Pak. J. Agri. Sci., Vol. 51(2), 373-381; 2014
ISSN (Print) 0552-9034, ISSN (Online) 2076-0906
http://www.pakjas.com.pk

PENALIZED SELECTION OF VARIABLE CONTRIBUTING TO ENHANCED SEED YIELD IN MUNGBEAN (*Vigna radiata* L.)

Muhammad Amin[1,*], Wang Xiaoguang[1], Lixin Song[1], Hidayat Ullah[2] and M. Yasin Ashraf[3,*]

[1]School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning 116023, China; [2]Department of Plant Breeding and Genetics, The University of Swabi, Khyber Pakhtunkhwa, Pakistan; [3]Crop Stress Management Group, Nuclear Institute for Agriculture and Biology, Pakistan.
[*]Corresponding author's e.mail: aminkaniu@gmail.com; niabmyashraf@gmail.com

Penalized regression methods for simultaneous variable selection and coefficient estimation have received a great deal of attention in recent years. Especially those based on the least absolute shrinkage and selection operator (LASSO), that involves penalizing the absolute size of the regression coefficients. The ordinary least square and LASSO methods were used for selection of most significant traits contributing towards seed yield in mungbean plants with 18 morphological and yield associated traits and to develop the prediction model . Bayesian information criterion was applied to choose minimum tuning parameter. Results indicated that dry weight biomass and harvest index were highly significant characters towards seed yield while days to maturity, days to flowering, number of nodes per plant, pods per plant and degree of indetermination had a significant affect on response variable. Based on the results, it was rational to conclude that high yield of mungbean crop could be obtained by selecting the breading materials with these important characters on seed yield.

Keywords: Seed yield, LASSO, least square, variable selection, mungbean

INTRODUCTION

Mungbean (*Vigna radiata* (L.) Wilczek) is an important pulse crop of Asia which is widely grown in India, Bangladesh, Sri Lanka, Thailand and Pakistan. In Pakistan, it is grown as a supplemental and cash crop on 245.9 thousand hectares with a total production of 177.7 thousand tons and an average yield of 636 kg ha^{-1}. Maximum average yield of 663 kg ha^{-1} from an area of 9.5 thousand hectare was obtained in the Khyber Pakhtunkhwa Province of Pakistan due to suitability and proper adaptation of mungbean to the agro-climatic conditions of the area (Anonymous, 2009). To improve the genetic architecture of mungbean plant, several efforts have been made. Resultantly, some improved mungbean cultivars have been developed with desirable yield related traits like total dry matter and harvest index. These parameters were given maximum importance in improving per unit area seed yield in mungbean as these were found to be positively associated with grain yield per plant (Sadiq *et al.*, 2000).

Various applied statistical techniques like correlation, path coefficient analysis, multivariate analysis are used for selection of most important traits towards yield for breeding programs (Massart *et al.*, 1997; Mohammad *et al.*, 2005; Mohammad *et al.*, 2008; Tejbir *et al.*, 2009; Hussain *et al.*, 2012). Most of the crop improvement programs are to realize a marked improvement in crop yield. Since yield is a complex character which is controlled by association of various traits therefore, information on association of yield attributes and their direct and indirect effects on seed yield are of great importance. In view of this, correlation and path coefficient analysis are important statistical tools to evaluate breeding programs for high yield, as well as to examine direct and indirect contribution of the yield variables (Mohamed, 1999).

The path analysis has the advantage to partition the correlation coefficient into two components, one component measures the direct effect while the second component is the indirect effect of a predictor variable on the response variable (Dewey and Lu, 1959). This technique has been used in agriculture by plant breeders to assist in identifying traits that are useful as selection criteria to improve crop yield (Milligan *et al.*, 1990; Surek and Baser, 2003; Ilahi *et al.*, 2009). There are two essential goals in statistical theory; discovery of relevant and most important predictive variables having high prediction accuracy. Variable selection is fundamental to statistical modeling, it can significantly increase the performance of the fitted model and is an important area in linear regression analysis. A perfect variable selection can lead to better risk assessment and model interpretation.

Numerous studies has been reported about variable selection. Identification of true significant estimates can enhance the prediction accuracy of the fitted model. Penalized likelihood framework to approach the problem of variable selection was proposed by Fan and Li (2001). In practice, a large number of predictors are usually included at the initial stage of modeling to get the possible modeling biases. But, to enhance predictability and to select significant variables, statisticians and researchers usually prefer to use stepwise deletion and subset selection methods. Six statistical methods including stepwise deletion method were permored for evaluation of signifcant variables on different wheat genotypes as reported by Pirdashti *et al.* (2012). Although they are practically useful, these selection procedures ignore stochastic errors inherited in the stages of variable selections. Hence, their theoretical properties are somewhat hard to understand. Furthermore, the best subset variable selection suffers from several drawbacks, the most severe of which is its lack of stability as analyzed by Breiman (1996a). Let us consider the matrix form of linear regression model

$$y = X\beta + \varepsilon \qquad (1)$$

where X $n \times p_n$ is design matrix representing data set of 18 predictors or independent variables, β is a $P_n \times 1$ vector of unknown coefficients, ε is a vector of identically independent distributed (i.i.d) random variables with means zero and finite variance σ^2 and y is value of response variable. Here it is assumed that the data are centered, therefore, intercept is not included in the regression model. The ordinary least square (OLS) method can be used to estimate the regression unknown coefficients. The OLS estimates are not preferred by the data analysts due to two main reasons. The first one related with prediction accuracy, these estimates have often low bias but large variance. The prediction accuracy sometimes can be improved by shrinking or setting some regression coefficients to zero. In this way one can sacrifice a little bias to reduce the variance of the predicted values and may improve the overall prediction accuracy. The second one is related to interpretation. When model has a large number of predictors, researchers often would like to determine a smaller subset that describes the strongest effects. The two standard techniques subset selection and ride regression were used to improve the OLS estimates, but both have drawbacks. Although subset selection method provides interpretable models but can be extremely variable due to its discreteness, regressors are either retained or dropped from the model (Breiman, 1996b; Fan and Li, 2001). Therefore small changes in the data can result in very different models being selected and this can reduce its prediction accuracy. In contrast ridge regression is a continuous process that shrink coefficients and hence more stable. But it can not set any coefficients to exactly zero. So, it does not give an easy interpretable model. These selection procedures ignore the stochastic errors or uncertainty during the stage of variable selection (Fan and Li, 2001; Shen and Ye, 2002). A new technique called least absolute shrinkage and selection operator (LASSO) was proposed by (Tibshirani, 1996). According to this some coefficients shrink and others are set to zero. It contains the good features of both subset and ride regression. The LASSO estimates can be defined as

$$\hat{\beta}(LASSO) = \arg\min_{\beta} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \beta_j \qquad (2)$$

Where λ is a nonnegative regularization parameter, the second term in the above equation is called "l_1 Penalty". The current study is carried out to determine the most effective yield components by LASSO method. This work is expected to help the crop breeders to find out the most important traits towards seed yield.

MATERIALS AND METHODS

The studies were conducted with thirty mungbean genotypes and three check cultivars in randomized complete block design (RCBD) with three replications. The seeds for this study were obtained from Nuclear Institute for Food and Agriculture, Peshawar, Pakistan. The experiments were carried out during two years (2007-2008). Plot size for a mungbean genotype in each replication was 3.6 m^2. Eight morpho-physiological traits and eleven yield associated traits were studied like X_1 = days to flowering (DF), X_2 = plant height (PH), X_3 = days to maturity (DM), X_4 = number of nodes plant^{-1} (NPP), X_5 = leaf area (LA), X_6 = pods plant^{-1} (PPP), X_7 = seeds pod^{-1} (SPP), X_8 = fresh weight of biomass (FW), X_9 =1000-seeds weight (TSW), X_{10} = harvest index (HI), X_{11} = dry weight of biomass (DW), X_{12} = degree of indetermination (DIT), X_{13} = number of leaves plant^{-1} (LPP), X_{14} = petiole length (PTL), X_{15} = peduncle length (PDL), X_{16} = clusters plant^{-1} (CPP), X_{17} = pods cluster^{-1} (PPC), X_{18} = pod length (PL), and Y = seed yield (SY). Proper management practices were adopted throughout the growing seasons to ensure good crop growth.

Residual analysis is an essential tool for checking whether the statistical models fit to meet the data underlying model assumptions, like normality of residuals, linearity of the regression model, homoscedasticity, autocorrelation and multicollinarity. It is often obsevred that some observations do not seem to fit in the overall pattern of the data. The leverage value and mahalanobis distance are good devices to asses the multivariate outliers with respect to other x's values. The leverage value of the ith observation is defined as:

$$h_{ii} = X_i'(X'X)^{-1}X_i \qquad (3)$$

where h_{ii} is the diagonal entry of the hat matrix H, which provides a measure of distance of the ith case from the centroid of the x observations. In general, If $h_{ii} > 2\overline{h}$ then h_{ii} observation is considered as outlying observation. Mahalanobis distance is defined as:

$$M.D = \sqrt{(X' - \bar{X})\Sigma^{-1}(X' - \bar{X})'} \qquad (4)$$

where $\sum^{-1}$ is inverse covariance matrix of independent variables. The outlying or extreme observations with respect to y values are detected by the studentized deleted residuals r_i^*, this method is preferred over the ordinary deleted residuals. The ith studentized deleted residual is defined as:

$$r_i^* = d_i / S(d_i) \qquad (5)$$

It is not enough to asses whether the observation is or not an outlier. The next step is to ensure whether or not these are influential observations. The methods like Cook's distance (D$_i$) and DFFITS$_i$ which is an abbreviation for "difference in fits" are used for the identification of such observations. It is obtained as:

$$D_i = (\hat{Y} - \hat{Y}_{(i)}) \big/ p \times MSE \qquad (6)$$

If $D_i > F_{(\alpha,P,n-p)}$ then relative observation is defined as influential. The values of DFFITS$_i$ is defined as:

$$DFFITS_i = (\hat{Y} - \hat{Y}_{i(i)}) \big/ \sqrt{MSE_{(i)} \times h_{ii}} \qquad (7)$$

As a rule of thumb the observation is considered influential if the $|DFFITS_i| > 1$ but this criteria is applicable only for small to medium data sets and for large data sets when $|DFFITS_i| > 2 \times \sqrt{p/n}$ then the observation will be considered as influential. Shapiro-Wilk Test is used to test the normality. The test statistics is defined as:

$$W = \left[\left[\sum \alpha_{n-i+1}(x_{(n-i+1)} - x_{(i)})\right] \big/ SD\sqrt{n-1}\right]^2 \qquad (8)$$

where n = total number of observations, SD = standard deviation, $x_{(i)}$ = ordered sample from smallest to largest, $x_{(n-i+1)}$ = ordered sample from largest to smallest and $\alpha_{(n-i+1)}$= coefficient for observed n.

A matrix of simple pearson's correlation coefficients was computed between seed yield and its associated traits of mungbean as proposed by Steel and Torrie (1987). Dewey and Lu (1959) proposed path coefficient anaylsis technique to compute direct and inderect effects of component traits on seed yield. This is the extension of the regression model which partion the simple correlation coefficients into direct and inderect effects.

Stepwise regression methods are traditional approaches which consider one covariate on each step. The resulting model selected by cross-validation or generalized cross-validation may have lower prediction error to the future observation. Information theoretic approaches such as Akaike Information Criterion (AIC) by (Akaike, 1974) and Bayesian Information Criterion (BIC) by Schwarz and Gideon (1978) can select the best model from all the candidate models. The relative new methods based on penalized likelihood, such as the LASSO (Tibshirani, 1996), mainly consider the computational efficiency and stability. To measure the estimation accuracy, we follow Tibshirani (1997) and summarize the standard errors for the nonzero coefficients. The covariance matrix of the estimates can be written as:

$$(X^T X + \lambda W^-)^{-1} X^T X (X^T X \lambda W^-)^{-1} \hat{\sigma}^2 \qquad (9)$$

where $(.)^-$ means the generalized inverse of a matrix, λ is the tuning parameter, W is a diagonal matrix with diagonal elements $|\hat{\beta}|$ and $\hat{\sigma}^2$ is the estimation of the error variance. The optimal tuning $(X^T X + \lambda W^-)^{-1} X^T X$ parameter λ was estimated by BIC method:

$$BIC(\lambda) = \log(SSE) + d * \log(n)/n \qquad (10)$$

where SSE is the sum of the residuals and d is the number of nonzero parameters.

RESULTS AND DISCUSSIONS

***Residual analysis*:** The values of h_{ii} of sixty observations were calculated using equation (3) and the mean of these values was computed to be 0.30. The value of $2\overline{h} = 0.6$ was also calculated to point out high leverage observations from the given observations. No value was found to be greater than the mean value so it can be inferred that there was no such outlying observation in data set. These results have been verified by the mahalanobis distance method defined in equation (4) as well. To set out the criteria about the outlying observations as $M.D > \chi^2_{(\alpha,p-1)}$, the value of $\chi^2_{tabulated} = 28.27$ was noted from its table using p = 18, where p is the number of fitted parameters including intercept and with $\alpha = 0.05$ as value of level of significance. Here again, no value greater than specified value of $\chi^2_{tabulated}$ was observed, therefore, no observation can be considered as outlier with respect to x's values. To access the outlying observations with respect to y values, studentized deleted residuals r_i^* were computed by equation (5). The detection criteria for outlying observation based on $r_i^* > t_{(1-\alpha,n-p-1)}$, the t$_{tabulated}$=2.01 was taken from t-sistribution table with $\alpha = 0.05$ and 41 degree of freedom. It was observed from the results that five observations were outliers. The observations numbered 2, 33, 49, 51 and 60 were outlying with respect to response variable. The cook's distance, D$_i$, values were calculated as formula defined in equation (6). The F = 1.80 value was taken from $F-distribution$ table. The values of DFFITS$_i$ were computed by equation (7). By applying the rule of thumb no observation was recorded having value exceeding 1. The value of $\sqrt[2]{p/n} = 1.25$ was calculated and the same results were observed. No observation was observed to be greater than the specified value, so it was concluded that no influential observation was present in our data set. Same results were obtained when the other test DFFITS$_i$ was used. It was clear from both tests that none of the five outlying observations are influential, so one may decide that their influence is not so strong to call for any remedial measures. Similar tests of residual analysis were applied on wheat data as reported by Ammara and Aslam (2010). The graphical representation of leverage values, mahalanobis distance values, studentized deleted residual values, cook's distance values and DFFITS values are shown in Figure 1. The outlying observations with respect to y values were observed outside the limits. Shapiro-Wilk test mentioned in equation (8) was used to test the normality of response variable y. The test statistics was calculated as w = 0.98 with P = 0.652. The test passed the assumption of normality that response variable follows normal distribution. This formula was published by Samuel Shapiro and Wilk (1965).

Simple correlation analysis**:** To get an idea about overall pattern of the data set used in current study, descriptive statistics (minimum and maximum values, mean and standard deviation) of all the estimated variables was calculated and is presented in the Table 1. The minimum and maximum seed yield over two years were recorded as 1.0

Table 1. Basic statistics (minimum and maximum values, mean and standard deviation) for estimated variables of mungbean genotypes

Variables	Minimum	Maximum	Mean	Std. Deviation
Days to flowering (X_1)	38	56	47.2	5.1
Plant height (X_2)	38	88	55.5	10.3
Days to maturity (X_3)	77	97	86.5	4.8
Number of nodes plant-1 (X_4)	8	17	10.5	1.4
Leaf area (X_5)	117	271	181.8	31.8
Pods plant^{-1} (X_6)	8	29	15.7	4.6
Seeds pod^{-1} (X_7)	8	12	10.2	1.1
Fresh weight of biomass (X_8)	21	42	29.9	4.9
1000-seed weight (X_9)	45	75	58.7	6.1
Harvest index (X_{10})	22	37	27.6	3.5
Dry weight of biomass (X_{11})	5	10	7.7	1.1
Degree of indetermination (X_{12})	12	71	43.0	12.0
Number of leaves plant-1 (X_{13})	7	14	8.1	1.1
Petiole length (X_{14})	9	21	13.8	2.1
Peduncle length (X_{15})	6	14	9.2	1.5
Clusters plant^{-1} (X_{16})	4	10	6.1	1.2
Pods cluster^{-1} (X_{17})	2	5	3.5	0.5
Pod length (X_{18})	7	10	9.1	0.6
Seed yield (Y)	1.0	3.0	2.1	0.4

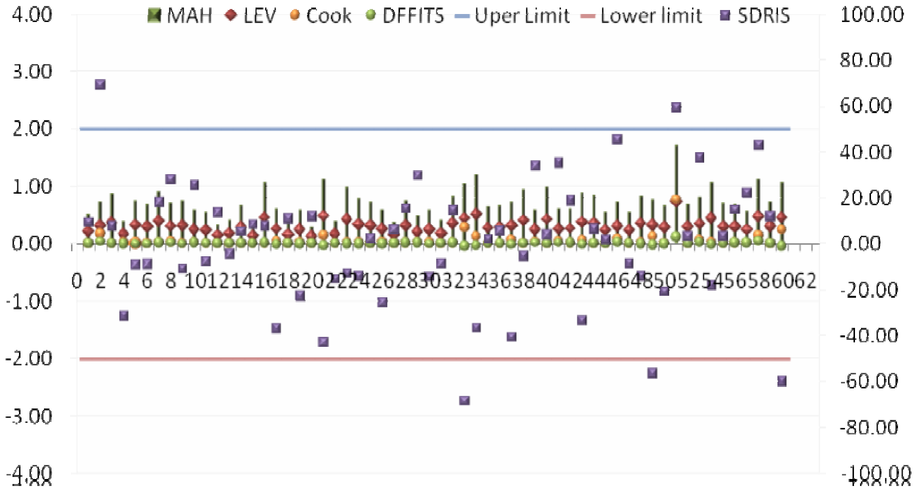

Figure 1. The plotted diagramme mahalanobis distance, studentized deleted residual, leverage values, DFFITS$_i$ are ploted, observations outside uper and lower limits are outlying observations

ton per ha and 3.0 tons per ha respectively with 0.4 value of standard deviation, though the yield was comparatively highier as compared to the national average yield but as stated eralier mungbean was not grown as a major crop in Pakistan. Moreover, the crop was not receiving good practices, fertilization and was grown on marginal or fallow land. The Pearson's correlation matrix for the estimated significant traits of mungbean are given in Table 2. The correlation coefficient *r* value requires both a magnitude and a direction of either positive or negative. It may take on a range of values $-1 \le r \le +1$. Results revealed that characters in the current study like days to flowering (0.29), pods per plant (0.44), seeds per pod (0.50), fresh weight of biomass (0.47), 1000-seeds weight (0.41), harvest index (0.63), dry weight of biomass (0.78), clusters per plant (0.37) and pod length (0.27), showed significantly positive correlation with seed yield of mungbean. Some positive and negative but weak correlations of different characters towards seed yield were observed which were not statistically significant because these traits have a little contribution in the selected germplasm. Seed yield was found to be strongly correlated with dry weight of biomass. Seed yield has positive and significant correlation with pods per plant and harvest index similar findings were reported by Mallikarjuna *et al.* (2006). A significantly positive association was recorded with number of pods per plant, number of seeds per pod, 100-seeds weight and harvest index towards seed yield as reported by Tijbir et al. (2009) contrary to negative and non significant correlation observed with seed yield. The study carried out by Mondal et al. (2011) points out that seed yield shows strong positive and significant correlation with total dry mass. In the currenct study we found that pod length and 1000-seeds weight have positive association but 1000-seeds weight was negatively correlated with seeds per pod as reported by Rohman et al. (2003) as well. Harvest index was found to be negatively correlated with plant height and days to maturity however correlation between harvest index and biomass yield was not significant, similar findings have been reported by Sharma and Smith (1986).

Path coefficient analysis: The correlation coefficients were partitioned into direct and indirect effects. Total, direct and indirect contribution of yield traits on seed yield of mungbean estimated through path coefficient analysis are shown in Figure 2. Direct (diagonal) and indirect (off-diagonal) effects among the yield traits are given in Table 3. The results show that dry weight of biomass and harvest index had high direct positive effects (0.79, 0.63) on seed yield. Positive direct effects of these traits on yield indicate their importance in determining these complex characters and therefore these factors should be kept in mind while practicing selection aimed at the improvement of seed yield. Similar results have been reported for maximum positive direct effect on yield by other workers (Tejbir et al., 2009). The dry weight of biomass and harvest index were shown to be strongly and significantly correlated with seed yield, and the direct effects of these two characters were high and positive. Munawar et al. (2001) have reported that dry biomass trait can be considered best yield component towards yield in mungbean, which supports our findings. This stresses that high yielding mungbean genotypes could be obtained by considering dry weight biomass and harvest index. The direct selection through mentioned traits can be effective. The highest positive indirect effects on seed yield of mungbean were recorded for fresh weight biomass (0.50) and seeds per pod (0.48) on seed yield of mungbean. Correlation coefficients between days to flowering, pods per plant, seeds per pod, fresh weight of biomass, 1000-seeds weight, clusters per plant and pod length were positive but their direct effects on seed yield showed negative or negligible impact. The results showed that seed yield was less associated with these traits. It can be concluded that direct selection through these characters would not be effective but these could be considered simultaneously as indirect causal factors. According to Sharma and Smith (1986) harvest index, the ratio of grain yield to total biomass yield, may be a useful selection trait for yield improvement, and this too supports our results.

Penalized regression analysis: The LASSO method defined in equation (2) was used to estimate and select regression

Table 2. Pearson's correlation coefficient (*r*) matrix for the estimated traits of mungbean genotypes

Traits	Days to flowering	Days to maturity	Nodes plant^{-1}	Pods plant^{-1}	Harvest index	Dry weight of biomass	Degree of indetermination
Days to maturity	0.39^{**}						
No. nodes plant^{-1}	0.02^{NS}	0.51^{**}					
Pods plant^{-1}	0.58^{**}	0.03^{NS}	-0.14^{NS}				
Harvest index	-0.01^{NS}	-0.23^{*}	-0.03^{NS}	0.18^{NS}			
Dry weight of biomass	0.37^{**}	0.04^{NS}	-0.28^{*}	0.41^{**}	0.01^{NS}		
Degree of indetermination	0.25^{*}	0.67^{**}	0.50^{**}	-0.21^{NS}	-0.22^{NS}	-0.13^{NS}	
Seed yield	0.29^{*}	-0.10^{NS}	-0.20^{NS}	0.44^{**}	0.63^{**}	0.78^{**}	-0.20^{NS}

*, ** = Significant at 5 and 1% probability level, respectively, NS=Non Significant

Table 3. Path coefficient analysis showing direct (diagonal) and indirect (off-diagonal) effect of the estimated yield components on seed yield of munbean

Var	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	Y
X_1	-0.06	0.00	0.02	0.00	0.00	0.04	0.00	-0.01	0.00	-0.01	0.29	0.01	0.00	0.00	0.00	-0.01	-0.01	0.00	0.26
X_2	-0.02	0.00	0.04	0.04	-0.01	0.00	0.00	0.00	0.00	-0.03	-0.08	0.04	-0.03	-0.01	0.00	0.00	0.00	0.00	-0.06
X_3	-0.02	0.00	0.05	0.03	-0.01	0.00	0.00	0.00	0.01	-0.14	0.00	0.03	-0.03	-0.01	0.00	0.00	0.00	0.00	-0.09
X_4	0.00	0.00	0.03	0.06	-0.01	-0.01	0.00	0.00	0.00	-0.02	-0.22	0.03	-0.04	-0.01	0.00	0.00	0.00	0.00	-0.19
X_5	-0.01	0.00	0.03	0.03	-0.03	0.00	0.00	0.00	0.00	-0.06	-0.01	0.03	-0.03	-0.01	0.00	0.00	0.00	0.00	-0.06
X_6	-0.03	0.00	0.00	-0.01	0.00	0.07	0.01	-0.01	0.00	0.11	0.32	-0.01	0.01	0.00	0.00	-0.02	0.00	0.00	0.44
X_7	-0.03	0.00	0.01	-0.01	0.00	0.04	0.01	-0.01	0.00	0.22	0.27	-0.01	0.01	0.00	0.00	-0.01	0.00	0.00	0.49
X_8	-0.02	0.00	0.01	-0.01	0.00	0.03	0.00	-0.02	0.00	-0.04	0.53	0.00	0.01	0.00	0.00	-0.01	0.00	0.00	0.48
X_9	0.01	0.00	-0.02	-0.01	0.00	0.00	0.00	0.00	-0.02	0.19	0.26	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.40
X_{10}	0.00	0.00	-0.01	0.00	0.00	0.01	0.00	0.00	-0.01	0.63	0.01	-0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.61
X_{11}	-0.02	0.00	0.00	-0.02	0.00	0.03	0.00	-0.01	-0.01	0.01	0.79	-0.01	0.01	0.00	0.00	-0.01	0.00	0.00	0.76
X_{12}	-0.01	0.00	0.03	0.03	-0.01	-0.01	0.00	0.00	0.00	-0.14	-0.10	0.05	-0.03	-0.01	0.00	0.01	0.00	0.00	-0.19
X_{13}	0.00	0.00	0.03	0.05	-0.01	-0.01	0.00	0.00	0.00	0.00	-0.21	0.03	-0.05	-0.01	0.00	0.00	0.00	0.00	-0.18
X_{14}	0.00	0.00	0.03	0.04	-0.02	-0.01	0.00	0.00	0.01	-0.10	-0.24	0.03	-0.03	-0.02	0.00	0.00	0.00	0.00	-0.31
X_{15}	0.00	0.00	0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00	-0.28	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	-0.28
X_{16}	-0.01	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.18	0.19	-0.01	0.00	0.00	0.00	-0.03	0.00	0.00	0.37
X_{17}	0.02	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.07	-0.07	0.00	0.00	0.00	0.00	0.00	0.02	0.00	-0.11
X_{18}	-0.01	0.00	0.00	0.01	-0.01	0.02	0.00	0.00	-0.01	0.28	-0.02	0.00	-0.01	0.00	0.00	-0.01	0.00	0.00	0.24

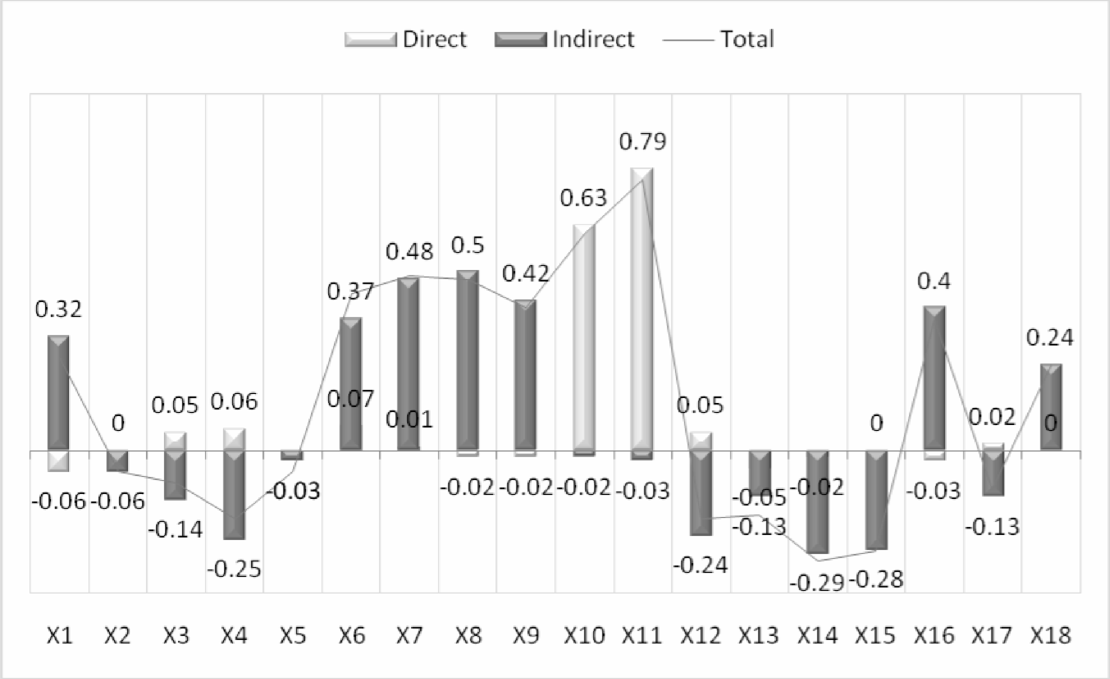

Figure 2. Direcct, indirect and total effect of estimated yield components on seed yield varaion of mungbean

coefficients among all predictors with different values of tuning parameter by using the R-language package (glmnet). The package was written by Friedman et al. (2008). A linear regression model with seed yield as response variable was fitted but first centering the predictors. The least square and LASSO estimates along their respective standard errors are given in Table 4. For every nonzero regression coefficient estimation, several values of tuning parameter were calculated. Initial or final value of λ for consideration of nonzero coefficient can be taken, and in this study we took initial value for this purpose. One nonzero estimate in the regression model was observed for the highest value of tuning parameter. The lowest value of tuning parameter was recorded when all the predictor variables are having nonzero coefficients. Dry weight of biomass was found to have high value of regression coefficient followed by harvest index.

Table 4. Results from ordinary leat squares and LASSO for seed yield of mungbean genotypes

Predictors	Least Square		LASSO		
	Coefficient	S. Error	Coefficient	S. Error	Decrease (%)
Days to flowering (X_1)	-0.0046*	0.0021	0	0.0017	18.42
Plant height (X_2)	-0.000039	0.0011	0	0.0009	17.92
Days to maturity (X_3)	-0.0044*	0.0024	0.0012	0.0020	18.15
Number of nodes plant-1 (X_4)	0.0180*	0.0087	0	0.0073	16.64
Leaf area (X_5)	-0.0033	0.0028	0	0.0002	91.73
Pods plant^{-1} (X_6)	0.0061*	0.0029	0.0007	0.0023	22.02
Seeds pod^{-1} (X_7)	0.0031	0.0088	0.0008	0.0050	43.38
Fresh weight of biomass (X_8)	-0.0015	0.0015	0	0.0013	14.21
1000-seeds weight (X_9)	-0.0016	0.0014	0	0.0011	18.21
Harvest index (X_{10})	0.0700**	0.0022	0.072	0.0018	18.06
Dry weight of biomass (X_{11})	0.2800**	0.0072	0.2709	0.0060	17.22
Degree of indetermination (X_{12})	0.0017*	0.0019	0	0.0008	57.81
Number of leaves plant-1 (X_{13})	-0.0170	0.0120	0	0.0101	15.77
Petiole length (X_{14})	-0.0031	0.0049	0	0.0041	16.61
Peduncle length (X_{15})	-0.0005	0.0051	0	0.0042	16.93
Clusters plant^{-1} (X_{16})	-0.0091	0.0086	0	0.0070	18.56
Pods cluster^{-1} (X_{17})	0.0140	0.0120	0	0.0103	13.98
Pod length (X_{18})	0.0012	0.0120	0	0.0094	21.79

*, ** = Significant at 5 and 1% probability level, respectively

The smallest value was observed regarding pods per plant and seeds per pod. Regression coefficients estimated through least square method showed that dry weight matter and harvest index were highly significant contributing traits towards seed yield. Days to flowering, days to maturity, number of nodes per plant, pods per plant and degree of indetermination were significant characters. The LASSO method gives nonzero coefficients to days to maturity, pods per plant, seeds per plant, harvest index and dry weight biomass, these are almost the same predictors selected throgh least square method. By noticing the standard errors for LASSO method that were estimated using defined criterion in equation (9), these standard errors were computed by fixing λ, by equation (10), at its optimal value 0.0078 for the original data set, a substantial percent decrease in standard errors was observed over least square method. It can be concluded that the results of LASSO are more reliable and give accurate prediction. Figure 3 shows the LASSO estimates as a function of the tuning parameter λ, it indicates that the absolute value of each coefficient tends to be 0 as the tuning parameter goes to infinity, the curves almost increase in a monotone fashion to their least square estimators. The scale of λ is changed to log λ to make picture more clear for users. The vertical broken line in Figure 3 shows the model for optimal value of tuning parameter, this line represents that five predictors are selected as important traits for improving the seed yield of mungbean. Tibshirani (1996) was concluded that for moderate size of predictors the LASSO give better results followed by ridge regression and then subset method, These findings supports our results.

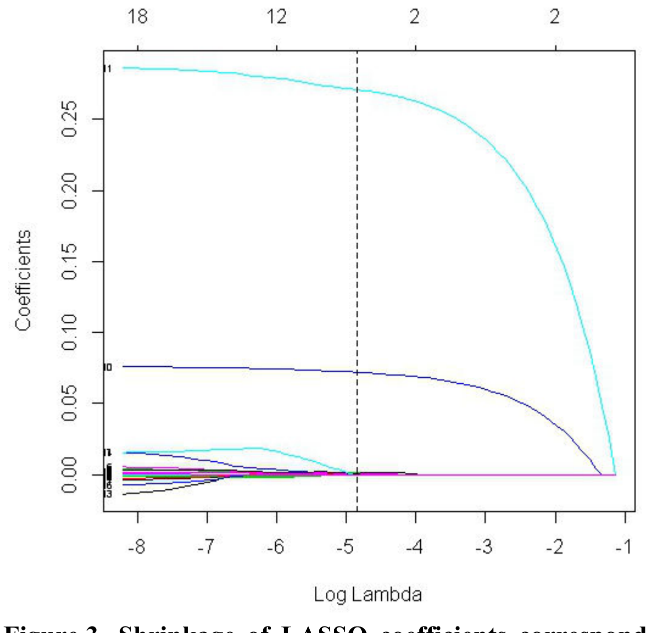

Figure 3. Shrinkage of LASSO coefficients corresponding to value of loglambda (logλ). Each curve represents estimated coefficient with its corresponding label. The dashed line shows the model for logλ=-4.85 selected by BIC

Conclusions: The path analysis and penalized regression analysis used in this study showed that dry weight of biomass and harvest index are the most important yield contributing components to be considered for selection of mungbean genotypes in later stages after gene fixation, as the additive effects cannot be eliminated. It is concluded that penalized regression techniques can give the best prediction results than ordinary least squares technique. Every statistical technique has its assumptions and constraints. The special care should be taken before using such techniques. In this study we introduced the LASSO method for variable selection in agricultural research. As this method defines the continuous shrinkage operation that can produce coefficients that are exactly zero. The beauty of this technique is that crop breeders can interpret their model very easily, this study will also be helpful to select plant traits that contribute more towards seed yield, and for the selection of best and stable model.

Acknowledgements: We are grateful to National Natural Science Foundation of China (Nos. 61175041 and 11371077) for their financial support. This work was also supported by the Fundamental Research Funds for the Central Universities in China (DUT12LK29). This research was sponsored by the Higher Education Commission (HEC), Islamabad Pakistan. Nuclear Institute for Food and Agriculture (NIFA), Peshawar Pakistan provided seeds of genotypes for this study.

REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control. 19:716-723.

Anonymous. 2009. Agriculture statistics of Pakistan. Ministry of Food, Agric. & Livestock, Econ Wing, Islamabad, Pakistan.

Breiman, L. 1996a. Heuristics of instability and stabilization in model selection. Ann. Stat. 24:2350-2383.

Breiman, L. 1996b. Better subset regression using the nonnegative garrotte. Technometrics 37:373-384.

Cheema, A.N. and M. Aslam. 2010. Regression Analysis of Wheat Model. Proceeding Two day International Conference on World Statistics Day “Statistics for Society”, held at Superior University Raiwind Road, Lahore Pakistan. pp.357-364.

Dewey, D.R. and K.H. Lu. 1959. A correlation and path co-efficient analysis of components of crested wheat grass and seed production. Agron J. 51:515-518.

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

Friedman, J., T. Hastie and R. Tibshirani. 2008. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33:1-22.

Hussain, M., S.K. Abdus, K. Ihsan and M. Muhammad. 2012. Correlation studies of some qualitative and quantitative traits with grain yield in spring wheat across two environments. Pak. J. Agri. Sci. 49:1-4.

Ilahi, F., H.N. T. Muhammad and S.A. Hafeez. 2009. Correlation and path coefficient analysis for achene yield and yield components in sunflower. Pak. J. Agri. Sci. 46:20-24.

Mallikarjuna, R.C., R.Y. Koteswara and R. Mohan. 2006. Genetic variablity and path analysis in mungbean. Legume Res. 29: 216-218.

Massart, D.L., B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi and J. Smeyers-Verbeke. 1997. Straight line regression and calibration. In Handbook of chemometrics and qualimetrics, Part A. Amsterdam, The Netherlands: Elsevier, pp.171-231.

Milligan, S.B., K.A. Gravois, K.P. Bischoff and F.A. Martin. 1990. Crop effects on genetic relationships among sugarcane traits. Crop Sci. 30:927-931.

Mohamed, N.A. 1999. Some statistical procedures for evaluation of the relative contribution for yield components in wheat. Zagazig J. Agric. Res. 26:281-290.

Mohammad, T., H. Sajjad, A. Muhammad, I.K. Muhammad and Z. Roshan. 2005. Path coefficient and correlation studies of yield and yield associated traits in candidate bread wheat (*Triticum Aestivum* L.) lines. Suran. J. Sci. Technol. 13:175-180.

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

Mondal, M.M.A., M.A. Hakim, A.S. Juraimi, M.A.K. Azad and M.R. Karim. 2011. Contribution of morpho-physiological attributes in determining the yield of mungbean. Afric. J. Biotechnol. 10:12897-12904.

Munawar, K., N. Khalid, N. Aminullah and S.B. Muhammad. 2001. Genetic variability and correlation studies in mungbean. J. Biol. Sci. 1:117-119.

Pirdashti, H., A. Ahmad, S. Fatemeh, H.J. Seyyed, S. Ataollah and A. Abolfazl. 2012. Evaluation of most effective variables based on statistically analysis on different wheat (*Triticum aestivum* L.) genotypes. Intl. J. Agric: Res & Rev. 2:381-388.

Rohman, M.M., A.S.M.I. Hussain, M.S. Arifin, Z. Akhter and M. Hasanuzzaman. 2003. Genetic variability, correlation and path analysis in mungbean. Asian J. Plant Sci. 2:1209-1211.

Sadiq, M.S., G. Sarwar and G. Abbas. 2000. Selection criteria for seed yield in mungbean (*Vigna radiata* L. Wilczek). J. Agric. Res. 38:7-12.

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461-464.

Sharma, R.C. and E.L. Smith. 1986. Selection for high and low harvest index in three winter wheat populations. Crop Sci. 26:1147-1150.

Shapiro, S.S. and M.B. Wilk. 1965. An analysis of variance test for normality (complete samples). Biometrika 52:591-611

Shen, X. and J. Ye. 2002. Adaptive model selection. J. Am. Stat. Assoc. 97:210-221.

Slavkovic, L., B. Skrbic, N. Miljevic and A. Onjia. 2004. Principal component analysis of trace elements in industrial soils. Environ. Chem. Lett. 2: 105-108.

Steel, R.G.D. and J.H. Torrie. 1987. Principles and Procedures of Statistics: A Biometrical Approach, 2nd Ed. McGraw Hill, USA, pp.272-277.

Surek, H. and N. Beser. 2003. Correlation and path coefficient analysis for some yield-related traits in rice (*Oryza Sativa* L.) under thrace conditions. Turkish J. Agric. 27:77-83.

Tejbir, S., S. Amitesh and A.A. Fayaz. 2009. Morpho-physiological traits as selection criteria for yield improvement in mungbean [*Vigna radiata* (L.) wilczek]. Legume Res. 32: 36-40.

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

Tibshirani, R.J. 1997. The LASSO method for variable selection in the cox model. Stat. Med. 16:385-395.