The Korean Journal of Applied Statistics
2021, Vol. 34, No. 5, 761–772

DOI: https://doi.org/10.5351/KJAS.2021.34.5.761
Print ISSN1225-066X / Online ISSN 2383-5818

Analysis of simulation results using statistical models

Ji-Hyun Kim[1,a], Bongseong Kim[a]

[a]Department of Statistics and Actuarial Science, Soongsil University

Abstract

Simulation results for the comparison of estimators of interest are usually reported in tables or plots. However, if the simulations are conducted under various conditions for many estimators, the comparison can be difficult to be made with tables or plots. Furthermore, for algorithms that take a long time to run, the number of iterations of the simulation is costly to to be increased. The analysis of simulation results using regression models allows us to compare the estimators more systematically and effectively. Since variances in performance measures may vary depending on the simulation conditions and estimators, the heteroscedasticity of the error term should be allowed in the regression model. And multiple comparisons should be made because multiple estimators should be compared simultaneously. We introduce background theories of heteroscedasticity and multiple comparisons in the context of analyzing simulation results. We also present a concrete example.

Keywords: heteroscedasiticity-consistent estimator, covariance matrix, simultaneous confidence intervals, conditional plots

1. 서론

통계연구에서 모의실험은 중요한 도구이다. 통계연구에서 추정하고 싶은 모수에 대한 추정량이 여러 개 있다고 하자. 이 추정량이 나무모형에 기반한 방법이나 부스팅 기법과 같은 비모수적 방법으로 얻어지는 경우 추정량의 성능을 이론적으로 비교하기 어려운데, 이때 모의실험을 통해 성능 비교가 이루어진다. 모의실험 결과를 보고할 때 그림이나 표를 많이 이용한다. 추정량의 성능을 비교하기 위해 제곱근평균제곱오차(root mean square error, RMSE)를 성능측도로 쓴다고 하자. 모의실험 결과를 보고할 때 각 추정량에 대해 반복실험에서 얻어지는 RMSE를 상자그림으로 나타내거나, RMSE의 평균과 표준오차를 표로 정리해 보고한다. 하지만 이런 방식의 보고가 충분하지 않은 경우가 있다.

모의실험 결과의 일반성을 확보하기 위해 다양한 상황 또는 실험조건에서 모의실험을 실시한다. 생존분석의 경우 중도절단비율(censoring rate)을 달리 하거나 비례위험가정(proportional hazards assumption) 성립 여부를 달리 하면서 모의실험을 실시한다. 만약 추정량 성능의 상대적 우위가 실험조건마다 달라진다면 결과가 복잡해지고 이를 그림이나 표로 판단하는 것이 적절하지 않게 된다. 이때 통계모형을 이용하면 모의실험 결과를 보다 체계적으로 분석하고 보고할 수 있다. 성능측도를 연속형 반응변수로 두고, 추정량의 종류와 실험조건을 각각 범주형 설명변수로 둔 회귀모형을 이용하면 된다. 회귀모형에 추정량의 종류와 실험조건에 대한 상호작용항을 포함시키면 다양한 실험조건에 따라 여러 추정량의 성능 우열이 어떻게 달라지는지를 객관적으로 판단할 수 있다. 한편 시간이 오래 걸리는 모의실험인 경우 반복횟수를 크게 하기 어려운데 회귀모형을 이용하면 모든 추정량의 정보를 동시에 이용하므로 동일한 조건에서 반복 관측값이 많지 않아도

[1] Corresponding author: Department of Statistics and Actuarial Science, Soongsil University, 369 Sangdo-Ro, Dongjak-Gu, Seoul 06978, Korea. E-mail: jxk61@ssu.ac.kr

Published 31 October 2021 / journal homepage: http://kjas.or.kr
© 2021 The Korean Statistical Society. All rights reserved.

분석이 가능하다. 따라서 회귀모형을 이용하면 체계적이면서도 효율적으로 여러 추정량의 성능을 비교할 수 있다.

회귀모형을 이용하여 모의실험 결과를 분석할 때 추가적으로 고려해야 할 사항이 있다. 모의실험에서 추정량의 종류와 실험조건에 따라 성능측도의 평균뿐만 아니라 분산도 달라질 수 있으므로 오차항의 이분산성(heteroscedasiticity)을 허용하는 모형을 써야 한다. 또한 추정량이 여러 개일 때 성능비교를 위해 다중비교(multiple comparison)를 해야 한다. 즉, 두 추정량의 성능 차이에 대한 신뢰구간을 구할 때 동시신뢰구간(simultaneous confidence intervals)을 구해야 한다.

본 연구에서 회귀모형을 이용해 모의실험 결과를 분석하는 방법을 제시하고자 한다. 회귀모형과 이분산성, 그리고 다중비교는 개별적으로 이론과 방법이 이미 정립된 주제이다 (White, 1980; MacKinnon과 White, 1985; Hothorn 등, 2008). 하지만 모의실험 결과에 대한 분석이라는 맥락에서 개별적 주제들을 종합해서 적용한다는 점에서 연구의 가치가 있다고 생각한다. 2절에서 문제를 설명하고 이분산성과 다중비교를 고려하는 방법을 제시하였다. 3절에서 구체적 예를 제시하여 복잡한 모의실험 결과를 체계적으로 분석하고자 하는 연구자들에게 실용적 도움을 주고자 하였다.

2. 분석 방법

모수 또는 피추정량(estimand)이라고 부르는 θ에 대한 여러 개의 추정량 중에서 어떤 추정량이 더 나은지를 판단하고 싶다고 하자. 좋은 추정량을 찾기 위해 추정량의 성능을 판단할 수 있는 기준인 성능측도를 정의하고 각 추정량의 성능측도를 이론적으로 계산할 수 있는지를 먼저 살펴본다. 대표적인 성능측도로 제곱근평균제곱오차(RMSE)를 들 수 있다. 추정량이 나무모형에 기반한 방법과 같이 비모수적 방법으로 얻어지는 경우 성능측도를 이론적으로 구하기 어렵다. 또한 점근적인(asymptotic) 성능측도를 이론적으로 구할 수 있다고 하더라도 유한 개의 표본에서 관측되는 성능측도와 어떤 차이가 있는지를 밝히기 어렵다. 이 때 모의실험으로 성능측도의 실현값을 반복적으로 구해 추정량의 성능을 비교한다.

i번째 모의실험 조건에서 j번째 추정량 $\hat{\theta}_{ij}$에 대해 k번째 모의실험에서 얻어진 성능측도 실현값을 y_{ijk}라고 하자. 모의실험 조건은 c개이고 비교하고 싶은 추정량 수는 m개이다. $\hat{\theta}_{ij}$를 얻기 위한 자료가 있었겠지만 여기서는 관심을 가질 필요가 없다. 모의실험 조건은 복수 개의 인자에 의해 정해지는데, 생존분석의 경우 중도절단비율, 비례위험가정 성립여부, 생존시간 분포 종류 등의 인자에 의해 실험조건이 정해진다. 만약 세 개 인자에 2개 수준씩 있다면 총 실험조건 수는 $2^3 = 8$개가 된다. 같은 조건과 같은 추정량에 대해 총 r번 반복적으로 행해지는 모의실험은 난수를 독립적으로 생성해서 이루어지므로 $k = 1, \ldots, r$에 대해 다음과 같은 모형을 가정할 수 있다.

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}.$$

ϵ_{ijk}는 모든 i, j, k에 대해 독립성을 가정한다(각 실험조건에서 모의실험을 반복할 때 독립적으로 난수를 생성하므로 i나 k가 다르면 독립성을 가정할 수 있다. 하지만 익명의 심사위원이 지적했듯이 i와 k가 고정이 되면 동일한 모의자료가 되므로 ϵ_{ijk}와 $\epsilon_{ij'k}$는 독립이 아닐 수 있다. 본 연구는 오차항의 독립성 가정에 실용적으로 문제가 없을 때 적용할 수 있는 방법이다). μ_{ij}는 추정량 $\hat{\theta}_{ij}$가 갖는 평균 성능이다. 추정량의 종류와 실험조건, 그리고 필요한 상호작용항을 나타내는 벡터 $\mathbf{x}_{ij}$를 써서 μ_{ij}를 ,

$$\mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$$

와 같이 나타낼 수 있다. 설명변수 벡터 $\mathbf{x}_{ij}$의 첫 번째 원소는 모수 추정의 식별성(identifiability)을 위해 항상 1로 정의한다. 추정량의 종류와 실험조건에 따라 성능측도 실현값의 변동이 달라질 수 있으므로 오차항 ϵ_{ijk}의 분산은 σ^2이 아니라 σ_{ij}^2이다.

2.1. 공분산행렬에 대한 이분산성일치추정량

앞에서 설명한 모형을 벡터와 행렬을 써서 다음과 같이 간결하게 표현할 수 있다.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.1}$$

$\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$이며 $\epsilon_i,\ i = 1, 2, \ldots, n$은 독립임을 가정한다. $n = cmr$이며 $\boldsymbol{\epsilon}$의 공분산행렬은 대각행렬인데 $\boldsymbol{\Omega}_{n\times n}$로 나타내기로 한다. $\mathbf{X}_{n\times p}$는 모의실험에서 얻어진 성능측도 $\mathbf{y}$의 평균을 표현하기 위한 설계행렬(design matrix)이다. 추정량의 성능이 추정량 종류와 모의실험의 조건에 따라 달라질 수 있으므로 추정량 종류와 실험조건을 나타내는 주효과항은 물론이고 상호작용항을 나타내는 열도 포함하게 된다. p는 필요한 항의 수로서 절편항을 포함한다. 예를 들어 추정량이 6개이면 5개의 가변수 또는 이항형 변수 $X_1, X_2, \ldots, X_5$가 필요하다. 한편 실험조건이 3개의 이항형 변수 X_6, X_7, X_8의 조합에 의해 정해진다고 하자. 그러면 설계행렬의 열은 추정량 종류와 실험조건을 나타내는 8개의 주효과항과 포함 가능한 모든 상호작용항들 중에서 반응변수인 추정량 성능의 평균을 효과적으로 예측하기 위해 필요한 항들로 이루어지게 된다(이 때 $X_1, X_2, \ldots, X_5$ 사이의 상호작용항이 포함되면 안 된다. 자료분석을 위해 R을 쓰는 경우 6개의 추정량 종류를 나타내는 변수를 하나의 인자형(factor-type) 변수로 정의해서 모형식(model formula)에 포함시키면 추정량 종류를 나타내는 가변수들 사이의 곱으로 표현되는 항이 설계행렬에 포함되는 것을 방지할 수 있다).

회귀계수 벡터 $\boldsymbol{\beta}$의 최소제곱(ordinary least square, OLS) 추정량을 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$라고 하자. White (1980)는 모형 (2.1)이 참일 때 $\hat{\boldsymbol{\beta}}$의 일치성(consistency)이 등분산성 모형에서와 마찬가지로 이분산성 모형에서도 만족된다는 것을 증명하였다. $\mathbf{x}_i^T$를 설계행렬의 i번째 행을 나타내는 $1 \times p$ 벡터라고 할 때 i번째 잔차는 $e_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ 이다. 잔차의 제곱을 대각원소로 갖는 대각행렬을,

$$\hat{\boldsymbol{\Omega}} = \text{diag}\left(e_1^2, e_2^2, \ldots, e_n^2\right)$$

라고 할 때, White (1980)는,

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\boldsymbol{\Omega}}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \tag{2.2}$$

가 $\hat{\boldsymbol{\beta}}$의 공분산행렬 $(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}$의 일치추정량이 됨을 증명하였다. 또한 몇 가지 정칙조건이 만족되면 $\hat{\boldsymbol{\beta}}$의 분포는 근사적으로 정규분포를 따름을, 즉

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N\left(\boldsymbol{\beta}, \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\boldsymbol{\Omega}}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right), \tag{2.3}$$

임을 증명하였다. 증명 과정에서 오차항의 정규성은 가정하지 않았다.

MacKinnon과 White (1985)는 추정량 (2.2)와 점근적으로 동등하면서(asymptotically equivalent) 유한표본(finite sample)에서 더 나은 성질을 갖는 추정량을 제안하였다. 이 추정량은 잭나이프(jackknife) 기법에 기반한 추정량이다. $\hat{\boldsymbol{\beta}}_{(i)}$를 i번째 관측값을 제외하고 구한 $\boldsymbol{\beta}$의 OLS 추정량이라고 할 때,

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i^T e_i^*$$

임이 알려져 있다. $e_i^* = e_i/(1 - h_{ii})$, h_{ii}는 행렬 $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$의 i번째 대각원소이다. MacKinnon과 White (1985)가 제안한 이분산성일치추정량(heteroscedasticity-consistent estimator)을 다음과 같은 $p \times p$ 행렬로 표현할 수 있다.

$$\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\beta}}_{(i)} - \frac{1}{n}\sum_{j=1}^{n}\hat{\boldsymbol{\beta}}_{(j)}\right)\left(\hat{\boldsymbol{\beta}}_{(i)} - \frac{1}{n}\sum_{j=1}^{n}\hat{\boldsymbol{\beta}}_{(j)}\right)^T.$$

MacKinnon과 White (1985)는 이 행렬이 다음 행렬과 같음을 유도하여 보다 빨리 계산할 수 있게 하였다.

$$\frac{n-1}{n}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{\Omega}^*\mathbf{X}-\frac{1}{n}\mathbf{X}^T\mathbf{e}^*\mathbf{e}^{*T}\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{2.4}$$

$\mathbf{\Omega}^* = \text{diag}\,(e_1^{*2}, e_2^{*2}, \ldots, e_n^{*2})$이고 $\mathbf{e}^* = (e_1^*, e_2^*, \ldots, e_n^*)^T$ 이다. MacKinnon과 White (1985)는 이 추정량을 HC3 추정량이라고 부르고, HC3 추정량을 썼을 때 White (1980)의 추정량이나 다른 몇 가지 수정된 추정량을 썼을 때보다 유한표본에서 $\boldsymbol{\beta}$의 신뢰구간에 대한 실제유의수준(actual significance level)이 명목유의수준(nominal significance level)에 보다 가깝다는 것을 모의실험으로 보였다. R 패키지 `sandwich` (Zeileis, 2004)의 `vcovHC` 함수를 이용하면 HC3 추정량을 얻을 수 있다.

2.2. 다중비교

통계적 모형을 이용하면 서로 다른 실험조건에서 서로 다른 추정량들의 평균성능을 나타내는 $E[y|\mathbf{x}]$를 보다 효율적으로 추정할 수 있다. 즉, 모의실험 반복횟수를 크게 하지 않고도 $E[y|\mathbf{x}]$을 효율적으로 추정할 수 있다. 추정량 성능은 모의실험 조건에 따라 달라지며 상대적 우위도 달라질 수 있다. 따라서 각 실험조건에서 추정량의 평균성능을 서로 비교해야 한다. 추정량 종류가 m개라면 각 실험조건에서 $m(m-1)/2$ 쌍의 추정량 평균성능을 다중비교해야 한다. 다중비교를 위해 동시신뢰구간을 구해야 하는데, 그 방법에 대해 알아보자.

모의실험이 이루어진 실험조건 수를 c개, 각 실험조건에서 성능을 비교하고 싶은 추정량 종류를 m개라고 하자. m개 추정량의 성능을 비교하기 위해 모의실험을 r번 반복해서 총 cmr개의 성능측도 y가 얻어진다. 실험조건 i와 추정량 종류 j가 같으면 벡터 $\mathbf{x}$는 같아져 모형으로부터 얻어지는 $\hat{y} = \mathbf{x}^T\hat{\boldsymbol{\beta}}$는 동일하다. 설계행렬 $\mathbf{X}_{(n\times p)}$에서 실험조건이 같은 행 $\mathbf{x}^T$은 mr개 있다. 이 중에서 서로 다른 추정량에 대응하는 m개의 행으로 이루어진 행렬을 $\mathbf{X}_s$라고 하자. 동일한 $\mathbf{X}_s$가 r개 존재하겠지만 그 중 어떤 것이든 상관없다(행렬 $\mathbf{X}_s$는 실험조건에 따라 달라지는 행렬이지만 복잡해서 기호에 이 사실을 나타내지 않았다). 그러면 $\hat{\mathbf{y}}_s = \mathbf{X}_s\hat{\boldsymbol{\beta}}$는 특정한 실험조건에서 m개 추정량의 평균 성능에 대한 추정값이 된다. m개 추정량의 모든 쌍에 대해 평균 차이에 대한 동시신뢰구간을 구하기 위해 $m(m-1)/2 \times m$ 행렬 $\mathbf{D}$를 다음과 같이 정의한다.

$$\mathbf{D} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ \vdots & & & & & \\ 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & & & & & \\ 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix} \tag{2.5}$$

벡터 $\mathbf{D}\mathbf{X}_s\boldsymbol{\beta}$는 특정 실험조건에서 모든 가능한 두 추정량의 성능 차이를 나타내며, $\mathbf{D}\hat{\mathbf{y}}_s = \mathbf{D}\mathbf{X}_s\hat{\boldsymbol{\beta}}$는 $\mathbf{D}\mathbf{X}_s\boldsymbol{\beta}$의 비편향추정량이 된다. 앞 절의 (2.3)에 의하여,

$$\mathbf{D}\hat{\mathbf{y}}_s \overset{a}{\sim} N\left(\mathbf{D}E(\mathbf{y}_s), \mathbf{D}\mathbf{X}_s\hat{\mathbf{\Omega}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}_s^T\mathbf{D}^T\right) \tag{2.6}$$

이다. $E(\mathbf{y}_s) = \mathbf{X}_s\boldsymbol{\beta}$는 특정 실험조건에서 m개 추정량의 평균성능을 나타내는 벡터이며 $\hat{\mathbf{\Omega}}_{\hat{\boldsymbol{\beta}}}$는 $\hat{\boldsymbol{\beta}}$의 HC3 추정량 (2.4)이다. 추정량의 성능을 비교하려면 $m(m-1)/2$ 쌍의 평균차이를 나타내는 벡터인 $\mathbf{D}E(\mathbf{y}_s)$에 대한 동시신뢰구간을 구해야 한다.

만약 벡터 $\mathbf{D}E(\mathbf{y}_s)$의 각 원소에 해당하는 한 쌍의 추정량 성능 차이에 대한 95% 신뢰구간을 구하는 작업을 모든 쌍에 대해 하게 되면 $m = 6$인 경우 제1종오류를 범할 확률이 최대 $15 \times 0.05 = 0.75$가 될 수 있다. 신뢰수준이 $1-\alpha$ 인 동시신뢰구간을 구하는 방법 중에 Bonferroni 방법은 각 신뢰구간을 구할 때 신뢰수준을 $1-\alpha/d$

로 두어 d개의 신뢰구간 중에서 적어도 한 개의 신뢰구간이 참값을 포함하지 않을 확률이 α를 넘지 않게 하는 방법이다. 개별 신뢰구간의 신뢰수준만 조정해주면 되므로 간단히 적용할 수 있다는 장점이 있다.

Hothorn 등 (2008)은 동시신뢰구간을 구하는 다른 방법을 제시하였는데, 본 연구의 문제에 맞는 기호로 다음과 같이 요약할 수 있다. $\mathbf{D}\hat{\mathbf{y}}_s$의 추정공분산행렬 $\mathbf{D}\mathbf{X}_s\hat{\mathbf{\Omega}}_{\hat{\beta}}\mathbf{X}_s^T\mathbf{D}^T$의 대각원소의 제곱근, 즉 $\mathbf{D}\hat{\mathbf{y}}_s$의 표준편차를 대각원소로 갖는 대각행렬을 $\mathbf{S}_n^{1/2}$라고 하자. 그리고 $\mathbf{D}\hat{\mathbf{y}}_s$를 표준화한 벡터를,

$$\mathbf{T}_n = \mathbf{S}_n^{-\frac{1}{2}} \left(\mathbf{D}\hat{\mathbf{y}}_s - \mathbf{D}E(\mathbf{y}_s)\right)$$

라고 하자. $\max|\mathbf{T}_n|$을 벡터 $\mathbf{T}_n$의 원소의 절대값 중 최대값을 나타낸다고 하면,

$$P\left(\max|\mathbf{T}_n| \le t\right) = 1 - \alpha$$

를 만족하는 t를 구하면 $\mathbf{D}E(\mathbf{y}_s)$에 대한 동시신뢰구간을 구할 수 있다. Hothorn 등 (2008)은 이 방법을 R 팩키지 `multcomp`로 구현하였다.

3. 적용 예

구체적 예를 통해 모의실험 결과를 통계적 모형으로 분석하는 방법에 대해 알아본다. 이 예는 본 연구를 하게 된 계기이기도 하다. 모의실험으로 얻은 결과가 본 연구에서는 다시 분석 대상 자료가 된다. 제시하고자 하는 사례의 이해를 돕기 위해 어떤 모의실험에서 얻어진 자료인지 먼저 간략히 설명한다.

연관성(association)이 아닌 인과성(causation)을 밝히고자 하는 인과연구(causal study)에서 개입(intervention) 또는 처리(treatment) 효과를 추정할 때, 모집단 전체에 미치는 평균처리효과(average treatment effect)보다 개별처리효과(individual treatment effect) 추정에 더 관심이 있을 때가 있다. 의학분야에서 정밀의료(precision medicine) 또는 개인맞춤형의료(personalized medicine)라고 부르는 영역은 사람마다 유전적 요인이나 환경, 생활양식이 다르므로 개인별로 적합한 치료 방법이 달라질 수 있음을 가정한다. 개별처리효과를 관측자료로부터 추정할 수 있기 위해서는 자료수도 많아야 하지만 고려해야 할 변수의 수도 많아야 한다. 인구학적 기본 변수는 물론이고, 유전적 정보, 생활습관, 혈액검사 결과 등 개인 건강과 관련된 다양한 변수와 함께 개인이 속해 있는 직장과 주거 환경과 관련된 변수도 모두 고려해야 하기 때문이다. 설명변수의 수도 많아지지만 설명변수와 처리변수, 설명변수와 반응변수 사이의 관계도 복잡해질 수 있으므로 처리효과를 추정할 때 비모수적 추정 방법이 많이 쓰인다 (Lu 등, 2018; Hu 등, 2020). 비모수적 방법들의 성능을 이론적으로 비교하기 어려우므로 성능비교는 대부분 모의실험으로 이루어진다. 비모수적 방법들에 대한 구체적 설명은 여기서 불필요해 보인다. 다만 여러 연구에서 보고된 것들 중에서 6개의 유력한 비모수적 추정 알고리즘으로 얻어지는 추정량의 성능을 비교하고자 모의실험을 실시했다는 것만 알아두자.

이미 실시한 모의실험의 결과로부터 여러 추정량의 성능을 비교하기 위해 보통 Figure 1과 같이 반복실험에서 얻어진 성능측도 실현값을 상자그림으로 그려 비교하거나 평균과 분산으로 보고한다. 하지만 확연한 성능 차이가 없거나 모의실험 조건별로 성능 우위가 달라진다면 그림으로 판단하기 어려워진다. 그리고 추정량 계산에 시간이 오래 걸리면 모의실험 반복횟수를 크게 하기 어렵고 따라서 추정량의 분산을 충분히 줄일 수 없어 평균과 분산만으로 비교하기 어렵다. 이 때 다양한 실험조건에서 이루어진 모의실험 전체 결과를 하나의 통계적 모형으로 분석한다면 이런 어려움을 극복하거나 완화할 수 있다.

이 절의 내용은 모의실험이 아니고 반복적으로 이루어진 모의실험에서 얻어진 결과를 한 번 분석하는 것이다. 모의실험에 쓰인 자료와 모의실험 결과 분석에 쓰이는 자료를 구분할 필요가 있다. 모의실험에는 다양한 실험조건에서 생성된 설명변수들과 이항형 처리변수, 중도절단된(right-censored) 수명을 나타내는 반응변수에 대한 자료가 쓰인다. 하지만 이 절에서 필요한 자료는 모의실험에 쓰인 자료가 아니라 모의실험

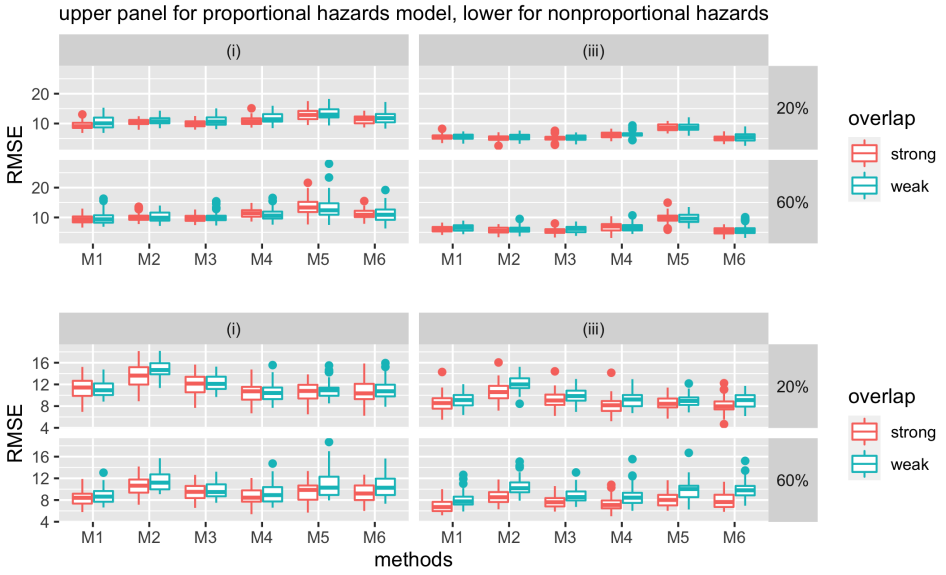


Figure 1: *Conditional boxplots to compare the performance of 6 methods in 16 experimental conditions.*

결과로 얻어지는 성능측도라는 변수와 모의실험에 쓰인 조건을 나타내는 변수에 대한 자료이다. i번째 실험조건에서 j번째 추정량에 대해 k번째 반복실험에서 얻어진 성능측도, 구체적으로 제곱근평균제곱오차(RMSE)의 실현값을 y_{ijk}라고 할 때, 이 절에서 반응변수는 y_{ijk}이다.

우리의 관심은 실험조건 i에서 6개 추정량의 성능측도 $E(y_{i1}), E(y_{i2}), \ldots, E(y_{i6})$를 비교하는 데에 있다. 특정한 추정량의 성능이 모든 실험조건에서 항상 우위를 보이는지, 아니면 실험조건에 따라 좋은 성능을 보이는 추정량이 달라지는지, 달라진다면 실험조건을 결정하는 요소 중에서 어떤 요소가 중요한지 등을 알아보고자 한다. 이런 사항들은 통계 모형을 이용하지 않고는 알아내기 힘들다.

개별처리효과에 대한 6개의 유력한 추정량이 있으며 16가지 다양한 실험조건에서 6개 추정량의 성능을 측정하는 모의실험이 반복적으로 50번 이루어졌다(50번의 반복횟수는 모의실험에서 크지 않은 수이다. 어떤 비모수적 추정량은 계산에 많은 시간이 소요되므로 반복횟수를 원하는 만큼 크게 하기 어려울 때가 있다). 따라서 y_{ijk}로 이루어진 벡터 $\mathbf{y}$의 원소 개수는 $16 \times 6 \times 50 = 4800$이다. 모의실험에서 다양한 상황을 만들기 위해 쓰인 16가지 실험조건은 4개의 이항형 변수에 의해 결정된다. 4개의 이항형 변수는 비례위험가정(proportional hazards assumption)의 성립여부, 생존시간을 생성하는 모형 두 개, 중도절단된 자료의 비율 두 가지, 처리효과를 추정할 수 있기 위한 기본가정 중의 하나인 중복가정(overlap assumption)의 만족 정도 두 가지를 각각 나타낸다(생존시간을 생성하는 두 개의 모형을 Figures 1과 4에서 (i)과 (iii)으로 표시하였다. 원래 모의실험에서는 모형에 포함되는 변수의 개수와 비선형 정도에 따라 세 개의 모형을 고려하였으나 여기서는 가장 다른 두 모형인 (i)과 (iii)만 고려하기로 한다).

추정량 성능을 비교하기 위한 모형 (2.1)에서 설계행렬 $\mathbf{X}$가 어떻게 구성되는지를 설명한다. 6개의 추정량 종류를 나타내는 다범주 변수를 R에서 하나의 인자형(factor-type) 변수로 정의한다. 16가지 실험조건을 나타내기 위한 4개의 이항형 변수도 각각 인자형 변수로 정의한다. 이 5개의 인자형 변수를 모형식(model formula)에 어떻게 포함시키느냐에 따라 모형과 설계행렬 $\mathbf{X}$가 정해진다. 추정량 종류를 나타내는 변수를 `C`라고 하고 실험조건을 나타내는 변수를 `E1, E2, E3, E4`라고 할 때, R에서 모형식을 `Y~C + E1 + E2 + E3 + E4`

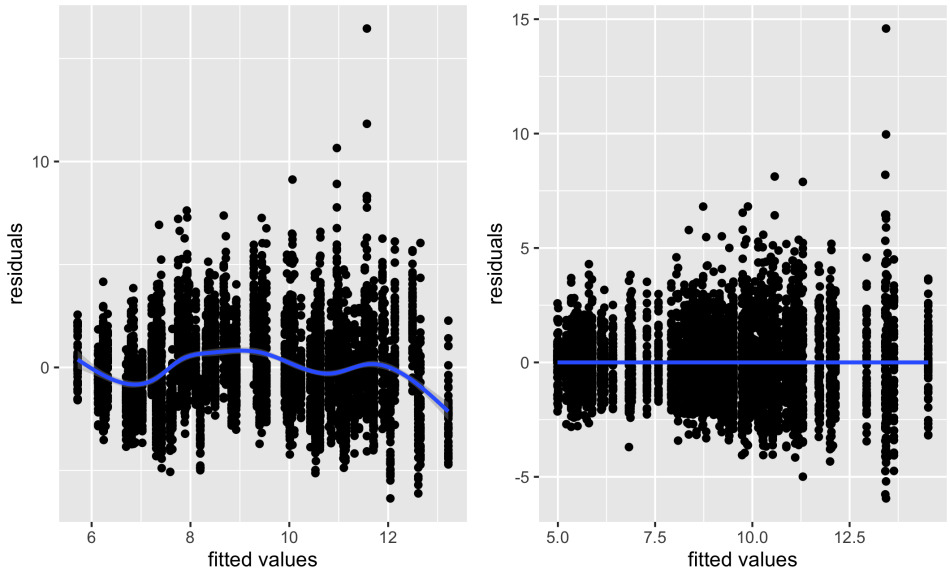
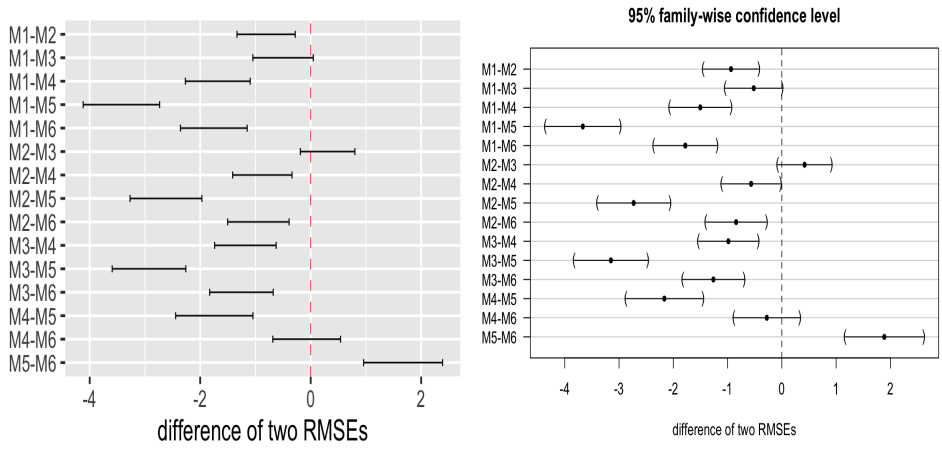


Figure 2: *Residuals vs. fitted values. Left for the model with main effects only, and right for the final model.*

라고 지정하면 추정량의 종류에 따른 주효과항을 나타내는 5개의 열과 실험조건의 주효과항을 나타내는 4개의 열이 생겨서 **X**는 절편항까지 포함해 10개의 열을 갖게 된다. 만약 모형식을 `Y~(C + E1 + E2 + E3 + E4)^5` 라고 지정하면, 주효과항 9개와 2차 상호작용항 26개, 3, 4, 5차 상호작용항과 절편항까지 포함하면 **X**는 총 96개의 열을 갖게 된다. 추정량 성능은 실험조건별로 달라질 수 있기 때문에 5개 변수의 주효과만 있는 최소모형부터 포화모형(saturated model)인 5차 상호작용항을 허용하는 모형 사이에서 단계적 변수선택(stepwise selection) 방법과 전진선택법, 후진선택법을 모두 적용해 보았는데 동일한 모형을 선택하였다. 선택된 변수의 수는 53개, 결정계수는 0.666, 수정결정계수는 0.662이었다. 여기에 53개의 항을 모두 나열할 필요는 없겠으나 추정량 종류와 실험조건과의 상호작용항, 그리고 실험조건들 사이의 상호작용항이 많이 선택되었음을 알린다. 예를 들어 추정량과 비례위험가정 성립여부, 생존시간 생성모형 사이의 3차 상호작용항이 선택되었으며 실험조건을 나타내는 4개의 이항변수 사이의 3차 상호작용항도 2개 선택되었다. 이는 추정량의 성능과 상대적 우위가 실험조건별로 달라진다는 것을 의미한다.

모형선택 또는 변수선택을 위해 R의 함수 `step`을 썼는데, `step` 함수는 모형선택을 할 때 이분산성을 고려하지 않는다. 모형선택 단계에서 중요한 것은 평균 $E[y|\mathbf{x}]$을 편향되지 않게 추정하는 모형을 찾는 것이므로 이분산성을 굳이 고려하지 않아도 된다(최종선택된 모형에서 각 항의 유의성을 판단할 때는 이분산성을 고려해야 한다. `multcomp::glht` 함수를 이용하면 된다). 평균 추정에 문제가 없는가를 평가하기 위해 적합값(fitted values)과 잔차(residuals)에 대한 그림을 이용할 수 있다. Figure 2의 왼쪽 그림은 주효과항만 있는 모형의 적절성을 진단하기 위해 그린 것이다. 벡터 **y**의 원소 개수는 4800개이지만 서로 다른 $\hat{y}$ 값의 개수는 $16 \times 6 = 96$개이다. 96개의 $\hat{y}$ 값에서 50개씩의 잔차가 얻어진다. 이 그림을 보면 잔차가 양의 방향이나 음의 방향으로 치우친 적합값이 96개 중에 여러 개 있음을 확인할 수 있다. $\hat{y}$이 평균 $E[y|\mathbf{x}]$의 추정값이므로 평균 추정에 문제가 없다면 잔차는 0을 중심으로 대칭 분포를 가져야 하는데 그렇지 못하므로 평균 추정에 문제가 있음을 의미한다. 반면에 Figure 2의 오른쪽 그림에는 이런 문제가 없다. 이 그림은 변수선택법으로 얻어진 최종 모형에 대한 진단그림이다.



(a) by Bonferroni method (b) by `multcomp` R package

Figure 3: *Simultaneous confidence intervals for multiple comparison.*

변수선택법으로 선택된 최종 모형을 적합하면 이분산성 외 다른 문제는 없는 것으로 보인다. Figure 2의 오른쪽 그림을 보면 적합값에 따라 잔차가 갖는 변동이 다름을 진단할 수 있다. 이분산성이 존재한다는 것을 통계적 검정을 통해서도 알아볼 수 있다. 다양한 검정 방법이 R에 구현되어 있는데, Levene (1960) 방법을 구현한 `car::leveneTest` 함수, Breusch와 Pagan (1979) 방법을 구현한 `lmtest::bptest` 등이 있다. 어떤 방법을 쓰든 이분산성이 존재한다는 결론을 얻었다. `leveneTest` 함수를 쓸 때 분산을 비교하고 싶은 그룹을 지정해야 하는데, 그룹을 추정량 6개로 정하든 추정량과 실험조건으로 결정되는 96개로 정하든 p값이 거의 0이었다. 이분산성 검정의 귀무가설은 등분산성이 성립한다는 것이다(표본크기가 크면 통계적 유의성이 꼭 실용적 유의성(practical significance)을 의미하지는 않는다는 가설검정의 한계에 주의해야 한다. 우리는 검정에만 의존하지 않고 그림으로 이분산성의 존재에 대한 진단을 했다).

우리의 주된 관심은 16개의 실험조건 각각에서 6개 추정량의 성능을 비교하는 것이다. 제일 나은 성능을 보이는 것이 있는지, 아니면 비슷한 성능을 갖는 추정량들이 여러 개 있는지를 객관적으로 판단하는 것이다. 이때 이분산성을 고려해야 하며 6개 추정량의 성능을 다중비교해야 한다. 모형선택 단계에서 절편항을 제외하고 53개의 항을 가진 모형이 선택되었는데, 이 모형을 $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$라고 하자. $\hat{\boldsymbol{\beta}}$의 공분산행렬 $(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}$의 일치추정량 (2.4)를 `sandwich::vcovHC` 함수를 이용하여 구하였다.

각 실험조건에서 6개 추정량의 성능을 모형으로부터 $\hat{\mathbf{y}}_{6\times 1} = \mathbf{X}_s\hat{\boldsymbol{\beta}}$와 같이 얻을 수 있다. 여기서 $\mathbf{X}_s$는 주어진 실험조건에서 6개의 추정량을 나타내는 행렬인데 행렬 $\mathbf{X}$에서 추출한 6개의 행으로 구성된다. 6개의 추정량에 대한 성능을 다중비교하기 위해 15×6 행렬 $\mathbf{D}$를 (2.5)와 같이 정의한다. 그러면 $\mathbf{D}\hat{\mathbf{y}}$는 6개의 추정량에서 모든 가능한 쌍의 추정량의 성능 차이에 대한 추정값이다. (2.6)을 이용하여 15개의 쌍에 대한 동시신뢰구간을 구할 수 있다. Figure 3 (a)는 특정한 하나의 실험조건에서 Bonferroni 방법으로 구한 동시신뢰구간으로서 직접 구현하였다. Figure 3 (b)는 $\max(|T_n|)$을 이용한 동시신뢰구간인데 R 팩키지 `multcomp`를 이용해서 그렸다. 두 그림이 매우 비슷하며 다중비교 결과도 동일하였다. 이런 결과가 이 자료에 국한된 것인지 보다 일반적인 것인지는 추가적인 연구가 필요하다. Figure 4는 총 16개의 실험조건 중에서 8개에 대한 동시신뢰구간을 한 화면에 그린 조건부그림(conditional plot)이다. 나머지 8개 실험조건에 대한 그림은 생략하였다. Figure 4와 Figure 1 상단에 있는 그림은 8개 실험조건에서 6개 추정량의 성능 차이에 대한 정보를 서로 다른 시각에서 보여주며 보완적이다.

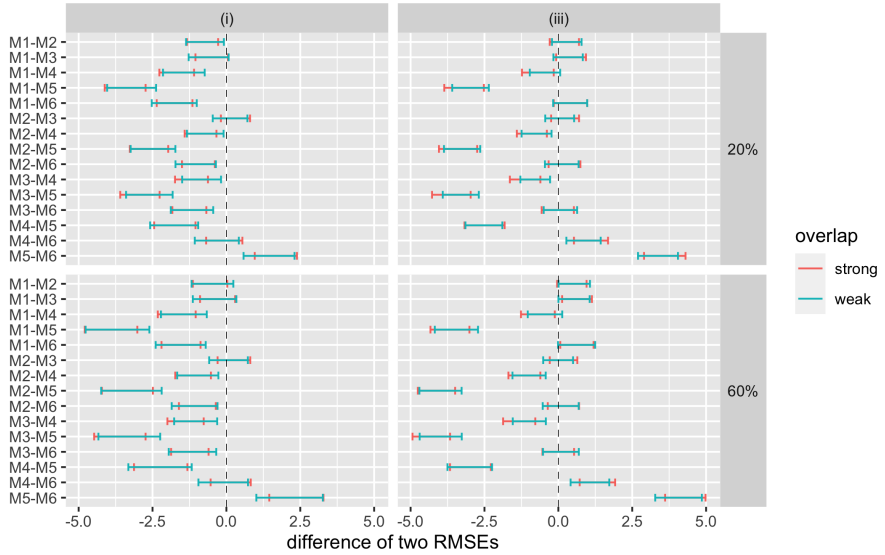


Figure 4: *Conditional plots of simultaneous confidence intervals for multiple comparison.*

동시신뢰구간 그림은 0을 기준으로 왼쪽이나 오른쪽에 위치한 신뢰구간을 바로 알아볼 수 있다는 장점이 있으나 하나의 추정량을 다른 추정량들과 비교할 때는 불편하다. 예를 들어 Figure 3에서 여섯 번째 추정량인 M6의 성능을 다른 방법들과 비교하려면 그림에서 M6가 포함된 신뢰구간 5개를 모두 찾아서 판단해야 한다. 신뢰구간에 대응하는 z값, 즉 성능차이를 표준오차로 나눈 값을 행렬로 나타내면 추정량별로 비교를 쉽게 할 수 있다. 아래 행렬 (3.1)은 Figure 3에 대응하는 z값을 계산한 행렬이다. Bonferroni 방법을 썼을 때 95% 신뢰수준에서 임계값은 $z_{1-0.025/15} \doteq 2.935$이다. 음수이면서 절대값이 임계값보다 크면 행에 있는 추정량의 RMSE가 열에 있는 추정량의 RMSE 보다 유의하게 작아서 성능이 더 좋음을 의미한다. 이 행렬을 보면 M6의 성능이 M1, M2, M3보다 떨어지고 M5보다 낫다는 것을 더 쉽게 알 수 있다. 필요하면 이 행렬을 16개의 실험조건에서 모두 출력하여 판단할 수 있다.

$$
\begin{array}{c|cccccc}
 & M1 & M2 & M3 & M4 & M5 & M6 \\
\hline
M1 & 0 & -4.493 & -2.682 & -8.406 & -14.539 & -8.518 \\
M2 & 4.493 & 0 & 1.826 & -4.776 & -11.827 & -4.997 \\
M3 & 2.682 & -1.826 & 0 & -6.225 & -12.889 & -6.398 \\
M4 & 8.406 & 4.776 & 6.225 & 0 & -7.320 & -0.346 \\
M5 & 14.539 & 11.827 & 12.889 & 7.320 & 0 & 6.875 \\
M6 & 8.518 & 4.997 & 6.398 & 0.346 & -6.875 & 0
\end{array} \tag{3.1}
$$

4. 요약 및 결론

향상된 컴퓨터의 계산 능력을 이용하는 통계적 방법들이 점점 늘어나고 있다. 비모수적 방법이라고 부르는 추정방법에 따라 얻어지는 여러 추정량의 성능을 비교할 때 모의실험을 많이 이용한다. 비교하고 싶은 추정량의 수가 많고 비교하고 싶은 모의실험 상황, 즉 실험조건 수가 많아지면 모의실험 결과를 단순히 그래프나 표로 해석하기가 어려워진다. 특히 실험조건과 추정량에 따라 추정량 성능의 변동이 다르고 추정량 성능의 상대적 우위가 다르면 통계적 모형을 이용하지 않고 분석하기 힘들다. 본 연구에서 통계적 모형을 이용하여

모의실험 결과를 분석하는 방법을 보였다. 통계적 모형을 이용하면 얻게 되는 장점은 다음과 같다.

1. 비교하고자 하는 추정량의 성능 우열이 실험조건에 따라 달라지는지 여부를 모형의 상호작용항 유의성 여부로 판단할 수 있어 보다 체계적이다.
2. 상자그림이나 평균과 분산 표로 비교할 때보다 적은 횟수의 모의실험으로 성능 비교가 가능하다.
3. 추정량과 실험조건을 결정하는 변수들이 모두 범주형이므로 평균 성능을 편향없이 추정할 수 있다(실험조건을 결정하는 정량적 변수가 있을 수 있으나 이 정량적 변수의 특정한 몇 개의 값에서 모의실험을 실시할 경우 범주형으로 간주할 수 있다).
4. 이분산성을 고려하면 평균성능 추정량의 분산에 대한 일치추정량을 얻을 수 있다.
5. 다중비교를 실시하면 여러 추정량의 성능 차이에 대한 유의성 판단을 객관적으로 할 수 있다

추정회귀계수 $\hat{\boldsymbol{\beta}}$에 대한 이분산성 일치추정량에 대한 이론과 동시신뢰구간에 대한 이론은 이미 정립되어 있다. 이를 구현하는 R 패키지도 이미 개발되어 있다. 본 연구에서는 이 두 주제를 결합하여 모의실험 결과를 체계적으로 분석할 수 있다는 것을 제안하였다. 그리고 적용 사례도 제시하여 모형을 선택하고 성능을 비교하는 절차를 보였으며 도움이 되는 그림과 출력도 제안하였다. 대규모 모의실험으로 여러 추정량의 성능을 다양한 조건에서 비교하고자 하는 연구자에게 이 연구가 도움이 되기를 바란다.

이 연구에서 제안한 방법은 오차항의 독립성을 가정한다. 오차항의 독립성에 대한 가정은 잔차로 진단할 수 있다. 진단 결과 문제가 있으면 본 연구에서 제안한 방법은 정확한 방법이 되지 못한다. 이분산성을 고려하면서 비독립성(non-independence)을 수용하는 방법에 대한 후속연구가 필요하다.

감사의 글

귀중한 시간을 할애해서 심사를 진행해주신 편집위원님과 정확하고 건설적인 심사 의견을 주신 두 분 심사위원님께 감사드립니다.

References

Breusch TS and Pagan AR (1979). A simple test for heteroscedasticity and random coefficient variation, *Econometrica*, **47**, 1287–1294.

Hothorn T, Bretz F, and Westfall P (2008). Simultaneous inference in general parametric models, *Biometrical Journal*, **50**, 346–363.

Hu L, Ji J, and Li F (2020). Estimating heterogeneous survival treatment effect in observational data using machine learning, https://arxiv.org/abs/2008.07044, under review.

Levene H (1960). Robust tests for equality of variances. In Ingram Olkin; Harold Hotelling; et al. (eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford University Press, 278-–292.

Lu M, Sadiq S, Feaster DJ, and Ishwaran H (2018). Estimating individual treatment effect in observational data using random forest methods, *Journal of Computational and Graphical Statistics*, **27**, 209–219.

MacKinnon JG and White H (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties, *Journal of Econometrics*, **29**, 305–325.

White H (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, **48**, 817–838.

Zeileis A (2004). Econometric computing with HC and HAC covariance matrix estimators, *Journal of Statistical Software*, **11**, 1–17.

Received June 21, 2021; Revised July 29, 2021; Accepted August 2, 2021

통계모형을 이용하여 모의실험 결과 분석하기

김지현[1,a], 김봉성[a]

[a]숭실대학교 정보통계보험수리학과

요 약

모의실험 결과를 보통 표나 그림으로 보고한다. 하지만 모의실험이 다양한 조건에서 실시되었고 실험조건마다 추정량의 성능 우위에 대한 결과가 다르면 표와 그림만으로 추정량의 성능을 비교하기가 어렵다. 또한 실행시간이 오래 걸리는 모의실험의 경우 모의실험 반복횟수를 크게 하기 어렵다. 회귀모형을 이용하여 모의실험 결과를 분석하면 보다 체계적이고 효과적으로 추정량의 성능을 비교할 수 있다. 이때 실험조건과 추정량에 따라 성능의 변동이 다를 수 있으므로 회귀모형에서 오차항의 이분산성을 허용해야 하며, 여러 개의 추정량을 동시에 비교해야 하므로 다중비교를 실시해야 한다. 모의실험 결과에 대한 분석이라는 맥락에서 이분산성과 다중비교에 대한 배경이론을 소개하고 예시를 통해 구체적 분석방법도 제시한다.

주요용어: 이분산성일치추정량, 공분산행렬, 동시신뢰구간, 조건부 그림

[1]교신저자: (06978) 서울시 동작구 상도로 369, 정보통계보험수리학과 숭실대학교. E-mail: jxk61@ssu.ac.kr