回歸分析에서의 觀測値統合과 有效性

鄭 基 俊*

〈目 次〉

I. 序 論

經濟現象의 分析에 利用되는 統計資料는 元來의 調査된 상태로서는 지나치게 細分되어 있거나 방대하기 때문에 어떤 방법으로든지 統合의 과정을 거치게 되는 것이 보통이다. 그러나 統合의 과정에서는 거의 불가피하게 원래의 資料가 갖고 있는 情報의 일부분을 상실하게 되며, 이 情報의 상실의 정도는 統合의 방법에 따라서 달라질 것이다.

이 統計資料의 統合의 문제를 線形回歸分析과 관련하여 생각해 보면, 이 情報喪失의 정도는, 母數推定量의 有效性에 의해서 表現될 수 있을 것이다. 觀測된 統計資料의 統合에 따르는 有效性의 상실의 문제는, 프레이즈와 에이치슨[3], 타일[4], 尹錫範[1] 등에서 다루어지고 있다. 그러나 그 問題를 보다 명쾌하게 이해하는 데는 旣存의 문헌이 不足함이 있다고 생각된다.

本所論에서는 觀測値統合에 따르는 有效性의 喪失을 새로운 방법으로 명확히 밝히고, 이를 근거로 하여, 바람직한 統合의 方法에 관해서 고찰하기로 한다.

II. 豫備的 考察—說明變數가 하나 뿐인 경우

觀測値의 統合에 따라서 발생하는 문제의 本質을 간단히 파악하기 위하여, 우선 가장 단순한 線形模型을 가지고 觀測値統合의 效果를 분석해 보기로 한다.

* 本硏究所 硏究員, 서울大學校 經濟學科 副敎授

설명변수가 하나 뿐인 線形模型,

$$y_i=\beta x_i+\varepsilon_i, \quad \varepsilon_i\sim\text{iidN}(0,\sigma^2), \quad i=1,\cdots,n \tag{1}$$

을 고려해 보자. 그리고 n개의 觀測値가 $n_1,n_2,\cdots,n_G$개씩 G個의 群으로 統合되는 경우의 有效性의 變化를 알아 보기 위하여 다음과 같이 群平均 $\bar{x}_g$ 等을 定義한다.

$$\bar{x}_g=\frac{1}{n_g}\sum_{i\in s_g} x_i, \quad g=1,\cdots,G.$$

$$\bar{y}_g=\frac{1}{n_g}\sum_{i\in s_g} y_i, \quad g=1,\cdots,G.$$

$$\bar{\varepsilon}_g=\frac{1}{n_g}\sum_{i\in s_g} \varepsilon_i, \quad g=1,\cdots,G.$$

단, $n_1+\cdots+n_G=n$이며, s_g는 g번째 群의 觀測値番號集合이다. 그리고 s_g 속에는 n_g개의 番號가 들어 있고, 集合 $\{1,2,\cdots,n\}$은 $s_1,\cdots,s_G$로 分割된다.

그러면 統合된 模型은 다음과 같이 된다.

$$\bar{y}_g=\beta\bar{x}_g+\bar{\varepsilon}_g, \quad \bar{\varepsilon}_g\sim\text{iidN}\left(0,\frac{\sigma^2}{n_g}\right), \quad g=1,\cdots,G. \tag{2}$$

이 統合된 模型에서, $\bar{\varepsilon}_g$의 分散은 $\frac{\sigma^2}{n_g}$으로 n_g가 모두 같지 않은 한(우리는 같다고 가정하지 않는다.) 이 模型은 異分散(heteroscedastic)模型이 된다. 즉 等分散模型인 원래의 模型 (1)에서는 OLS를 적용하여 β의 有效推定量을 얻을 수 있으나, 模型 (2)에는 GLS를 적용해야만 β의 有效推定量 $\hat{\beta}$을 얻을 수 있다. 그런데 模型 (2)와 같은 異分散模型의 경우에는 GLS는 WLS와 같으므로 이 模型의 兩邊에 加重値 $\sqrt{n_g}$를 곱해 보면 다음과 같다.

$$\sqrt{n_g}\ \bar{y}_g=\beta\sqrt{n_g}\,\bar{x}_g+\sqrt{n_g}\bar{\varepsilon}_g, \quad \sqrt{n_g}\bar{\varepsilon}_g\sim\text{iidN}(0,\sigma^2), \quad g=1,\cdots,G. \tag{3}$$

즉 模型 (2)의 GLS 推定量 $\hat{\beta}$은 模型 (3)의 OLS 推定量과 같고 따라서 $\hat{\beta}$은 다음과 같이 된다.

$$\hat{\beta}=\frac{\sum_{g=1}^{G} n_g\bar{x}_g\bar{y}_g}{\sum_{g=1}^{G} n_g\bar{x}_g^{\,2}}. \tag{4}$$

그리고 $\hat{\beta}$의 分散은 다음과 같이 된다.

$$\mathrm{var}(\hat{\beta})=\frac{\sigma^2}{\sum_{g=1}^{G} n_g \bar{x}_g{}^2}. \tag{5}$$

그런데 群으로 統合하지 않은 模型 (1)에 OLS를 적용하여 얻는 β의 推定量 b와 그 分散은 各各 다음과 같다.

$$b=\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i{}^2}. \tag{6}$$

$$\mathrm{var}(b)=\frac{\sigma^2}{\sum_{i=1}^{n} x_i{}^2}. \tag{7}$$

그러면 여기서 式 (7)로 정의되는 b의 分散과 式 (5)로 定義되는 $\hat{\beta}$의 分散을 比較해 보자. 이 比較를 위해서는 그 두 式의 右邊의 分子는 共通이므로, 分母를 比較해 보면 된다. 그리하여 이 分母의 差를 구해보면 다음과 같다.

$$\sum_{i=1}^{n} x_i{}^2-\sum_{g=1}^{G} n_g \bar{x}_g{}^2=\sum_{g=1}^{G}\sum_{i\epsilon s_g} x_i{}^2-\sum_{g=1}^{G} n_g \bar{x}_g{}^2=\sum_{g=1}^{G}\{\sum_{i\epsilon s_g} x_i{}^2-n_g \bar{x}_g{}^2\}$$

$$=\sum_{g=1}^{G}\{\sum_{i\epsilon s_g}(x_i-\bar{x}_g)^2\}\geqq 0$$

여기서 첫째 等式은 n개의 項을 G개의 群으로 나눈 것을 고려하면 쉽게 理解될 것이며 세째 等式은 큰 괄호 내의 表現들 사이에 성립하는 恒等關係 때문에 성립한다. 그리고 마지막 不等式은 모든 $x_i(i\epsilon s_g)$가 $\bar{x}_g$와 같을 때에만 等式이 되고, 그렇지 않으면 强不等式이 된다. 이 關係로부터

$$\mathrm{var}(\hat{\beta}) \geqq \mathrm{var}(b) \tag{8}$$

의 關係를 얻는다. 즉 모든 $g(g=1,\cdots,G)$에 관하여, s_g 속의 모든 i에 관한 x_i가 $\bar{x}_g$와 一致할 때는 $\hat{\beta}$의 분산과 b의 分散이 一致하지만 그렇지 않은 경우에는 언제나 $\hat{\beta}$의 分散은 b의 分散보다 크다. 즉 觀測値의 統合에 의해서 有效性을 상실한다.

以上의 內容을 一般化하기 위한 豫備作業으로서, 그 內容을 行列로 나타내면 다음과 같다. 즉 線形模型 (1)은

$$\underset{n\times 1}{y} = \underset{n\times 1}{x\beta} + \underset{n\times 1}{\varepsilon}, \qquad \varepsilon\sim \mathrm{N}(0,\sigma^2 I_n)$$

으로 되며, 統合된 模型 (2)는

$$\bar{y}=\bar{x}\beta+\varepsilon, \qquad \varepsilon\sim N(0,\sigma^2JJ')$$

로 된다. 단,

$$\bar{y}=[\bar{y}_1,\cdots,\bar{y}_G]'=Jy,$$

$$\bar{x}=Jx$$

$$\varepsilon=J\varepsilon$$

이며 $G\times n$ 行列 J는

$$J=\begin{pmatrix} \frac{1}{n_1}\iota_1' & & 0 \\ & \frac{1}{n_2}\iota_2' & \\ & & \ddots \\ 0 & & \frac{1}{n_G}\iota_G' \end{pmatrix} \tag{9}$$

로 정의된다. 단, $\iota_g'=[1,\cdots,1]$인 n_g차원 벡터이다. 그리고 JJ'는

$$JJ'=\begin{pmatrix} \frac{1}{n_1^2}\iota_1'\iota_1 & & 0 \\ & \frac{1}{n_2^2}\iota_2'\iota_2 & \\ & & \ddots \\ 0 & & \frac{1}{n_G^2}\iota_G'\iota_G \end{pmatrix}=\begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \frac{1}{n_2} & \\ & & \ddots \\ 0 & & \frac{1}{n_G} \end{pmatrix} \tag{10}$$

로 됨을 알 수 있다.

III. 一般線形模型에서의 觀測値統合

위의 설명은 다음과 같이 一般化될 수 있다. 즉 一般線形模型을

$$\underset{n\times 1}{y}=\underset{n\times k,}{X}\underset{k\times 1}{\beta}+\underset{n\times 1}{\varepsilon}, \qquad \varepsilon\sim N(0,\sigma^2I) \tag{11}$$

라 하고, 이를 처음부터 $n_1,\cdots,n_G$개씩 통합하여 그 平均에 관한 模型을 만들면,

$$\underset{G\times 1}{Jy}=\underset{G\times k,}{JX}\underset{k\times 1}{\beta}+\underset{G\times 1}{J\varepsilon}, \qquad J\varepsilon\sim N(0,\sigma^2JJ') \tag{12}$$

가 된다. 단, J는 前節의 (9)로 定義된 $G\times n$行列이다.

이 變形된 模型에 GLS를 적용하여 구한 β의 推定量을 $\hat{\beta}$이라 하면, 이는 다음과 같이 표현된다.

$$\begin{aligned}\hat{\beta}&=[X'J'(JJ')^{-1}JX]^{-1}X'J'(JJ')^{-1}Jy\\&=\beta+[X'J'(JJ')^{-1}JX]^{-1}X'J'(JJ')^{-1}J\varepsilon.\end{aligned} \tag{13}$$

그리하여 $\hat{\beta}$의 分布는

$$\hat{\beta}\sim N(\beta,\ \sigma^2[X'J'(JJ')^{-1}JX]^{-1}) \tag{14}$$

로 나타낼 수 있다.

그런데 統合을 하지 않은 상태에서 原模型 (11)의 LS推定量을 b라 하면,

$$b=(X'X)^{-1}X'y=\beta+(X'X)^{-1}X'y \tag{15}$$

가 되고, b의 分布는

$$b\sim N(\beta,\sigma^2[X'X]^{-1}) \tag{16}$$

로 나타낼 수 있다. 여기서 b와 $\hat{\beta}$의 有效性을 比較하는 다음 사실을 증명하고자 한다. 즉

「β의 推定量으로서 b는 $\hat{\beta}$보다 有效하다.」

는 것을 보이고자 한다.

이는 다음과 같은 定理의 형태로 나타낼 수 있다.

定理 : 式(13) 및 (15)로 定義되는 β의 推定量 $\hat{\beta}$과 b에 관해서는 $\hat{\beta}$의 共分散行列에서 b의 共分散行列을 뺀 것이 陽半定符號行列이라는 의미에서 $\hat{\beta}$은 b보다 有效하다.

이 定理를 증명하기 위해서는 다음과 같은 일련의 補助定理들을 고려하는 것이 편리하다.

補助定理 1 : $I-J'(JJ')^{-1}J$는 陽半定符號行列이다.

補助定理 2 : A가 陽半定符號行列이면 $X'AX$는 陽半定符號行列이다.

補助定理 3 : A와 B가 모두 陽定符號行列이고, $A-B$가 陽半定符號行列이면, $B^{-1}-A^{-1}$는 陽半定符號行列이다.

證明 : 먼저 補助定理 1에서는 $I-J'(JJ')^{-1}J$가 對稱冪等行列임을 쉽게 確認할 수 있다. 그리고 대칭멱등행렬의 特性根은 반드시 0 또는 1이므로, 그 行列은 陽半定符號行列이다. 補助定理 2는 A가 陽半定符號行列이면, $A=B'B$를 充足하는 行列 B가 반드시 존재한다는 사실에 의해서 증명할 수 있다. 즉,

$$y'X'AXy=y'X'B'BXy=Z'Z\geq 0,\ \text{단 } Z=BXy.$$

즉 모든 벡터 y에 관하여 $y'X'AXy\geq 0$가 성립하므로, $X'AX$는 陽半定符號行列이다.

補助定理 3의 證明은 드라입즈[2, p. 583]를 참고할 수 있다. (그러나 드라입즈는 $A-B$가 陽定符號行列일 때, $B^{-1}-A^{-1}$가 陽定符號인 경우를 證明하고 있다.) 먼저 A가 陽定符號行列이면,

$$A=WW'$$

를 충족하는 非特異行列 W가 存在한다는 사실을 認定하는 것으로부터 出發하자. 그러면,

$|\lambda A-B|=0$를 충족하는 特性根 λ는 어떤 性質을 가질 것인가를 보기로 하자. 이 特性方程式에 A대신 WW'를 代入하고, W가 非特異行列임을 고려하면,

$$0=|\lambda A-B|=|\lambda WW'-B|=|W(\lambda I)W'-WW^{-1}BW^{-1\prime}W'|$$
$$=|W\{\lambda I-W^{-1}BW^{-1\prime}\}W'|=|W|^2|\lambda I-W^{-1}BW^{-1\prime}|.$$

그런데 $|W|\neq 0$이므로,

$$|\lambda I-W^{-1}BW^{-1\prime}|=0$$

이다. 그런데 B가 陽定符號行列이면, $W^{-1}BW^{-1\prime}$ 역시 陽定符號行列이며, 그 逆도 成立하므로, B가 陽定符號行列이면, λ는 陽數이다. 이러한 λ를 對角元素로 하는 對角行列을 Λ라 하고, 그에 대응하는 特性벡터로 이루어지는 直交行列을 C라 하면,

$$W^{-1}BW^{-1\prime}=C\Lambda C'$$

의 관계가 항상 성립한다. 이를 變形하면,

$$B=WC\Lambda C'W'$$

로 쓸 수 있다. 그리고,

$$WCC'W'=WIW'=WW'=A$$

의 관계가 성립한다. 그러므로,

$$W^*=WC$$

라고 정의하면, W^*는 非特異行列이며,

$$A=W^*W^{*\prime},\quad B=W^*\Lambda W^{*\prime}$$

의 관계가 成立한다. 여기서 A와 B의 差를 만들어 보면,

$$A-B=W^*(I-\Lambda)W^{*\prime}$$

가 된다. 그러므로, $A-B$가 陽半定符號行列이 되려면,

$$0<\lambda_i\leq 1$$

임을 알 수 있다. 앞의 不等號는 앞에서 이미 증명한 바 있고, 뒤의 不等號는 $A-B$가 陽半定符號라는 데서 나오는 조건이다. 또 이 命題의 逆도 成立하는 것을 보일 수 있다.

다음은 B^{-1}와 A^{-1}의 差를 알아보기 위하여, B^{-1}와 A^{-1}를 각각

$$B^{-1}=W^{*\prime-1}\Lambda^{-1}W^{*-1}$$

$$A^{-1}=W^{*\prime-1}W^{*-1}$$

라고 表現해 보자. 그러면,

$$B^{-1}-A^{-1}=W^{*\prime-1}(\Lambda^{-1}-I)W^{*-1}.$$

그런데 $0<\lambda_i\leq 1$이므로, Λ^{-1}의 對角元素 $\frac{1}{\lambda_i}$은,

$$1 \leq \frac{1}{\lambda_i} < \infty$$

가 되며, 따라서, $A^{-1}-I$의 對角元素는 모두 非陰數이다. 그러므로, $A^{-1}-I$는 陽半定符號行列이고, 따라서 $B^{-1}-A^{-1}$도 陽半定符號行列이다. 이로써 補助定理 3의 證明이 끝났다.

本定理의 證明은 앞의 補助定理들을 綜合하여 얻을 수 있다. 즉 $X'X$ 및 $X'[J'(JJ')^{-1}J]X$는 모두 陽定符號行列인데, 그 差 즉 $X'X-X'[J'(JJ')^{-1}J]X$는 陽半定符號行列이다. 따라서 補助定理 3에 의하면, $\{X'[J'(JJ')^{-1}]X\}^{-1}-(X'X)^{-1}$는 陽半定符號行列이다. 그런데 우리는 $\hat{\beta}$과 b의 共分散行列이 각각

$$\mathrm{var}(\hat{\beta}) = \sigma^2[X'J'(JJ')^{-1}JX]^{-1}$$

$$\mathrm{var}(b) = \sigma^2[X'X]^{-1}$$

임을 알고 있으므로, 결국

$$\mathrm{var}(\hat{\beta}) - \mathrm{var}(b) = \sigma^2\{[X'J'(JJ')^{-1}JX]^{-1} - (X'X)^{-1}\} = \sigma^2 D \qquad (17)$$

는 陽半定符號行列이다. 이로써 증명이 完了되었다.

IV. 有效性의 상실 없는 統合

그러면 어떤 종류의 統合이 有效性의 상실을 막을 수 있을까? 이를 위해서는 $D=0$가 되어야 할 것이며, $D=X'M_JX$로 쓸 수 있으므로 우선

$$X'X-X'[J'(JJ')^{-1}J]X = X'[I-J'(JJ')^{-1}J]X$$

에서 $M_J=I-J'(JJ')^{-1}J$를 分析하는 것이 편리하다.

式 (10)을 고려하면서, M_J를 變形하면 다음과 같다.

$$M_J = I-J'(JJ')^{-1}J$$

$$= \begin{pmatrix} I_1 & & 0 \\ & I_2 & \\ & & \ddots \\ 0 & & I_G \end{pmatrix} - \begin{pmatrix} \frac{1}{n_1}\iota_1 & & 0 \\ & \frac{1}{n_2}\iota_2 & \\ & & \ddots \\ 0 & & \frac{1}{n_G}\iota_G \end{pmatrix} \begin{pmatrix} n_1 & & 0 \\ & n_2 & \\ & & \ddots \\ 0 & & n_G \end{pmatrix} \begin{pmatrix} \frac{1}{n_1}\iota_1' & & 0 \\ & \frac{1}{n_2}\iota_2' & \\ & & \ddots \\ 0 & & \frac{1}{n_G}\iota_G' \end{pmatrix}$$

$$= \begin{pmatrix} I_1 & & 0 \\ & I_2 & \\ & & \ddots \\ 0 & & I_G \end{pmatrix} - \begin{pmatrix} \frac{1}{n_1}\iota_1\iota_1' & & 0 \\ & \frac{1}{n_2}\iota_2\iota_2' & \\ & & \ddots \\ 0 & & \frac{1}{n_G}\iota_G\iota_G' \end{pmatrix} = \begin{pmatrix} A_1 & & 0 \\ & A_2 & \\ & & \ddots \\ 0 & & A_G \end{pmatrix}.$$

단, A_g는 다음과 같이 定義되는 $n_g \times n_g$行列이다.

$$A_g = I_g - \frac{1}{n_g} \iota_g \iota_g'.$$

그리고 I_g는 $n_g \times n_g$ 單位行列이다.

여기서 行列 D도 X를 적절히 分割하여 다음과 같이 표현해 보자.

$$D = X'M_J X = [X_1' \; X_2' \cdots X_G'] \begin{pmatrix} A_1 & & & 0 \\ & A_2 & & \\ & & \ddots & \\ 0 & & & A_G \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_G \end{pmatrix}$$

$$= X_1'A_1X_1 + X_2'A_2X_2 + \cdots + X_G'A_GX_G.$$

여기서 X_g는 $n_g \times k$行列이다. 그리고

$$X_g'A_gX_g, \quad g = 1, \cdots, G$$

를 다시 分割하여 表現해 보면, $n_g \times k$行列 X_g의 j번째의 벡터를 x_{gj}라 할 때,

$$X_g'A_gX_g = \begin{pmatrix} x_{g1}' \\ \vdots \\ x_{gk}' \end{pmatrix} A_g [x_{g1} \; \cdots \; x_{gk}] = \begin{pmatrix} x_{g1}'A_gx_{g1} & x_{g1}'A_gx_{g2} \cdots x_{g1}'A_gx_{gk} \\ \vdots & \\ x_{gk}'A_gx_{g1} & x_{gk}'A_gx_{g2} \cdots x_{gk}'A_gx_{gk} \end{pmatrix}$$

가 된다. 그런데

$$\frac{1}{n_g} x_{gj}'A_gx_{gj}, \qquad g = 1, \cdots, G, \quad j = 1, \cdots, k$$

는 g번째 群에서 j번째 變數의 標本分散으로 解釋할 수 있고,

$$\frac{1}{n_g} x_{gi}'A_gx_{gj}, \qquad g = 1, \cdots, G, \quad i, j = 1, \cdots, k$$

는 g번째 群에서 i번째 變數와 j번째 變數의 標本共分散이라 解釋할 수 있다. 그러므로, $k \times k$ 行列인

$$\frac{1}{n_g} X_g'A_gX_g, \qquad g = 1, \cdots, G$$

는 g번째 群의 k개의 變數들의 分散共分散行列이 된다. 그런데 유효성의 상실이 없으려면, $X'M_JX = 0$ 즉, $X_g'A_gX_g = 0$이어야 한다. 이것은 b와 $\hat{\beta}$을 비교하여 $\hat{\beta}$이 b에 비하여 有效性을 잃지 않기 위해서는 統合된 모든 群에서 變數들의 分散 및 共分散이 모두 0이 되어야 한다는 말이 되는 셈이다.

그런데 g群의 j變數의 分散이 0이면,

$$x_{gj}'A_gx_{gj} = (A_gx_{gj})'(A_gx_{gj}) = 0$$

에서,

$$A_g x_{gj}=0, \quad g=1,\cdots,G, \quad j=1,\cdots,k \tag{18}$$

이다. 따라서, 共分散도

$$x_{gi}' A_g x_{gj}=0$$

이 된다. 그러면 式 (18)의 의미는 무엇인가? 이 式에서 A_g는 $n_g \times n_g$行列로서, 그 位數(rank)는 n_g-1이다. 그리고 그 零空間은 ι_g에 의해서 생성되는 1次元이다. 그러므로, 式 (18)이 성립하려면, x_{gj}가 ι_g와 선형종속의 관계에 있어야 한다. 이 말은, x_{gj}가 동일한 원소로 된 n_g차원 벡터임을 의미한다.

그러므로 統合된 모형에서의 β의 추정량 $\hat{\beta}$이 b에 비하여 有效性을 상실하지 않기 위해서는, 統合된 各群 內部에서 각 설명변수가 각각 同一한 값을 가져야 한다고 말할 수 있다.(1)

參 考 文 獻

〔1〕 尹錫範, 『計量經濟學』, 法文社, 서울, 1978.

〔2〕 Dhrymes, P.J., *Econometrics: Statistical Foundations and Applications*, Harper and Row, New York, 1970.

〔3〕 Prais, S.J., and J. Aitchison, "The Grouping of Observations in Regression Analysis," *Review of the International Statistical Institute*, 22 (1954), pp. 1-22.

〔4〕 Theil, H., *Principles of Econometrics*, New York: Wiley, 1971.

(1) 尹錫範교수 〔1〕는 統合된 模型에서의 β의 GLS推定量 $\hat{\beta}$과, 그 模型에서의 β의 OLS推定量의 有效性을 비교해 주고 있다. 그러나 문제의 本質上 통합된 모형에서의 GLS추정량과 통합되기 전의 모형에서의 OLS추정량(이는 GLS추정량이기도 하다.)을 비교하는 것이 더 큰 의의가 있을 것이다.