Journal of the Korean Data & Information Science Society
2012, **23**(4), 643–648

http://dx.doi.org/10.7465/jkdi.2012.23.4.643
한국데이터정보과학회지

잭나이프 및 붓스트랩 방법을 이용한 임상자료의 회귀계수 타당성 확인[†]

손기철[1] · 신임희[2]

[12]대구가톨릭대학교 의과대학

접수 2012년 5월 10일, 수정 2012년 5월 29일, 게재확정 2012년 6월 4일

요약

여러 임상자료를 이용하여 반응변수와 설명변수간의 관계를 규명하는 분석이 많이 이루어지고 있다. 이를 위해서 회귀분석이 흔히 사용되고 있으며, 이를 통해 설명변수가 반응변수를 얼마나 설명하는지 또한 모형이 얼마나 자료에 적합한지에 대해 분석하고 있다. 그러나 임상자료로 분석된 회귀모형에 대한 타당성 확인은 대부분 분석된 회귀모형이 얼마나 자료를 설명하는가를 나타내는 결정계수만을 살펴보는 것에 그치고 있다. 결정계수 이외의 다른 방법으로도 분석된 회귀모형의 회귀계수에 대한 타당성을 확인할 필요가 있다. 따라서 본 논문에서는 잭나이프 회귀분석과 붓스트랩 회귀분석을 이용하여 임상자료로 분석한 회귀모형의 회귀계수에 대한 타당성을 확인하는 방법을 소개하고자 한다.

주요용어: 붓스트랩, 잭나이프, 회귀계수, 회귀모형, 회귀분석.

1. 서론

회귀분석은 두 개 또는 그 이상의 설명변수와 하나의 반응변수간의 관계를 규명하기 위한 통계적 분석기법이다. 많은 분야에서 이를 이용하여 각각의 설명변수가 반응변수에 어떠한 영향을 미치고 있는지를 분석하고, 사용된 회귀모형이 자료를 얼마나 설명하는지를 파악한다. 특히 임상자료에서는 회귀모형을 이용하여 환자의 상태에 영향을 미치는 유전자나 요인이 무엇인지도 살펴보고 이를 통해 향상된 치료제나 치료법을 개발하고 있다. 그러나 수집된 자료로 적합된 회귀모형은 자료를 얼마나 설명하는가를 나타내는 결정계수 외에는 달리 타당성에 대한 확인을 하지 않고 있다. 따라서 본 논문에서는 회귀계수 값의 타당성을 확인하는 잭나이프 회귀분석 (Freedmanh과 Peters, 1984)과 붓스트랩 회귀분석 (Sahinler과 Topuz, 2007)을 소개하고 이를 임상자료에 적용시켜 보았다.

회귀모형에서 뿐만 아니라 모집단에서 표본을 추출하여 모수를 추정할 때에도 연구자들은 추출된 표본 또는 표본에서 계산된 통계량에 대한 타당성을 확인하여야 한다. 이에 대하여 표본을 모집단으로 간주하고 반복하여 같은 크기의 부표본을 수 없이 재추출하는 방법인 재표집 방법들 (Good, 2005)이 많이 연구되었다. 대표적으로 잭나이프 방법, 붓스트랩 방법, 임의순열 검증, 교차 검증 등이

[†] 본 연구는 보건복지부 보건의료연구개발사업 (과제고유번호 : A084177)의 지원과 보건복지부 통합의료 진료지침 개발과 연구사업 (과제고유번호 : 3033-320)의 지원에 의하여 이루어진 것임.
[1] (705-718) 대구광역시 남구 대명4동 3056-6, 대구가톨릭대학교 의과대학, 전임강사.
[2] 교신저자: (705-718) 대구광역시 남구 대명4동 3056-6, 대구가톨릭대학교 의과대학 의학통계학 교실, 교수. E-mail: ihshin@cu.ac.kr

있는데 이들은 모두 계산 집약적인 통계적 방법들이다. 이런 방법들은 여러 분야와 방법들에 접목시켜 사용되고 있으며 회귀분석도 예외는 아니다. 흔히 회귀분석은 회귀계수 $\boldsymbol{\beta}$를 추정하고 이를 이용하여 최종모형을 설정한다. 선형회귀모형을 $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$이라고 가정하자. 여기서 $\boldsymbol{Y}$는 반응변수, $\boldsymbol{X}$는 설명변수, $\boldsymbol{\beta}$는 회귀계수 그리고 $\boldsymbol{\epsilon}$는 오차항이다. 그러면 $\boldsymbol{\beta}$는 최소제곱추정법에 의해 $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$으로 추정되고, 이에 대한 분산은 $Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$로 주어진다. 그러나 $\boldsymbol{\beta}$의 값이 어느 정도 타당한지에 대한 확인은 하지 않는다. 따라서 회귀모형에 대한 타당성 확인 (Dette와 Munk, 1998)이 필요하고 이를 위해 잭나이프 방법, 붓스트랩 방법 등과 같은 재표집 방법과 회귀분석이 접목되어, 각각 새로운 회귀계수 (가령 잭나이프 방법과 붓스트랩 방법으로 구한 회귀계수를 $\widehat{\boldsymbol{\beta}}^{(J)}$과 $\widehat{\boldsymbol{\beta}}^{(B)}$라고 하자)를 추정하고 이를 최소제곱법으로 구한 $\widehat{\boldsymbol{\beta}}$와 비교하여 타당성을 확인 하게 되는 것이다. 이 논문은 다음과 같이 구성하였다. 2절에서 잭나이프 회귀분석과 붓스트랩 회귀분석을 소개한다. 그리고 3절에서는 2절에서 소개한 방법을 임상자료에 적용하여 보고, 4절에서 결론을 기술한다.

2. 타당성 확인 방법

2.1. 잭나이프 회귀분석

재표집 관측값을 기본으로 하는 잭나이프 회귀분석은 고정적인 설명변수를 가지는 자료로부터 회귀모형이 설정될 때 적용된다. 잭나이프 재표집은 두 가지가 있는데 하나는 관측된 자료에서 하나의 자료를 제외하는 것에 기반을 하고 있고 나머지 하나는 관측된 자료에서 두 개 이상의 자료를 제외하는 것에 기반을 두고 있다 (Efron과 Gong, 1983; Wu, 1986; Shao와 Tu, 1995). Y_i를 i번째 자료의 반응변수 값이라 하고 X_{ij}를 i번째 자료의 j번째 설명변수 값이라 하자. 여기서 $i = 1, \cdots, n$이고 $j = 1, \cdots, p$이다. 또한 $\boldsymbol{Y} = (Y_1, \cdots, Y_n)'$으로 하고 $\boldsymbol{X} = (X_1, \cdots, X_n)$으로 하자. 그러면 하나의 자료를 제외하는 잭나이프 회귀분석은 다음과 같은 단계로 계산된다.

단계1. i번째 자료를 제외하고 $n-1$개의 자료로 구성된 표본을 n개 생성한다.

단계2. 단계1에서 생성된 n개의 표본으로 회귀분석을 통하여 각 표본에 대한 회귀계수 $\widehat{\boldsymbol{\beta}}^{(J_i)} = (\widehat{\beta}_0^{(J_i)}, \widehat{\beta}_1^{(J_i)}, \cdots, \widehat{\beta}_p^{(J_i)})$, $i = 1, \cdots, n$를 추정한다.

단계3. 단계2에서 얻은 $\widehat{\boldsymbol{\beta}}^{(J_i)}$의 평균을 계산하여 잭나이프 회귀계수 $\widehat{\boldsymbol{\beta}}^{(J)}$를 계산한다.

단계4. 따라서 하나의 관측값을 제외한 잭나이프 회귀방정식은 $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(J)}$이다.

다음으로 두 개 이상 즉 $d \geq 2$개의 관측값을 제외한 잭나이프 회귀분석은 다음과 같은 단계로 수행된다.

단계1. n개의 자료 중 d개의 자료를 제외하고 $n-d$개의 자료로 구성된 표본을 $S = {}_nC_{n-d}$개 생성한다.

단계2. 단계1에서 생성된 S개의 표본으로 회귀분석을 통하여 각 표본에 대한 회귀계수 $\widehat{\boldsymbol{\beta}}^{(J_s)} = (\widehat{\beta}_0^{(J_s)}, \widehat{\beta}_1^{(J_s)}, \cdots, \widehat{\beta}_p^{(J_s)})$, $s = 1, \cdots, S$를 추정한다.

단계3. 단계2에서 얻은 $\widehat{\boldsymbol{\beta}}^{(J_s)}$의 평균을 계산하여 잭나이프 회귀계수 $\widehat{\boldsymbol{\beta}}^{(J)}$를 계산한다.

단계4. 따라서 d개의 관측값을 제외한 잭나이프 회귀방정식은 $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(J)}$이다.

2.2. 붓스트랩 회귀분석

붓스트랩 회귀분석 방법은 두 가지의 접근법이 있다. 하나는 재표집 관측값에 근거한 방법이고 다른 하나는 재표집 오차에 근거한 방법이다 (Efron, 1979; Efron과 Tibshirani, 1993). 재표집 관측값에 근거한 붓스트랩 회귀분석은 우선 반응변수가 랜덤일 때 적용되는 방법이고 다음과 같이 계산된다.

단계1. n개의 자료가 표집될 확률을 동일하게 하여 복원추출로 n개의 관측값으로 구성된 표본인 $\boldsymbol{Y}^{(b1)} = (Y_1^{(b1)}, \cdots, Y_n^{(b1)})'$과 $\boldsymbol{X}^{(b1)} = (X_1^{(b1)}, \cdots, X_n^{(b1)})$를 하나 생성한다.

단계2. 단계1에서 생성된 표본으로 회귀분석을 통하여 회귀계수 $\widehat{\boldsymbol{\beta}}^{(b1)} = (\boldsymbol{X}^{(b1)'}\boldsymbol{X}^{(b1)})^{-1}\boldsymbol{X}^{(b1)'}\boldsymbol{Y}^{(b1)}$를 추정한다.

단계3. 단계1과 2를 B번 반복하여 $\widehat{\boldsymbol{\beta}}^{(b)} = (\widehat{\boldsymbol{\beta}}^{(b1)}, \cdots, \widehat{\boldsymbol{\beta}}^{(bB)})$를 구한다.

단계4. 단계3에서 얻은 $\widehat{\boldsymbol{\beta}}^{(b)}$의 평균을 계산하여 붓스트랩 회귀계수 $\widehat{\boldsymbol{\beta}}^{(J)}$를 계산한다.

단계5. 따라서 d개의 관측값을 제외한 붓스트랩 회귀방정식은 $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(J)}$이다.

다음으로 재표집 오차에 근거한 붓스트랩 회귀분석은 반응변수가 고정일 때 적용되는 방법으로 실험설계에서 사용되어지고 다음과 같이 계산된다.

단계1. n개의 자료로 회귀분석을 적합한다.

단계2. Y_i와 $\hat{Y}_i$의 차이로 잔차 $e_i = Y_i - \widehat{Y}_i$를 계산한다.

단계3. 단계2에서 계산한 e_i를 확률을 동일하게 하여 복원추출로 n개의 표본인 $\boldsymbol{e}^{(b)} = (e_1^{(b)}, \cdots, e_n^{(b)})'$을 하나 생성한다.

단계4. 원래의 회귀모형식인 $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$에 $\boldsymbol{e}^{(b)}$를 더하여 $\boldsymbol{Y}^{(b)}$를 계산한다.

단계5. 단계4에서 구한 $\boldsymbol{Y}^{(b)}$를 사용하여 $\boldsymbol{\beta}^{(b1)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^{(b)}$를 계산한다.

단계6. 단계3에서 단계5까지를 B번 반복하여 $\widehat{\boldsymbol{\beta}}^{(b)} = (\widehat{\boldsymbol{\beta}}^{(b1)}, \cdots, \widehat{\boldsymbol{\beta}}^{(bB)})$를 구한다.

단계7. 단계6에서 얻은 $\widehat{\boldsymbol{\beta}}^{(b)}$의 평균을 계산하여 붓스트랩 회귀계수 $\widehat{\boldsymbol{\beta}}^{(J)}$를 계산한다.

단계8. 따라서 d개의 관측값을 제외한 붓스트랩 회귀방정식은 $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(J)}$이다.

앞에서 소개한 잭나이프와 붓스트랩 회귀모형을 사용하여 원 회귀모형의 회귀계수에 대한 타당성을 확인하는 방법은 다음과 같이 이루어진다. 먼저 원 회귀모형의 회귀계수값을 추정하고 잭나이프와 붓스트랩 회귀모형의 회귀계수 값을 추정한다. 각각의 설명변수에 해당되는 세 가지 회귀계수 값의 차이가 나지 않는 경우 원 회귀모형의 회귀계수 값이 타당하다고 할 수 있다. 그러나 자료에서 편차가 큰값이나 이상치가 있는 특별한 경우는 제시한 방법과는 다른 방법으로 회귀계수에 대한 타당성을 확인하여야 할 것이다.

3. 예제

예제로 사용된 자료는 Table 3.1에 주어진 빈혈환자 자료로써 30명으로 구성되어 있으며 약물 투여 이전의 Hb (pre Hb), 몸무게 (weight; 단위 kg), 약의 투여용량 (dose; 단위 mg), 약물 치료 이후의 Hb (post Hb)의 변수가 있다 (신임희, 2008). 이 자료를 이용하여 빈혈 환자의 이전 Hb와 몸무게로 약물 치료 후 Hb를 예측하는 회귀분석을 수행하고자 한다. 분석한 결과는 Table 3.2와 같이 회귀계수가 구해진다. 즉, 원 자료를 이용한 원 회귀모형 (raw regression model)은 post Hb = 15.430 -0.156 ×pre Hb -0.011 ×몸무게 이다.

Table 3.1 Data of anemia patient

No.	post_Hb	pre_Hb	weight	dose	No.	post_Hb	pre_Hb	weight	dose
1	13	10.5	70	65	16	14	8.2	46	95
2	14	11.2	53	70.5	17	13.6	11.6	58	75.5
3	13.5	9.8	68	56	18	13.8	9.6	48	80.5
4	12.5	8.5	73	83.5	19	14.1	8.2	49	58
5	12.7	11.3	55	66	20	13.9	9.4	69	90.5
6	13	10.6	48	91.3	21	12.7	10.2	54	75.5
7	12.6	9.3	78	84.5	22	12.9	10.1	51	95.5
8	11.8	9.4	47	73.5	23	12.8	11.5	60	80.5
9	13.2	10.2	75	90.5	24	14.5	8.6	58	55
10	14.3	9.6	45	55	25	14	8.7	56	58.5
11	13.7	11.5	66	80.5	26	13	10.3	50	90.5
12	12.8	8.7	71	60.5	27	12.5	8.2	74	75
13	14.1	10.7	52	93.5	28	12.2	11.9	46	50.5
14	13.7	8.3	75	57.5	29	12.6	10.7	85	85
15	13.7	10.5	61	75.3	30	11.7	10.4	48	75

다음으로 2.1 절에서 소개한 잭나이프 회귀분석을 이용하여 구한 잭나이프 회귀계수와 그에 따른 표준오차 값은 Table 3.2에 제시하였다. 여기서는 하나의 자료를 제외한 잭나이프 회귀분석을 사용하였다. 즉, 잭나이프 회귀분석을 이용한 회귀모형 (jackknife regression model)은 post Hb = 15.430 -0.155 ×pre Hb -0.011 ×몸무게 이다. 마지막으로 붓스트랩 회귀분석을 이용하여 구한 붓스트랩 회귀계수와 그에 따른 표준오차 값도 Table 3.2에 제시하였다. 이도 마찬가지로 붓스트랩 회귀모형 (bootstrap regression model)은 post Hb = 15.469 -0.150 ×pre Hb -0.013 ×몸무게이다. 이 결과는 원 회귀모형의 회귀계수에 대한 타당성을 확인 할 수 있게 해준다. 다시 말해서 원 회귀모형에 대한 회귀계수가 잭나이프 회귀모형에 대한 회귀계수나 붓스트랩 회귀모형에 대한 회귀계수와 차이가 많이 난다면 원 회귀모형의 회귀계수는 타당성이 있다고 할 수 없다.

Table 3.2 Regression coefficients by three regression models

regression coefficient	raw regression model	jackknife regression model	bootstrap regression model
	$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}^{(J)}$($S.E(\hat{\boldsymbol{\beta}}^{(J)})$	$\hat{\boldsymbol{\beta}}^{(B)}$($S.E(\hat{\boldsymbol{\beta}}^{(B)})$)
intercept	15.430	15.426(0.0093)	15.469(0.0375)
pre Hb	-0.156	-0.155(0.0007)	-0.150(0.0031)
weight	-0.011	-0.011(0.0001)	-0.013(0.0003)

그러나 빈혈 환자 자료로 분석한 결과 세 가지 모형에 대한 회귀계수가 크게 차이나지 않음을 확인 할 수 있다. 따라서 분석된 원 회귀모형의 회귀계수 값이 타당하다고 할 수 있다. 물론, 원 회귀모형의 회귀계수와 잭나이프 회귀모형의 회귀계수는 아주 미세한 차이가 나지만 붓스트랩의 회귀계수는 원 회귀모형의 것과 약간 더 큰 차이가 있어 보인다. 그러나 이 차이도 아주 미세하다고 판단 할 수 있다. Figure 3.1은 세 가지 모형에 대한 각각의 적합값으로 그린 그림이다. 세 가지 선 모두 거의 같은 값을 가지는 것을 확인 할 수 있다.

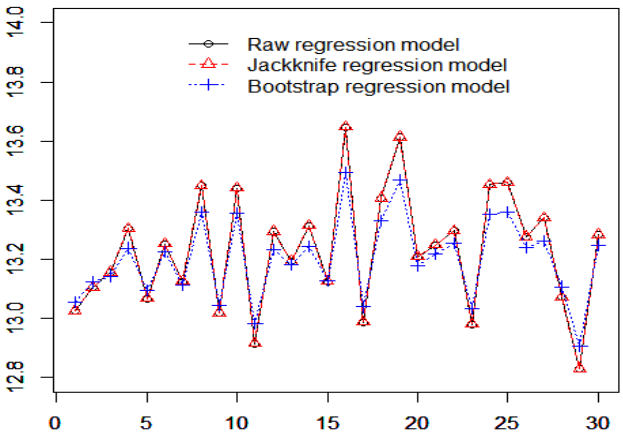

Figure 3.1 Fitted values by three regression models

4. 결론

본 논문에서는 잭나이프 회귀분석과 붓스트랩 회귀분석을 이용하여 회귀모형의 회귀계수에 대한 타당성을 확인하는 방법을 살펴보았다. 임상자료에 적용한 결과 원 회귀모형, 잭나이프 회귀모형 그리고 붓스트랩 회귀모형에서 각 회귀계수 값의 차이는 거의 없음을 확인할 수 있었다. 또한 각 회귀모형으로 계산한 세 가지의 적합값도 차이가 없음을 확인하였다. 이를 근거로 분석된 원 회귀모형의 회귀계수 값이 타당하다고 할 수 있다. 여러 가지 재표집 방법 중 잭나이프와 붓스트랩을 회귀분석에 접목시킨 방법은 몇 가지의 장점을 가진다. 첫 번째 회귀모형에 대한 설명력을 나타내는 결정계수 이외의 방법으로 분석된 회귀모형에 따르는 회귀계수 값의 타당성을 확인할 수 있다. 또한 여러 통계 소프트웨어 프로그램의 발달로 인해 이 방법들을 쉽게 구현할 수 있다. 따라서 회귀모형을 자료에 적합시킨 후 결정계수 만으로 회귀계수 값의 타당성을 판단할 것이 아니라 앞에서 제시한 잭나이프와 붓스트랩 회귀분석을 통하여 회귀계수 값의 타당성을 확인할 필요가 있는 것으로 사료된다.

참고문헌

Dette, H. and Munk, A. (1998). Validation of linear regression models. *The Annals of Statistics*, **2**, 778-800.
Efron, B. (1979). Bootstrap method: Another look at the jackknife. *Annals of Statistics*, **7**, 1-26.
Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, **37**, 36-48.
Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*, Chapman & Hall, New York.
Freedman, D. A. and Peters, S. C. (1984). Bootstrapping a regression equation: Some empirical results. *Journal of the American Statistical Association*, **79**, 9-106.
Good, P. (2005). *Introduction to statistics through resampling methods and R/S-PLUS*, John Wiley & Sons, New York.
Sahinler, S. and Topuz, D. (2007). Bootstrap and jackknife resampling algorithm for estimation of regression parameters. *Journal of Applied Quantitative Methods*, **2**, 188-199.
Shao, J. and Tu, D. (1995). *The jackknife and bootstrap*, Springer, New York.
Shin I. H. (2008). *Solution medical statistics series 1*, Koonja, Seoul.
Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, **14**, 1343-1350.

Journal of the Korean Data & Information Science Society
2012, **23**(4), 643–648

http://dx.doi.org/10.7465/jkdi.2012.23.4.643
한국데이터정보과학회지

Check for regression coefficient using jackknife and bootstrap methods in clinical data[†]

Ki Cheul Sohn[1] · Im Hee Shin[2]

[12]School of Medicine, Catholic University of Daegu

Received 10 May 2012, revised 29 May 2012, accepted 4 June 2012

Abstract

There are lots of analysis to determine the relation between dependent variable and explanatory variables. Often the regression analysis is used to do this, and we can analyze the how much the explanatory variable can be related with dependent variable and how much the regression model can explain the data. But the validation check of regression model is usually determined by coefficient of determination. We should check the validation of regression coefficient with different methods. This paper introduces the method for validation check the regression coefficient using the jackknife regression and bootstrap regression in clinical data.

Keywords: Bootstrap, jackknife, regression analysis, regression coefficient, regression model.

[†] This study was supported by the grant of Korea Ministry of Health & Welfare, Republic of Korea (Project No: A084177 and Project No: 3033-320).

[1] Full time instructor, School of Medicine, Catholic University of Daegu 705-718, Republic of Korea.
[2] Corresponding author: Professor, Department of Medical Statistics, School of Medicine, Catholic University of Daegu 705-718, Republic of Korea. E-mail: ihshin@cu.ac.kr.