오차분산의 추정에 대한 고찰

김 종태*, 고 정환**

*712-714)경북 경산시 진량면 내리동 대구대학교 자연과학대학 통계학과 조교수

**(760-749)경북 안동시 송천동 388 안동대학교 자연과학대학 통계학과 부교수

요 약

비모수 회귀모형에 있어서의 오차분산을 추정하는 방법들 중 차분에 기저한 방법(difference-based methods)을 이용한 기존의 추정량들을 비교 분석하는데 목적이 있다. 특히 점근적인 최적 이차차분에 기저한 Hall과 Kay, Titterington(1990)의 HKT 추정량에 대한 그들의 추정량에 대한 문제점들을 제시하고, HKT추정량과, GSJS 추정량, Rice 추정량에 대하여 모의실험을 이용하여 모수에 대한 수렴속도를 비교 분석하였다. 또한 GSJS 추정량에 대한 일치성과 수렴 속도를 보였다.

1. 서 론

오차분산의 추정량이 통계적 자료분석과 통계적 추론에 미치는 영향력을 생각해 볼 때 회귀함수 f의 추정에 대한 연구는 모수적, 비모수적인 많은 연구들에 비하면 오차분산의 추정의 발전에 대한 연구의 노력은 상대적으로 미약하였다. 비모수 회귀모형에 있어서의 차분에 기저한 방법에 관한 오차분산에 대한 연구는 Rice(1984)의 연구에 의하여 처음 시작이 된 후 Gasser, Sroka와 Jennen-Steinmetz(1986)과Hall, Kay와 Titterington(1990,1991),Buckely와 Eagleson(1989), Thompson, Kay와 Titterington(1991)등의 학자들에 의해 연구되었다. 만약 회귀함수 f가 모수적 모형을 가질 경우에는 자연적으로 최소제곱 추정량을 사용할 수 있지만, 차분에 기저한 오차분산 추정의 장점은 자료의 회귀분석적인 역할 뿐 아니라 적합도 검정의 적용과 자료를 분석하기 위한 회귀함수의 스무드(smoothing) 정도를 선택하는 문제에도 중요한 역할을 한다.

이 연구의 목적은 차분에 기저한 오차분산 추정량으로 잘 알려진 Rice(1984)의 추정량과 Gasser와 Sroka, Jennen-Steinmetz(1986), (GSJS)의 추정량, Hall과 Kay, Titterington (1990), (HKT)의 추정량을 이용하여 위의 회귀모형에 대하여 어느 추정량이 오차분산을 얼마나 정확하게 잘 추정하는가에 대한 비교분석을 통해 실제 사용에 있어서 어떻게 적용을 하여야 하는가하는 방법을 제공하는데 있다. 특히 이 연구의 모의실험 과정에서 발견한 점근적 최적 차분에 기저한 HKT의 추정량이 소표본에서는 Hall과 Kay, Titterington(1990)의 주장과는 달리 GSJS 추정량보다 좋지 않다는 것이다.

우리가 다룰 회귀 모형은 다음과 같다.

$$y_j = f(t_j) + \varepsilon_j \quad (j=1, \ldots, n),$$

여기서 f는 미지의 함수이며, 오차 ε_j는 평균 0과 분산 σ^2을 가지는 독립적이고 동일한 분포를 가지는 확률변수들이라고 하자.

2. 오차분산의 추정

반응변수 $\boldsymbol{y}=\{y_1,\cdots,y_n\}^T$, $y_i=y(t_i)$은 다음의 회귀모형을 가진다고 가정하자.

$$\boldsymbol{y}=\boldsymbol{f}+\boldsymbol{\varepsilon}\ . \qquad (2.1)$$

이때 고정된 값 $t_1<t_2<\cdots<t_n$에 대하여 회귀함수 $\boldsymbol{f}=\{f(t_1),\cdots,f(t_n)\}^T$는 미지의 스무드(smooth) 함수이며, 오차 (residuals) $\boldsymbol{\varepsilon}=\{\varepsilon_1,\cdots,\varepsilon_n\}^T$는 다음의 성질들을 갖는다.

$$E(\varepsilon_i)=E(\varepsilon_i^3)=0,$$

$$Var(\varepsilon_i)=\sigma^2 \text{와} \quad E(\varepsilon_i^4)<\infty\ .$$

Rice (1984)는 오차의 분산 σ^2에 대한 일차 차분에 기저한 추정량을 다음과 같이 제시하였다.

$$\hat{\sigma}_R^2=\frac{1}{2(n-2)}\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2 \qquad (2.2)$$

$$=\frac{1}{2(n-2)}\sum_{i=1}^{n-1}\left(\frac{y_{i+1}-y_i}{t_{i+1}-t_i}\right)^2(t_{i+1}-t_i)^2.$$

추정량들의 성질에 있어서 고차 차분 (higher order differences)들을 사용함으로 추정량의 편의(bias)를 감소시킬 수 있다. 이때 물론 분산의 값이 증가하는 결과는 감수해야한다. 분산의 추정 뿐 아니라 모든 추정에 있어서 우리는 편의를 최소화하고 분산도 최소화하는 추정량을 가지길 원한다. Gasser와 Sroka, Jennen-Steinmetz(1986)의 연구에서, 그들이 제시한 2차 차분 (second order differences)을 이용한 오차분산의 추정량은 편의와 분산을 줄이는 작용을 하였다. Gasser와 Sroka, Jennen-Steinmetz(1986)이 제시한 추정량을 GSJS 추정량이라고 이름을 짓고 $\hat{\sigma}^2_{GSJS}$ 이라고 표기하겠다.

GSJS의 추정량을 공부하기 앞서, 먼저 차분에 기저한 분산 추정에 있어서 의사오차 (pseudo-residuals), $\tilde{\varepsilon}_i$에 대하여 살펴보자. 의사오차 $\tilde{\varepsilon}_i$는 t_{i-1}, t_i, t_{i+1}의 계획된 점들(design points)의 연속적인 삼중관계 (continuous triples)에 의해 얻어지는데 일직선상의 두 개의 외곽 관찰값 (outer observations)들을 결합하고 난 다음이 일직선과 그 중앙의 관찰값 $y(t_i)$ 사이의 차분을 계산한다. 수식적으로 표현하면 다음과 같다.

$$\tilde{\varepsilon}_i=\frac{t_{i+1}-t_i}{t_{i+1}-t_{i-1}}y(t_{i-1})$$
$$+\frac{t_i-t_{i-1}}{t_{i+1}-t_{i-1}}y(t_{i+1})-y(t_i) \qquad (2.3)$$
$$=a_iy(t_{i-1})+b_iy(t_{i+1})-y(t_i),$$
$$i=2,\cdots,n-1.$$

GSJS 추정량의 기본적인 개념은 위의 오차들이 $\boldsymbol{f}=0$일 때 $\tilde{\varepsilon}_i^2$이 σ^2에 대한 불편성을 갖게 하기 위하여 정규화 (normalization)를 시키고 이러한 정규화 된 오차의 평균을 σ^2의 추정량으로서 사용한다는 것이다. 다음은 의사 오차를 행렬로 표현하여 오차의 분산에 대한 GSJS 추정량을 구한 것이다.

그러므로 의사오차 $\tilde{\boldsymbol{\varepsilon}}$ 는 다음과 같은 행렬로서 표현 되어진다.

$$\tilde{\boldsymbol{\varepsilon}}=\begin{bmatrix} c_2a_2y_1 & -c_2y_2 & +c_2b_2y_3 \\ c_3a_3y_2 & -c_3y_3 & +c_3b_3y_4 \\ \vdots & & \\ c_{n-1}a_{n-1}y_{n-2} & -c_{n-1}y_{n-1} & +c_{n-1}b_{n-1}y_n \end{bmatrix}.$$

그러면 의사오차 $\tilde{\boldsymbol{\varepsilon}}$의 오차분산의 추정량 $\hat{\sigma}^2_{GSJS}$은 다음과 같이 제시되어진다.

$$\hat{\sigma}^2_{GSJS}=\frac{1}{(n-2)}\sum_{i=2}^{n-1}\tilde{\varepsilon}_i^2, \qquad (2.4)$$

여기서

$$\tilde{\varepsilon}_i=c_i(a_iy_{i-1}+b_iy_{i+1}-y_i).$$

다음의 정리는 GSJS 추정량의 σ^2에 대한 일치성과 수렴의 정도를 조사하였다.

정리 2.1. 회귀모형 (2.1)에 있는 조건에 대하여, 만약 의사 오차 $\tilde{\boldsymbol{\varepsilon}}=\boldsymbol{D}\,\boldsymbol{y}$가 정의되어 진다면

$$\hat{\sigma}^2_{GSJS}-\sigma^2=O_p\left(\frac{1}{\sqrt{n}}\right).$$

Hall과 Kay, Titterington은 1990년에 비모수 회귀분석에 있어서 점근적으로 최적화된 차분에 기저한 오차분산 추정량인 HKT 추정량, $\hat{\sigma}^2_{HKT}$를 다음과 같이 제시하였다.

표 2.1. HKT 오차분산 추정에 있어서 m의 값에 따른 최적화 차분 수열 d의 값

m	$(d_0, \cdots, d_m)$
1	(0.7071, -0.7071)
2	(0.8090, -0.5, -0.3090)
3	(0.1942, 0.2809, 0.3832, -0.8582)
4	(0.2708, -0.0142, 0.6909, -0.4858, -0.4617)
5	(0.9064, -0.2600, -0.2167, -0.1774, -0.1420, -0.1103)
6	(0.2400, 0.0300, -0.0342, 0.7738, -0.3587, -0.3038, -0.3472)
7	(0.9302, -0.1965, -0.1728, -0.1506, -0.1299, -0.1107, -0.0930, -0.0768)
8	(0.2171, 0.0467, -0.0046, -0.0348, 0.8207, -0.2860, -0.2453, -0.2260, -0.2879)
9	(0.9443, -0.1578, -0.1429, -0.1287, -0.1152, -0.1025, -0.0905, -0.0792, -0.0687, -0.0588)
10	(0.1995, 0.0539, 0.0104, -0.0140, -0.0325, 0.8510, -0.2384, -0.2079, -0.1882, -0.1830, -0.2507)

$$\hat{\sigma}^2_{HKT} = \frac{1}{n-m} \sum_{k=1}^{n-m} \left(\sum_{j=0}^{m} d_j y_{j+k} \right)^2 \qquad (2.5)$$

여기서 d_j는 수열로서 $\sum d_j = 0$과 $\sum d_j^2 = 1$를 만족하며 만약 $j < 0$과 $j > m$이면 $d_j = 0$이다. 그들은 점근적으로 오차평균자승합 $E(\hat{\sigma}^2_{HKT} - \sigma^2)^2$를 최소화시키는 d를 선택함으로서 "최적화"의 기본적 개념을 도입하였다. 이러한 최적화 차분 수열 d의 값들을 $m = 1, \ldots, 10$에 대하여 위의 표 2.1에서 제시하여 놓았다.

3. 추정량들의 비교와 비판

Rice의 추정량은 일차 차분에 기저한 추정량이었다. 이차 차분에 기저한 Gasser와 Sroka, Jennen-Steinmetz의 추정량인 (2.4)의 GSJS 추정량은 비록 분산의 값이 증가하는 결과를 초래하지만 편의면 에서는 훨씬 더 감소하는 결과를 얻어 므로서 Rice의 추정량보다 더 좋은 장점을 가진다. Hall과 Kay, Titerington(1990)은 그들의 추정량이 위의 두 추정량들보다 효율성 (efficiency) 측면에서 훨씬 우수하다고 주장하였다.

그러나 Hall과 Kay, Titterington(1990)의 논문에서 HKT 추정량의 다양한 m의 값들 중에 서, 적절한 m을 선택하는 방법을 전혀 제시하지 않았다. 이것은 실제적인 문제에 있어서 m값을 선택함에 있어서 경험적으로 선택하여야 하는 어려움이 있다.

추정량의 효율성을 생각해 볼 때 Hall과 Kay, Titterington의 주장은 HKT 추정량이 GSJS 추정량보다 점근적인 성질에 있어서의 낫다고 주장하였다. 그러나 다음절의 모의실험의 결과에서 보듯이 자료의 표본의 크기가 적을 경우에는 오히려 GSJS의 추정량이 HKT의 추정량보다 좋은 추정량임을 알 수 있다.

4. 모의실험에 의한 추정량들의 비교

추정량들의 비교를 위하여 (2.1)의 회귀모형에서 평활함수 f에 대하여 다음의 네 개의 모형들,

모형 A, $f(t) = t + a_1 \exp\{-a_2(t-a_3)^2\}$, $(a_1 = 0.5, a_2 = 50, a_3 = 0.5)$.

모형 B, $f(t) = a_1 \exp(-a_2 t)$, $(a_1 = 5, a_2 = 5)$.

모형 C, $f(t) = \exp(-t/a_1)\cos(t/a_2)$, $(a_1 = 2.0,\ a_2 = 1.0)$.

모형 D, $f(t) = \exp(-t/a_1)\cos(t/a_2)$, $(a_1 = 0.5,\ a_2 = 0.05)$.

그림 4.2는 위의 모형 A에 있는 실험 함수 f를 사용하여 오차의 분산 추정량들에 대한 모의 실험을 하였다. 그림 4.2의 (a)에서는 오차의 분포가 N(0, 0.005)를 따르는 경우이다. GSJS 추정량은 자료의 크기 $n=30$일 경우에 모분산에 가까웠고, Rice 추정량은 $n=200$이 되어서야 모분산에 가까워졌으며, HKT 추정량은 $n=250$일 때 비로서 모분산에 가까워졌음을 알 수 있다. 그림 4.2의 (b) - (d)의 결과에서 이러한 모분산에 가까워지는 수렴 속도는 주어진 모분산 값이 점점 커질수록 모분산에 수렴되도록 요구되는 자료의 크기가 적어짐을 알 수 있다. 특히 그림 (d)의 모분산 값 0.1에 대해서는 자료의 크기 $n=10$에서도 오차의 추정량들이 잘 적합이 됨을 알 수 있다. 그러나 모분산 값이 커질수록 각 추정량들에 대한 편의는 줄어들지만 추정량들에 대한 표준

그림 4.1. 모의실험을 위한 실험 함수 f들에 대한 분포.

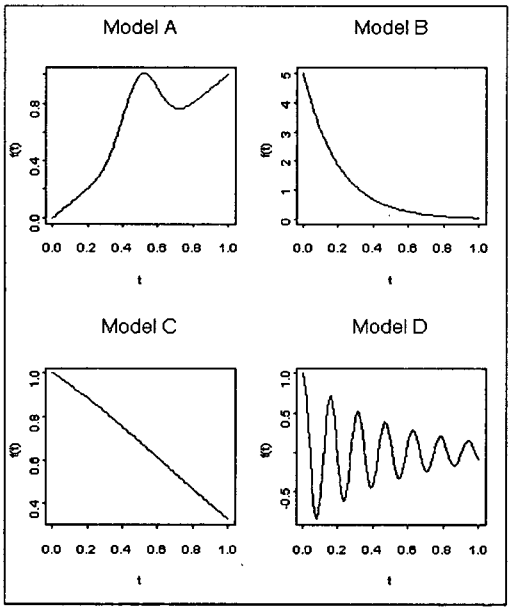


그림 4.2 모형 A에 대한 분산 추정량

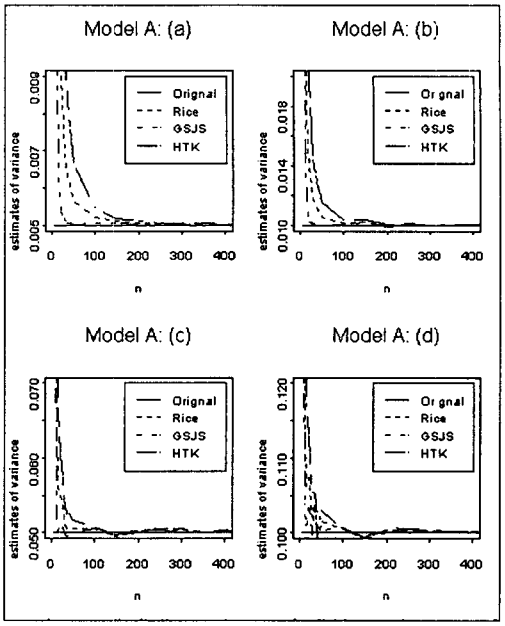
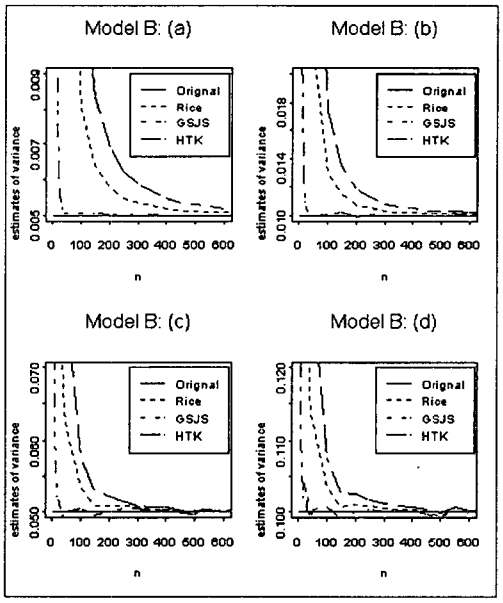


그림 4.3 모형 B에 대한 분산 추정량



※ 그림 4.2와 4.3의 (a)-(d)는 오차분포가 정규분포로서 평균 0이고, 분산이 각각 0.005, 0.01

그림 4.4 모형 C에 대한 분산 추정량

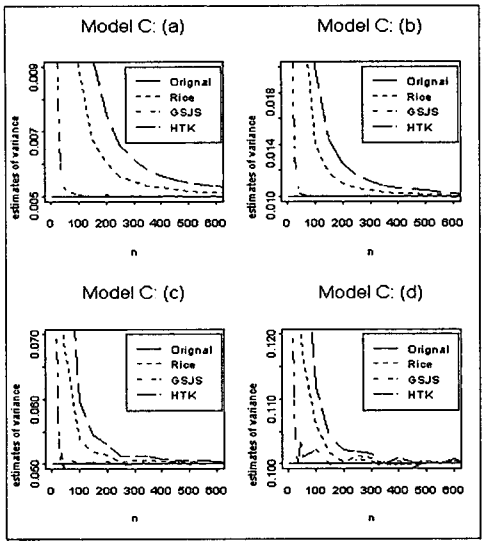
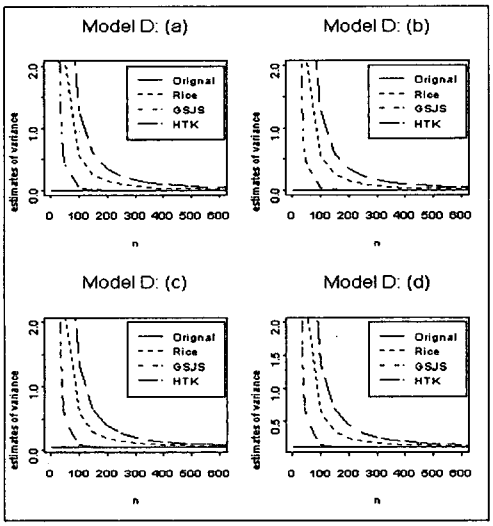


그림 4.5 모형 D에 대한 분산 추정량



※ 그림 4.4와 4.5의 (a)-(d)는 오차분포가 정규분포로서 평균 0이고, 분산이 각각 0.05, 0.1.

편차들은 커지게 된다. 그러므로 그림 (d)의 선들이 다소 매끄럽지 못하다.

그림 4.3과 그림 4.4에서는 그림 4.1에서 실험함수의 모양이 다소 유사한 것과 같이 비슷한 결과가 나타났다. 그러나 모형 B와 모형 C와 같은 경우, 즉 자료가 진동수(frequence)에 영향을 받지 않는 경우에 있어서는 모형 A나 모형 D보다 다음과 같은 심각한 결과들이 발생한다. 그림 4.3의 (a)인 경우 GSJS 추정량은 $n=40$일 경우 이미 모분산에 가깝게 수렴하였지만 Rice 추정량과 HKT 추정량은 자료의 크기가 $n=600$에 이르도록 모분산 값에 수렴하지 못한다는 것이다. 모분산 값이 점점 커질수록 모분산에 수렴하는 자료의 크기는 점점 작아지지만 추정량의 표준편차는 점점 커지게 되어 모형 A의 (d)의 결과 보다 더 매끄럽지 못한 결과들을 보였다.

그림 4.5는 그림 4.1의 (d)에 나타난 모형 D에 관한 추정량에 대한 비교 결과이다. GSJS 추정량은 자료의 크기가 $n=100$일 때 모분산 값에 거의 가깝게 수렴하고 Rice의 추정량은 $n=300$이 되어서야 모분산 값에 가까워진다. 그리고 HKT 추정량은 $n=300$ 보다 큰 자료의 값에서 모분산의 값에 수렴함을 알 수 있다.

5. 결 론

이절에서는 앞절의 모의실험의 결과를 기초로 다음의 질문에 답함으로서 본 연구의 결론을 맺는다. "그러면 Hall과 Kay, Tillterington (1990)의 HKT 추정량이 GSJS 추정량 보다 효율성이 좋다는 주장은 거짓인가?" 이 물음에 대한 답은 먼저 "거짓이 아니다." 그들의 최적화된 이차 차분에 기저한 HKT추정량은 점근적인 추정량의 성질을 가지므로 아래의 표 5.1의 결과에서 자료의 크기가 커지면 HKT의 추정량은 효율성에 있어서 다른 추정량들 보다 좋아진다. 표 5.1에서는 오차분산 σ^2에 대하여 표본의 크기 $n=5,50,500$일 때, 500번의 모의 실험을 이용하여 각 오차분산 추정량의 값과 각 추정량에 대한 표준편차와 편의를 보였다.

그러나 자료의 크기가 작은 경우들에 있어서는 앞 절의 모의 실험의 결과들에서 보듯이 HKT 추정량이나 Rice 추정량을 사용하는 것은 오차분산의 추정에 있어서 많은 오류를 범할 것 있다.

표 5.1. 모형 A에서 $N(0, 0.01)$의 분포를 갖는 오차분산의 추정량과 추정량들에 대한 편의와 표준편차 모의실험의 값

σ^2	n	$\hat{\sigma}^2_R$	$\hat{\sigma}^2_{GSJS}$	$\hat{\sigma}^2_{HKT}$
0.01	5	.7843E-01	.6799E-01	.1373
편의		.6843E-01	.5799E-01	.1273
표준편차		.2812E-01	.4077E-01	.4064E-01
0.01	50	.1058E-01	.9967E-02	.1156E-01
편의		.5823E-03	.3274E-04	.1560E-02
표준편차		.2520E-02	.2882E-02	.2311E-02
0.01	500	.9992E-02	.9993E-02	.9990E-02
편의		.8191E-05	.7468E-05	.1043E-04
표준편차		.7936E-03	.9149E-03	.7081E-03

앞으로의 연구에서는 오차의 분포를 정규 분포에 국한시키지 않고 비대칭 분포에 대한 연구와 회귀함수를 평활 스플라인추정이나 커널함수추정 등을 이용한 비교분석과 모수-비모수적 (semiparametric) 회귀 모형에서의 오차의 분산 추정에 대한 연구가 가능할 것이다.

참고 문헌

[1] Buckley, M. J., Eagleson, G. K. (1989). A graphical method for estimating the residual variance in nonparametric regression. Biometrika Vol. 76, 2, 203-10.

[2] Gasser, T., Sroka, L. & Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. Biometrika Vol. 73, 625-33.

[3] Hall, P., Kay, J. W. & Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika Vol. 77, 521-28.

[4] Hall, P., Kay, J. W. & Titterington, D. M. (1991). On estimation of noise variance in two-dimensional signal processing. Advanced Applied Probability. Vol. 23. 115-123.

[5] Hall, P. & Marron, J. S. (1990). On variance estimation in nonparametric regression. Biometrika Vol. 77, 415-9.

[6] Rice, J. (1984). Bandwidth choice for nonparametric kernel regression. Annals of Statistics, Vol. 12, 1215-30.

[7] Thompson, A. M., Kay, J. W. and Titterington, D. M. (1991). Noise estimation in signal restoration using regulation. Biometrika, Vol. 79, 475-488.