CME LIVE OPEN ACCESS

Journal of
Clinical and Medical Engineering

Research Article

Investigating the Prediction of Adaptable Behaviors of Wheat under Environmental Variables

Mehari Gebre Teklezgi*and Yared Tbebu Gebru

Lecturers at Department of Statistics, Adigrat University,

Corresponding Author: Mehari Gebre Teklezgi Lecturers at Department of Statistics, Adigrat University, And P.O.Box: 50, Adigrat, Tigray, Ethiopia.

Received: 2023 Sep 15 **Accepted:** 2023 Sep 21 **Published:** 2023 Oct 16

Abstract

Durum wheat is the 10th most essential crop in the world, which covers about 10% of the world›s wheat. The study is aimed to investigate the predictive value of various types of environmental variables on the future values of different traits durum Wheat. Ordinary multiple linear regression with stepwise variable selection method on the complete data set, and multiple linear regression models with predictors selected by penalized methods with mean square error cross-validation, were used. Findings showed that there are some predictors which affect positively and some others affect negatively for Plant Height and Grain Weight. Model with predictors selected by Elastic net method seem to have good prediction on the Plant Height for both OLS and WLS estimation methods, while the prediction from the lasso based model is not that much reasonable. In conclusion, inferences and predictions by the ordinary MLR models are not trusted due to the presence of multicollinearity, and violation of some model assumptions. However, predictions using the models with predictors selected by the shrinkage as well as WLS methods were better as the effects of the variability on these methods are minimal. Better predictions were found on the Plant Height and grain Weight.

Keywords: Statistics Education Research; Cross-Validation Mean Square Error; Multiple Linear Regression; Penalized Methods; Bias-Variance Trade-Off; Least Square Methods.

1. Introduction

Wheat is the most important grown crop, and regularly used for food for billions of people in the world [01-12]. Durum wheat is the 10th most essential crop in the world, which covers about 10% of the world's wheat, and is an economically important because of its unique rheological characteristics and the varieties of industrial end-products that can be derived from it, such as pasta and several types of flat breads [10]. The world's farming systems are facing mounting challenges that require our crop plants to yield significantly more, using less nutrients, land and water, under increasingly harsh and variable conditions. To meet this challenge, ongoing and efficient plant breeding, which is underpinned by access to and utilization of appropriate genetic variations for key plant traits will be required. Thus, increasingly breeders will be forced to seek the variation they require from genetic resource collections conserved in genebanks. Therefore, it is very important that natural diversity for traits related to drought adaptation and climate change in general should be recognized and kept in genebanks which ensures the long-term conservation of genetic resources to be readily available for use by breeders, researchers and other users. Genebanks are the most noticeable storehouses of plant genetic resources to look for important traits, providing the raw material for crop improvement, and is the most important preservation method for species producing orthodox seeds that withstand dehydration to low moisture contents and storage at very low temperatures [18]. As a result, they play a key role in contributing to the sustainable development of agriculture, helping to increase food production and thus to overcome hunger and poverty by maintain to high standards of survival and quality of the germplasm under their care. The preceding 25 years have seen notable growth in assembling and conserving these resources. However, many genebanks now facing major problems of size and organization [23, 24]. It is also well known that crop productivity is highly dependent on climatic changes and variability. In different studies about the adaptive traits, almost similar results were found. Eight field assessments were carried out in different temperature regimes in Spain, as stated by [4]. Grain Yield of durum wheat under Mediterranean environments is regularly limited by high temperature. It was also declared that different moisture regimes was mainly linked with differences in spikes per square meter and kernels per spike, these differences

may in turn contribute to significant Grain yield differences. Besides, studied Grain Weight of durum wheat with a two-way anova, and found that durum wheat exposed to high temperatures significantly decreased its Grain Weight [5]. A variance study for Grain Yield and yield components held by in Sardinia during the period between December and June in the years 1989 and 1990, and revealed that these characters were affected mostly by temperature and moisture [8]. Another study was carried out from 13 Mar, 2007 through 12 May, 2009 at the University of Arizona Maricopa Agricultural Center, Maricopa by and suggested that promising increases in overall temperature have a negative effect on spring durum wheat yield [17]. Moreover, a field study was carried out on the tolerance of durum wheat to high temperatures using analysis of variance at Elvas, Portuguese by and stated that Grain yield and individual grain weight were considerably affected by temperature increase [13]. A study by evaluated phonological traits of durum wheat such as Plant Height in highly different rainfall conditions in Mediterranean countries (Italy, Morocco, Spain, Syria, and Tunisia), and others [14, 15]. And was stated that all the investigated traits have values varies across the different environments depending on the rainfall availability and very low Grain yield attributed to low rainfall. It was also assessed the relationships between the critical environmental factors and the phenotypic traits by means of correlation analysis and stated that water input in the vegetative phase was significantly related to Plant Height and Thousand Kernel Weight. The main objective of this study is to investigate the predictive value of various types of long-term agro-climatic variables on the future values of the some adaptive traits of durum wheat as well as the association between these traits and those of the different agro-climatic characteristics. Besides, other specific objectives are also present as assessing the predictive value of the agro-climatic variables on the future observations of Plant Height of the durum wheat landraces, and to study their association; investigating the predictive value of the climatic variables on the Thousand Kernel Weight of the durum wheat landraces, and to study their association. It has been also investigated the predictive value of the agro-climatic variables on the future observations of Grain Weight of the durum wheat landraces, and to study their association.

Data Description

238 durum wheat landraces were chosen from the International Center for Agricultural Research in the Dry Areas (ICARDA) genebanks, and collected from 9 different countries; Turkey, Iran, Iraq, Spain, Italy, Syria, Jordan, Greece and Palestine. These landraces were evaluated at the ICARDA station TelHady, Syria for three different response variables. 1. Plant Height (PHT): is the height of the plant from ground to top of spike measured in centimeter, excluding awns. 2. Thousand Kernel Weight (TKW): This is the weight in grams of 1000 well-developed whole grains, dried to 13% moisture content 3. Grain Weight (GRY): is weight of grains that was harvested, and registered on a scale of kilogram per hectare. In this study, 57 environmental variables including geographic coordinates: longitude and latitude were used. 36 out of the 55 are monthly long term averages for minimum, maximum temperature and for precipitation. The remaining 19 variables are derived from the monthly temperature and rainfall values in order to generate more biologically meaningful variables. These bio-climatic variables represent annual trends (e.g., mean annual temperature, annual precipitation), seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors (e.g., temperature of the coldest and warmest month, and precipitation of the wet and dry quarters). Data Availability: This data was taken from the Agricultural Research Company which is called International Center for Agricultural Research in the Dry Areas (ICARDA) in Microsoft excel, but it is not available online. It is available with me as Microsoft excel. Conflict of interest: The authors declare that there is no conflict of interest regarding the publication of this paper. This research was not funded by anybody. We did it without any fund.

2. Methodology

2.1. Multiple Linear Regression: There are crucial targets in regression analysis; such as making certain predictions and dealing with hypothesis tests [26]. In order to attain these goals, multiple linear regression models are used, which are among the most commonly applied statistical techniques for relating a set of two or more predictor variables, with a continuous response variable, with the restriction that the conditional mean of the response is linearly related to the predictor variables. This has the form:

$$Yi = \beta 0 + \sum_{j=1}^{p} \beta jxij + ei \qquad (2.1)$$

Where, n and pare the number of observations and the number of predictors, respectively. Yi is the response for the ith observation (i=1, 2, 3 ...238). Xij is the jth predictor for the ith observation, is the intercept. B0 is the effect parameter of the jth predictor. ei are independent and identically normally distributed with mean 0 and constant variance δ2. This model is applied for the two response variables (Days to Healing and Days to Maturity), independently. It is important to make sure that the assumptions of the model are satisfied. Violation of any of the model assumptions might possibly have an impact on the model's performance that is due to the inclusion of predictor variables that should not have been included or the exclusion of important predictor variable that were considered but rejected for inclusion in the model. Assumptions such as constant variance, linearity, outliers and normality should be checked. Violation of some of these assumptions might not have bad effect on the predictions. However, for the inferences (hypothesis testing), violation of any of these assumptions might be found misleading test statistics (p-values) and this might lead us to bad conclusions. As the predictors are expected to be correlated, there is a need for other parameter estimation methods that cope better with multicollinearity of course, there are also more general reasons why we might consider an alternative to the ordinary multiple linear regressions [21]. The first reason is prediction: the least-squares estimators frequently have small bias but large variance, and prediction can occasionally be improved by introducing bias in the estimates

of the regression coefficients, because it often comes with a reduction of their variability. This may improve the overall prediction performance (measured by mean-squared error (MSE)). The other motivation is for interpretation. With a large number of predictors, we often would like to identify a smaller subset of these predictors that demonstrate the strongest effects. In this case, model fitting was done using ordinary least squares, with stepwise selection criteria (explained more lately).

2.2. Penalized Regression Methods

Penalized regression methods are examples of modern approaches to model selection. Because they produce more stable results for correlated data, they are often preferred to traditional selection methods. Statistical model selection process based on such shrinkage methods work in such a way that it computes the prediction performance of various models in order to choose the approximate best model for the given data based on their predictability [7]. Usual model selection techniques such as stepwise selection methods achieve simplicity, but they have been revealed to yield models that have low prediction accuracy, especially in the presence of correlated predictors or when there are many predictors: - Penalized estimation methods may help as they are known to give better prediction accuracy; they received quite some attention over the last decade [9]. Shrinkage methods estimate the regression coefficients by minimizing the residual sum of squares (RSS), which is the same as that of the ordinary least squares, but with a penalty term added to put a constraint on the magnitude of the estimates of regression coefficients. These constraints cause the coefficient estimates to be biased, but it improves the overall prediction performance of the model by reducing the variance of the coefficient estimates [7]. These estimation methods and their relation to prediction performance, rely on the bias-variance trade-off [9]. Penalized estimation methods yield a sequence of models, each associated with a specific value of one or more penalty parameters. The researcher needs to apply a method to find the optimal value of the penalty parameter(s). This optimal value should correspond to an optimal model, that is, the model that has the smallest mean squared error. For this reason, K-fold cross-validation was used as it is recommended by [7]. With this method, and e.g. with K=10, the training data is partitioned into ten subsets (folds) consisting of observations (1, 11, 21 ...), (2, 12, 22 ...), and so on. Nine of these folds are used for model fitting, with a given value of the penalty parameter, and with the resulting fitted model the responses in the left-out fold are predicted and the corresponding prediction errors are computed. This process is repeated for each of the ten folds. At last, the prediction errors are squared and averaged, resulting in the cross-validation mean square error (MSECV), which measures the model predictive performance. It is computed as follows. First, calculate for each fold j,

$$MSEj(\lambda) = 1/nk \sum_{i\epsilon jth\ part}(yi - \hat{y}i^{-k}(\lambda))2 \qquad (2.2)$$

Where $\hat{y}i^{-k}$ is the predicted value from the fitted model without the observations in the kth left out part, and nk is the number of observations in the kth group. Finally, the CV estimate of the MSE is computed as

$$MSECV(\lambda) = 1/k \sum_{j=1}^{k} MSE(\lambda)j,$$

This is done for many values of λ and chooses the value of λ which gives the smallest MSECV(λ). Based on this, the model with minimum MSECV is selected as the best model. The main reason to use the shrinkage methods is that it works in such a way that the reduction in variance is of greater magnitude than the bias induced in the estimators [4]. Therefore, the net effect gives better predictions (the resulting model would have smaller MSE than the unbiased OLS model fit). After model fitting, in order to assure the validity of these fitted models, their different assumptions and overall goodness of fit test were assessed. In order to check the homogeneity of the variance of error terms, the white test is used. It jointly tests whether the error terms have homogeneous variance and whether they are independent and identically distributed [2]. Besides, residual versus predicted plots are constructed to reveal outlying observations as well to see whether the linearity assumption is fulfilled.

Bias-Variance Trade-off: It indicates the exchange of bias and variance, i.e by introducing bias in to the OLS estimators, the variance may reduce substantially. The bias-variance trade-off can be best explained by the mean square error (MSE) of a model, which is basically its expected prediction error. For a model M with regression coefficients β^, The MSE of a model is the sum of the variance of the predictions and the squared bias [3]. And it is given by:

$$MSE(M) = E\left(Ynew - (\hat{\beta}o + \sum_{j=1}^{p-1} \hat{\beta}j\, Xnew,j)\right)2 = Var\left(Ynew - (\hat{\beta}o + \sum_{j=1}^{p-1} \hat{\beta}j\, Xnew,j)\right) + Bias(\hat{\beta})2 \qquad (2.3)$$

Where Y_{new} and X_{new} represents a new data that are not used to obtain the coefficient estimates β^. In addition, the MSE of a linear model with regression coefficients β^ can be estimated by the average square error (ASE), as given by the following formula.

$$ASE(M) = \frac{\sum_{i=1}^{n}\left(Ynew-(\hat{\beta}o+\sum_{j=1}^{p-1}\hat{\beta}jXnew,j)\right)2}{n} \qquad (2.4)$$

In this study, different shrinkage methods were employed and are given as follow.

Lasso regression: Lasso (Least Absolute Shrinkage and selection operator) is a penalized estimation method that was first formulated by [20]. This method adds the sum of the absolute values of the coefficients to the sum of squared errors criterion. In particular, parameter estimators are defined as

$$\beta lasso = argmin\_\beta \sum_{i=1}^{n}(yi - \beta0 - \sum_{j=1}^{\nu}\beta jxij)2 + \lambda \sum_{j=1}^{\nu}|\beta j| \qquad (2.5)$$

Where λ≥0.

In this method, the parameter estimates are shrunken towards zero with increasing penalty parameter. However, some parameter estimates become exactly zero when the penalty parameter becomes sufficiently large. A zero parameter estimate implies that the corresponding predictor is no longer in the model, and, hence, lasso regression may be

looked simultaneously as an estimation method and model selection method. In other words, selecting an appropriate value of the penalty parameter is strongly related to model selection. In practice, this tuning parameter (λ) controls the strength of the penalty, and has a great importance. Indeed when λ is sufficiently large then some coefficients are forced to be equal to zero, this way reducing the dimensionality. The larger the parameter λ, the more coefficients are shrunken to zero. On the other hand if λ = 0, we have the ordinary least squares regression. There are many advantages, but also some limitations in using the lasso method. First of all, the lasso can provide a very good prediction accuracy of the fitted prediction models, because shrinking and removing coefficients can reduce variance without a substantial increase of the bias, resulting in a decreased MSE due to the variance-bias trade-off. Moreover, it helps to increase the model interpretability by eliminating irrelevant predictors that are not sufficiently related to the response variable, reducing over-fitting [6]. However, it also has its own limitations; when it is applied to high dimensional data (p>>>n), it gives at most n non-zero parameter estimates, and if there is a group of variables with high pair-wise-correlations among them, then this method tends to select only one variable from them, and doesn't care which one is selected (the model can't do group selection) [9]. In order to overcome these limitations, other method; elastic net method may be used.

Elastic net: This shrinkage method is an extension of lasso regularized regression method that linearly combines the lasso and ridge penalties. It reduces some of the limitations of the lasso method. For a high-dimensional predictor (p>>>n), unlike the lasso, it can give more than n non-zero parameter estimates. If there are grouped variables (highly correlated among one another), this method tends to select more than one predictor variable (it performs group selection) [9]. The coefficients of the elastic net method are estimated by minimizing the following penalized residual sums of squares. In particular, the estimate is given by following penalized residual sums of squares. In particular, the estimate is given by

$$\hat{\beta} = argmin\_\beta \sum_{i=1}^{n}(yi - \beta 0 - \sum_{j=1}^{p}\beta jxij)2 + \lambda 2\sum_{j=1}^{p}\beta j2 + \lambda 1 \quad (2.6)$$

Where $\lambda 2\sum_{j=1}^{p}\beta j2 \quad and \quad \lambda 1\sum_{j=1}^{p}|\beta j|$
are the penalties with λ2, λ1 ≥0. The lasso part of this penalty performs variable selection by setting some coefficients to exactly 0; whereas the ridge part of the penalty encourages the group selection by shrinking the coefficients of correlated variables toward each other, and stabilizes the lasso regularization path [27].

2.3. Post Model Selection Data Analysis Methods

The least square methods involve in estimating parameters by minimizing the squared differences between observed responses, and their corresponding model based predictions. In this study, Ordinary least square and weighted least square estimation methods are used.

Ordinary Least Square (OLS): Ordinary least squares are probably the most popular estimation methods of the parameters in a linear regression model. Their estimators are consistent and optimal in the class of linear unbiased estimators (LUE), when there is constant variance and independence of the observations. They are computed by minimizing the residual sums of squares, which is given below. However, the estimators may result in high variable estimates of the regression coefficients in the presence of multicollinearity [22].

$$\sum_{i=1}^{n}(yi - \beta 0 - \sum_{j=1}^{p}\beta jxij)2 \quad (2.7)$$

Weighted Least Square Estimation Method: One of the general assumptions underlying the majority of modeling methods is that each observation provides equally precise information about the deterministic part of the total process variation. Hence, it is assumed that the standard deviation of the error term is constant over all values of the predictor variables [19]. When the data does not meet these model assumptions, the parameter estimators will not be the most efficient estimators. Every term in the WLS encompasses an extra weight that indicates how much each data point in the data set affects the final parameter estimates. Less weight is given to the less precise observations and more weight to more precise data points during parameter estimation, and therefore using weights which are inversely proportional to the variance at every data point yields more precise parameter estimates [28]. During estimation, the weights compensate for the distorting effect of heteroscedasticity as well as down-weighting the influence of outliers [16]. Moreover, the estimates are calculated as a result of minimizing the weighted residual sum of squares (WRSS) [25]. The weighted least squares criterion is given by:

$$\sum_{i=1}^{n} wi(yi - \beta 0 - \sum_{j=1}^{p}\beta jxij)2 \quad (2.8)$$

Where WI is the weight of the ith observation. WLS residuals are given by $\sqrt{wi}$ (yi-ŷ) where wi= $1/\delta i^2$, δi^2 is error variance for observation i. The error variance is calculated as follow. Firstly, residuals (ei) are calculated, and then a model with the response variable squared residual (ei2) is fitted. From this model, predicted value of squared residual($\hat{e}i^2$) is estimated. Therefore, this predicted residual is the consistent estimator of δi^2. Due to this reason, WLS estimates may be more efficient comparing to the OLS estimates.

3. Results and Discussion

Summary statistics of the three response variables are presented. It is revealed that the total number of observations is 238 for all the variables, with no missing data. The variability among the measurements of the GRY (609.53 Std deviation) is higher as compared to PHT (11.52 Std deviation) and TKW (4.59 Std deviation). For PHT, the tallest accession (118) has almost double height of the shortest one (61). Similar pattern can be observed for TKW where the Grain Weight was almost double for some accessions compared to others. The accessions presented a high variability for GRY showing a difference of almost 3 ton/hr between the low and the high yielding accessions. In addition to this, heat map was constructed to visualize at the co-linearity among the 57 predictor variables, see Figure 1. It showed that the predictors can be characterized in to 5 distinct clusters in addition to few predictors that are not assigned to any of these clusters.

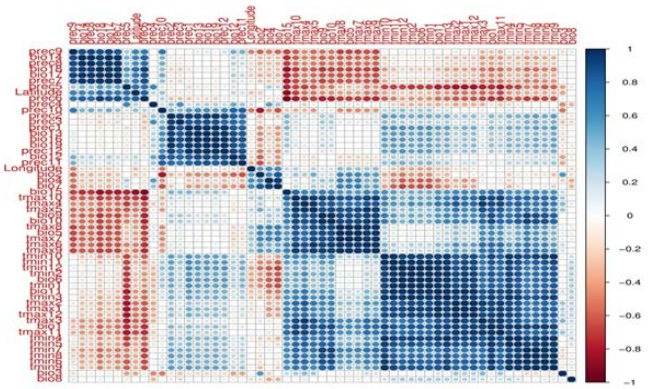

Figure 3.1: Heat map of the correlations between all the 57 predictors. The red color indicates the par-wise negative correlation whereas the blue color indicates pair-wise positive correlation. The white color is for no correlation. The largest one contained all the monthly predictors for minimum temperature plus monthly maximum temperature during winter time (tmax11, 12, 1, 2, 3) and three bio-climatic predictors related to temperatures (bio1, bio6 and bio11). The second cluster has variables related to moisture during summer time such as precipitation during May, June, July, August and September; and bio14, bio17 and bio18. The third cluster contains variables such as the precipitation during January, February, March, November and December. Besides, bio12, bio13, bio16 and bio19 are included in this cluster. The fourth cluster includes some monthly predictors for maximum temperature (tmax4, 5, 6, 7, 8, 9, 10) and some bio-climatic variables such as bio5, bio9, bio10 and bio15. The fifth cluster has some bio-climatic variables such as bio2, bio4 and bio7. In general, it can be said that there is high positive as well as negative correlations, which indicates the existence of high multicollinearity.

To further examine the multicollinearity, the variance inflation factors (VIF) were computed from OLS fit from the model with all the predictors included. It shows that the VIF is high (VIF>10) for all the predictors. This is an indication of high correlation among the predictors, and then high multicollinearity. It is noted that variables bio7 and prec12 have no VIF, because they are linear combination of the other variables (they have been set to 0). A graphical representation of the VIFs is given by the histogram in Figure 2. Only 19 predictors have a VIF smaller than 1000; the other have even larger VIFs. From this it should be noted that most of the predictors have VIF>1000, which is an indication of high multicollinearity. This suggests that the methods which are going to be used in this study should certainly be methods that work well in the presence of multicollinearity.

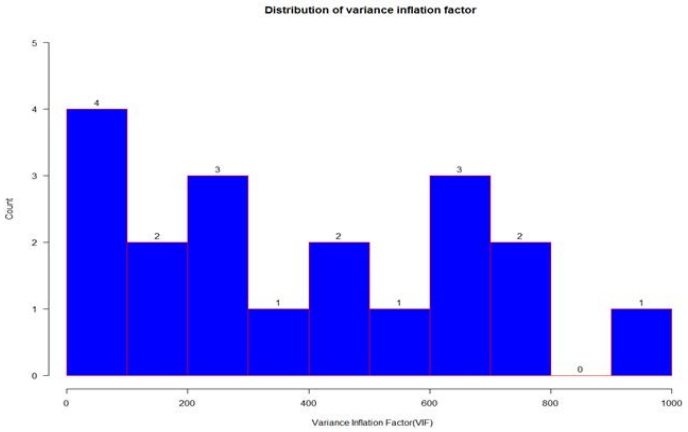

Figure 3.2: Histogram of Variance Inflation Factors (only the VIF ≤1000 of 19 predictors are shown). As revealed in figure 2, the numbers on each bar are the number of predictors those their VIF are within the interval. This suggests that the methods which are going to be used in this study should certainly be methods that work well in the presence of multicollinearity.

Model Building

Model fitting were done using OLS, Lasso and Elastic net methods. The OLS method was used in combination with the stepwise selection method for model building. This process consists of a series of alternating forward selection and backward elimination steps. Forward selection adds variables to the model if the variable is significant at the 0.15 significance level, whereas backward elimination removes variables from the model if a variable is not significant at 0.15 levels. As a result, the final predictors included in the ordinary MLR model are selected based on this criterion. The respective fitted models are given in Table 1 with their respective RMSE, and Table 3&4 with all included predictors. On the other hand, in order to select the optimal models based on the shrinkage methods, cross-validation (CV) with mean square error (MSE) as a model evaluation criterion were used. Firstly, random partitioning was used to split the available data into training set and test set. The model was fitted on the training set, including the selection of the penalty parameter, and validated using the test set. As it can be revealed (Table 1), four different partitions were used for each response; lasso and elastic net methods were applied for each partitioning.

Table 3.1: Comparison of partitions for the shrinkage based MLR models and comparison of predictive performance of all the three MLR models.

Variables	PHT			
Partitions	20-80	30-70	35-65	40-60
Methods(RMSE)	Lasso(10.369) Enet(10.418)	Lasso(10.530) Enet(10.530)	Lasso(10.654) Enet(10.649)	Lasso(10.730) Enet(10.730)
Variables	**GRY**			
Partitions	20-80	30-70	35-65	40-60
Methods (RMSE)	Lasso(624.32) Enet(624.32)	Lasso(579.17) Enet(579.17)	Lasso(631.38) Enet(631.38)	Lasso(568.42) Enet(568.42)
Variables	**TKW**			
Partitions	20-80	30-70	35-65	40-60
Methods (RMSE)	Lasso(4.456) Enet(4.456)	Lasso(4.571) Enet(4.571)	Lasso(4.605) Enet(4.605)	Lasso(4.542) Enet(4.542)

MSECV= mean square error based on cross-validation, PHT=Plant height, TKW= Thousand Kernel Weight, GRY-=Grain weight. The selected partitions and respective methods are in bold letters. For each partition, root mean square errors (RMSEs) were presented for all the models. Based on this, the partitions in bold letter were selected for each response since the models within these partitions have smaller RMSEs. The selected predictors for all the fitted models based on the shrinkage methods are given below (table 3), for each response. Model assumptions were checked after model fitting. It is revealed from Table 2 of the normality test for the complete (original) data, that the residuals find from regression models fitted for GRY and TKW are normality distributed, whereas for PHT are not normally distributed, all at the 5% significance level. For the test data set, the residuals for PHT and TKW are normally distributed, while for GRY are not normally distributed, all at the 5% level of significance. It should be noted that the normality assumption is needed only for the OLS fitted models.

Table 3.2: Results for normality, homogeneity of variance and Goodness of fit (GOF) tests.

Test(P-value)	Ordinary MLR Models using original data set			
	PHT	GRY		TKW
Normality(P-val)	0.0005	0.055*		0.078*
White(P-val)	0.118 *	0.214 *		0.878*
GOF(P-val)	0.297*	0.689 *		0.194*
Shrinkage based Models using test data set				
Test(P-value)	Lasso(PHT)	Enet(PHT)	Lasso(GRY)	Lasso(TKW)
White(P-val)	0.466*	0.954*	0.461*	0.144*
GOF(P-val)	0.575*	0.574*	0.135*	0.038

Original data was used for Ordinary MLR model, and test data was used for all the shrinkage based MLR models. Normality and Homogeneity of variance tests are based on Shapiro-Wilk and white test, respectively, PHT=Plant height, TKW= Thousand Kernel Weight, GRY=Grain weight. Tests with

* showed that error terms are identically, independently and normally distributed, have constant variance and the model has no lack of fit at 5% level of significance.

In all the ordinary MLR models and shrinkage based MLR models of all the data sets, the homogeneity of variance test showed that there is constant variance at 5% level of significance. Results of the goodness of fit test for the ordinary MLR models based on the original data set indicated that no model shows lack of fit. However, for the shrinkage based MLR models, the model for TKW showed lack of fit. This may happen due to the reason that the relationship between the response and predictor variables is not linear.

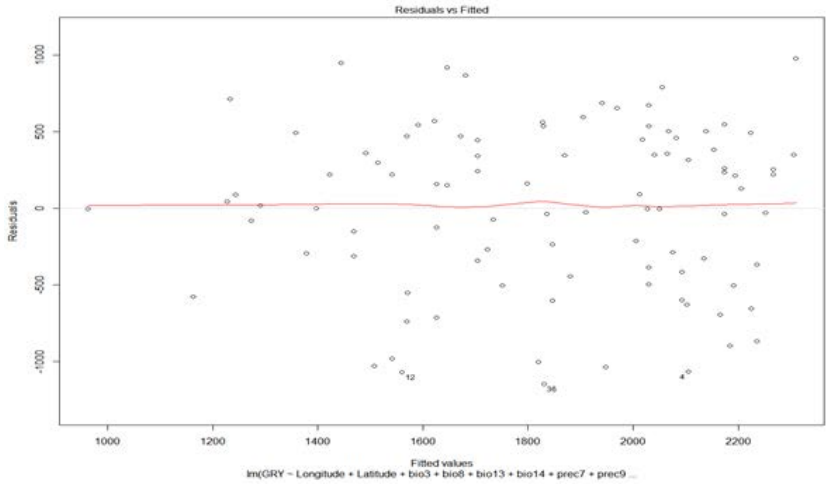

Figure 3.3: Plot of residuals versus predicted values for test data set (40% of the data set) of Grain Weight, lasso based MLR model. It is observed from Figure 3 that the red line is a smoothed high order polynomial curve to show the pattern of residual movement in order to assess linearity. Moreover, observations 12, 36 and 91 are identified as influential outliers. In this case, it should be noted that there is no noticeable deviation from the linearity. Of course this is given as a sample for one of the three response variables. For the other shrinkage based MLR models using the test data set, observations 43, 45 and 49, and observations 8, 19 and 28 are identified as outliers for PHT and TKW, respectively. In these figures, it is observed that there seems some deviations from the linearity, especially for TKW.

Inference Post Model Selection

For the shrinkage methods, for responses of GRY and TKW, the elastic net results coincided with the results of the lasso method, and hence only results for the models fitted by the lasso method are presented here. However, for PHT results for both the Lasso and the Elastic net are given. Parameter estimates based on both OLS and WLS estimation methods of the ordinary MLR models for the five responses are given Table 3. Based on the WLS estimation method, it is noticeable that prec5 and prec9 has positive significant effect, while using OLS method, prec5 and bio3 have positive significant effect on the Plant Height. Moreover, using the WLS method, bio7 and bio13 have increasing significant effect, while bio16 has decreasing significant effect on the Grain Weight. By the OLS estimation method, bio7 and bio13 have increasing significant effect on this response. Longitude from both WLS and OLS estimation methods has increasing significant effect on the Thousand Kernel Weight.

Table 3.3. Estimated values for OLS and WLS estimation methods of ordinary MLR models, using complete data set.

Plant Height(PHT)						
Intercept	69.20378	8.14286	<.0001*	73.44546	8.88475	<.0001*
prec5	0.21588	0.04950	<.0001*	0.17496	0.05550	0.0018*
prec9	0.11041	0.05666	0.0525	0.11922	0.05071	0.0196*
bio3	0.48946	0.20038	0.0153*	0.40551	0.22116	0.0680
Grain Weight(GRY)						
Intercept	255.78127	311.57512	0.4125	179.99244	314.18510	0.5673
bio7	3.82040	0.79024	<.0001*	4.08747	0.77727	<.0001*
bio13	17.85035	8.64917	0.0401*	19.90808	8.18516	0.0158*
bio16	-5.42633	3.21529	0.0928	-6.28979	3.03851	0.0396*
Thousand Kernel Weight(TKW)						
Intercept	31.13630	1.00523	<.0001*	30.77325	0.91403	<.0001*
Longitude	0.08395	0.02906	0.0042*	0.08953	0.02676	0.0010*

bio3= Isothermality, bio7= Temperature Annual Range, bio9= Mean Temperature of Driest Quarter, bio12= Annual Precipitation, bio13= Precipitation of Wettest Month, bio14= Precipitation of Driest Month, bio15= Seasonality precipitation, bio16= Precipitation of Wettest Quarter, bio18= Precipitation of Warmest Quarter, ,bio19= Precipitation of Coldest Quarter, preci= Precipitation of ith month, tmini= Minimum temperature of ith month, tmaxi= Maximum temperature of ith month (i=1,2,3,...,12), Par.Est=Parameter estimate, Std.Er= Standard Error.
* Indicated significance at 5% level of significance.

Besides, to evaluate the predictability of these models, see Figure 4 for both WLS and OLS estimation methods for Plant height. It is noticed that there is some variability in the residuals. Although the predicted value continuously increases as a function of the Plant height, the variability seems to need some concerns. From this Figure, our model seems to have two subsections of performance. Besides, the prediction level between the weighted and ordinary least square estimation methods seems similar. The prediction level also seems consistent though the variability is not.

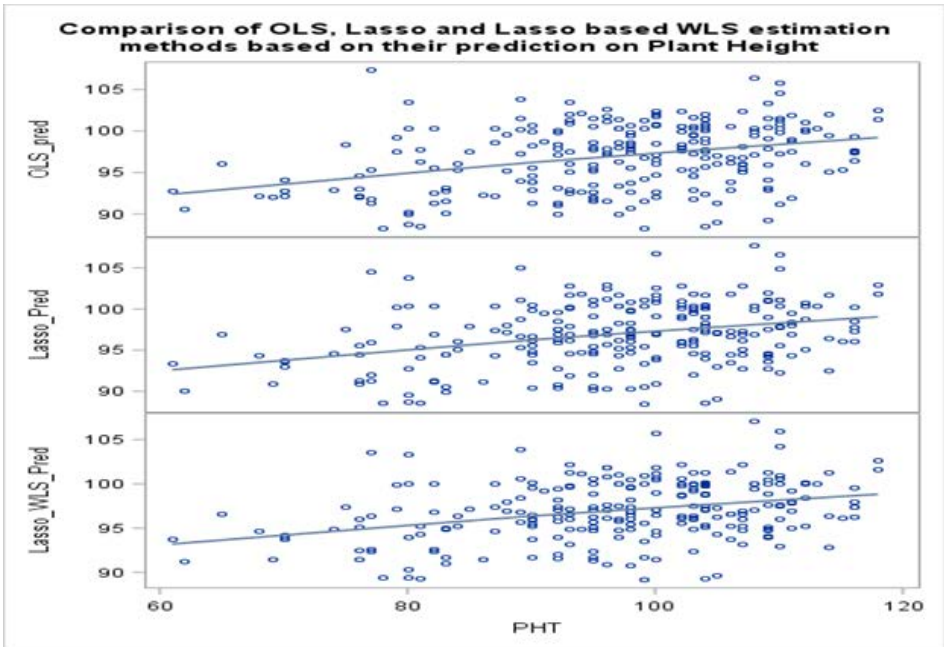

Figure 3.4: Predicted versus actual value for Plant height of the complete data set. Furthermore, as the predictive Figures shown, the WLS methods seem slightly to perform better prediction than the OLS methods, however the difference is not that much visible. The RMSE of the models used WLS estimation method are less than that of the models used OLS estimation method in all the models, which is suggesting that the estimates from the WLS estimation method might be more sensible and precise results. The models used the WLS estimation method might have better predictability may be due to the fact that this method minimizes the effect of variability. Moreover, parameter estimates by the MLR models with the predictors selected by shrinkage methods are given for GRY Table 4. Based on the WLS estimation method in Grain Weight, bio3 and tmin11 have decreasing significant effect, whereas tmax11 has increasing significant effect. Holding constant the other predictors in the model, a unit increases in bio3 and tmin11, results a decreasing for the grain Weight by 122.892 and 13.986 kg/hectare, respectively. Tmin11 and bio3 have decreasing significant effects on the Grain Weight as the OLS estimation method showed. The Grain Weight decreases by 19.202 and 102.662 kg/hectare as the tmin11 and bio3 increased by a unit measure, respectively.

Table 3.4: Parameter Estimates of MLR models with the predictors selected by lasso and elastic net methods using test data set.

Grain Weight (GRY)							
Effect	Pen.Est	OLS estimation method			WLS estimation method		
		Par.Est	Std.Er	P-value	Par.Est	Std.Er	P-value
Intercept	-277.299	3793.887	2530.737	0.1376	3054.933	2244.606	0.1772
Longitude	7.761610	-13.6342	11.41222	0.2356	-9.94215	9.87285	0.3169
Latitude	63.27940	18.69395	43.10231	0.6656	59.74043	32.61262	0.0706
bio3	-41.8961	-102.662	50.53256	0.045*	-122.89213	47.90010	0.0121*
bio8	0.221600	0.80888	2.01095	0.6885	0.14471	2.19654	0.9476
bio13	3.718134	3.82237	2.22535	0.0895	3.42343	2.25237	0.1323
bio14	13.07108	60.72441	38.97756	0.1230	58.70185	37.69593	0.1232
prec7	3.415454	-51.4843	34.34114	0.1376	-67.40306	35.79828	0.0632
prec9	-14.2640	-11.3503	15.88892	0.4770	-3.17688	12.90839	0.8062
tmin9	-6.96310	-6.20974	8.50869	0.4675	-14.02547	8.16688	0.0896

tmin11	-0.66160	-19.2018	6.19670	0.002*	-13.98567	5.08390	0.0073*
tmax11	10.15632	19.44664	10.15798	0.0590	24.01414	9.61972	0.0145*
MLR model based on lasso for plant Height(PHT)							
Intercept	79.86821	76.99277	23.70949	0.0024*	71.72450	21.59442	0.0020*
bio2	0.059033	-0.02680	0.16233	0.8697	0.03554	0.15075	0.8149
bio3	0.028272	0.31676	0.82552	0.7032	0.25107	0.76565	0.7448
bio12	0.004314	0.01090	0.01443	0.4544	0.02297	0.01506	0.1355
bio18	0.012706	0.16811	0.20120	0.4084	-0.01030	0.18748	0.9565
prec4	0.047435	-0.46107	0.20369	0.0291*	-0.45347	0.20499	0.0330*
prec5	0.013244	0.95650	0.33894	0.0074*	0.70691	0.34804	0.0493*
prec6	0.155099	-0.63164	0.50575	0.2190	-0.12980	0.50336	0.7979
Elastic net based MLR model for Plant Height(PHT)							
Intercept	87.596328	120.18742	21.76889	<.0001*	128.8806	21.57749	<.0001*
bio2	0.016853	-0.02597	0.16601	0.8765	0.05299	0.16765	0.7536
bio3	0.043455	-0.36131	0.77691	0.6443	-0.80393	0.78435	0.3115
bio18	0.071392	0.09044	0.07659	0.2443	0.12496	0.06942	0.0794
prec4	0.074432	-0.04310	0.11542	0.7107	-0.12933	0.11989	0.2872
tmin11	-0.017790	-0.10937	0.07914	0.1743	-0.08849	0.06922	0.2085
Lasso based MLR model for TKW							
Intercept	33.736	30.70027	1.50524	<.0001*	31.03786	1.89625	<.0001*
Longitude	0.0174	0.06480	0.04336	0.1418	0.05421	0.05033	0.2871

bio8= Mean Temperature of Wettest Quarter, bio2= Mean Diurnal Range (Mean of monthly (max temp - min temp)), Pen. Est=Penalized coefficient estimates, Par.Est=Parameter estimate.

* Indicated significant at 5% level of significance.

Furthermore, parameter estimates for Plant Height, for both lasso and elastic net based models and Thousand Kernel Weight, for lasso based model, are given in Table 4. By the WLS estimation method, as prec5 increased by a unit measure, Plant Height increases by 0.707 centimeters. While prec4 showed a unit increase, Plant Height might decrease (reduced) by 0.453 centimeters, by holding constant all the other predictors within the models. On the other hand using the OLS method, Plant Height increases by 0.956 centimeters as prec5 showed a unit increment. When prec4 increases by one unit, Plant Height decreases by 0.461 centimeters. Note that the negative effect of the some predictors on the Plant Height showed that in some cases the predictors have no importance in growing the height of the durum wheat, on another cases, the height of the plant might be shrunken (become short) as a result of these negative effects. Based on as table 4, in most of the parameters, these penalized estimates are somehow smaller in magnitude than the un-penalized coefficient estimates (estimates from post model selection). However, in some parameters the penalized estimates are larger in magnitude. This indicates that on the process of shrinking some of the parameters forced to have smaller magnitude whereas others to have larger values.

To evaluate the predictability of the MLR models with predictors selected by shrinkage methods, see the elastic net based model (fig 5.2); there seems continuously increasing of the predicted value as a function of the actual value, however there seems high variability. Observations are not close to the diagonal line, and this might indicate prediction is questionable. On the other hand, the lasso based model (fig 5.1) seems to have three subsections of performance. The first one is where actual values between about 70 and 85. Within this subsection, the diagonal line seems straight with small dispersed data points. The second one is when actual values between 85 and 105. Within this subsection, there are ups and downs with random moves. The third case is where actual values above 105. In this zone, the prediction seems better comparing to the other subsections. However, in all cases our model seems random, less predictive. It is important to note that the prediction is more sensible for the WLS estimation methods than that of the OLS methods in all the models, but not that much visible in the predictive plots. This may happen due to the fact that the data is highly random dispersed.

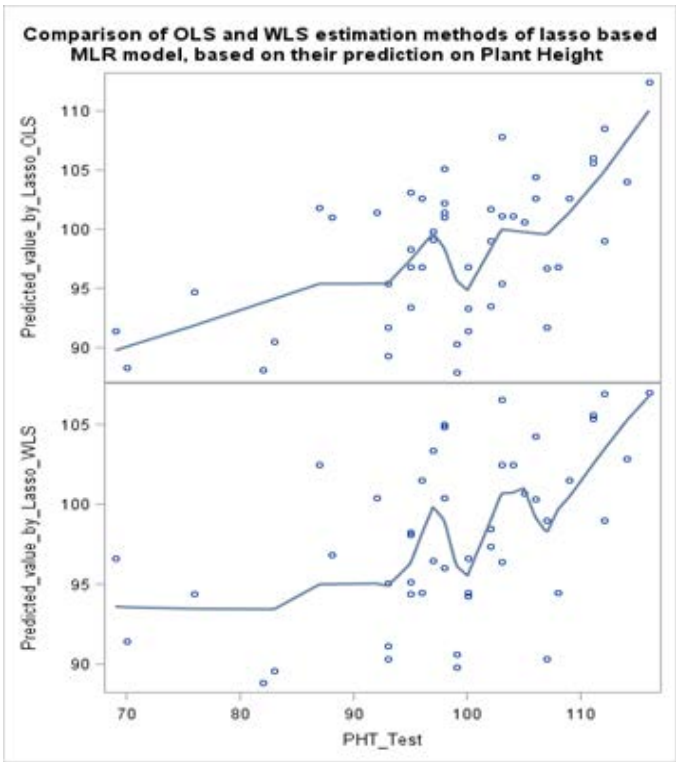

fig 3.5.1

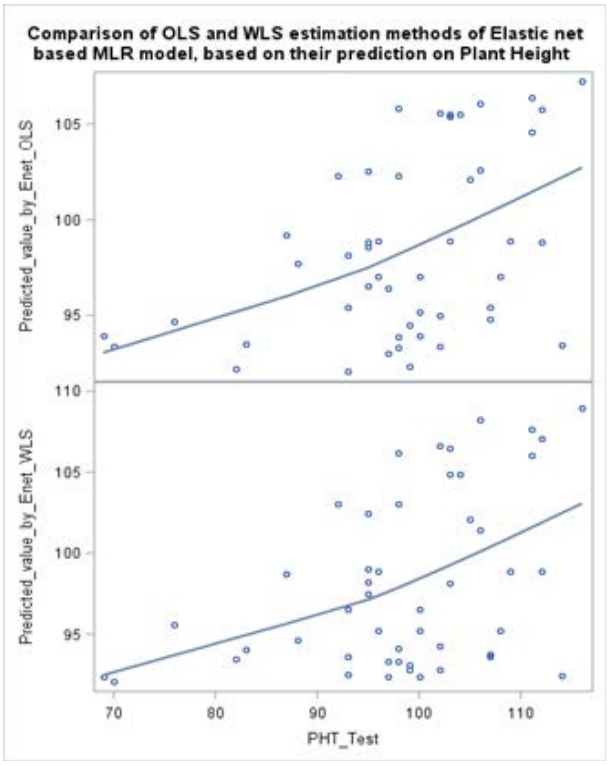

fig 3.5.2

Figure 3.5: Actual versus predicted values for Plant Height of both lasso and elastic net based MLR models. In general, the parameter estimates from the ordinary MLR models are not sensible as the fitted models based on this are questionable due to the multicollinearity problem. Especially for prediction these models are not advisable. Differently, the estimates from the MLR models with predictors selected by penalized methods are more reasonable since these methods are not that much affected by variability, and are more important for prediction, thanks to the bias-variance trade-off method. Moreover, due to the violation of some model assumptions, p-values might be disturbed, and then the inference (hypothesis testing) may be questionable. However, these assumptions may not be that much important for the prediction, it may not be affected even with violations of some of them. Besides, the estimates from WLS estimation methods might also be more efficient than the estimates from the OLS estimation methods. This might be due to the reason that the OLS estimation method is easily affected by the model assumptions. In addition to this, the RMSE of the WLS estimation methods in all the models and the response variables are smaller than the OLS methods, which indicates there is better prediction by the WLS estimation methods. Therefore, the most sensible predictions may be made by the shrinkage method based MLR models with WLS estimation methods.

4. Conclusion

The WLS estimation methods of shrinkage based models revealed that Bio3 and temperature minimum of November (tmin11) have decreasing significant effect, while maximum temperature of November (tmax11) has increasing significant effect on the Grain Weight based on WLS method. Minimum temperature of November (Tmin11) and bio3 have negative significant effects on the Grain Weight as OLS method showed. Precipitation of month May (Prec5) has increasing, but precipitation of month April (prec4) has decreasing significant effect on plant Height using both OLS and WLS methods of the Lasso based model. However, there is no predictor with significant effect by the Elastic net based model on the Plant height. The ordinary MLR models on Grain Weight and Plant Height seem to have continually increasing relationship of the predicted values as a function of the actual values, but predictions are questionable since there is considerable variability. The models on Thousand Kernel Weight also showing that predicting using these models is not trustful. From models with predictors selected by shrinkage methods, it is revealed that Elastic net based model seems to have a little bit good prediction on the Plant Height for both OLS and WLS estimation methods though there is considerable variability and outliers, while the prediction from the Lasso based model is not that much reasonable. Furthermore, for the Grain Weight showed that there seems sensible prediction as their predicted value increase continuously as a function of the actual values, but we should also note that there is sounding variability which may make the prediction uncertain. The Lasso based model used for Thousand Kernel Weight is not predicting well.

In summary, our results suggested that inferences and predictions by the ordinary MLR models are not trusted due to the effect of multicollinearity. Not only that, as there are some violated model assumptions, the test statistics (p-values) are not believable, as a result, the inferences (hypothesis tests) may not be dependable. However, predictions using the models with penalized methods are more reasonable as

the effects of the variability on these methods are minimal. Moreover, the WLS methods give more sensible estimates and predictions than the OLS estimation methods. Although there is substantial variability, better predictions are observed on the Plant Height and grain Weight, especially by the weighted least squares estimation methods.

As a recommendation, it is better if further study on this topic is done using nonlinear and robust methods.

References

1. Brown, A. H. D. (1989). Core collections: a practical approach to genetic resources management. Genome, 31(2), 818-824.
2. Christensen, L. A. (1997, March). Introduction to building a linear regression model. In Proceedings of the Twenty-Second Annual SAS Users Group International Conference.
3. Cohen, R. A. (2006, March). Introducing the GLMSELECT procedure for model selection. In Proceedings of the Thirty-First Annual SAS Users Group International Conference (pp. 4770-4792). Citeseer.
4. Del Moral, L. G., Rharrabti, Y., Villegas, D., & Royo, C. (2003). Evaluation of grain yield and its components in durum wheat under Mediterranean conditions: an ontogenic approach. Agronomy Journal, 95(2), 266-274.
5. Dias, A. S., & Lidon, F. C. (2009). Evaluation of grain filling rate and duration in bread and durum wheat, under heat stress after anthesis. Journal of Agronomy and Crop Science, 195(2), 137-147.
6. Fonti, V., & Belitser, E. (2017). Feature selection using lasso. VU Amsterdam research paper in business analytics, 30, 1-25.
7. Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning: data mining, inference, and prediction. Springer.
8. Giunta, F., Motzo, R., & Deidda, M. (1993). Effect of drought on yield and yield components of durum wheat and triticale in a Mediterranean environment. Field Crops Research, 33(4), 399-409.
9. Gunes, F. (2015). Penalized regression methods for linear models in SAS/STAT®. In Proceedings of the SAS Global Forum 2015 Conference. Cary, NC: SAS Institute Inc.
10. Kabbaj, H., Sall, A. T., Al-Abdallat, A., Geleta, M., Amri, A., et al (2017). Genetic diversity within a global panel of durum wheat (Triticum durum) landraces and modern germplasm reveals the history of alleles exchange. Frontiers in plant science, 8, 1277.
11. Khan, M. H., Hassan, G., Khan, N., & Khan, M. A. (2003). Efficacy of different herbicides for controlling broadleaf weeds in wheat. Asian Journal of Plant Sciences (Pakistan).
12. Khazaei, H., Street, K., Bari, A., Mackay, M., & Stoddard, F. L. (2013). The FIGS (Focused Identification of Germplasm Strategy) approach identifies traits related to drought adaptation in Vicia faba genetic resources. PloS one, 8(5), e63107.
13. Maçãs, B., Gomes, M. C., Dias, A. S., & Coutinho, J. (2000). The tolerance of durum wheat to high temperatures during grain filling. Options Méditerranéennes. Durum wheat improvement in the Mediterranean region: new challenges, 257-261.
14. Maccaferri, M., Sanguineti, M. C., Demontis, A., El-Ahmed, A., Garcia del Moral, L., et al (2011). Association mapping in durum wheat grown across a broad range of water regimes. Journal of experimental botany, 62(2), 409-438.
15. Mackay, M., Von Bothmer, R., & Skovmand, B. (2005). Conservation and utilization of plant genetic resources-future directions. Czech Journal of Genetics and Plant Breeding, 41(Special Issue), 335-344.
16. Nishida, K. (2021). Skewing methods for variance-stabilizing local linear regression estimation. Communications in Statistics-Simulation and Computation, 50(7), 2089-2106.
17. Ottman, M. J., Kimball, B. A., White, J. W., & Wall, G. W. (2012). Wheat growth response to increased temperature from varied planting dates and supplemental infrared heating. Agronomy Journal, 104(1), 7-16.
18. Rao, N. K. (2006). Manual of seed handling in genebanks (No. 8). Bioversity International.
19. Romano, J. P., & Wolf, M. (2017). Resurrecting weighted least squares. Journal of Econometrics, 197(1), 1-19.
20. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.
21. Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations. CRC press.
22. Van der Kooij, A. J., & Meulman, J. J. (2008). Regularization with ridge penalties, the lasso, and the elastic net for regression with optimal scaling transformations. Submitted for publication.
23. Van Hintum, T. J., Brown, A. H. D., & Spillane, C. (2000). Core collections of plant genetic resources (No. 3). Bioversity International.
24. Wu, W., May, R., Dandy, G. C., & Maier, H. R. (2012). A method for comparing data splitting approaches for developing hydrological ANN models.
25. Yaffee, R. A. (2002). Robust regression analysis: some popular statistical package options. Statistics, social science, and mapping group academic computing services information technology services, 1-12.
26. Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.
27. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301-320.
28. https://www.azdhs.gov/documents/preparedness/state-laboratory/lab-licensurecertification/technical/resources/calibration-training/11-weighted-least-squaresregression-calib.pdf. Accessed on May 5, 2018.