 국민 1인당 GNP에 대하여 회귀분석을 수행해 본다. 괄호 안의 자료(1980~1993)는 실제로 경제활동인구와 비경제활동인구로 분리되어 있는 자료이지만 제안한 모델 적용을 위하여 두 자료를 합한 통합된 자료로만 사용하였고 실제 값은 결과 평가시에 사용하였다. 절편이 있는 회귀분석 모형을 분리형 자료, 결합형 자료에 적용한 결과는 각각 다음과 같다.

(1) 분리된 자료만 사용할 때

$y_1 = 17,132 + 0.339 \cdot x$ (표준편차 : 0.2961), $y_2 = 6,481 + 0.637 \cdot x$ (표준편차:0.2581)

(2) 분리된 자료에 통합 자료를 첨가하여 사용할 때

$y_1 = 15,472 + 0.5924 \cdot x$ (표준편차 : 0.0420) , $y_2 = 9,150 + 0.3878 \cdot x$ (표준편차:0.0413)

로 적합된다. 두 식을 각각 비교하면 분리된 자료만 사용할 때 보다 통합 자료를 첨가하여 사용한 자료에 적합된 $\hat{\beta}_1$과 $\hat{\beta}_2$의 분산의 추정치가 크게 감소되었음이 보인다. 이때 한 가지 언급할 것은 중심화 상수를 결정할 때 통합형 자료를 분리형 자료와 함께 사용하는 방법인데 이 적용에서는 분리형도 일단 통합형으로 간주하여 반응변수의 총평균을 구하고, 이것에 분리형 자료에서 나타나고 있는 각 변수의 평균의 상대적 크기를 곱하여 각 변수의 중심화 상수로 하였다.

표 3.3: 한국의 경제활동가능인구

연도	경제활동가능인구(천명)	경제활동인구(천명)	비경제활동인구(천명)
1980	24,463	(14,431)	(10,032)
1981	25,100	(14,683)	(10,417)
1982	25,638	(15,032)	(10,605)
1983	26,212	(15,118)	(11,094)
1984	26,861	(14,997)	(11,865)
1985	27,553	(15,592)	(11,961)
1986	28,225	(16,116)	(12,109)
1987	28,995	(16,873)	(12,082)
1988	29,602	(17,305)	(12,298)
1989	30,265	(18,023)	(12,242)
1990	30,887	(18,539)	(12,348)
1991	31,422	(19,048)	(12,374)
1992	31,898	(19,426)	(12,472)
1993	32,400	(19,803)	(12,597)
1994	32,939	20,236	12,614
1995	33,558	20,797	12,761
1996	34,182	21,188	12,994
1997	34,736	21,604	13,132

자료출처 : 통계로 본 대한민국 50년의 경제사회상 변화(1998)

한편 위의 두 추정식을 통하여 1980년부터 1993년까지의 설명변수인 GNP에 대한 경제활동인구와 비경제활동인구를 추정한 결과가 실제 자료 중 경제활동인구의 차이를 잔차1, 비경제활동인구와의 차이를 잔차2 로 , 그리고 두 인구를 합한 경제활동가능인구와의 차이를 통합잔차라고 하면 다음의 표 3.4와 같다. 특별히 통합잔차 부분을 보면 80년에서 84년까지는 오히려 분리형 자료만 사용하였을 때가 잔차가 작게 나타났고, 그 이후에서는 역시 결합형자료를 사용했을 때의 잔차가 거의 50% 수준으로 작게 나타났다. 그러나 잔차1과 잔차2를 각각 비교하면 결합형자료를 사용하였을 때가 두 변수 모두 압도적으로 좋은 결과를 나타냄으로써 통합형 자료의 첨가 효과를 잘 보여주고 있다.

4. 결론

본 논문에서는 이변량 정규분포를 가정한 이변량 자료의 분석시 이 두 변량의 합으로 작성된 자료들(통합형 자료)을 이용하는 과정과 이 자료가 첨가되었을 때 나타나는 효과, 즉 모수추정에 있어서 어느 경우에 얼마 정도의 도움을 주는지를 회귀분석 모형에서 조사하였다. 그 첨가 효과의 척도로는 최우추정량의 평균제곱오차의 제곱근을 이용하였다.

설명변수가 하나인 단순회귀모형에서 과거의 통합 자료 첨가효과를 β_1, β_2의 최우추정량에 대한 점근 분산의 비(x_0에서 추정치 $\tilde{\beta}_i x_0 (= \hat{\beta}_i \cdot x_0)$의 분산의 비)로 모의실험을 통하여 조사해 보면 상관계수 $\rho = -0.5$에서 첨가효과가 최소였고, 특히 분산이 작은 자료의 첨가효과는 거의 없었다. 반대로 상관계수 $\rho = \pm 1$에서 첨가효과가 최대로 나타났다. 또, 소표본에서의 모의실험결과는 분산이 큰 자료에 통합자료를 첨가했을 때 첨가효과가 크게 나타나서 대표본 성질과 동일함을 보였고, 실제 자료에 절편을 포함한 회귀모형을 적합시켰을 때 분리된 자료만 사용하였을 때에 비해서 훨씬 작은 잔차를 갖는 것을 보였다.

표 3.4: 실제 분리된 자료와 추정 회귀식으로 구한 분리된 자료 및 잔차

연	분리형 자료			결합형 자료		
도	잔차 1	잔차 2	통합잔차	잔차 1	잔차 2	통합잔차
1980	-3,821	2,532	-1,289	-1,987	262	-1,725
1981	-3,617	2,826	-791	-1,820	591	-1,229
1982	-3,300	2,954	-346	-1,526	743	-783
1983	-3,275	3,329	54	-1,547	1,162	-385
1984	-3,455	3,989	534	-1,770	1,866	96
1985	-2,878	4,050	1,172	-1,208	1,941	733
1986	-2,465	3,990	1,525	-877	1,962	1,085
1987	-1,928	3,549	1,621	-505	1,683	1,178
1988	-1,861	3,079	1,218	-711	1,482	771
1989	-1,453	2,440	987	-535	1,071	536
1990	-1,166	2,117	951	-418	916	498
1991	-949	1,593	644	-419	607	188
1992	-653	1,537	884	-185	611	426
1993	-444	1,346	902	-102	544	442

참고문헌

[1] Marquardt (1963), An Algorithm From Least-Squares Estimation of Non-Linear Parameters. SIAM *Journal of Applied Mathematics*, Vol 11, 431-441.

[2] Park, C. J. and Lee, S. H. (1996), *A Note on Estimation of Multinominal Cell Probabilities When Some frequency Counts are Merged*, Unpublished manuscript.

[3] Pulter D. S., Kalyanam K. and J. S. Hodges (1996), A Bayesian Approach for Estimating Target Market Potential with Limited Geodemograpgic Information, *Journal of Marketing Research*, 134-149.

[4] 이경아 (1997), 이변량 자료에 합산된 일변량 자료를 첨가시킬 때 효과. 충남대학교 통계학과 석사학위논문.

[5] 이석훈 (1998a), 확률적으로 분류된 개체들의 판별분석. 한국분류학회 Vol. 2, 7-19.

[6] 이석훈, 이경희 (1998b), 분류기준의 세분화에 의하여 불완전해진 구학습표본을 포함하는 판별분석. Unpublished manuscript.

[7] 통계로 본 대한민국 50년의 경제사회상 변화(1998), 97, 통계청.

[1999년 2월 접수, 1999년 3월 최종수정]

부록

2.1절의 로그 우도함수에서 구한 $p=1$인 경우의 이차 편도함수는 아래와 같다.

$g_{11}=\frac{\partial^2 l_2}{\partial\beta_1^2}=-C_2\cdot SSX-D_2$, $g_{12}=g_{21}=\frac{\partial^2 l_2}{\partial\beta_1\partial\beta_2}=C_4\cdot SSX-D_2$,

$g_{13}=g_{31}=\frac{\partial^2 l_2}{\partial\sigma_{11}\partial\beta_1}=-\sigma_{22}F_2-F_1$, $g_{14}=g_{41}=\frac{\partial^2 l_2}{\partial\sigma_{22}\partial\beta_1}=F_5-\sigma_{11}F_2-F_1$,

$g_{15}=g_{51}=\frac{\partial^2 l_2}{\partial\sigma_{12}\partial\beta_1}=F_4+2\sigma_{12}F_2-2F_1$, $g_{22}=\frac{\partial^2 l_2}{\partial\beta_2^2}=-C_3\cdot SSX-D_2$,

$g_{23}=g_{32}=\frac{\partial^2 l_2}{\partial\sigma_{11}\partial\beta_2}=-F_4-\sigma_{22}F_3-F_1$, $g_{24}=g_{42}=\frac{\partial^2 l_2}{\partial\sigma_{22}\partial\beta_2}=-\sigma_{11}F_3-F_1$,

$g_{25}=g_{52}=\frac{\partial^2 l_2}{\partial\sigma_{12}\partial\beta_2}=-F_5+2\sigma_{12}F_3-2F_1$, $g_{33}=\frac{\partial^2 l_2}{\partial\sigma_{11}^2}=\frac{n}{2}C_2^2+H_1-G_1C_2$,

$g_{34}=g_{43}=\frac{\partial^2 l_2}{\partial\sigma_{11}\partial\sigma_{22}}=\frac{n}{2}C_4^2+H_1+C_4G_3$

$g_{35}=g_{53}=\frac{\partial^2 l_2}{\partial\sigma_1 1\partial\sigma_{12}}=-nC_2C_4+2H_1+2C_2^2C_4S_{11}-\frac{C_2P_1S_{12}}{(DET)^2}+\frac{P_2S_{22}}{(DET)^3}$,

$g_{44}=\frac{\partial^2 l_2}{\partial\sigma_2^2}=\frac{n}{2}C_3^2+H_1-C_3G_2$,

$g_{45}=g_{54}=\frac{\partial^2 l_2}{\partial\sigma_{22}\partial\sigma_{12}}=-nC_3C_4+2H_1+\frac{P_2S_{11}}{(DET)^3}-\frac{C_3P_1S_{12}}{(DET)^2}+2C_3^2C_4S_{22}$,

$g_{55}=\frac{\partial^2 l_2}{\partial\sigma_{12}^2}=n(C_2C_3+C_4^2)+4H_1-\frac{C_2P_1S_{11}}{(DET)^2}+6C_2C_3C_4S_{12}+2C_4^3S_{12}-\frac{C_3P_1S_{22}}{(DET)^2}$

위에서 $D_1=(VarZ)^{-1}\{SSZW-SSW\cdot(\beta_1+\beta_2)\}$, $D_2=(VarZ)^{-1}\cdot SSW$,

$A_{11}=\sum_{i=1}^n y_{1i}\cdot x_i$, $A_{12}=\sum_{i=1}^n y_{2i}\cdot x_i$, $SSX=\sum_{i=1}^n x_i^2$, $SSW=\sum_{i=1}^k w_i^2$,

$E_1=\sigma_{22}A_{11}-\sigma_{12}A_{12}-SSX(\sigma_{22}\beta_1-\sigma_{12}\beta_2)$, $E_2=-\sigma_{12}A_{11}+\sigma_{11}A_{12}-SSX(-\sigma_{12}\beta_1+\sigma_{11}\beta_2)$,

$F_1=(VarZ)^{-1}D_1$, $F_2=(DET)^{-2}E_1$, $F_3=(DET)^{-2}E_2$,

$F_4=(DET)^{-1}(-A_{12}+SSX\cdot\beta_2)$, $F_5=(DET)^{-1}(A_{11}-SSX\cdot\beta_1)$,

$H_1=\frac{k}{2(VarZ)^2}-\frac{SSZ}{(VarZ)^3}$, $P_1=\sigma_{11}\sigma_{22}+3\sigma_{12}^2$, $P_2=\sigma_{11}\sigma_{22}\sigma_{12}+\sigma_{12}^3$이다.

The effect of adding the summed univariate data to the bivariate data in regression model*

Nae-Hyun Park[1] Sukhoon Lee[2] No-Mahn Kim[3]

ABSTRACT

This research deals with the problem of using the summed univariate data of originally bivariate data as well as the bivariate data in the regression model.

The model with one explanatory variable is considered to invetigate the relationship between the effect of adding the summed data and the bivariate model. Here the effect is measured by the size of the mean squared error.

The large sample properties are drawn through the maximum likelihood estimator property and also a simulation study is carried out for the small sample. Finally, an analysis with a real data is shown as an illustration.

* The authors wish to acknowledge the financial support of the Korea Research Foundation made in the program year of 1997.

1) Department of Statistics, College of Natural Science, Chungnam National University, Taejon 305-764, Korea
2) Department of Statistics, College of Natural Science, Chungnam National University, Taejon 305-764, Korea
3) Agency For Defence Development, P.O Box 35, Yusung Post Office