Доклади на Българската академия на науките
Comptes rendus de l'Académie bulgare des Sciences
Tome 71, No 9, 2018

AGRICULTURAL SCIENCES
Plant breeding

ADAPTIVE LASSO ANALYSIS FOR GRAIN YIELD AND YIELD COMPONENTS IN TWO-ROWED BARLEY UNDER RAINFED CONDITIONS

Suna Akkol, Diğdem Arpalı*, Mehmet Yağmur**

(*Submitted by Academician A. Atanassov on March 7, 2018*)

Abstract

The goal of this study was to determine the yield components related to grain yield in order to improve barley yield under rainfed conditions of Turkey (Van). Stepwise and Adaptive Lasso methods were performed for selection of most significant yield components. As cohesion criteria to compare Stepwise and Adaptive Lasso methods, the adjusted coefficient of determination and Akaike Information Criterion were used. Results revealed that when there were dependencies between independent variables stepwise and Adaptive Lasso achieved the same results. It has been determined that spike number per m^2 and grain weight per spike can be used as the most effective selection criteria for barley breeding studies due to their significant effects on grain yield.

Key words: lasso regression, variable selection, shrinkage methods, two-rowed barley, grain yield

1. Introduction. Barley used as animal feed and malt material in our country has a production area of 2.7 million hectares and production value 6.7 million tons [1]. Yield is under the influence of genetic structure and environmental factors. Grain yield in cereals can be improved by understanding the interrelationships among yield, yield components, vegetative growth, and growth durations. The success of plant breeding programmes is increased by accurately determining selection criteria related to grain yield [2].

Multiple linear regression (MLR), correlation and path analysis have been used in agriculture by plant breeders in order to assist in identifying components

DOI:10.7546/CRABS.2018.09.17

as selection criteria to increase crop yield [3–5]. Variable selection is important in statistical modelling because it leads to better model interpretation and high prediction accuracy. In MLR Ordinary Least Square (OLS) estimates have two important problems: prediction accuracy and interoperation. The first problem means that OLS estimates have low bias but large variance. The second problem means that analysis having a large number of predictors cannot achieve smaller and stronger predictors because of dependencies in independent variables [6]. A new technique called adaptive least absolute shrinkage and selection operator (Lasso) [6] was used by researcher to provide prediction accuracy and interoperation [7]. It was shown that LASSO does not possess oracle properties [8] and an evolution of the Lasso was offered named Adaptive Lasso [9]. Adaptive Lasso regression is proposed in order to improve the lack of oracle properties of the Lasso [9]. This method has a procedure which successfully makes prediction and simultaneously variable selection as Lasso regression.

This study aimed to estimate the relationship between barley grain yield and yield components and to determine most important components for barley breeders in order to improve yield under rainfed conditions of Turkey (Van). MLR, stepwise and Adaptive Lasso techniques were performed in order to achieve this goal.

2. Materials and methods. ***Materials.*** Tokak 157/37, Tarm-92, Orza-96, Bülbül-89, Yesevi-93, Aydanhanım, Zeynelağa, Kalaycı-97, Cildir-02, Karatay-94, Efes-3 two rowed barley varieties were used in the study. These varieties have been developed by The Field Crops Central Research Institute, The Anatolian Agricultural Research Institute, The Bahri Dagdas International Agricultural Research Institute and Anadolu Beer Malt and Food Organisation. The field experiments were conducted for two years in 2004–2006 in the province of Van, east of Turkey ($38 \circ 55'$N, $42 \circ 05'$E, 1.725 m above sea level).

The region has continental climate with limited precipitation in summer and low temperature in winter. The soil of the experimental field has a clayey structure and is slightly alkaline. It is insufficient in organic matter and nitrogen, and moderate in phosphorus [10]. The average rainfall of the first year was 357.9 mm and the second year's was 416.9 mm.

Methods. The plot size was 9.6 m^2 in the study. Seeding was done in October with 8 rows of each plot and row spacing was 20 cm. The seeding rate was adjusted for a density of 500 viable seeds m^{-2}. Plots were fertilized at seeding time with 64 kg P_2O_5 ha^{-1} and 25 kg N ha^{-1} (18% N and 46% P_2O_5, diammonium phosphate) and 35 kg N ha^{-1} (21% N, ammonium sulphate) was applied at booting stage.

The data included days to heading (days), spike number per m^2 (number), spike length (cm), grain number per spike (number), grain weight per spike (g), thousand grain weight (g), harvest index (%) and grain yield (kg da^{-1}) measured across two years.

Statistical analysis. The regression model for p explanatory variables and n observations is given as follows:

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i + e_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

$$Y = \mu 1_n + X\beta + e. \tag{2}$$

The above equations are equivalent to each other. Here $Y = [y_1, y_2, \ldots, y_n]^T$ is the vector of dependent variables, 1_n is a column vector of n ones, μ is intercept (β_0), $X = [x_{ij}]$ is an $n \times p$ design matrix of explanatory variables, $\beta = [\beta_1, \ldots, \beta_p]$ is the vector of regression coefficients and $e = [e_1, \ldots, e_2]$ is the vector of the residual with NIID$(0, I\sigma_e^2)$. OLS estimates can be written as an optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2 . \tag{3}$$

Adaptive Lasso regression estimates the regression coefficients by the following equation in the Lagrangian form [8, 9]:

$$\hat{\beta}(alasso) = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|, \quad j = 1, 2, \ldots, p, \tag{4}$$

where $\left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2$ is the loss function, $\lambda \geq 0$ is a shrinkage parameter which regulates strength of penalty and is important for success of Adaptive LASSO, $\hat{\omega}_j$ states a data-defined weight, $\hat{\omega}_j = 1/\left|\hat{\beta}_J^{\text{ini}}\right|^{\gamma}$, γ is a positive constant and $\hat{\beta}_j^{\text{ini}}$ is initial consistent estimator of β obtained from ordinary least square [9] and $\sum_{j=1}^{p} |\beta_j| = \|\beta\|_1$ is called the l_1-norm penalty on β and $\hat{\beta}(\lambda_n)$ shows parameter estimation

$$\hat{\beta}(\lambda_n) = \arg\min \left\{ \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \sum_{j=1}^{p} \lambda_n \hat{\omega}_{jn} |\beta_j| \right\}. \tag{5}$$

If $\lambda_n \to \infty$ and $\lambda_n/\sqrt{n} \to 0$, the adaptive Lasso has oracle property [8, 9].

Model selection. The adjusted coefficient of determination ($\overline{R}^2$) and Akaike Information Criterion (AIC) were used as cohesion criteria to compare Stepwise and Adaptive Lasso

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{6}$$

where R^2 is the determination coefficient, n is the sample size and p is the total number of explanatory variables in the mode not including the constant. AIC [11] is as follows

$$\text{AIC} = -2ll + 2p, \tag{7}$$

where ll is the log-likelihood and p is the parameter number.

The statistical analyses were performed by using `GLMSELECT` procedure in SAS program [12]. *AIC* is used in choose sub-option and *none* is used in stop for the model selection.

3. Results and discussion. OLS results in MLR analysis are shown in Table 1. The table contains regression coefficients, its standard errors, t value, probability of the estimated variables in predicting barley grain yield and Variance Inflation Factor (VIF) values for explaining multicollinearity problem. T-test applying to regression coefficient revealed that spike number per m^2 and grain weight per spike had significantly contributed towards grain yield. However, five variables named days to heading, spike length, grain number per spike, thousand grain weight and harvest index did not affect the grain yields. All grain yield components in the current study explained 88% of total variation. The dependency assumption in independent variables was investigated by performing MLR analysis. VIF values of all independent variables were found smaller than 10. The results prove that there are no dependencies between independent variables, so there is no multicollinearity problem in the current data [13].

Stepwise and Adaptive Lasso methods were used to estimate and select variables among all grain yield components and the results are presented in Table 2.

T a b l e 1

OLS results in MLR analysis and VIF values

Variables	β's	SE	t	Prob $>\lvert t\rvert$	VIF
Days to heading (X_1)	0.98517	0.54687	1.80	0.0748	1.48
Spike number per m^2 (X_2)	0.49907	0.02345	21.28	< 0.0001	1.45
Spike length (X_3)	3.18719	4.07989	0.78	0.4366	2.57
Grain number per spike (X_4)	-0.68084	1.56932	-0.43	0.6654	4.43
Grain weight per spike (X_5)	141.27694	30.92938	4.57	< 0.0001	5.04
Thousand grain weight (X_6)	0.76617	0.87355	0.88	0.3826	2.58
Harvest index (X_7)	-0.03957	0.48484	-0.08	0.77835	1.29
R^2	0.8811				
$\overline{R}^2$	0.8725				

β – coefficient of regression, SE – Standard Error, R^2 – coefficient of determination, $\overline{R}^2$ – adjusted coefficient of determination

T a b l e 2

Results of Stepwise and Adaptive Lasso methods used for selection and estimation of variables in predicting barley grain yield and model selection criteria

Stepwise				
Variables	β's	Std. β's	Partial R^2	Model R^2
Spike number per m^2 (X_2)	0.500102	0.902246	0.719	0.719
Grain weight per spike (X_5)	155.735895	0.397881	0.154	0.873
Days to heading (X_1)	0.805702	0.063084	0.002	0.875
Number of variables	3			
$\overline{R}^2$	0.88			
AIC	694.97			
Adaptive Lasso				
Spike number per m^2 (X_2)	0.493874	0.891009	0.695	0.695
Grain weight per spike (X_5)	153.691031	0.392656	0.178	0.873
Days to heading (X_1)	0.527991	0.041340	0.002	0.875
Number of variables	3			
$\overline{R}^2$	0.88			
AIC	695.93			

β – coefficient of regression, Std. β – standardized coefficient of regression, $\overline{R}^2$ – adjusted coefficient of determination, AIC – Akaike Information Criterion

The table contains the coefficients of regression, standardized coefficients of regression, partial and cumulative R^2 and AIC values for the methods used in the current study. The variable selection was made using AIC because the statistical analyses were performed by using AIC in choose sub-option. Graphical representation of variable selection according to AIC was demonstrated in Fig. 1 for Adaptive Lasso.

In Table 2, stepwise regression result shows that three variables named spike number per m^2, grain weight per spike and days to heading have significant effect in grain yield prediction. 71.9% of the total variation was explained by spike number per m^2 and, respectively, 15.4% by grain weight per spike and 0.2% by days to heading. So, 87.5% of total variation in barley seed has been disclosed by all variables in regression models. According to Adaptive Lasso results given in Table 2 spike number per m^2, grain weight per spike and days to heading were found to be significant. The amount of explained in the total variation by the variables has been 69.5%, 17.7% and 0.2%, respectively. When performing Adaptive Lasso regression, 87.5% of total variation in barley grain yield could be attributed to three variables like in stepwise regression.

Information criteria for Stepwise and Adaptive Lasso are presented in Table 2. These techniques produced similar results for $\overline{R}^2$ and AIC value (88% and 695,

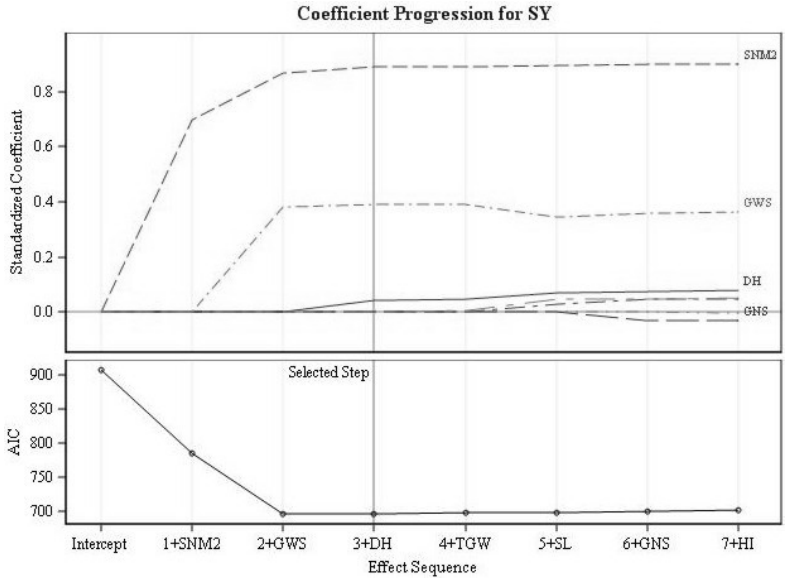

Fig. 1. Standardized coefficient and AIC for estimation and variable selection. Legend: SNM2 – spike number per m^2, GWS – grain weight per spike, DH – days to heading, TGW – thousand grain weight, SL – spike length, GNS – grain number per spike, HI – harvest index

respectively). The same independent variables were selected for both methods. The independent variables were grain weight per spike, spike number per m^2 and days to heading, respectively, the amount of effect in the model.

It was reported that the number of seeds per spike, thousand grain weight and the number of spikes per m^2 were the most important yield components on barley grain yield in the application of stepwise regression analysis [14]. In our study, number of seeds per spike and thousand grain weight were not important even though number of spikes per m^2 was an important predictor. It was found in previous studies that spike length, number of seeds per spike, seeds weight per spike and thousand grain weight were positively correlated with grain yield and these yield components were demonstrated as appropriate selection criteria in breeding programmes [15,16]. In this study correlation between grain weight per spike and grain yield was not analysed but it can be seen in Table 3 that grain weight per spike effected positively the grain yield. The duration of grain filling in late-flowering varieties is shortened in the generative period, especially due to the decrease in precipitation after heading and the increase in temperatures in the arid and semiarid climatic zones in our country [17]. The selection of early-flowering varieties over arid conditions would lead to an increase in the grain filling times and an increase in dry matter accumulation in the grain. For this reason, the character of earlier flowering in cool seasons cereals was identified as an important selection criterion for drought tolerance [18]. It has also been reported that grain

Table 3

The results of OLS, Stepwise and Adaptive Lasso in MLR analysis

Variables	OLS	Stepwise	Adaptive Lasso
Days to heading (X_1)	0.98517	0.805702	0.527991
Spike number per m^2 (X_2)	0.49907	0.500102	0.493874
Spike length (X_3)	3.18719		
Grain number per spike (X_4)	−0.68084		
Grain weight per spike (X_5)	141.27694	155.735895	153.691031
Thousand grain weight (X_6)	0.76617		
Harvest index (X_7)	0.98517		
Number of variables	7	3	3
R^2	0.8811	0.8786	0.8782
$\overline{R}^2$	0.8725	0.8749	0.8745

weight in arid conditions is largely dependent on the environmental conditions after flowering stage [19]. The results of our study also revealed that days to heading is an important predictor for grain yield. This finding was consistent with the literature [17–19].

OLS, stepwise and Adaptive Lasso results are presented in Table 3. OLS had seven independent variables with $\overline{R}^2 = 0.8725$. Stepwise and Adaptive Lasso regressions which performed the variable selection had the same three independent variables with $\overline{R}^2 = 0.8749$ and 0.8745, respectively.

The variable selection is important for better model interpretation and high prediction accuracy [6]. For this purpose the stepwise method has been frequently used by different researchers [4, 5, 20]. The results obtained in this study show that OLS estimates did not have any problem about prediction accuracy and interoperation because of provided independence assumption. Lasso technique for variable selection in agricultural research was introduced and concluded that Lasso gives better results for mungbean [7]. In the current study we applied adaptive Lasso technique and found out that the results of Stepwise and Adaptive Lasso were the same because independent variables had dependencies assumption.

Recently, erratic and insufficient rainfall due to climatic changes may cause drought stress during grain filling periods in barley production areas of our country. It is necessary to establish new approaches against the drought and especially to take the yield components which are effective over grain yield into barley breeding programmes. As a result of the research, it was determined that the number of spikes per square meter and the grain weight per spikes were the most important criteria to be considered in arid areas where large amounts of barley are grown.

4. Conclusions. The stepwise and Adaptive Lasso analysis used in the current study showed that days to heading, spike number per m^2 and grain weight

per spike are the most important yield components in two-rowed barley under rainfed conditions in Van. We introduced the Adaptive Lasso method for variable selection in agricultural research, as an alternative to Stepwise. In this study we revealed that Stepwise and Adaptive Lasso gave the same results because the assumption of independence was provided.

REFERENCES

[1] Türkiye İstatistik Kurumu (TÜİK) (2016) `http//tuik.gov.tr` (Erişim Tarihi: 26.08.2016).

[2] Karasu A., M. Öz (2010) A study of coefficient analysis and association between agronomical characters in dry bean (*Phaseolus vulgaris* L.), Bulgarian J. Agricultural Sci., **16**, 203–211.

[3] Mohammad T., A. Muhammad, S. Fazale, I. K. Muhammad, J. K. Abdul (2008) Identification of traits in bread wheat genotypes (*Triticum aestivum* L.) contributing to grain yield through correlation and path coefficient analysis, Pak. J. Bot., **40**, 2393–2402.

[4] El-Mohsen A. A. A., M. A. A. El-Shafi (2014) Regression and path analysis in Egyptian bread wheat, J. Agri-Food and Applied Sc., **2**(5), 139–148.

[5] Nasri R., A. Kashani, F. Paknejad, S. Vazan, M. Barary (2014) Correlation, path analysis and stepwise regression in yield and yield component in wheat (*Triticum aestivum* L.) under the temperate climate of Ilam province, Iran, Indian J. Fund. Appl. Life Sci., **4**(4), 188–198.

[6] Tibshirani R. (1996) Regression shrinkage and selection via the Lasso, J. Royal Stat. Soc. Series B, **58**, 267—288.

[7] Amin M., W. Xiaoguang, L. Song, H. Ullah, M. Y. Ashraf (2014) Penalized selection of variable contributing to enhanced seed yield in mungbean (*Vigna radiata* L.), Pakistan J. Agricult. Sci., **51**(2), 328–391.

[8] Fan J., R. Li (2001) Variable selection via nonconcave penalized likelihood and its oracle properties, J. Am. Stat. Assoc., **96**, 1348–1360.

[9] Zou H. (2006) The adaptive lasso and its oracle properties, J. Am. Stat. Assoc., **101**, 1418–1429.

[10] Kacar B. (1995) Bitki ve Toprağın Kimyasal Analizleri. III. Toprak Analizleri, No 3 (Chemical Analyses of Plant and Soil. II. Soil Analyses), Ankara Universitesi Ziraat Fakültesi Egitim Arastırma ve GelistirmeVakfı Yayınları Ankara-Türkiye.

[11] Akaike H. (1974) A new look at the statistical model identification, IEEE Trans. on Automatic Control, **19**(6), 716–724.

[12] SAS: SAS/STAT (2014) Cary, NC, USA, SAS Institute Incorporation.

[13] Burger M., J. Repiský (2012) Problems of linear least square regression, In: Proc. ARSA (Advanced Research in Scientific Areas), 257–262.

[14] Ataei M. (2006) Path analysis of barley (*Hordeum vulgare* L.) yield, Tarım Bilimleri Dergisi, Ankara Üniv. Ziraat Fak. Dergisi, **12**, 227–232.

[15] Shendy M. Z. (2015) Gene action and path coefficient studies for yield and yield components of some barley crosses, Egypt J. Plant Breed, **19**(4), 1155–1166.

[16] Kundalia S. K., P. P. Singh, S. Solomon (2006) Interrelationship of yield and associated characters in hull-less barley (*Hordeum vulgare* L.), New Botanist Int. J. Plant Sci. Res., **33**, 1–4.
[17] Genç İ., A. C. Ülger, T. Yağbasanlar, Y. Kirtok, M. Topal (1988) A comparative study on the yield and yield parameters of triticale, wheat and barley under the conditions of Çukurova, J. Agric. Fac. Ç. Ü., **3**(2), 1–14.
[18] Rana V. K., S. C. Sharma (1987) Correlation among some morpho-physiological characters associated with drought tolerance in wheat, Crop Improv., **24**(2), 194–199.
[19] Öztürk A., Ş. Akten (1999) Some morphophysiological characters and grain yield effect in winter wheat, Turkish J. Agriculture and Forestry, **23**, 409–422.
[20] Siahbidi M. M. P., A. P. Aboughadareh, G. R, Tahmasebi, M. Teymoori, M. Jasemi (2013) Evaluation of genetic diversity and interrelationships of agro-morphological characters in durum wheat (*Triticum durum* Desf.) lines using multivariate analysis, Int. J. Agriculture, **3**(1), 184.

Department of Animal Science,
Biometry and Genetic Unit
Faculty of Agriculture
Van Yüzüncü Yıl University
Van, Turkey
e-mail: sgakkol@yyu.edu.tr

*Department of Field Crops
Faculty of Agriculture
Van Yüzüncü Yıl University
Van, Turkey

**Department of Field Crops
Faculty of Agriculture
Ahi Evran University
Kırşehir, Turkey