Journal of the Korean
Data & Information Science Society
1998, Vol. 9, No. 2, pp. 95 ~ 103

회귀모형의 선형성에 대한 커널붓스트랩검정 [1]

백장선 [2] · 김민수 [3]

요약

회귀모형의 선형성을 검정하는 방법으로서 Azzalini와 Bowman은 회귀모형의 오차항이 정규분포를 따른다는 가정하에서 커널회귀추정량을 이용한 유사우도비 검정이라는 비모수적 방법을 제안하였다. 붓스트랩(bootstrap)기법을 도입하여 그들의 검정방법을 변형한 커널붓스트랩검정이라는 새로운 검정법을 제시하고 모의실험을 통해 검정력을 살펴보았다. 제안된 방법은 오차항의 분포가 정규분포가 아닌 경우에도 적용이 가능하였다.

주제어 : 선형성 검정, 커널 추정량, 붓스트랩

1. 서론

종속변수와 설명변수에 대한 관측자료 $\{(y_i, x_i)\}_{i=1}^n$에 대하여 회귀모형 $y_i = m(x_i) + \epsilon_i, \quad i = 1, 2, \cdots, n$ 을 고려해보자. 이 때 ϵ_i는 평균이 0이고 분산이 σ^2인 독립적인 오차항이다. 그리고 x_i들은 편의상 $x_1 \leq x_2 \leq \cdots \leq x_n$인 순서로 정리되어 있다고 가정하자. 선형모형의 타당성을 검정한다는 것은 다음의 가설에 대한 검정이라고 할 수 있다.

$$H_0 : m(x) = \beta_0 + \beta_1 x$$
$$H_1 : m(x)\text{는 매끄러운 곡선이다.}$$

선형모형의 타당성을 검정하는 기존의 방법에는 그래프를 이용하는 방법, 모수적 방법, 그리고 비모수적인 방법등이 알려져 있다. 그래프를 이용하는 방법은 보통 자료를 선형모형에 적합시킨 다음 잔차를 도식화하는 방법이 대표적이고, 모수적 방법에는 적합결여 검정(lack-of-fit test) 과 일반적으로 시계열변수의 자기상관을 검사하는데 쓰이는 것으로 더 알려진 더빈-왓슨 검정 (Durbin- Watson test) 등이 있다. 최근에 비모수적인 방법들이 많이 연구되고 있는데, 비모수적인 방법들의 대부분은 검정통계량을 모수적 모형으로부터

[1]이 논문은 1997년 한국학술진흥재단의 공모과제 연구비 1997-003-D00051의 일부에 의하여 연구되었음.
[2]광주시 북구 용봉동 전남대학교 자연과학대학 통계학과 조교수
[3]광주시 북구 용봉동 전남대학교 자연과학대학 통계학과 박사과정

추정한 잔차를 다시 비모수적인 추정방법으로 재추정함으로써 유도된다. 비모수적인 방법에는 Cox와 Koh(1989), Azzanlini, Bowman과 Härdle(1989), Eubank와 Spiegelman(1990), Azzalini와 Bowman(1993)등이 연구되어있다. 또한 Kim, Hong과 Jeong(1996)은 모수적 회귀모형의 적합성 검정을 할 수 있는 여러 가지 비모수적 방법들에 대한 검정력을 모의실험을 통하여 비교하였다. Azzalini와 Bowman(1993)의 유사우도비 검정(pseudolikelihood ratio test)은 $\epsilon_i \sim N(0, \sigma^2)$ 의 가정하에서 즉, 모형의 오차항이 정규분포를 따른다는 가정하에 확률변수의 2차형식으로 표현되는 검정통계량이 χ^2분포에 근사한다는 것을 이용해 귀무가설을 기각할 수 있는 임계치(critical value)를 결정하여 검정하였다.

본 논문에서는 유사우도비 검정의 통계량을 약간 수정하고 임계치를 결정함에 있어 커널추정 량(kernel estimator)의 이용과 붓스트랩기법을 도입하여 정규분포에 대한 가정이 없을 때도 사용할 수 있는 검정을 개발함으로써 보다 확장되고 일반화된 해결책을 제시한다. 본 논문의 2절에서는 이 논문에서 제시하는 커널붓스트랩 검정을 설명한다. 3절에서는 커널붓스트랩 검정의 검정력을 모의실험을 통해 살펴본다. 4절에서는 결론을 기술한다.

2. 커널붓스트랩 검정

2.1 문제제기

Azzalini와 Bowman은 H_0하에서 최소제곱추정량 $\hat{\beta}_0, \hat{\beta}_1$을 이용하여 회귀함수 추정량을 $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$로 구성하고, H_1하에서 $y = (y_1, y_2, \cdots, y_n)'$이라 할 때 Nadaraya-Watson 커널 추정량을 이용하여 $\hat{m}(x) = Wy$로 추정하였다. (이때 W는 평할모수 h, 커널 K로 이루어진 $n \times n$행렬이다.) 그런 다음 모형의 적합성 검정에 흔히 사용되는 검정통계량 $F = (RSS_0 - RSS_1)/RSS_1$에 착안하였다. 이때 RSS_0와 RSS_1은 각각 선형모형과 비모수적모형에 적합한 후 계산된 잔차평방합이다. 만약 계획행렬(design matrix) X에 대하여 $M_0 = I - X(X'X)^{-1}X'$를 정의하고, $M_1 = (I - W)'(I - W)$ 라 하면 위의 F검정통계량은 $F = (y'M_0y - y'M_1y)/y'M_1y$임을 알 수 있다. 그런데 잔차들이 $e = M_0y$이므로 원래의 가설들은 다음의 가설들과 동등함을 알 수 있다.

H_0^* : 모든x_i 에 대하여$E(e) = 0$ 이다.

H_1^* : $E(e)$는 어떤 매끄러운 곡선이다.

이에 상응하는 검정통계량은 $F^* = (e'e - e'M_1e)/e'M_1e$이며 그들은 정규확률변수의 2차형식이 χ^2분포에 근사한다는 것을 이용하여 검정통계량의 임계치를 계산하였다.

Härdle과 Mammen(1993)에서 언급되었듯이 대부분 비모수적 검정통계량의 점근분포의 수렴속도가 매우 느리므로 실제로는 Monte Carlo 근사나 붓스트랩에 의해 검정통계량의 임계치를 계산 한다. 따라서 본 연구에서는 $F^* + 1 = e'e/e'M_1e$를 검정통계량으로 삼고 이것의 귀무가설하에서의 분포에 대한 임계치를 붓스트랩으로 추정하고자 한다. 이렇게 함으로써 점근분포의 수렴속도를 빠르게 할수 있으며 또한 오차항이 정규분포를 따른다는 가정으로부터도 벗어날 수 있다. 그리고 검정통계량의 $e'M_1e$를 계산하는 평할방법으로는

Nadaraya-Watson 커널추정량 대신 국소선형커널추정량(local linear kernel estimetor)를 이용하였다.

국소선형 커널추정량은 경계지역(boundary region)에서도 다른 커널추정량들 보다 좀 더 추정을 잘하는 것으로, 또한 등간격의 관측점이 아니라고 하더라도 수행력(performance)이 뛰어난 것으로 알려져 있다(Wand 와 Jones (1995)).

2.2 검정방법

하나의 설명변수에 대한 회귀모형을 고려하면 다음과 같다.

$$y_i = m(x_i) + \epsilon_i, \qquad i = 1, 2, \cdots, n.$$

여기서 ϵ_i는 평균이 0이고 분산이 σ^2인 독립인 확률변수이다. 결국, 우리의 목적은 과연 위의 모형이 단순회귀모형

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \cdots, n.$$

으로 축소될 수 있는지를 검정하는데 있다. 즉, 가설 H_0와 H_1과, 이미 언급한 바와같이 이와 동 등한 가설 H_0^*와 H_1^*에 대하여 본 논문에서는 다음과 같이 검정통계량을 제안한다.

$$T = \frac{e'e}{e_K'e_K}$$

이때 e_i를 선형모형의 최소제곱추정결과에서 나온 잔차라 할 때 $e'e = \sum_{i=1}^n e_i^2$이며, $\hat{E}_K(e)$를 잔 차들에 대한 커널추정치 $\hat{E}_K(e) = We$라 하며 $e_K = e - \hat{E}_K(e)$로서 $e_K'e_K = e'M_1e = \sum_{i=1}^n (e_i - \hat{E}_K(e_i))^2$ 이다. 구체적인 $\hat{E}_K(e_i)$ 계산은 에파네치니코프 커널 $K(u) = (3/4)(1 - u^2)I(|u| \le 1)$ 과 국소선형 커널추정량을 이용하여 다음과 같이 수행하였다.

$$\hat{E}_K(e_i) = \hat{E}_K(e_i; h) = \frac{1}{n}\sum_{j=1}^{n} \frac{\hat{s}_2(x_i;h) - \hat{s}_1(x_i;h)(x_j - x_i)K_h(x_j - x_i)e_j}{\hat{s}_2(x_i;h)\hat{s}_0(x_i;h) - \hat{s}_1(x_i;h)^2}.$$

이때 $\hat{s}_r(x_i;h) = n^{-1}\sum_{j=1}^n (x_j - x_i)^r K_h(x_j - x_i)$, $K_h(u) = h^{-1}K(u/h)$이다.

여기서 검정통계량 T는 큰 값에 대하여 유의하게 된다는 것을 알 수 있다. e는 최소제곱법에 의해서 추정된 회귀직선에서 나온 잔차이므로 $e'e$는 잔차제곱합이고 $e_K'e_K$는 다시 잔차들을 국소선형 커널추정량으로 추정한 회귀선에서 나온 잔차들의 제곱합 이다. 그래서 만약 선형모형이 맞는다면 최소제곱법으로 추정한 회귀선이나 커널추정법으로 추정한 회귀선이나 비슷할 것이므로 잔차제곱합들도 비슷하게 되어 T값은 1에 가까울 것이지만 비선형모형이 맞는다면 최소제곱법으로 추정한 후에 나온 잔차제곱합이 커널추정법으로 추

정한 후에 계산된 잔차제곱합 보다 더 큰 값을 갖게 되어 T값은 1보다 커지게 된다. 그러면 이제 문제는 귀무가설을 기각할 수 있는 임계치를 찾는데 있다고 할 수 있다. 그 문제에 대한 해결책으로는 붓스트랩을 이용하는 방법을 제안한다.

붓스트랩을 이용하는 방법이란 표본으로부터 복원으로 자료를 추출하는 것을 말하는데, 그 목적은 표본와 비슷한 성질을 가진 자료를 생성하여 많은 횟수에 걸쳐 실험하므로써 모집단의 분포 를 알아내고자 하는데 있다. 여기서는 먼저 자료를 최소제곱법으로 추정한 후에 나온 잔차들에서 T통계량 값을 구하고 다시 그 잔차들과 비슷한 성질을 가진 붓스트랩자료들로부터의 T통계량 값에서 임계치를 계산하므로써 검정을 할 수 있다. 그런데 이 자료에 대한 붓스트랩기법에 있어서는 먼저 일반적인 붓스트랩기법으로 자료들을 얻고 그 자료들의 평균을 계산하여 다시 빼주므로 써 붓스트랩표본들의 평균을 0으로 만들어주었다. 이는 귀무가설하에서는 $E(e)=0$이므로 생성된 붓스트랩표본이 동일한 성질을 갖도록 하기 위해서이다. 또한 커널추정량의 계산시에 사용되는 평활모수는 다음과 같은 Craven과 Wahba(1979)의 일반화 교차타당성함수 $GCV(h)$가 최소가 되는 값을 계산해서 사용되었다.

$$GCV(h)=\frac{(1/n)RSS(h)}{(1-K(0)/(nh))^2}.$$

물론 일반화 교차타당성함수에 의한 평활모수 선택법이 최적으로 알려진 것은 아니지만 계산의 속도가 빠르므로 이 방법을 사용하였다.

다음은 커널붓스트랩검정에서 유의수준이 α일 때 귀무가설 H_0하의 임계치를 추정하는 알고리즘이다.

붓스트랩 검정 알고리즘

1단계

1.1 최소제곱 추정치 $\hat{\beta}_0, \hat{\beta}_1$ 를 계산
1.2 $\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$에 의해 $\hat{Y}_i$를 추정
1.3 $e_i=Y_i-\hat{Y}_i$
1.4 $\{e_i\}_{i=1}^n$으로부터 T_0를 계산
  1.4.1 $e'e=\sum_{i=1}^n e_i^2$
  1.4.2 $e_K'e_K=\sum_{i=1}^n(e_i-\hat{E}_k(e_i))^2$,
  여기서 $\hat{E}_k(e_i)$는 $E(e_i)$에 대한 커널추정치이다.
  1.4.3 $T_0=e'e/e_K^{'}e_K$

2단계

$j=1,2,\cdots,M$ (붓스트랩 표본 생성 횟수)
2.1 $\{e_i\}_{i=1}^n$으로부터 붓스트랩자료 $e_{j1}^B, e_{j2}^B, \cdots, e_{jn}^B$추출

2.2 $\{e_{ji}^B\}_{i=1}^n$으로부터 T_{jB}를 계산

2.2.1 $e'_{jB}e_{jB} = \sum_{i=1}^n (e_{ji}^B)^2$

2.2.2 $e'_{jBK}e_{jBK} = \sum_{i=1}^n (e_{ji}^B - \hat{E}_K(e_{ji}^B))^2$,

여기서 $\hat{E}_K(e_{ji}^B)$는 j번째 붓스트랩표본의 $E(e_{ji}^B)$에 대한 커널추정치이다.

2.2.3 $T_{jB} = e'_{jB}e_{jB}/e'_{jBK}e_{jBK}$

3단계

3.1 $T_{.B} = T_{jB}$를 오름차순으로 정렬한 집합

3.2 $T_{.B}$의 $(1-\alpha)$분위수를 계산

임계치 $\leftarrow T_{.B}$의 $(1-\alpha)$분위수

검정결론

T_0와 임계치를 비교하여,

T_0 >임계치 이면 H_0를 기각한다.

$T_0 \leq$임계치 이면 H_0를 채택한다.

3. 검정력 모의실험

검정력을 추정해보기에 앞서 선형모형이 진실인 경우 붓스트랩에 의해 계산된 임계치를 이용 하여 귀무가설에 대한 기각율을 추정했을 때 그것이 과연 미리 정한 유의수준과 같은지 검정해 보기로 하자. 왜냐하면 정해진 유의수준과 제안된 검정법으로부터 구한 추정값이 같아야만 추정된 검정력이 의미가 있기 때문이다. 자료는 다음의 모형에서 얻어 유의수준 α=0.05를 실험해 보았다.

$$Y_i = 1 + x_i + \epsilon_i, \qquad i = 1, \cdots, n, \qquad \epsilon_i \sim N(0, 0.025^2).$$

먼저 등간격으로 0부터 1까지 x의 관측점을 25개(M=499) 잡아서 500번 모의실험한 결과 기각율 $\hat{\alpha} = 0.056$, 그리고 관측점을 50개(M=299) 잡았을 때는 기각율 $\hat{\alpha} = 0.054$가 계산되었다 (M은 붓스트랩 표본 생성횟수). 물론 추정된 기각율이 미리 설정한 $\alpha = 0.05$보다는 약간 크지만 500개의 자료에서 0.05에 대한 95%신뢰구간이 (0.0308, 0.0691)이므로 유의수준 $\alpha = 0.05$와 동일하다고 할 수 있다. 위에서 붓스트랩의 표본생성회수를 $n = 25$일 때에 $n = 50$일 때보다 많게 한 이유는 표본이 작은 경우에 붓스트랩한 자료는 아무리 좋은 기술을 이용한다해도 표본이 많은 경우보다 원 표본의 정보량에 덜 접근할 것이기 때문에 그것을 조금이라도 보완하기 위해서이다.

다음으로 추정된 회귀선이 직선에서 멀어질수록 커널붓스트랩 검정의 검정력도 높아지는지를 알아 보기 위해 다음과 같은 모형에 대해 오차항이 정규분포를 따르는 경우와 그렇지않은 경우(여기서는 t분포를 따르는 경우)로 나누어 모의실험해 보았다.

$$Y_i = \{5 + \tfrac{5}{1+\sqrt{1-\gamma}}[(x_i - \delta) - \sqrt{(x_i - \delta)^2 + \gamma}]\} + \epsilon_i,$$

(1) $\epsilon_i \sim N(0, \sigma^2)$, (2) $\epsilon_i \sim t_{df}$, 여기서 $\delta = \sqrt{1-\gamma}$이다.

x_i는 0과 4사이에서 등간격으로 50개(M=299)의 관측점을 추출하여 γ는 0.01, 0.1, 0.5, 1에 대하여 실험을 실시하였다. γ의 변화에 따라서 위의 모형의 회귀함수를 살펴보면 값이 감소할수록 선형에서 멀어진다(그림3.1 참조). γ의 변화에 따른 검정력의 추정값을 표3.1에 정리하였다. 그런데 표3.1에서 γ값이 감소할수록 검정력이 증가함을 볼 수 있으므로 추정된 회귀선이 선형에서 멀어질수록 (1)과 (2)의 오차항 분포에 대하여 검정력이 높아짐을 알 수 있다. 그리고 t 분포의 경우는 자유도가 작을수록 꼬리부분이 두터우므로 모의실험결과 자유도가 작은 경우가 큰 경우보다 검정력이 작게 나타났는데 이것은 타당하게 여겨진다. 또한 t분포의 경우 자유도가 커질수록 표준정규분포에 근사해 가는데 검정력 역시 비슷해짐을 알 수 있다.

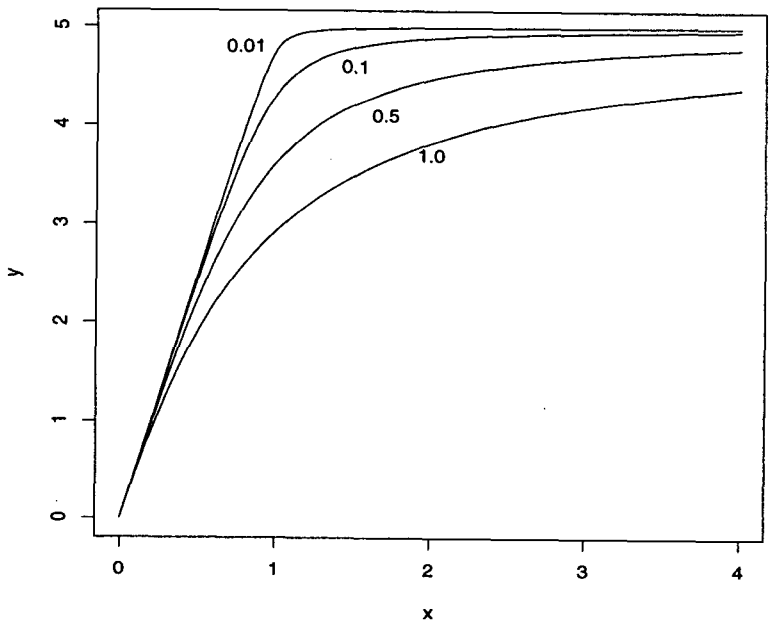

그림*3.1* $\gamma = 0.01, 0.1, 0.5, 1.0$의 변화에 따른 $m(x) = \{5 + \frac{5}{1+\sqrt{1-\gamma}}[(x-\delta) - \sqrt{(x-\delta)^2+\gamma}]\}$ 곡선

		γ			
		0.01	0.1	0.5	1.0
오차항의 분포	$N(0, 0.5^2)$	1.000	1.000	1.000	0.896
	$N(0, 0.75^2)$	0.998	0.984	0.836	0.502
	$N(0, 0.1^2)$	0.840	0.732	0.484	0.252
	t_{30}	0.826	0.700	0.504	0.244
	t_{10}	0.722	0.612	0.402	0.204
	t_6	0.642	0.530	0.316	0.150

표*3.1* γ의 변화에 따른 여러가지 회귀함수와 오차항 분포 모형에 대한 검정력 추정값

4.결론

모형의 선형성을 검정하는 문제에 있어서 기존의 검정법들은 대부분 오차항의 정규분포에 대한 가정하에서 쓸 수 있는 것들이었다. 이 논문은 정규분포의 가정에 영향을 받지않는 즉, 분포의 가정에 무관한 검정을 찾고자 Azzalini와 Bowman의 비모수적 검정법에 그 형태의 변화와 커널추정량의 수정, 붓스트랩기법등을 도입하여 오차항의 분포에 독립적인 커널붓스트랩 검정이라는 새로운 검정법을 제시하였다. 표3.1에서 확인할 수 있듯이 회귀식이 직선에서 멀어질수록 오차항의 분포에 관계없이 커널붓스트랩 검정의 검정력이 증가하였다. 그래서 특히 자료의 관측점이 많거나 오차항의 분포가 정규분포를 따르지 않는 경우에 모형의 선형성을 검정하려할 때에는 이 논문에서 제시하는 커널붓스트랩 검정을 이용하면 효과적일 것이다.

참 고 문 헌

1. Azzalini, A., Bowman, A. W. (1993). On the use of nonparametric regression for checking linear relationships, *Royal Statistical Society*, B, 55, No 2, 549-557.

2. Cox, D. D., Koh, E.(1989). A Smoothing Spline Based Test of Model Adequacy in Polynomial Regression, *Annals of the Institute of Statistical Mathematics*, 41, 383-400.

3. Craven, P. and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions, *Numerische Mathematik*, 31, 377-403.

4. Durbin, J. and Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression, *I. Biometrika*, 37, 409-428.

5. Eubank, R. L. and Spiegelman, C. H. (1990). Testing the Goodness of Fit of a Linear Model via Nonparametric Regression Techniques. *Journal of American Statistical Association*, 85, 387-392.

6. Härdle, W. and Mammen, E.(1993). Comparing nonparametric regression versus parametric regression fits, *The Annals of Statistics*, 21, 1926-1947.

7. Kim, C., Hong, C., Jeong, M.(1996). Comparisons between Goodness-of-Fit Test for Parametric Model via Nonparametric Fit, 한국통계학회논문집, Vol. 3, No. 3, 39-46.

8. Raz, J. (1990). Testing for No Effect when Estimating a Smooth Function by Nonparametric Regression : A Randomisation Approach. *Journal of American Statistical Association*, 85, 132-138.

9. Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman & Hall.

A Bootstrap Test for Linear Relationship by Kernel Smoothing [4]

Jangsun Baek [5] and Minsoo Kim [6]

Abstract

Azzalini and Bowman proposed the pseudo-likelihood ratio test for checking the linear relationship using kernel regression estimator when the error of the regression model follows the normal distribution. We modify their method with the bootstrap technique to construct a new test, and examine the power of our test through simulation. Our method can be applied to the case where the distribution of the error is not normal.

Key Words : Test for linear relationship, kernel estimator, bootstrap technique

[4]The authors wish to acknowledge the partial financial support of the Korea Research Foundation (1997-003-D00051) made in the program of 1997.

[5]Assistant Professor, Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea

[6]Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea