농업생명과학연구 48(3) pp.147-155 http://dx.doi.org/10.14397/jals.2014.48.3.147
Journal of Agriculture & Life Science 48(3) pp.147-155

R을 이용한 벌점화 축소추정 기법 비교연구: 요크셔 돼지 산자수와 SNP에 대한 적용 사례

박민수[1] · 김태헌[2] · 조은석[2] · 김희발[3] · 오희석[1*]

[1]서울대학교 통계학과, [2]농촌진흥청 국립축산과학원, [3]서울대학교 식물동물생명공학부

접수일(2014년 3월 19일), 수정일(2014년 4월 28일), 게재확정일(2014년 5월 28일)

A comparative Study of Regularized Regression Approaches using R: Application to SNP and Litter Size of Yorkshire Pigs

Minsu Park[1] · Tae-Hun Kim[2] · Eun-Seok Cho[2] · Heebal Kim[3] · Hee-Seok Oh[1*]

[1]Department of Statistics, Seoul National University, Seoul, 151-747, Korea
[2]National Institute of Animal Science, R.D.A., Suwon, 441-706, Korea
[3]Department of Agricultural Biotechnology and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, 151-747, Korea

Received: MAR. 19. 2014, Revised: APR. 28. 2014, Accepted: MAY. 28. 2014

초록

본 연구는 R을 기반으로 하여 벌점화 축소추정 기법을 활용한 유전체 선발 방법론들을 소개하고 이를 활용하여 요크셔 돼지의 산자수를 추정하고 비교 분석하고자 하였다. 비교하고자 하는 7가지의 벌점화 축소추정 기법은 다음과 같다: 능형회귀, 능형회귀 BLUP, LASSO, adaptive LASSO, fused LASSO, elastic net 그리고 SCAD. 농촌진흥청에서 제공한 519마리 암퇘지의 2,000개 유전자형 자료와 종속변수로 사용된 산자수가 연구에 사용되었다. 연구 결과의 비교 평가를 위해서 사용된 예측오차와 Pearson 상관계수에 따르면, 예측오차는 LASSO와 elastic net이 각각 0.6884과 0.6876으로 가장 작은 값을 가졌으며, 상관계수는 0.7466과 0.7388로 7가지 방법론 중에서 가장 좋은 예측력을 나타내었다. R을 활용한 이러한 연구 결과를 토대로 LASSO와 elastic net이 유전체 선발 기법 중 좋은 방법론임을 알 수 있었다.

검색어 – 유전체 선발, 암퇘지, 산자수, R 소프트웨어, 벌점 회귀모형

ABSTRACT

The aim of this study was to introduce the genomic selection methods via regularized regression in R packages for estimating litter size using dense molecular markers of Yorkshire pigs. The predictive performance of seven regularized linear methods such as ridge regression, ridge regression BLUP, LASSO, adaptive LASSO, fused LASSO, elastic net and SCAD were comparatively evaluated through the simulation study. After implementation of quality control and normalization, the whole dataset provided by the Rural Development Administration remains 47,112 SNPs and litter size as the dependence variable on 519 pigs. The randomly selected 2,000 SNPs are used in simulation study by reason of reducing computation costs. According to results of the prediction error and the Pearson correlation used as the measure of comparing the comprehensive performance, LASSO and elastic net have 0.6883 and 0.6876 for the average prediction error and 0.7407 and 0.7388 for the average correlation, respectively. From these reasons, it is concluded that LASSO and elastic net is a good approach to genomic selection using R.

Key words - Genomic selection, Sow, Litter size, R software, Regularized regression

*Corresponding author: Hee-Seok Oh
Tel: +82-2-880-2660
Fax: +82-2-883-6144
E-mail: heeseok@stats.snu.ac.kr

Ⅰ. 서론

분자적 표지인자 기술을 통해 동물의 유전적 육종가(genomic breeding value)를 추정하여 향상시키고자 하는 노력들은 꾸준히 발전되어 왔다. 표지인자와 관련된 큰 효과를 가진 대립유전자의 조절에 있어서는 효과적인 마커도움선발(marker assisted selection)은 작은 효과를 지닌 많은 대립유전자와 관련된 형질에 대해서는 신뢰할만한 효과를 찾아내지 못하였다. 또한 GWAS (genome-wide association study)를 통한 유전적인 변이를 조사하는 방법도 유의한 유전자좌 영역의 위치를 탐색하지만 결과에 대한 해석상의 어려움 및 주변 마커들간의 관계성을 배제하고 분석한다는 단점을 지닌다(Meuwissen *et al.*, 2001). 이러한 문제점의 해결책은 큰 효과를 갖는 양적형질좌위(quantitative trait loci) 위치의 소수 표지인자를 찾는 것이 아니라, 낮은 효과를 지닌 많은 표지인자들을 찾아낼 수 있는 방법을 개발하는 것에 있다. 이를 위한 방법 중에 하나가 유전체 육종가에 근거한 선발이다. 유전체 선발 기법은 유전형 및 표현형 정보를 가진 개체들을 통해서 모형을 설립하고, 새로운 개체에 대한 추정된 반응변수 값을 도출한다. SNP (single nucleotide polymorphism) 자료는 대부분 대용량 자료의 형태를 지니고 있고, 주변 변수들이 상관성을 지니고 있기에 계산상의 효율과 추정치의 정확도를 동시에 향상하고자 하는 방법론들이 다양하게 개발되고 있다 (Endleman, 2011; Ogutu *et al.*, 2012).

본 연구에서는 돼지의 어미 돼지의 생산성과 관련있는 산자수와 유전자형 자료를 데이터 분석을 위한 통계 및 그래픽스를 지원하는 통계 오픈소프트웨어인 R (www.r-project.org)을 통해 최근에 변수선택법으로 사용되고 있는 통계적 방법론의 내용 및 사용 방법 등을 소개하고 비교 분석하고자 한다.

Ⅱ. 재료 및 방법

본 연구에서 사용된 자료는 농촌진흥청(Rural Development Administration)에서 제공받았으며, 702두 돼지의 Illumina Porcine 60K SNP Beadchip 유전자형 자료와 산자수, 산차수 그리고 임신기간 등과 같은 표현형 자료로 구성되어 있다. 돼지의 개체수는 702두이었으며, 표현형을 가진 돼지 4,163마리 중에 유전자형을 지닌 돼지와 부합하고 중복된 개체를 제거하여 분석에 이용하였다.

2.1 자료의 전처리 과정

농촌진흥청에서 제공받은 702두의 자료에 대한 전처리 과정을 진행하였다. 본 연구에서 Illumina 칩을 사용한 60K SNP 마커의 702두에 대한 유전자형 정보에 대하여 실제 분석에 사용할 마커만을 선별하기 위한 전처리 과정(quality control)과 결측치 정보에 대한 예측(imputation) 과정을 수행하였다. 전처리 과정에는 PLINK (Purcell *et al.*, 2007) 프로그램을 사용하였으며 그 기준은 각 마커별 소수 대립유전자 빈도(minor allele frequency)〈0.01, 결측 유전자형(missing genotype)〉0.5 포함 그리고 하디-와인버그 평형 검정(Hardy-Weinberg equilibrium test): p〈0.0001의 기준에 반하는 자료는 제거하였다. 전처리 과정의 마무리는 분석에 사용할 수 있도록 유전자형에 대하여 각 마커별로 소수유전자와 다수유전자의 빈도를 계산하여 소수 동형 유전자형을 0, 다수 동형유전자형을 2 그리고 이형유전자형을 1로 나타내었다.

돼지의 특징을 나타내는 표현형 자료에는 다음의 정보가 포함되어 있었다. 1회 분만으로 출산한 돼지의 수인 산자수, 한 마리 돼지의 과거분만 횟수를 나타내는 산차수, 출산된 돼지의 생존과 사산된 숫자, 그리고 임신기간까지 다양한 변수들이 돼지의 출산과 관련된 변수들로 포함되어 있었다. 이 중에서 본 연구에서 중점적으로 다룰 산자수와 다른 변수들간의 상관관계를 비교해 보았을 때 통계적으로 유의하지 않은 변수들이 존재하며, 또한 해석상의 어려움을 가진 변수들이 포함되어 있기에 종속변수로 사용할 변수는 산자수와 산차수만을 고려하기로 하였다. 산차수는 1회부터 12회까지 존재하고 산차

수가 커질수록 해당 산차에 포함되는 개체는 급격히 줄어들었기에 산차별 산자수가 동일한지를 확인하기 위해서 Tukey HSD 검정을 시행하였으며, 통계적으로 유의하지 않았으므로 산차별 산자수는 동일하다고 판단하여 돼지의 개체가 급격히 줄어들지 않는 범위 내에서 평균값을 계산하여 평균 산자수를 구하여 사용하였다. 종속변수로 평균 산자수를 사용하기 위해서는 정규성 가정(Gaussianity assumption)을 만족시키는지를 확인하여야 한다. R 패키지 {MASS} 안에 boxcox 함수를 사용하여 박스콕스변환(Box-Cox transformation)을 하여 자료가 정규성을 만족시키기 위한 변환을 취하였으며, 변환된 자료가 정규성을 만족하는지는 확인하기 위해서 내장 패키지인 {stats}의 shapiro.test 함수를 통해 정규성 검정 방법 중의 하나인 Shapiro-Wilk 정규성 검정(Royston, 1982) 시행하여 통계량 값은 0.9856이고 유의확률은 0.3으로 결과를 도출하였으며, 이를 통해 변환된 평균 산자수가 정규성 가정을 만족한다는 사실을 알 수 있었다.

이상의 과정을 통하여 702두의 61,177개의 SNP 마커 중에 519두의 47,112 SNP 마커를 추출하였으며, 본 연구의 목적이 변수선택 방법론들의 비교이기 때문에 결과해석의 용이함을 위해서 2,000 SNP만을 임의추출하여 분석에 사용하였다.

2.2 통계적 방법론

2.2.1 모형

돼지의 산자수에 영향을 미치는 SNP의 효과를 추정하기 위해서 본 연구에서는 여섯 가지 선형 회귀 방법이 사용하여 비교 분석하였다. 기본 모형은 다음과 같이 표현할 수 있다.

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \quad (1)$$

여기서,

$\boldsymbol{y}$: 산자수에 대한 벡터

$\boldsymbol{\beta}$: SNP 효과에 관한 벡터

X : 유전자형을 나타내는 행렬

$\boldsymbol{\epsilon}$: 정규분포 을 따르는 i.i.d. 랜덤 오차.

우리의 목표는 주어진 자료를 통계 소프트웨어인 R을 통해 산자수에 영향을 주는 SNP를 추정하고 다양한 모형을 비교 대조하여 SNP효과를 가장 잘 설명해주는 최적의 모형을 선택하는데 있다. 모형을 통해서 추정되는 $\hat{\boldsymbol{\beta}}$은 변수선택과 추정을 동시에 할 수 있는 regularization 기술로 얻어지게 되며, 본 연구에서 사용되는 방법론들은 이어서 설명하고자 한다.

2.2.2 능형회귀(Ridge regression)

기존의 모형 추정 방법론들의 마커 선발에서 회귀계수를 과잉 추정하는 문제를 피하기 위해서 회귀계수의 평균 쪽으로 줄어들게 고안된 방법론이 능형회귀이다(Hoerl and Kennard, 1970; Whittaker *et al.*, 2000). 이러한 수축(shrinkage)은 모든 회귀계수의 효과들에 대하여 동시에 적용된다. 능형회귀의 추정치는 제약조건(constraint) $\sum_{j=1}^{p}\beta_j^2 \le t^2$ 하에서

$$\min_{\beta}\left(\sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$

로 표현할 수 있다. 여기서 $t \ge 0$이고, $t = \infty$이면 모형은 최소제곱법과 동일하다. 능형회귀는 라그랑즈 승수법(Lagrange multiplier)에 의해서 벌점함수로 다음과 같이 표현할 수 있다.

$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\beta} \| \boldsymbol{y} - X\boldsymbol{\beta} \|^2 + \lambda \| \boldsymbol{\beta} \|_2^2$.

여기서, $\lambda \ge 0$는 능형회귀의 조율모수(tuning parameter)이며, λ를 포함하고 있는 식을 능형회귀의 벌점함수이라 한다. 조율모수 t 또는 λ는 교차확인법(cross validation)을 통해 흔히 선택되며, 이를 통해 얻어진 추정치는 최소제곱의 추정치보다 축소되어 편의(bias)는 있지만 분산(variance)을 줄이면서 추정된 모형으로부터 얻을 수 있는 예측오차를 줄일 수 있다는 장점을 지닌다. 본 연구의 실제 분석에서는 R 패키지 {glmnet}을 사용하였으며, glmnet 함수 옵션 중에 alpha=0을 이용하면 능형회귀를 추정할 수 있고 종속변수가 정규성 가정을 만족하였기 때문에 family 옵션을 아래와 같이 나타낼 수 있다(Park *et al.*, 2011). 교차분석법을 사용하기

위해서는 패키지 내 cv.glmnet 함수를 사용한다.

```
library(glmnet)          # load glmnet
glmnet(x=X, y=y, family="gaussian", alpha=0)
                         # fitting ridge regression
```

2.2.3 능형회귀 BLUP (Ridge regression best linear unbiased prediction: RR-BLUP)

능형회귀 BLUP은 조율모수를 추정하여 추정량을 구하는 능형회귀와 달리 변수를 고정효과가 아닌 임의효과로 다루는 모형을 기본으로 한다(Piepho, 2009). 모형 (1)에서 $\boldsymbol{X\beta}$ 대신에 임의효과를 표현하는 $\boldsymbol{Zu}$를 사용한다. 여기서, $\boldsymbol{u}$는 $\boldsymbol{u} \sim N(0, A\sigma_u^2)$의 공분산 구조를 가진 모형에서 여러 유전자의 효과에 대한 벡터이다. 혈연 계수 행렬(numerator relationship matrix) A는 집단의 혈통으로부터 구성할 수 있고, σ_u^2는 여러 유전자 효과에 대한 분산이다. 능형회귀 BLUP은 {rrBLUP} 패키지를 사용하였으며 다음의 mixed.solve 함수를 사용하여 분석하였다.

```
library(rrBLUP)          # load rrBLUP
mixed.solve(y=y, Z=X)    # fitting ridge
                              regression BLUP
```

2.2.4 LASSO (Least absolute shrinkage and selection operator)

회귀분석에서 회귀계수를 추정할 때 일반적으로 사용되는 최소자승법의 단축된 형태인 LASSO는 능형회귀와 다르게 축소와 변수선택을 동시에 실행함으로써 예측력을 향상시키고 적은 수의 변수만을 선택하여 추정된 모형에 대한 해석 또한 용이하게 할 수 있다(Tibshirani, 1996). 이 방법은 회귀 계수의 절대값 합을 줄이기 위해 잔차 제곱 합을 최소화한다. LASSO에서 SNP 효과는 임의효과로 가정하며 추정치는 제약조건 $\sum_{j=1}^{p}|\beta_j| \le t$ 하에서

$$\min_{\beta}\left(\sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$

로 표현되고 $t \ge 0$이다. 상수 t는 일부 추정된 SNP 효과에 대해서 정확히 0을 가능하게 한다. LASSO에서 사용되는 벌점함수는 $\lambda \| \beta \|_1$이고, 이는 능형회귀가 모든 마커 효과를 추정하는 0이 아닌 값으로 추정하는 것과 다르게 일부의 SNP마커만을 선발하여 종속변수를 추정하게 된다. LASSO는 R 패키지 {glmnet}을 사용하였고, 능형회귀와 같은 glmnet 함수에 옵션을 alpha=1로 바꾸면 LASSO 추정으로 바뀌게 된다. LASSO를 추정하기 위한 패키지는 다양하게 제공되고 있지만 최적화 문제를 푸는 방법이 서로 다르기 때문에 패키지 내의 함수에 따라서 계산속도가 다르게 된다. 다양한 패키지들을 비교해 볼 때 glmnet 함수가 계산속도가 다른 함수들에 비해 효과적임을 확인하였다. 사용 방법은 능형회귀의 alpha에 대한 옵션을 제외하곤 동일하다.

```
library(glmnet)          # load glmnet
glmnet(x=X, y=y, family="gaussian", alpha=1)
                         # fitting LASSO
```

2.2.5 Fused LASSO

SNP를 자료를 통한 변수선택법은 변수들간의 성질들에 대해서 고려해야 한다. 인접한 SNP간에는 비슷한 성질을 지닌 경우가 많으며 실제 자료상에서는 코딩 결과가 비슷하다는 것을 확인할 수 있다. 즉, 순서화된 변수들을 지닌 자료에 대해서는 인접 변수들 간의 상관성을 고려한 fused LASSO (Tibshirani *et al.*, 2005)가 효과적이다. Fused LASSO의 벌점함수는 다음과 같다.

$$\lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=2}^{p}|\beta_j - \beta_{j-1}|.$$

λ_1은 LASSO에서 사용한 조율모수와 같으며 인접 회귀계수를 조율하는 λ_2가 추가된다. Fused LASSO는 강한 양의 상관관계가 있는 변수들을 지닌 자료에 대해서 효과적인 반면 계산량이 많다는 단점이 있다. Fused LASSO를 구현할 수 있는 패키지는 다른 방법론에 비해 다양하지 않으며 계산속도 또한 다른 방법론에 비해 현저하게 많다는 어려움이 있다. 계산상의 효율이 높은 방법론을 MATLAB에서는 제공되고 있지만 아직 R에서는 제공되지 않는다(Liu

et al., 2010). 사용 가능한 패키지 중에는 {lqa}가 있으며 옵션 중에 method를 변화시킴으로 다양한 방법론들을 구현할 수 있다. 교차분석법을 사용하기 위해서는 cv.lqa를 사용한다.

```
library(lqa)                    # load lqa
lqa(y~X, family=gaussian(), penalty=fused.lasso
(c(lambda1, lambda2)))
                                # fitting fused LASSO
```

2.2.6 Adaptive LASSO

LASSO의 경우 추정량의 경우 편의가 발생하는 대신 분산을 줄이는 방법을 사용하고 있으며, 이의 편의를 줄이기 위해 제한된 방법론이 adaptive LASSO이다(ALASSO; Zou, 2006). LASSO와 다른 점은 벌점함수에 다음의 가중치를 사용한다는 것이다.

$$\omega_j = \left(\left|\hat{\beta}_j^{con}\right|\right)^{-\gamma},\ j=1,...,p.$$

여기서, $\gamma > 0$이고 $\hat{\beta}_j^{con}$는 일치성(consistency)을 만족하는 추정량으로 주로 최소제곱법이나 능형회귀를 통해서 추정치를 구할 수 있다. 설명된 가중치를 사용한 ALASSO의 벌점함수는 $\lambda\sum_{j=1}^{p}\omega_j|\beta_j|$이다. 변수의 차원의 수가 적을 경우에는 최소제곱법에 의한 추정량이 좋은 결과를 도출하지만, 차원이 큰 경우엔 최소제곱 추정량이 좋은 추정량이 되지 못하므로 ALASSO는 차원이 큰 경우에 적용하기 어렵고, 최초 추정치의 선택이 어렵다는 단점이 있다(Park *et al.*, 2011). ALASSO의 경우 R 패키지 {lqa}를 사용 가능하지만 {parcor} 패키지를 사용하면 계산속도와 결과의 안정성 면에서 이점을 가질 수 있다. 패키지 내의 adalasso 함수를 이용하고 함수 내에 교차분석법을 같이 사용할 수 있다.

```
library(parcor)                 # load parcor
adalasso(X=X, y=y)              # fitting ALASSO
```

2.2.7 Elastic net

능형회귀와 LASSO의 벌점함수를 절충안으로 사용하는 elastic net은 자료의 크기에 비하여 높은 상관도를 지닌 많은 추정 변수가 있을 때 유용하게 사용되는 방법론이다(Zou and Hastie, 2005). Elastic net의 추정치는 제약조건 $(1-\alpha)\sum_{j=1}^{p}|\beta_j|+\alpha\sum_{j=1}^{p}\beta_j^2 \le t$하에서 $\min_{\beta}\left(\sum_{i=1}^{n}y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2$로 표현되며 $t \ge 0$이고, 여기서 α는 다음과 같이 표현된다.

$$\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}.$$

Elastic net의 벌점은 능형회귀의 벌점($\alpha=1$)과 LASSO의 벌점($\alpha=0$) 사이의 절충값을 갖는다. LASSO의 경우 많은 상관성 있는 변수들 중에서 하나의 변수만을 선택한다는 단점을 보완하여 상관관계에 있는 모든 변수들을 선택하도록 벌점함수를 구성하였다. Elastic net은 능형회귀와 LASSO의 절충안이므로 {glmnet}을 사용하여 분석 할 수 있다. 옵션 내에 alpha 값은 0과 1사이의 값을 0.01로 나누어 모든 값들을 비교하면서 찾는 방법을 선택하였다.

```
library(glmnet)                 # load glmnet
glmnet(x=X, y=y, family="gaussian", alpha=(0,1)
중에 선택)                      # fitting elastic net
```

2.2.8 SCAD (Smoothly clipped absolute deviation)

LASSO의 벌점함수가 볼록성을 가지며 편의가 생기는 점을 줄이기 위한 방법으로 비볼록 벌점함수를 사용한 방법론이 SCAD이다(Fan and Li, 2001). SCAD의 벌점함수는 다음과 같다.

$$\lambda I(|\beta| \le \lambda) + \frac{(\alpha\lambda-|\beta|)_+}{(\alpha-1)\lambda} I(|\beta| > \lambda),\ \alpha > 2.$$

SCAD는 LASSO보다 절대값이 큰 회귀계수에 대해서 축소되는 양을 줄이는 벌점함수를 사용하고 있으며, 벌점함수가 볼록 벌점함수를 지니지 않으므로 최적화 문제의 계산이 어렵다(Park *et al.*, 2011). R에서는 {ncvreg} 패키지를 사용하며 SCAD와 비슷한 방법론들이 같이 포함되어 있어서 분석하기 용이하다. 옵션 중에 penalty를 SCAD로 세팅하고 분석한다. 교차분석법을 사용하려면 cv.ncvreg 함수를 통

해서 최적의 모수를 추정할 수 있다.

```
library(ncvreg)                    # load ncvreg
ncvreg(X=X, y=y, family="gaussian",
penalty="SCAD")                    # fitting SCAD
```

2.3 모형비교

본 연구에서는 모형들 간의 비교 분석을 위해 5개 묶음 2단계 교차분석법을 사용하였다. 1단계의 교차분석법은 모형들을 비교하기 위해서 사용되며, 전체 자료를 5개의 배반된 묶음의 자료로 먼저 나눈다. 5개의 묶음 중에 4개의 묶음은 모형을 설립하는데 사용되고 나머지 한 묶음은 추정된 모형에 새로운 자료를 적용시켰을 때에 실제 값과 얼마나 차이가 있는지를 확인하기 위하여 사용된다. 모형을 비교하기 위해서 논문에서 사용한 도구는 실제 값과 추정된 값의 예측 오차(prediction error)와 상관계수(correlation)를 통해서 추정된 모형이 잘 적합한지를 판단하기로 한다. 예측 오차는 다음과 같이 정의한다.

$$\mathrm{PE} = \sqrt{\frac{\| y_{val} - X_{val}\hat{\beta}_{tra} \|^2}{n_{val}}}.$$

여기서, val은 1단계 test-set을 의미하며, tra는 training-set으로부터 얻어진 예측값을 의미한다.

모형을 추정하기 위해서는 모형 내의 구성요소로서 추정해야 하는 다양한 모수들이 존재하기 때문에 2단계에서는 1단계에서 모형을 추정하기 위해서 추출된 4묶음을 다시 5개의 묶음으로 나뉘어서 교차분석법을 사용하여 모수를 추정하게 되며, 모수 후보군 중에 가장 작은 교차확인오차를 갖는 모수 추정값을 선택하게 된다.

Ⅲ. 결과

본 연구에서는 다양한 변수선택법을 비교 분석함으로 SNP 자료를 통한 결과를 살펴보고 이를 통해 어떠한 방법론들이 유용한지를 알아보고자 한다. 사전분석에서 모든 방법론들의 5묶음 교차분석법에 대한 정확도의 결과가 상대적으로 낮은 값들로 측정되어서, 각 분석 방법 별 정확도를 향상시켜 결과의 구분을 명확하게 하고자 부스트랩 샘플을 통해 보완하고자 한다. 부스트랩은 주어진 자료로부터 복원추출을 통해 분석에 필요한 만큼의 자료를 만들 수

Table 1. Pearson correlation as accuracy and average correlation value of each fold for regularization linear models in simulation study

Fold	Regularization linear models						
	Ridge	RR-BLUP[1]	LASSO	Fused LASSO	Adaptive LASSO	Elastic net	SCAD
1	0.755	0.770	0.788	0.784	0.627	0.783	0.578
2	0.637	0.687	0.733	0.695	0.702	0.694	0.578
3	0.700	0.675	0.668	0.634	0.575	0.700	0.508
4	0.758	0.768	0.788	0.733	0.644	0.779	0.542
5	0.707	0.751	0.756	0.708	0.655	0.738	0.552
Ave corr[2]	0.712^{ab}	0.730^{ab}	0.747^{a}	0.710^{ab}	0.641^{bc}	0.739^{a}	0.552^{c}

a, b, and c Means in the same row with different superscirpts differ significantly ($p<0.05$).
[1]RR-BLUP: Ridge regression best linear unbiased prediction.
[2]Ave corr: average correlation.

Table 2. Prediction error (PE) of each fold and average prediction error for regularization linear models and MFPE

Fold	MFPE[1]	Regularization linear models						
		Ridge	RR-BLUP[2]	LASSO	Fused LASSO	Adaptive LASSO	Elastic net	SCAD
1	1.074	0.718	0.722	0.657	0.703	0.842	0.661	0.890
2	1.037	0.799	0.806	0.706	0.821	0.742	0.748	0.866
3	1.057	0.767	0.835	0.790	0.893	0.870	0.751	0.913
4	0.973	0.648	0.664	0.609	0.726	0.746	0.611	0.819
5	0.987	0.706	0.679	0.680	0.741	0.746	0.667	0.825
Ave PE[3]	1.025	0.728^{b}	0.741^{ab}	0.688^{b}	0.777^{ab}	0.789^{ab}	0.688^{b}	0.863^{a}

a, b, and c Means in the same row with different superscirpts differ significantly ($p<0.05$).
[1]MFPE: mean fitting prediction error (the root mean squared error with respect to fitted mean of litter size from the first-deep training set).
[2]RR-BLUP: Ridge regression best linear unbiased prediction.
[3]Ave PE: average prediction error.

있는 방법이다. 최초에 사용한 암퇘지는 519마리로 이를 부스트랩 샘플을 이용하여 700두를 사용하였으며, 이 중 중복된 자료는 35%이다.

방법론들을 비교하기 위한 결과값으로는 상관계수와 예측 오차를 사용하기로 한다. 각 방법론의 예측 오차와 비교할 수 있는 객관적인 평가 요소인 종속변수의 평균값으로 예측 오차를 계산한 평균적합 예측 오차(mean-fit prediction error: MFPE)를 비교하고자 한다. 평균적합 예측 오차는 앞선 예측 오차의 정의에서 $\mathbf{X}_{val}\hat{\boldsymbol{\beta}}_{tra}$을 training-set의 종속변수 평균값인 $\bar{\mathbf{y}}_{tra}$으로 대신하여 정의된다.

Table 1은 각 묶음에 따른 상관계수를 나타낸 표이다. 상관계수는 추정된 값과 실제 값과의 상관성을 확인하는 것이므로 높은 값을 갖는 방법론이 좋은 모형이라고 할 수 있다. 표에 따르면 각 묶음별로 LASSO와 elastic net의 총 묶음의 평균 상관계수 값이 adaptive LASSO나 SCAD에 비해 유의적($p<0.05$)으로 높게 나타났다. Adaptive LASSO는 차원의 수가 개체 수보다 클 경우엔 최소제곱법에 의한 추정량이 좋은 추정량이 되기 어렵기 때문에 LASSO에 비해 좋지 않음을 볼 수 있다. Fused LASSO는 능형회귀와 비슷한 결과를 보였는데, 이는 분석 방법의 특성상 주변의 인접한 변수들의 관계를 고려하여 유의한 많은 변수들을 선택하는 방법을 택하고 있는데 계산상의 편의를 위해서 4만 여개의 SNP 원자료 중에 2,000개만을 랜덤하게 선택하여 분석한 결과이어서 elastic이나 LASSO보다 낮은 상관관계를 보였지만 통계적으로 유의한 수준의 차이는 아니었다.

Table 2는 예측 오차 및 평균적합 예측 오차를 표현하였다. 상관계수와 비슷한 결과를 보였으며, 오차가 작을수록 추정된 모형이 원자료를 잘 표현한 결과이므로 다른 모형에 비해 상대적으로 작은 예측 오차를 갖는 모형이 변수선택과 추정에 효과적임을 나타낸다. 결과 값에 비추어 보면 능형회귀, LASSO 그리고 elastic net의 예측 오차 값이 SCAD 보다는 유의적($p<0.05$)으로 낮게 나타났으며, SCAD는 상관계수의 결과 뿐 아니라 예측 오차를 비교한 결과에서도 가장 좋지 않은 결과를 보였다.

Table 3은 각 묶음에 따른 모형의 0이 아닌 회귀계수의 개수를 나타낸다. 즉, 선택된 변수의 개수를 나타낸다. 먼저, 능형회귀 분류는 축소 추정을 할 뿐

Table 3. Number of non-zero estimated coefficients by regularization linear models of each fold (total SNPs: 2,000)

Fold	Regularization linear models						
	Ridge	RR-BLUP[1]	LASSO	Fused LASSO	Adaptive LASSO	Elastic net	SCAD
1	2,000	2,000	184	2,000	60	293	70
2	2,000	2,000	96	2,000	77	304	80
3	2,000	2,000	167	2,000	66	996	73
4	2,000	2,000	135	2,000	82	266	87
5	2,000	2,000	167	2,000	74	789	87
Ave num[2]	2,000	2,000	149.8	2,000	71.8	529.6	79.4

[1]RR-BLUP: Ridge regression best linear unbiased prediction.
[2]Ave num: average number of estimated coefficients.

변수선택은 하지 않으므로 본 연구에서 사용한 2,000개의 변수가 모두 추정되었다. Fused LASSO 또한 2,000개의 변수를 모두 선택하여 추정하였는데 이는 변수선택을 위한 모수 추정이 굉장히 작은 값을 선택하여 모형을 세워서 변수선택이 거의 일어나지 않았음을 확인할 수 있다. SCAD와 adaptive LASSO가 적은 변수만을 사용하여 모형을 세운 반면 elastic net은 평균적으로 500여개의 변수를 사용하여 모형을 설립하였다. Elastic net의 경우 LASSO에 비해 상대적으로 많은 변수를 선택하였는데, 이는 주변의 영향력 있는 변수들을 함께 선택하는 모형의 특징으로 인해서 더 많은 변수 선택이 있었음을 확인 할 수 있다.

Ⅳ. 결론

본 연구에서는 돼지의 산자수에 영향을 주는 SNP의 효과를 확인하기 위해서 변수선택법과 추정에 주로 사용되는 다양한 방법론에 대해서 설명하고 모형의 결과를 통해서 어떠한 모형들이 잘 적합 하는지를 확인하고자 하였다. 설명된 모든 방법론들은 통계 소프트웨어인 R을 통해서 구현하였으며, 각 방법별로 사용한 패키지와 함수를 간략하게 설명하여 방법론을 사용하고자 하는 분들에게 소개하고자 하는 것이 목적이다. 도출된 결과의 명확한 구분을 위해서 519두의 돼지를 부스트랩 기법을 사용하여 700두의 샘플로 사용하였으며, SNP는 분석 가능한 47,112개의 자료 중에 분석에 용이한 2,000개의 변수만을 랜덤하게 선택하고 순서화시켜서 방법론들을 비교하는데 사용하였다. 모형의 비교를 위해서 상관계수, 예측오차 그리고 0이 아닌 선택된 회귀계수 개수를 계산하였으며, 본 연구에서 사용한 자료를 통해서 얻어진 결과에 따르면 LASSO와 elastic net이 일부 변수만을 사용하여 가장 좋은 예측력을 나타냈으며, 가장 낮은 예측 오차를 갖는다는 사실을 알 수 있었다. 아직 R에서는 임의효과를 고려한 모형 분석 방법들이 패키지로 제공되지 않기에 본 연구에서는 능형회귀 BLUP만을 고려하였지만 현재 발전하고 있는 혼합모형을 고려한 모형들에 대한 분석 또한 앞으로 연구하고자 하는 방향이다.

Ⅴ. 감사의 글

본 논문은 농촌진흥청 차세대 바이오그린21사업

(과제번호:PJ008068)의 지원에 의해 연구된 것임.

≫ Literature cited

Endelman, J. B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 4:250-255.

Fan, J. and R. Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Statist. Assoc. 96:1348-1360.

Hoerl, A. E. and R. W. Kennard. 1970. Ridge regression: biased estimation for nonorthogonal problems. Technomterics. 12:55-67.

Liu, J., L. Yuan and J. Ye. 2010. An efficient algorithm for a class of fused LASSO problems. In Advances in Neural Information Processing Systems (NIPS). Curran Associates, Inc., Vancouver, BC.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819-1829.

Ogutu, J. O., T. Schulz-Streeck and H. P. Piepho. 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc. 6(Suppl. 2):S10.

Park, C., Y. Kim, J. Kim, J. Song and H. Choi. 2011. Data mining with R. pp.297-319. Kyowoo Pub. Co., Seoul, Korea.

Piepho, H. P. 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49:1165-1176.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas and M. Ferreira. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559-575.

Royston, P. 1982. An extension of Shapiro and Wilk's *W* test for normality to large samples. Appl. Stats. 31:115-124.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. Series B. 58:267-288.

Tibshirani, R., M. Saunders, J. Zhu and K. Knight. 2005. Sparsity and smoothness via the fused LASSO. J. R. Stat. Soc. Series B. 67:91-108.

Whittaker, J. C., R. Thompson and M. C. Denham. 2000. Marker-assisted selection using ridge regression. Genet. Res. 75:249-252.

Zou, H. 2006. The adaptive lasso and its oracle properties. J. Am. Statist. Assoc. 101:1418-1429.

Zou, H. and T. Hastie. 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B. 67:301-320.