Journal of the Korean Data & Information Science Society
2009, **20**(2), 251–260

한국데이터정보과학회지

단순 선형회귀 모형에서 자기공분산에 근거한 최적 추정 방법

박철용[1]

계명대학교 통계학과

접수 2009년 1월 20일, 수정 2009년 3월 13일, 게재확정 2009년 3월 18일

요 약

이 논문에서는 단순 선형회귀 모형에서 회귀 계수의 최적 추정량을 구할 수 있는 자기공분산에 근거한 추정 방법을 제시하였다. 이 방법이 직관적으로 매혹적이지는 않지만 이 최적 추정량이 해당 회귀 계수의 불편추정량이 된다. 설명변수가 0과 1사이의 균등간격의 값을 가지면, 오차가 자기회귀이동평균 모형을 따르면 성립하는 조건 하에서 이 최적 추정량이 최소제곱 추정량과 점근적으로 동일한 분포를 가진다는 것을 보였다. 추가적으로 똑 같은 조건 하에서 이 최적 추정량이 해당 회귀 계수에 확률상 수렴한다는 것을 자체적으로 입증하였다.

주요용어: 단순 선형회귀, 최소제곱법, 최적 자기공분산법.

1. 머리말

단순 선형회귀 모형 $Y_i = \alpha + \beta x_i + \epsilon_i,\ i = 1, 2, \ldots, n$에서 회귀 계수를 추정하는 문제는 너무나 많이 다뤄진 문제일 것이다. 고전적인 회귀 모형에서 부과하는 오차 ϵ_i에 대한 가정은 평균이 0이고 무상관이라는 것이다. 이 때 오차 ϵ_i에 대한 분포 가정 없이 회귀 계수의 추정 방법으로 가장 쉽게 접근할 수 있는 것이 최소제곱법이다. 이 최소제곱법에 의한 추정량은 가우스-마르코프 정리에 의해 최량 선형불편추정량(best linear unbiased estimator)이 된다고 알려져 있다 (Seber, 1977). 또한 정규분포 모집단을 가정하고 구할 수 있는 최대우도 추정량과 같아지게 되어 정규분포 모집단을 가정할 경우 최대우도 추정량이 가지는 여러 가지 (점근적) 우수성들을 보유하게 된다. 특정 모형에서 최소추정법이나 기타 추정법과의 비교 연구로는 이우동 (1996), 강희정과 김순영 (2000), Rahman과 Pearson (2003) 등이 있다.

이 연구는 단순한 호기심에서 출발하였다. 만약 (x_i, Y_i)가 시계열 자료라면 오차의 무상관성은 당연히 위배될 가능성이 높으며, 이럴 경우 무상관성을 보장해주는 방향으로 회귀 계수를 추정하게 되면 최소제곱법에 의한 추정량과 비슷해지지 않을까하는 궁금증이다. 그래서 가장 먼저 시도해 본 것이 1차 자기공분산(first order autocovariance)에 기초한 추정 방법이다.

당연히 가장 먼저 시도한 방법은 자기공분산이 0이 되도록 만들어주는 추정량이었다. 이 방법은 단순선형회귀 모형에서는 너무 과도하게 자기공분산이 0이 되도록 조정된 추정량을 얻어 만족스런 결과를 얻을 수 없었다. 그러나 이 개념을 발전시켜 비모수 회귀에 적용하여 Kim 등 (2004), Park 등 (2006) 같은 연구결과를 얻을 수 있었다.

다음에 시도한 것이 우리가 쉽게 접근할 수 최소 혹은 최대 기법이다. 다시 말해 자기공분산이 최소 혹은 최대가 되는 추정량을 구하는 것이다. 자기공분산이 회귀 계수의 값에 상관없이 항상 양 혹은 음인

[1] (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 교수. E-mail: cypark1@kmu.ac.kr

경우 각각 자기공분산이 최소 혹은 최대가 되는 추정량이 자기공분산을 0에 가깝게 만들어 주어 적절한 추정량이라고 생각될 수 있다. 그러나 자기공분산이 0을 만족하는 회귀 계수값이 존재할 경우에 사용하는 것은 직관에 반하는 것이라 무리가 있으리라 생각되었지만, 자기공분산이 0이 되는 추정량이 만족스런 결과를 제시하지 못했기 때문에 이 방법을 시도해 보았다. 그 결과 우선 두 가지 중요한 결과를 얻을 수 있었다. 하나는 자기공분산에 근거한 추정량이 회귀 계수의 불편추정량이라는 결과이며 다른 결과는 아주 특별한 설명변수 값 $x_i = i/n$을 가질 경우 오차가 자기회귀이동평균(ARMA; autoregressive and moving average) 모형을 따르면 성립되는 조건 하에서 이 추정량이 최소제곱법에 의한 추정량과 점근적으로 동일한 분포를 가진다는 결과이다. 추가적으로 똑 같은 조건 하에서 이 최적 추정량이 해당 회귀계수에 확률상 수렴(convergence in probability)한다는 것을 자체적으로 증명하였다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 자기공분산에 근거한 최적 추정량의 도출과 그 성질들을 입증하였다. 3절에서는 간단한 모의실험에 의해 2절에서 유도한 자기공분산에 근거한 최적 통계량의 성질이 소표본 하에서도 성립한다는 것을 보였다. 마지막으로 4절에서는 이 연구결과를 정리하고 결론을 내린다.

2. 자기공분산에 근거한 최적 추정량과 그 성질

이 절에서는 자기공분산에 근거한 최적 추정량의 유도와 그 성질을 살펴보게 된다. 구체적으로 2.1에서 이 최적 추정량을 유도하고, 2.2에서 이 추정량의 불편성을 보이고, 설명변수가 $x_i = i/n$의 값을 가지는 경우 오차가 자기회귀이동평균 모형을 따르면 만족되는 조건 하에서 이 최적 추정량이 최소제곱 추정량과 근사적으로 동일하다는 것을 보이게 된다. 추가적으로 똑같은 조건 하에서 이 최적 추정량이 회귀 계수에 확률상 수렴한다는 것을 자체적으로 입증하게 된다.

2.1. 자기공분산에 근거한 최적 추정량

자기공분산에 근거한 최적 추정량 유도에 앞서 몇 가지 표기법을 정의하도록 하자.

정의 2.1 회귀 모형 $Y_i = \alpha + \beta x_i + \epsilon_i$, $i = 1, 2, \ldots, n$에서 다음을 정의한다.

$$\tilde{\bar{Y}} = \sum_{i=1}^{n-1} (Y_i + Y_{i+1}) / \{2(n-1)\}, \quad \tilde{\bar{x}} = \sum_{i=1}^{n-1} (x_i + x_{i+1}) / \{2(n-1)\}.$$

정의 2.2 두 개의 벡터 $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{y} = (y_1, \ldots, y_n)$에 대해서 다음을 정의한다.

$$CP(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n-1} (x_i - \tilde{\bar{x}})(y_{i+1} - \tilde{\bar{y}}).$$

상기 정의 2.2를 이용하면 이 논문에서 주로 사용할 표기법인 $CP(\mathbf{x}, \mathbf{x}), CP(\mathbf{x}, \mathbf{Y})$ 및 $CP(\mathbf{Y}, \mathbf{x})$가 다음과 같이 정의되는 것을 알 수 있다.

$$CP(\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n-1} (x_i - \tilde{\bar{x}})(x_{i+1} - \tilde{\bar{x}}),$$

$$CP(\mathbf{x}, \mathbf{Y}) = \sum_{i=1}^{n-1} (x_i - \tilde{\bar{x}})(Y_{i+1} - \tilde{\bar{Y}}),$$

$$CP(\mathbf{Y}, \mathbf{x}) = \sum_{i=1}^{n-1} (Y_i - \tilde{\bar{Y}})(x_{i+1} - \tilde{\bar{x}}).$$

정의 2.3 회귀 모형 $Y_i = \alpha + \beta x_i + \epsilon_i,\ i = 1, 2, \ldots, n$에서 자기공분산에 근거한 회귀 계수에 대한 최적 추정량은

$$Q(\alpha, \beta) = \sum_{i=1}^{n-1} \epsilon_i \epsilon_{i+1} = \sum_{i=1}^{n-1} (Y_i - \alpha - \beta x_i)(Y_{i+1} - \alpha - \beta x_{i+1}) \tag{2.1}$$

을 α, β에 대해 편미분한 값이 0이 되는 $\tilde{\alpha}, \tilde{\beta}$이다.

상기 정의 2.3에 의해 구해지는 추정량은 직관적인 면에서 '최적'이라고 표현하기 힘든 면이 있다. 물론 $Q(\alpha, \beta)$가 항상 양수이거나 항상 음수인 경우에는 자기공분산이 0에 가깝도록 하는 추정량이 선택되기 때문에 최적이라 할 수 있다. 그러나 $Q(\alpha, \beta)$가 양수와 음수가 동시에 가능한 경우에는 자기공분산이 0이 되는 해인 두 개 추정량의 평균이 정의 2.3에 의한 추정량이 된다는 것 외에는 직관적인 설명이 어렵다. 그러나 나중에 나오는 정리에 의해 이 추정량이 여러 가지 좋은 성질을 만족하기 때문에 최적이라는 표현을 사용하였다.

상기 정의에서 자기공분산을 표본 자기공분산 형태로 $\sum_{i=1}^{n-1}(\epsilon_i - \tilde{\bar{\epsilon}})(\epsilon_{i+1} - \tilde{\bar{\epsilon}})$를 $Q(\alpha, \beta)$로 놓고 추정량을 구할 수도 있을 것이다. 그러나

$$\sum_{i=1}^{n-1}(\epsilon_i - \tilde{\bar{\epsilon}})(\epsilon_{i+1} - \tilde{\bar{\epsilon}}) = \sum_{i=1}^{n-1} \left\{ Y_i - \tilde{\bar{Y}} - \beta(x_i - \tilde{\bar{x}}) \right\} \left\{ Y_{i+1} - \tilde{\bar{Y}} - \beta(x_{i+1} - \tilde{\bar{x}}) \right\}$$

가 성립하여 절편에 대한 추정량을 구할 수 없는 문제가 발생하여 정의 2.3을 사용하였다. 그런데 β의 추정량은 두 방법에 의해 똑 같은 결과를 얻게 된다.

또한, $Q(\alpha, \beta)$를 α만의 함수로 고려하였을 때 이차항의 계수가 1인 이차함수가 되어 최소 자기공분산 추정량으로 해석될 수 있다. 그러나 $Q(\alpha, \beta)$를 β만의 함수로 고려하였을 때 $\sum x_i x_{i+1}$이 이차항의 계수가 되어 이것의 부호에 따라 최소 혹은 최대 자기공분산 추정량으로 해석될 수 있다. 이러한 점 때문에 최소 혹은 최대 추정량이라는 표현보다 최적 추정량이라는 표현을 사용하였다.

상기 정의들에 근거하여 다음과 같이 단순 선형 회귀 모형에서 자기공분산에 근거한 회귀 계수의 최적 추정량을 구할 수 있다.

정리 2.1 회귀 모형 $Y_i = \alpha + \beta x_i + \epsilon_i,\ i = 1, 2, \ldots, n$에서 자기공분산에 근거한 회귀 계수에 대한 최적 추정량은 다음과 같다.

$$\tilde{\alpha} = \tilde{\bar{Y}} - \tilde{\beta}\tilde{\bar{x}}, \quad \beta = \frac{CP(\mathbf{x}, \mathbf{Y}) + CP(\mathbf{Y}, \mathbf{x})}{2CP(\mathbf{x}, \mathbf{x})}.$$

증명: 먼저 $Q(\alpha, \beta)$의 α에 대한 편미분을 $\tilde{\alpha}, \tilde{\beta}$에서 계산한 $\partial Q(\tilde{\alpha}, \tilde{\beta})/\partial\alpha$가 0이 되어야 한다. 따라서 다음이 성립한다.

$$\frac{\partial Q(\tilde{\alpha}, \tilde{\beta})}{\partial \alpha} = -\sum_{i=1}^{n-1}(Y_i - \tilde{\alpha} - \tilde{\beta}x_i) - \sum_{i=1}^{n-1}(Y_{i+1} - \tilde{\alpha} - \tilde{\beta}x_{i+1}) = 0$$
$$\Leftrightarrow \tilde{\bar{Y}} - \tilde{\alpha} - \tilde{\beta}\tilde{\bar{x}} = 0 \ \Leftrightarrow\ \tilde{\alpha} = \tilde{\bar{Y}} - \tilde{\beta}\tilde{\bar{x}}.$$

마찬가지로 $Q(\alpha, \beta)$의 β에 대한 편미분을 $\tilde{\alpha}, \tilde{\beta}$에서 계산한 $\partial Q(\tilde{\alpha}, \tilde{\beta})/\partial\beta$가 0이 되어야 한다. 따라서

다음이 성립한다.

$$\begin{aligned}\frac{\partial Q(\tilde{\alpha},\tilde{\beta})}{\partial \beta} &= -\sum_{i=1}^{n-1} x_i(Y_{i+1}-\tilde{\alpha}-\tilde{\beta}x_{i+1}) - \sum_{i=1}^{n-1} x_{i+1}(Y_i-\tilde{\alpha}-\tilde{\beta}x_i) = 0 \\ &\Leftrightarrow \; -\sum_{i=1}^{n-1} x_i\left(Y_{i+1}-\tilde{\tilde{Y}}-\tilde{\beta}(x_{i+1}-\tilde{\tilde{x}})\right) - \sum_{i=1}^{n-1} x_{i+1}\left(Y_i-\tilde{\tilde{Y}}-\tilde{\beta}(x_i-\tilde{\tilde{x}})\right) = 0 \\ &\Leftrightarrow \; -CP(\mathbf{x},\mathbf{Y}) + \tilde{\beta}CP(\mathbf{x},\mathbf{x}) - CP(\mathbf{Y},\mathbf{x}) + \tilde{\beta}CP(\mathbf{x},\mathbf{x}) = 0 \\ &\Leftrightarrow \; \tilde{\beta} = \frac{CP(\mathbf{x},\mathbf{Y})+CP(\mathbf{Y},\mathbf{x})}{2CP(\mathbf{x},\mathbf{x})}\end{aligned}$$

이것으로 증명이 마무리 된다. □

식 (2.1)에 $\tilde{\alpha}$값을 대입하여 $Q(\tilde{\alpha},\beta)$를 구해보면 다음과 같다.

$$\begin{aligned}Q(\tilde{\alpha},\beta) &= \sum_{i=1}^{n-1}\left\{(Y_i-\tilde{\tilde{Y}})-\beta(x_i-\tilde{\tilde{x}})\right\}\left\{(Y_{i+1}-\tilde{\tilde{Y}})-\beta(x_{i+1}-\tilde{\tilde{x}})\right\} \\ &= CP(\mathbf{Y},\mathbf{Y}) - \beta\{CP(\mathbf{x},\mathbf{Y})+CP(\mathbf{Y},\mathbf{x})\} + \beta^2 CP(\mathbf{x},\mathbf{x}).\end{aligned}$$

따라서 $CP(\mathbf{x},\mathbf{x})>0$이면 $\tilde{\beta}$가 $Q(\tilde{\alpha},\beta)$를 최소화시키며, $CP(\mathbf{x},\mathbf{x})<0$이면 $\tilde{\beta}$가 $Q(\tilde{\alpha},\beta)$를 최대화시키는 것을 알 수 있다.

2.2. 자기공분산에 근거한 최적 추정량의 성질

먼저 자기공분산에 근거한 최적 추정량 $\tilde{\alpha},\tilde{\beta}$가 각각 회귀 계수 α,β의 불편추정량이 된다는 다음의 정리를 증명하고자 한다.

정리 2.2 회귀 모형 $Y_i=\alpha+\beta x_i+\epsilon_i,\ i=1,2,\ldots,n$에서 $E(\epsilon_i)=0$이라는 조건만 있으면 $E(\tilde{\alpha})=\alpha,\ \ E(\tilde{\beta})=\beta$를 만족한다.

증명: 간단한 계산에 의해 다음이 성립하는 것을 알 수 있다.

$$\begin{aligned}E(\tilde{\tilde{Y}}) &= \frac{\sum_{i=1}^{n-1}\{E(Y_i)+E(Y_{i+1})\}}{2(n-1)} = \frac{2\alpha+\beta\sum_{i=1}^{n-1}(x_i+x_{i+1})}{2(n-1)} = \alpha+\beta\tilde{\tilde{x}}, \\ E\{CP(\mathbf{x},\mathbf{Y})\} &= \sum_{i=1}^{n-1}(x_i-\tilde{\tilde{x}})E(Y_{i+1}-\tilde{\tilde{Y}}) = \sum_{i=1}^{n-1}(x_i-\tilde{\tilde{x}})[\beta(x_{i+1}-\tilde{\tilde{x}})] = \beta CP(\mathbf{x},\mathbf{x}), \\ E\{CP(\mathbf{Y},\mathbf{x})\} &= \sum_{i=1}^{n-1}E(Y_i-\tilde{\tilde{Y}})(x_{i+1}-\tilde{\tilde{x}}) = \beta CP(\mathbf{x},\mathbf{x}).\end{aligned}$$

따라서 다음이 성립한다.

$$\begin{aligned}E(\tilde{\beta}) &= \frac{E\{CP(\mathbf{x},\mathbf{Y})\}+E\{CP(\mathbf{Y},\mathbf{x})\}}{2CP(\mathbf{x},\mathbf{x})} = \frac{2\beta CP(\mathbf{x},\mathbf{x})}{2CP(\mathbf{x},\mathbf{x})} = \beta, \\ E(\tilde{\alpha}) &= E(\tilde{\tilde{Y}}) - E(\tilde{\beta})\tilde{\tilde{x}} = \alpha+\beta\tilde{\tilde{x}}-\beta\tilde{\tilde{x}} = \alpha.\end{aligned}$$

그러므로 두 추정량의 불편성이 입증되었다. □

만약 ϵ_i가 기댓값이 0이고 무상관이라면 가우스-마르코프 정리에 의해 최소제곱 추정량이 최량선형 불편추정량이 된다는 것이 잘 알려져 있다. 따라서 통상적인 회귀 분석의 가정 하에서는 자기공분산에 근거한 최적 추정량이 최소제곱 추정량보다 분산이 크게 나타나게 된다.

그래서 자기공분산에 근거한 최적 추정량이 최소제곱 추정량과 근사적으로 동일하게 되는 조건을 찾아보게 되었다. 그 중에 쉽게 생각해 볼 수 있었던 조건이 설명변수가 $x_i = i/n, i = 1, 2, \ldots, n$를 만족하는 경우이다. 이 경우 오차가 자기회귀이동평균 모형을 따르면 만족하는 조건 하에서 표본크기가 커짐에 따라 자기공분산에 근거한 최적 추정량과 최소제곱 추정량이 점근적으로 동일하다는 것을 다음과 같이 보일 수 있다.

정리 2.3 회귀 모형 $Y_i = \alpha + \beta x_i + \epsilon_i,\ i = 1, 2, \ldots, n$에서 $x_i = i/n, i = 1, 2, \ldots, n$이고 ϵ_i가 $E(\epsilon_i) = 0$, $\sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty$이며 정상적(stationary)이라면

$$\tilde{\alpha} = \hat{\alpha} + O_p(n^{-1}), \quad \tilde{\beta} = \hat{\beta} + O_p(n^{-1})$$

을 만족한다. 단, 여기서 $\gamma(k) = Cov(\epsilon_1, \epsilon_{1+k}) = Cov(\epsilon_n, \epsilon_{n-k}) = \gamma(-k)$, $\hat{\alpha}, \hat{\beta}$는 각각 α, β의 최소제곱 추정량이고, $O_p(n^{-1})$은 $O_p(n^{-1})/n^{-1}$이 확률상 유계(bounded in probability)인 확률변수이다.

증명: 먼저 간단한 계산에 의해 다음이 성립함을 유도할 수 있다.

$$\tilde{\bar{x}} = \bar{x}\frac{n}{n-1} - \frac{x_1 + x_n}{2(n-1)} = \bar{x} + O(n^{-1}),\ \tilde{\bar{Y}} = \bar{Y}\frac{n}{n-1} - \frac{Y_1 + Y_n}{2(n-1)} = \bar{Y} + O_p(n^{-1}).$$

이것을 이용하여 다음이 성립하는 것을 알 수 있다.

$$\begin{aligned}
CP(\mathbf{x}, \mathbf{x}) &= \sum_{i=1}^{n-1} x_i x_{i+1} - (n-1)\tilde{\bar{x}}^2 = \sum_{i=1}^{n-1} x_i \left[x_i + (x_{i+1} - x_i)\right] - (n-1)\tilde{\bar{x}}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 1 + (n-1)/(2n) - (n-1)\tilde{\bar{x}}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + O(1), \\
CP(\mathbf{x}, \mathbf{Y}) &= \sum_{i=1}^{n-1} x_i Y_{i+1} - (n-1)\tilde{\bar{x}}\tilde{\bar{Y}} = \sum_{i=1}^{n-1} \left[x_{i+1} + (x_i - x_{i+1})\right] Y_{i+1} - (n-1)\tilde{\bar{x}}\tilde{\bar{Y}} \\
&= \sum_{i=1}^{n} x_i Y_i - x_n Y_n - \bar{Y} + \frac{Y_1}{n} - (n-1)\tilde{\bar{x}}\tilde{\bar{Y}} = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) + O_p(1).
\end{aligned}$$

단, 여기서 $O(1)$는 유계(bounded)인 수열이다.

마찬가지 방법으로 $CP(\mathbf{Y}, \mathbf{x}) = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) + O_P(1)$임을 쉽게 보일 수 있다. 이 결과들을 종합하면 다음이 성립하게 된다.

$$\begin{aligned}
\tilde{\beta} &= \frac{CP(\mathbf{x}, \mathbf{Y}) + CP(\mathbf{Y}, \mathbf{x})}{2CP(\mathbf{x}, \mathbf{x})} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})/n + O_p(n^{-1})}{\sum (x_i - \bar{x})^2/n + O(n^{-1})} \\
&= \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})/n}{\sum (x_i - \bar{x})^2/n} + O_p(n^{-1}) = \hat{\beta} + O_p(n^{-1}), \\
\tilde{\alpha} &= \tilde{\bar{Y}} - \tilde{\beta}\tilde{\bar{x}} = \bar{Y} - \hat{\beta}\bar{x} + O_p(n^{-1}) = \hat{\alpha} + O_p(n^{-1}).
\end{aligned}$$

이것으로 증명이 마무리 된다. □

위의 정리 2.3에서 사용한 $x_i = i/n, i = 1, 2, \ldots, n$보다 훨씬 완화된 조건을 제시할 수도 있을 것이다. 이 문제는 4절에서 잠시 언급되겠지만 엄격하게 완화된 조건을 찾아내는 것이 그리 간단한 문제가 아니라고 판단되기 때문에 여기에서는 더 이상 추구하지 않고 추후과제로 남겨 놓기로 한다.

또한 ϵ_i가 $E(\epsilon_i) = 0$, $\sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty$이며 정상적(stationary)이라는 가정은 표본크기가 커지면 추정량 $\tilde{\alpha}, \tilde{\beta}$(혹은 $\hat{\alpha}, \hat{\beta}$)가 α, β에 수렴할 수 있도록 세운 강한 조건이다. 그러나 ARMA 모형에서도 성립되는 조건이기 때문에 (Brockwell과 Davis, 2006) 독립표본이라는 가정보다는 훨씬 완화되었음을 알 수 있다. 추정량의 수렴성은 다음의 정리에서 증명된다.

정리 2.4 정리 2.3의 조건 하에서 $\tilde{\alpha} = \alpha + o_p(1)$, $\tilde{\beta} = \beta + o_p(1)$이다. 단, 여기서 $o_p(1)$은 표본크기가 커짐에 따라 0에 수렴하는 확률변수이다.

증명: 정리 2.3에 의해 $\hat{\alpha} = \alpha + o_p(1)$, $\hat{\beta} = \beta + o_p(1)$을 보이면 된다. 먼저 $\hat{\beta}$를 정리하면 다음과 같이 된다.

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})\left[\beta(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon})\right])}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

그런데 $\bar{x} = 1/2 + O(n^{-1}), \sum x_i^2 = n/3 + O(1)$이기 때문에 $\sum x_i\epsilon_i/n = o_p(1)$, $\bar{\epsilon} = o_p(1)$을 보이게 되면 $\hat{\beta} = \beta + o_p(1)$을 입증하게 된다.

먼저 다음이 성립한다.

$$\begin{aligned} Var(\sum_{i=1}^n x_i\epsilon_i/n) &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n x_i x_j \gamma(i-j) \le \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n |\gamma(i-j)| \\ &= \frac{1}{n}\sum_{k=-(n-1)}^{n-1}(1 - k/n)|\gamma(k)| \le \frac{1}{n}\sum_{k=-(n-1)}^{n-1}|\gamma(k)|. \end{aligned} \tag{2.2}$$

따라서 표본크기가 커짐에 따라 $Var(\sum x_i\epsilon_i/n)$은 0으로 수렴하고 따라서 $\sum x_i\epsilon_i/n = o_p(1)$이 성립된다. 또한 $Var(\bar{\epsilon})$의 계산에서는 상기 유도식 (2.2)에서 $x_i = x_j = 1$을 대입하면 되기 때문에 바로 $\bar{\epsilon} = o_p(1)$을 유도할 수 있다. 이것으로 $\hat{\beta} = \beta + o_p(1)$이 입증되었다.

다음으로 $\hat{\alpha}$를 정리하면 $\hat{a} = \bar{Y} - \hat{\beta}\bar{x} = \alpha + \beta\bar{x} + \bar{\epsilon} - \hat{\beta}\bar{x} = \alpha + (\hat{\beta} - \beta)\bar{x} + \bar{\epsilon}$가 된다. 그런데 앞의 계산에서 $\hat{\beta} = \beta + o_p(1), \bar{\epsilon} = o_p(1)$을 보였기 때문에 $\hat{\alpha} = \alpha + o_p(1)$이 입증되었다. 이것으로 증명이 마무리 된다. □

3. 모의실험

이 절에서는 2절에서 얻은 자기공분산에 근거한 최적 추정량의 소표본 하에서의 성질을 확인하는 간단한 모의실험을 수행한다. 우리가 오차 ϵ_t의 분포로서 고려하는 것은 정리 2.3에 의해 자기공분산에 근거한 최적 추정량과 최소제곱 추정량이 점근적으로 동일한 분포를 가지는 ARMA 모형으로 R의 관례에 따라 다음과 같이 나타낼 수 있다.

$$\epsilon_t = \sum_{i=1}^p a(i) \times \epsilon_{t-i} + \delta_t + \sum_{j=1}^q m(j) \times \delta_t - j.$$

단, 여기서 δ_t는 평균이 0이고 무상관인 확률변수이다.

이 모의실험에서는 간단하게 독립표본, AR(1), MA(1), ARMA(1,1) 네 가지 모형만 고려하였다. 또한 표본크기 n으로는 시계열 자료에서는 그리 크다고 할 수 없는 100을 사용하였으며, $\alpha = \beta = 1$이며 $\{\delta_t\}$의 분포로는 표준정규분포에서의 확률표본이 사용되었다. 이렇게 $x_i = i/n, i = 1, 2, \ldots, n$인 회귀 모형 $Y_i = \alpha + \beta x_i + \epsilon_i,\ i = 1, 2, \ldots, n$에서 표본을 500개 생성하여 추정량들을 계산하였다.

이 모의실험의 결과를 표로서 정리한 것이 표 3.1이다.

표 3.1 자기공분산 추정량과 최소제곱 추정량의 평균 비교

모형	자기공분산 추정량	최소제곱 추정량
독립표본	0.9977(0.1948) 0.9920(0.3402)	0.9964(0.1942) 0.9920(0.3402)
a(1)=0.5	0.9669(0.4020) 1.0466(0.7021)	0.9682(0.3985) 1.0466(0.7021)
m(1)=-0.5	1.0029(0.1064) 0.9969(0.1833)	1.0021(0.1057) 0.9969(0.1833)
a(1)=0.5, m(1)=-0.5	1.0075(0.1956) 0.9777(0.3347)	1.0068(0.1947) 0.9777(0.3347)

이 표에는 자기공분산에 기초한 추정량 $\tilde{\alpha}, \tilde{\beta}$및 최소제곱 추정량 $\hat{\alpha}, \hat{\beta}$의 표본평균과 괄호 안에 표준오차가 나타나 있다. 먼저 두 방법의 평균 및 표준오차를 비교하면 최대 소숫점 셋째 자리에서의 차이가 있을 뿐이다. 또한 불편성의 관점에서 살펴보면 '|추정량-모수|'값이 '표준오차/$\sqrt{100}$'범위 안에 존재하는 것을 알 수 있다.

이 모의실험의 결과를 그림으로 정리한 것이 그림 3.1, 그림 3.2이다.

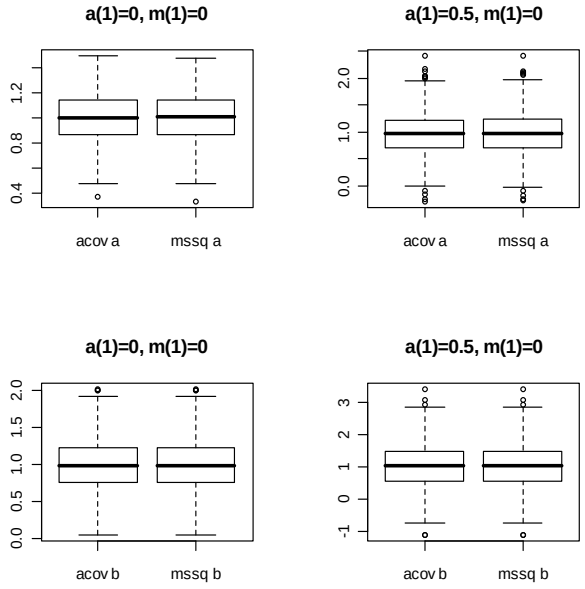

그림 3.1 독립표본과 AR(1) 모형에서의 비교

각 그림에는 두 개의 상자그림이 그려져 있는데, 왼쪽은 자기공분산에 근거한 추정량, 오른쪽은 최소

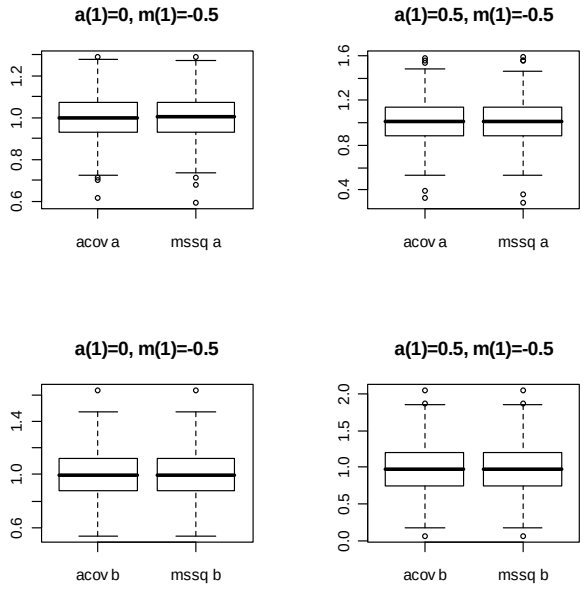

그림 3.2 MA(1)과 ARMA(1,1) 모형에서의 비교

제곱 추정량에 해당된다. 각 그림에는 자기공분산법 및 최소제곱법을 각각 acov, mssq라고 표시하고 있으며 절편과 기울기 추정량은 각각 a, b라고 표시하고 있다. 이 두 그림에 의하면 두 추정 방법에 의한 추정량의 분포가 거의 일치하고 있음을 알 수 있다.

4. 결론

이 연구에서는 단순 선형회귀 모형에서 회귀 계수를 추정하는 하나의 방법으로서 1차 자기공분산을 이용하는 방법을 제안하고 있다. 구체적으로 오차의 자기공분산을 각 회귀 계수에 대해 미분한 값이 0이 되도록 만들어 주는 추정량을 최적의 추정량으로 사용하게 된다. 이 추정량이 회귀 계수의 불편추정량이 되는 것을 보였다. 또한 설명변수가 $x_i = i/n,\ i = 1, 2, \ldots, n$이면 오차가 자기회귀이동평균 모형을 따르면 만족하는 조건 하에서 이 추정량이 최소제곱 추정량과 점근적으로 동일한 분포를 가지며 해당 회귀 계수로 확률상 수렴한다는 것을 보였다. 또한 자기회귀이동평균 모형에서의 간단한 모의실험을 통해 이 추정량의 성질들을 소표본 하에서 확인하였다.

이 연구를 지속하는 방향으로 다음의 두 가지를 생각할 수 있을 것이다. 먼저 이 추정량과 최소제곱 추정량의 점근적 동질성을 만족하는 완화된 충분조건을 찾는 것이다. 지금 현재 생각할 수 있는 완화된 조건은 증명과정에서 필요했던 최소한의 조건이다. 설명변수에 대해 필요한 조건은

$$\sum x_i^2 = c_1 n + O(1),\quad \bar{x} = c_2 + O(n^{-1}),\quad \sum x_i(x_{i+1} - x_i) = c_3 + O(n^{-1})$$

이다. 단, 여기서 c_1, c_2, c_3은 상수이다. 또한 반응변수에 대해 필요한 조건도

$$\sum_{k=-(n-1)}^{n-1} |\gamma(k)|/n = o(1)$$

으로 완화시킬 수 있으리라 생각된다. 이 조건은 k가 커짐에 따라 $\gamma(k)$가 0으로 수렴하면 만족되는 조건이다 (Wei, 1990). 연구의 다른 방향으로 자기공분산에 근거한 추정량이 최소제곱 추정량보다 더 좋은 결과를 보여주는 조건을 찾아보는 것도 좋은 추후연구과제가 될 수 있으리라 생각한다.

참고문헌

강희정, 김순영 (2000). 자귀회귀모형에서의 로버스트한 모수 추정방법들에 관한 연구. <한국데이터정보과학회지>, **11**, 1-8.

이우동 (1996). 임의중단모형에서 최소제곱법을 이용한 와이블분포의 모수 추정. <한국데이터정보과학회지>, **7**, 263-272.

Brockwell, P. J. and Davis, R. A. (2006). *Time series: Theory and methods*, 2nd Ed., Springer, New York.

Kim, T. Y., Kim, D., Park, B. U. and Simpson, D. (2004). Nonparametric detection of correlated errors. *Biometrika*, **91**, 491-496.

Park, B. U., Lee, Y. K., Kim, T. Y. and Park, C. (2006). A simple estimator of error correlation in nonparametric regression models. *Scandinavian Journal of Statistics*, **33**, 451-462.

Rahman, M. and Pearson, L. M. (2003). A note on estimating parameters in the two-parameter Weibull distribution. *Journal of the Korean Data & Information Science Society*, **14**, 1091-1102.

Seber, G. A. F. (1977). *Linear regression analysis*, John Wiley & Sons, New York.

Wei, W. W. S. (1990). *Time series analysis-univariate and multivariate methods*, Addison Wesley, New York.

Journal of the Korean Data & Information Science Society
2009, **20**(2), 251–260

한국데이터정보과학회지

An estimation method based on autocovariance in the simple linear regression model

Cheolyong Park[1]

Department of Statistics, Keimyung University

Received 20 January 2009, revised 13 March 2009, accepted 18 March 2009

Abstract

In this study, we propose a new estimation method based on autocovariance for selecting optimal estimators of the regression coefficients in the simple linear regression model. Although this method does not seem to be intuitively attractive, these estimators are unbiased for the corresponding regression coefficients. When the exploratory variable takes the equally spaced values between 0 and 1, under mild conditions which are satisfied when errors follow an autoregressive moving average model, we show that these estimators have asymptotically the same distributions as the least squares estimators. Additionally, under the same conditions as before, we provide a self-contained proof that these estimators converge in probability to the corresponding regression coefficients.

Keywords: Least squares method, optimal autocovariace method, simple linear regression.

[1] Professor, Department of Statistics, Keimyung University, Daegu 704-701. E-mail: cypark1@kmu.ac.kr