Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2019;5(2): 120-136

Original Article

Quantifying the Relationship between Adaptive Traits and Agro-climatic Conditions

Mehari Gebre

Department of Statistics, College of Natural and Computational Science Adigrat University, Adigrat, Tigray, Ethiopia.

ARTICLE INFO

Received 20.01.2018
Revised 15.04.2018
Accepted 06.10.2018
Published 01.02.2019

Key words:
Cross-validation;
Mean Square Error;
MLR;
Penalized Methods;
Lasso;
Elastic net;
Bias-Variance Trade-off;
Weighted Least Square

ABSTRACT

Background & Aim: Durum wheat is an economically important and regularly eaten food for billions of people in the world. In the International Center for Agriculture Research in the Dry Areas (ICARDA), genbanks are using Focused Identification of the Germplasm Strategy (FIGS) to find out and quantify relationships between agro-climatic conditions and the presence of specific traits. Hence, the study is aimed to investigate the predictive value of various types of long-term agro-climatic variables on the future values of different traits.

Method: 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 as a model selection criterion, are used to analyze 238 durum wheat landraces. Each of the models are fitted on Days to Heading and Days to Maturity response variables with 57 predictor variables, independently. Ordinary least square and weighted least square estimation methods were used.

Result: Findings implied that there is high multicollinearity among the predictor variables. It is found that there are some predictors which affect positively and some others affect negatively for both Days to Heading and Days to Maturity using both ordinary and shrinkage based models. It is revealed that the prediction from the lasso based model is not that much reasonable. Furthermore, for the Days to Heading showed that there seems better prediction as their predicted value increase continuously as a function of the actual values though there is considerable variability.

Conclusion: 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 methods may be better as the effects of the variability on these methods are minimal. Moreover, the WLS methods might give more sensible predictions than the OLS estimation methods. Better predictions were found on the Days to Heading.

Introduction

Wheat is a routinely eaten food for billions of people in the world; used to make flour for leavened, different types of breads, cookies, cakes, pasta, noodles and couscous; for fermentation making beer and alcohol [11]. Triticum durum(Durum wheat) is the only tetraploid form of wheat broadly being used these days, and is the 10th most essential crop in the world, which covers about 10% of the world's wheat. Durum wheat 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; however in the preceding century only part of the genetic variety accessible for this species has been captured in modern varieties through breeding [10]. Wheat breeders over the past century have increased the productivity and adaptability via strong selection applied to genes controlling agronomical important traits, and genotypic stability to be able to grow wheat, in a range of climatic

Email Address: *meharistat@gmail.com*

Please cite this article in press as Gebre M. Quantifying the Relationship between Adaptive Traits and Agro-climatic Conditions. J Biostat Epidemiol. *2019; 5(2): 120-136*

zones varying from warm and dry to cool and wet environments which are mostly located in areas subject to alternating favorable and stressed conditions[10].Therefore, genetic improvement via breeding for tolerance to biotic and a biotic stresses remains a strategic practice to improve its productivity and stability. In the last decades, many durum wheat varieties have been developed based on field assessment for higher yield, disease resistance, stress tolerance and good seed quality [10].

Several methods were developed to overcome the size problem of genebanks. The most widely used is the concept of core collections introduced by [1]. Core collection is a subset of a collection capturing the majority of genetic variation in a genebank with little genetic redundancy. To develop a core collection, one can use passport, environmental, phenotypic or molecular data. The International Center for Agricultural Research in the Dry Areas (ICARDA) in collaboration with Australian partners has developed an alternative approach for better targeting adaptive traits over the past 10 years. The Focused Identification of Germplasm Strategy (FIGS) is a trait-based approach allowing the identification of sought traits with high probability, and was designed to get better efficiency with which specific adaptive traits are identified from genetic resource collections. It is based on the principle that adaptive traits displayed by an accession will reflect the selection pressures of the surroundings from which it was originally sampled [15]. In the international center for Agriculture Research in the Dry Areas (ICARDA), the genbank is using the Focused Identification of the Germplasm Strategy (FIGS) to find out and quantify relationships between collection site agro-climatic conditions and the presence of specific traits, such as disease resistance or heat tolerance, as a result this approach led to the discovery of previously undiscovered genes and useful variations of known genes for resistance to serious pests and diseases. The FIGS approach uses both trait and environmental data to develop a best bet set with high probability of finding adaptive trait [12]. 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]. 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 Days to Heading. 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 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 Days to Heading 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 Day to Maturity of durum wheat landraces, and 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 two different response variables. 1. Days to Heading (DHE): is the number of days required for the inflorescence (head of plant) to emerge from the flag leaf of a plant or a group of plants in a study. 2. Days to Maturity (DMA): this is the number of days required for the plant from seeding to seed/grain ripening. 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).

Methodology

Multiple Linear Regressions: 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:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i$$

Where, n and p are 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. β0 is the effect parameter of the jth predictor. ei are independent and identically normally distributed with mean 0 and constant variance σ^2. 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).

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,

$$MSE_j(\lambda) = 1/nk \sum_{i\epsilon jth\, part}(y_i - \hat{y}_i^{-k}(\lambda))^2 \quad .$$

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

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

This is done for many values of λ and chooses the value of λ which gives the smallest $MSECV(\lambda)$. 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 $\hat{\beta}$, The MSE of a model is the sum of the variance of the predictions and the squared bias [3]. and it is given by:

$$MSE(M) = E\left(\mathrm{Y_{new}} - (\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j \,\mathrm{X_{new,j}})\right)^2$$

$$= Var\left(\mathrm{Y_{new}} - (\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j \,\mathrm{X_{new,j}})\right) + \mathrm{Bias}\left(\hat{\beta}\right)^2$$

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

$$ASE(M) = \frac{\sum_{i=1}^{n}\left(\mathrm{Y_{new}} - (\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j \,\mathrm{X_{new,j}})\right)^2}{n}$$

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

$$\hat{\beta} lasso = argmin\_\beta \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\lambda \geq 0$.

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

$$\hat{\beta} = argmin\_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

where $\lambda_2 \sum_{j=1}^{p} \beta_j^2$ and $\lambda_1 \sum_{j=1}^{p} |\beta_j|$ are the penalties with $\lambda_2, \lambda_1 \geq 0$.

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].

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 by:

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

However, the estimators may result in high variable estimates of the regression coefficients in the presence of multicollinearity [22].

Weighted least Square (WLS) 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 heteroskedasticity 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

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

where wi is the weight of the ith observation.

WLS residuals are given by $\sqrt{w_i}\,(y_i - \hat{y})$ where $w_i = 1/\sigma_i^2$, σ_i^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 (e_1^2) 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 σ_i^2. Due to this reason, WLS estimates may be more efficient comparing to the OLS estimates.

Results and Discussion

Summary statistics of the two response variables were presented. It is revealed that the total number of observations is 238 for both the variables, with no missing data. The variability among the measurements of DMA(3.06 std deviation) is smaller as compared to the DHE(4.17 std deviation). For both the variables, there is 21 and 25 days respectively between the earliest and latest accession(differences between the maximum and the minimum values). Besides, 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.

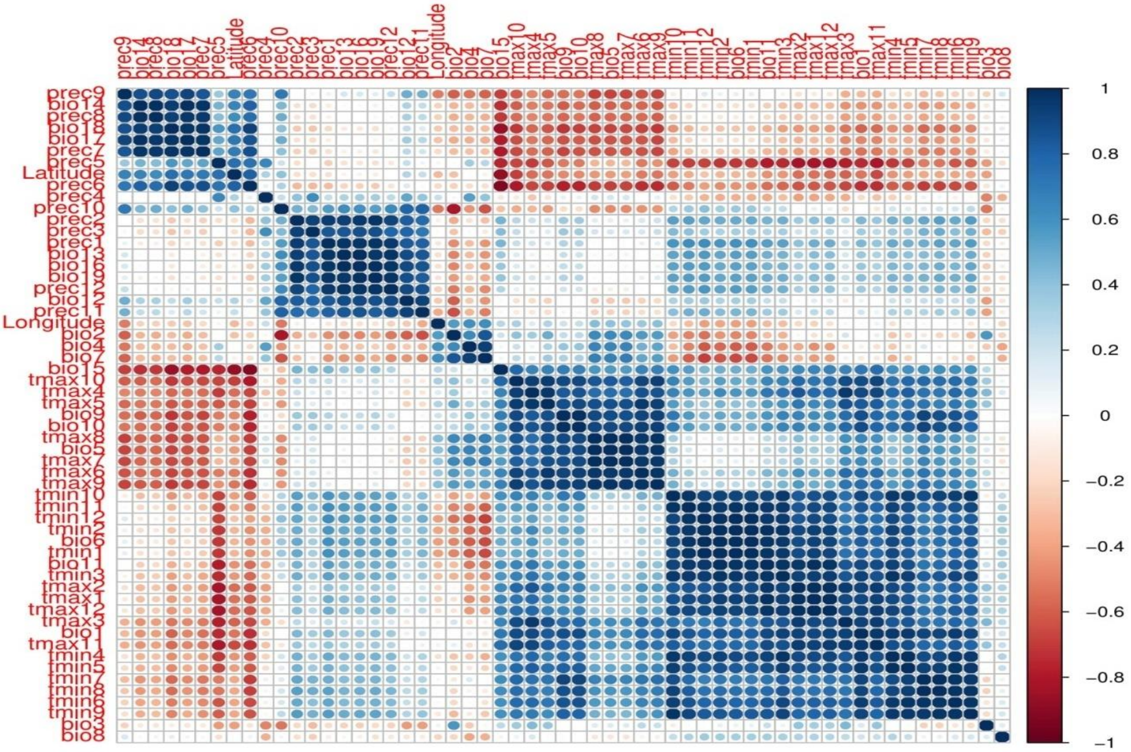

Figure 1: Heat map of the correlations between all the 57 predictors. The red color indicates the par-wise negative correlation whereas the blue color indicates pair-wise positive correlation. The white color is for no correlation.

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 and the response used here was Days to Heading. The result showed 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.

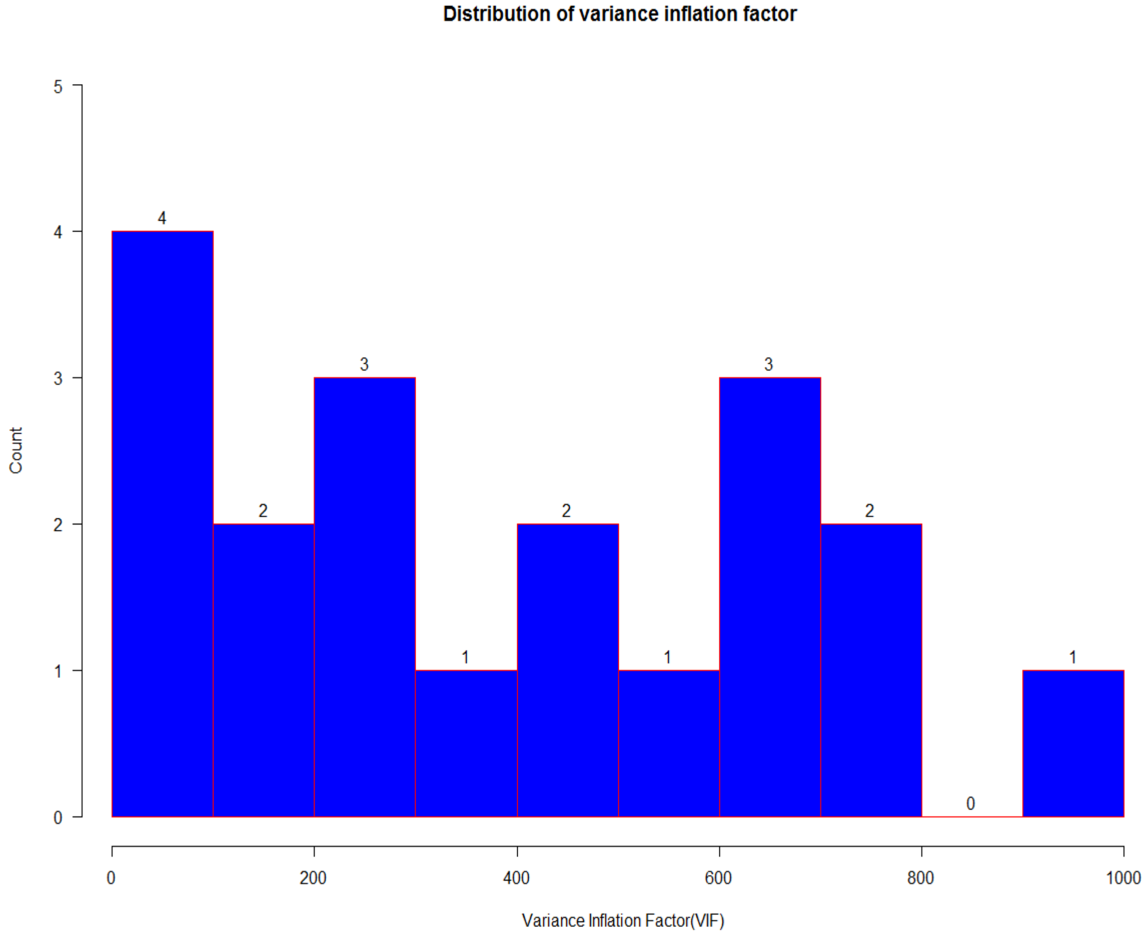

Figure 2: Histogram of Variance Inflation Factors (only the VIF ≤1000 of 19 predictors are shown). The numbers on each bar are the number of predictors those their VIF are within the interval.

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 level. 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. 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, four different partitions were used for each response; lasso and elastic net methods were applied for each partitioning.

Table 1: Comparison of partitions for the shrinkage based MLR models in order to select the best partition which gives the optimal models, and comparison of predictive performance of all the three MLR models, based on RMSE.

Variables	DHE			
Partitions	20-80	30-70	35-65	40-60

Methods(RMSE)	Lasso(3.323
	)
	Enet(3.323)
		Lasso(3.159)
		Enet(3.158)
			Lasso(3.538)
			Enet(3.538)
				Lasso(3.310)
				Enet(3.310)

Variables	DMA			
Partitions	20-80	30-70	35-65	40-60

Methods
(RMSE)
	Lasso(2.780
	)
	Enet(2.787)
		Lasso(3.010)
		Enet(3.010)
			Lasso(2.501)
			Enet(2.506)
				Lasso(2.904)
				Enet(2.904)

MSE=MSECV= mean square error based on cross-validation, DHE=Days to heading, DMA=Days to maturity. N.B. 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 both the fitted models based on the shrinkage methods are given below.

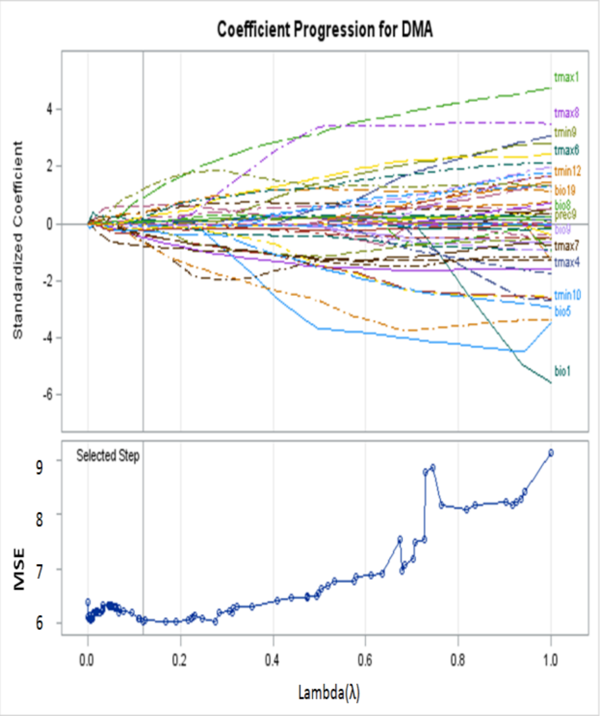

fig 3.1

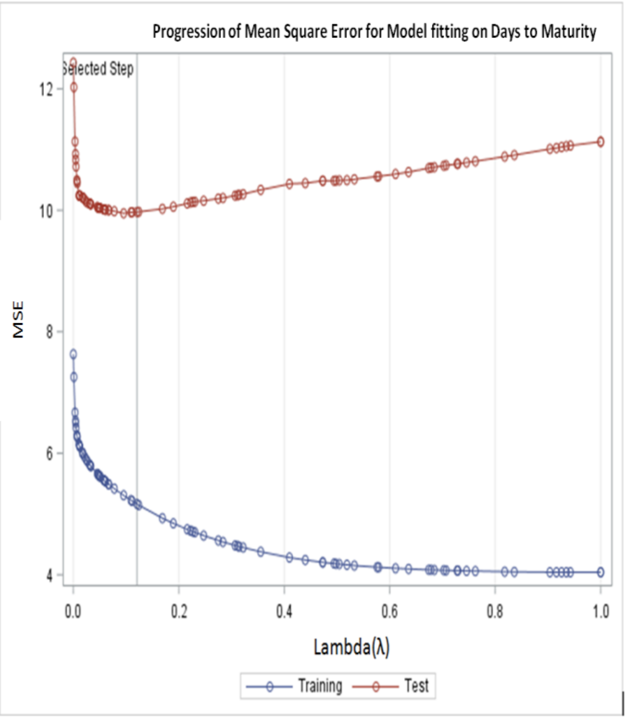

fig 3.2

Figure 3: Forward variable selection process based on lasso method, vertical axis is MSE, horizontal axis is the tuning parameter (λ). Figure 3.1 is variable selection process, whereas Figure 3.2 is comparison between training and test sets.

For better understanding of the model fitting, Figure 3 is presented. It relates to the model fitting process using lasso method for the Days to Maturity (DMA). Figures 3.1 showed that some predictors change their directions because of an entrance of other predictors. Moreover, we can observe that the mean square errors (MSE) in Figure 3.1 and in Figure 3.2 for the test set increase on average as model complexity increases, whereas the MSE for training decreases monotonically as the model becomes more complex. The parsimonious model is selected at about lambda 0.11, where the The MSE has minimum value. It should be noted that Figure 3 is given as a sample for this response only, but for the others the graphs are not presented as they are similar.

Model assumptions were checked after model fitting. It is revealed from Table 2 of the normality test for the complete (original) data, and revealed that the residuals find from regression models fitted for DHE are normality distributed, whereas for DMA are not normally distributed, all at the 5% significance level. For the test data set, the residuals for DHE are normally distributed, while those of DMA 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 2: Results for normality, homogeneity of variance and Goodness of fit test (GOF) tests, for Ordinary MLR model using the original data, and all the shrinkage based MLR models using test data. Normality and Homogeneity of variance tests are based on Shapiro-Wilk and white test, respectively.

Variables	Test(P-value) using Ordinary MLR Models using original data set		
	Normality test	White test	GOF test
DHE	0.098*	0.509*	0.405*
DMA	0.0001	0.628*	0.676*
	Test(p-value) using Shrinkage based Models using test data set		
	White test	GOF test	
DHE(Lasso)	0.917*	0.352*	
DMA(Lasso)	0.640*	0.411*	

DHE=Days to heading, DMA=Days to maturity. Tests with * showed that error terms are identically and independently normally distributed, have constant variance and the model has no lack of fit at 5% level of significance. In both 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.

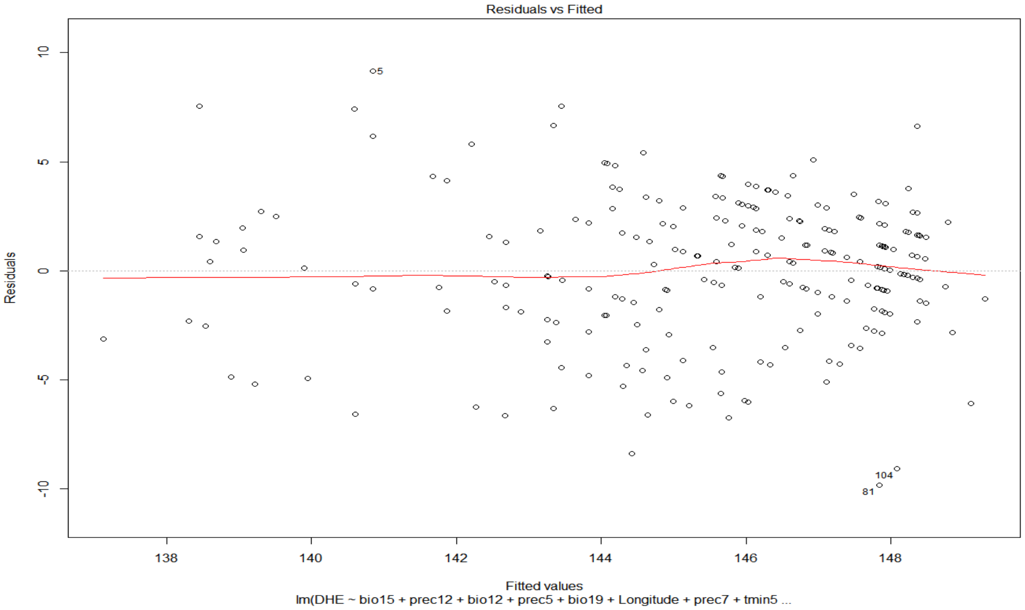

Figure 4: Plot of residual versus predicted values for the complete data set of Days to Heading, for the ordinary MLR model. The most extreme observations are labeled with the row numbers of the data in the data set. Figure 4 shows that observations 5 (from Iraq), 81 and 104 (from Turkey), are identified as outliers. The red line is a smoothed high order polynomial curve to provide us some suggestion on the pattern of residual movement in order to assess the linearity. In this case, we can observe that there is no that much visible deviation from the linearity. Note that Figure 4, which is related to the ordinary MLR model on DHE. Moreover, it is observed that there seems some deviations from the linearity for both Days to Heading and Days to Maturity.

Inference Post Model Selection: For the shrinkage methods, for both DHE and DMA, 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. Based on the WLS estimation method, it is noticeable that prec12 and tmin6 have positive significant effect, while bio15 and tmin5 have negative significant effect on the Days to Heading of the plant. As prec12 and tmin6 increase by a unit measure making constant the other predictors in the model, the mean value of days to heading of the plants increases by 0.11988 and 0.09643 days, respectively. On the other hand, as bio15 and tmin5 increase by a unit measure, in average Days to Heading decreases by 0.15027 and 0.12427 days, respectively. Based on the OLS estimation method, bio15 and tmin5 have negative significant effects whereas longitude and prec12 have positive significant effects on the Days to Heading. On Days to Maturity, tmax8 and bio18 have positive significant effect, whereas bio14 has negative significant effect based on WLS estimation method. From the OLS estimation method, tmax8 has positive, while bio14 has negative effect on Days to Maturity.

Table3. Estimates for OLS and WLS estimation methods of ordinary MLR models, for both the responses using complete data set

Days to Heading(DHE)						
	OLS estimation method			WLS estimation method		
Variable	Par.Est	Std.Er	P-value	Par.Est	Std.Er	P-value

Intercept	151.83868	3.31052	<.0001*	153.08083	3.13188	<.0001*
bio15	-0.13693	0.03593	0.0002*	-0.15027	0.03539	<.0001*
prec12	0.12734	0.03866	0.0011*	0.11988	0.04081	0.0036*
bio12	-0.00581	0.00910	0.5239	-0.01029	0.00910	0.2591
prec5	0.01382	0.03750	0.7129	0.01263	0.03569	0.7237
bio19	-0.02417	0.01548	0.1198	-0.01479	0.01576	0.3491
Longitude	0.07246	0.03207	0.0248*	0.06547	0.03417	0.0566
prec7	-0.08729	0.06016	0.1482	-0.07947	0.05842	0.1751
tmin5	-0.09472	0.04421	0.0332*	-0.12427	0.04347	0.0047*
prec3	-0.03110	0.03595	0.3880	-0.02612	0.03872	0.5005
tmin6	0.07192	0.03813	0.0605	0.09643	0.03753	0.0108*
Days to Maturity(DMA)						
Intercept	172.02020	2.49376	<.0001*	171.77239	2.16838	<.0001*
prec6	0.06455	0.04407	0.1443	0.03576	0.03782	0.3454
prec9	-0.01394	0.03911	0.7218	-0.05703	0.03609	0.1155
tmax8	0.02655	0.01329	0.0470*	0.02250	0.01092	0.0405*
bio14	-0.23068	0.08468	0.0069*	-0.28157	0.07392	0.0002*
bio18	0.06898	0.03943	0.0815	0.11396	0.03603	0.0018*
bio9	-0.01623	0.01578	0.3049	-0.00996	0.01355	0.4632

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), P-values indicated by * are significant at 5% level of significance. Par.Est=Parameter estimate, Std.Er= Standard Error.

Besides, to evaluate the predictability of these models, see Figure 5 for both WLS and OLS estimation methods for Days to Heading. It is noticed that there is some variability in the residuals. Although the predicted value continuously increases as a function of the Days to Heading, the variability seems to need some concerns. From this Figure our model seems to have two subsections of performance. The first one is where actual values between about 130 and 145. within this zone, the variability seems to be higher, while prediction may be low. The second one is when actual values between 145 and 155, and within this zone variability may be lower comparing to the first case, and then model's predictability might be better.

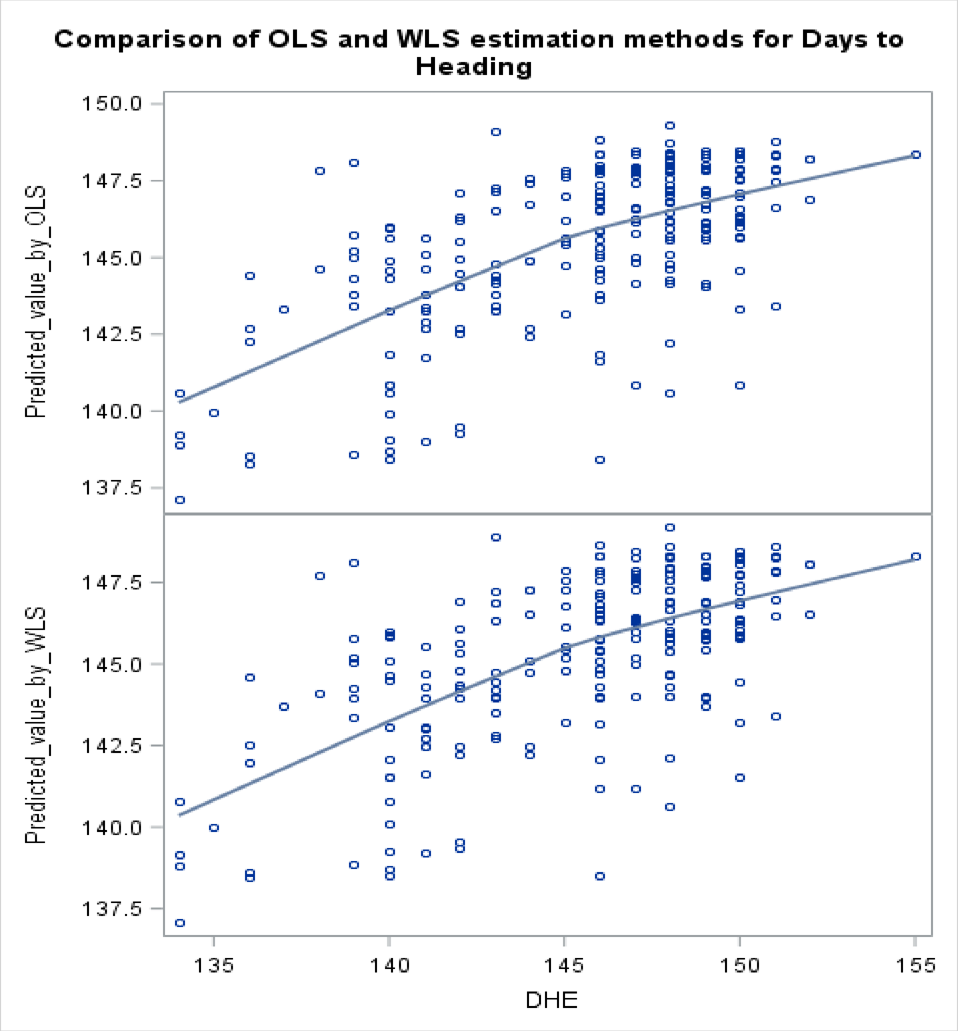

Figure 5: Predicted versus actual value for Days to Heading. Horizontal axis is actual value and vertical axis is predicted value from both WLS and OLS estimation methods for the complete data set.

Furthermore, as the predictive Figures shown, the WLS methods seem slightly to perform better prediction than the OLS methods. 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 DHE. From the WLS estimation method, prec1, prec11 and tmin10 have increasing significant effect, while bio8, bio15 and prec10 have decreasing significant effect on the Days to Heading. Making fixed other predictors within the model, a unit increase on prec1, prec11 and tmin10, the mean value of Days to Heading increases by 0.137, 0.097 and 0.152 days, respectively. Differently, the mean value of Days to Heading decreases by 0.025, 0.313 and 0.214 days as a unit increase in bio8, bio15 and prec10, respectively. Based on OLS method, bio15 and prec10 have decreasing significant effect. Days to Heading decreases by about 0.274 and 0.197 days as bio15 and prec1o showed a unit increase, respectively.

Table 4: Parameter Estimates from OLS and WLS estimation methods in MLR models with the predictors selected by lasso, for DHE using test data set.

Days to Heading(DHE)							
Effect	Pen.Est	OLS estimation method			WLS estimation method		
		Par.Est	Std.Er	P-value	Par.Est	Std.Er	P-value
Intercept	161.91713	171.32559	23.80617	<.0001*	158.53771	20.99306	<.0001*
Longitude	0.022710	0.04292	0.06524	0.5135	0.02579	0.06450	0.6909
bio3	-0.191243	0.48053	0.67458	0.4794	0.84005	0.62064	0.1817
bio8	0.007160	-0.01978	0.01537	0.2035	-0.02503	0.01136	0.0320*
bio9	-0.038843	-0.07145	0.12962	0.5838	-0.07521	0.10469	0.4758
bio15	-0.085114	-0.27406	0.07134	0.0003*	-0.31320	0.06780	<.0001*
prec1	-0.024071	0.08982	0.07311	0.2247	0.13752	0.06808	0.0486*
prec2	-0.029153	0.03076	0.05484	0.5772	-0.03616	0.05458	0.5106
prec3	-0.029588	-0.10133	0.05530	0.0725	-0.05166	0.05465	0.3489
prec7	-0.031512	-0.16280	0.12258	0.1898	-0.24869	0.12512	0.0521
prec9	-0.018249	-0.09091	0.14495	0.5332	-0.02098	0.13324	0.8755
prec10	-0.009609	-0.19705	0.06278	0.0028*	-0.21366	0.06073	0.0009*
prec11	-0.013724	0.07749	0.05641	0.1753	0.09711	0.04532	0.0368*
prec12	0.084874	-0.01333	0.05845	0.8205	-0.03935	0.05347	0.4651
tmin5	-0.034460	0.07267	0.13406	0.5900	0.04670	0.11241	0.6795
tmin7	0.047139	0.08428	0.11005	0.4472	0.08432	0.09397	0.3737
tmin10	-0.016990	0.08010	0.07340	0.2801	0.15225	0.06245	0.0182*
tmax1	0.018678	-0.07661	0.07925	0.3381	-0.12334	0.07782	0.1191
tmax5	0	-0.11930	0.12573	0.3470	-0.10763	0.10488	0.3095

bio8= Mean Temperature of Wettest Quarter, P-values indicated by * are significant at 5% level of significance. Pen.Est=Penalized coefficient estimates, Par.Est=Parameter estimate

For Days to Maturity (Table 5) from the WLS method, prec11 and tmax3 have increasing significant effects, while tmax12 has decreasing significant effect. From the OLS estimation method observed that the prec11 and longitude have positive significant effect, whereas tmax12 and bio8 have negative significant effects. Using WLS method, holding constant the other predictors within the models, a unit increase in prec11 and tmax3, the number of Days to Maturity increases by 0.214 and 0.300, respectively, while a unit increment in tmax12 results in a decrease by 0.533 units in Days to Maturity. As per the OLS method, a one unit increment on each prec11 and longitude, it shows an increment by 0.175 and 0.260 days respectively, on the Days to Maturity. Whereas a unit increase in bio8 and tmax12, resulted in a decrement on the Days to Maturity by 0.05 and 0.46 days, respectively.

Table 5: Parameter Estimates from OLS and WLS estimation methods in MLR model with the predictors selected by lasso, for DMA using test data set.

Days to Maturity(DMA)							
Intercept	193.568035	263.54831	35.47217	<.0001*	216.95360	56.30371	0.0004*
Longitude	0.011990	0.25948	0.11874	0.0330*	0.37873	0.21507	0.0865
Latitude	-0.230676	-1.24861	0.73622	0.0953	0.07897	1.08401	0.9423
bio3	-0.347769	-1.02314	0.65161	0.1219	-1.12271	0.83599	0.1875
bio7	-0.000166	-0.22187	0.11686	0.0627	-0.13515	0.22662	0.5546
bio8	0.004902	-0.05015	0.02141	0.0227*	-0.02947	0.03070	0.3434
bio14	-0.286960	-0.56957	0.34311	0.1024	-0.89000	0.52256	0.0969
bio15	-0.021809	0.03390	0.13067	0.7963	0.14818	0.15686	0.3510
bio16	-0.001836	0.00883	0.05973	0.8830	-0.03110	0.06749	0.6476
bio18	0.056089	0.30091	0.18264	0.1049	0.50567	0.28249	0.0816
prec1	0.001517	-0.06810	0.10156	0.5052	-0.14258	0.15022	0.3487
prec2	-0.013558	0.03612	0.11291	0.7502	0.12270	0.12185	0.3205
prec3	-0.018030	-0.11676	0.07747	0.1373	-0.13668	0.08905	0.1333
prec6	0.021903	0.14676	0.24179	0.5463	-0.35663	0.38318	0.3580
prec7	0.123920	-0.25564	0.32801	0.4390	-0.28429	0.43566	0.5181
prec10	-0.008705	-0.13467	0.08336	0.1117	0.00234	0.11392	0.9837
prec11	0.017590	0.17509	0.08270	0.0386*	0.21395	0.09029	0.0231*
prec12	0.026439	0.00661	0.07746	0.9323	0.08323	0.08929	0.3573
tmin5	-0.041847	-0.01374	0.08862	0.8773	-0.10780	0.11010	0.3339
tmin7	-0.031242	-0.02880	0.13140	0.8273	-0.12690	0.19178	0.5123
tmin10	-0.020572	-0.10522	0.12268	0.3947	-0.00390	0.19817	0.9844
tmax1	0.060931	0.25261	0.15948	0.1187	0.30826	0.23108	0.1904
tmax3	0.018064	0.14905	0.10365	0.1559	0.30030	0.14097	0.0399*
tmax6	0.013713	-0.06106	0.11107	0.5846	-0.10530	0.12521	0.4058
tmax7	-0.041863	0.16162	0.15099	0.2889	0.09066	0.19939	0.6520
tmax9	0.084502	0.16837	0.10901	0.1280	0.16070	0.16856	0.3466
tmax12	-0.049345	-0.45946	0.16730	0.0081*	-0.53323	0.22980	0.0259*

P-values indicated by * are significant at 5% level of significance. Pen.Est=Penalized coefficient estimates, Par.Est=Parameter estimate

The penalized coefficient estimates are presented in Tables 4 and 5 for In most of the parameters(penalized coefficient estimates, tables 4&5), there 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.

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.

Conclusion

The WLS estimation methods of shrinkage based models revealed that precipitations of January and November, and October minimum temperature have increasing significant effect, while bio8, bio15 and October precipitation (prec10) have decreasing significant effect on the Days to Heading. From WLS method, Precipitation of November (prec11) and maximum temperature of March (tmax3) have increasing, while maximum temperature of December (tmax12) has decreasing significant effects on Days to Maturity of the durum wheat. From the OLS method observed that the precipitation of November (prec11) and longitude have increasing significant effect, whereas December maximum temperature (tmax12) and bio8 have decreasing significant effects on the Days to Maturity. The ordinary MLR models on Days to Heading seemed 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 Days to Maturity also showed that predicting using these models is not trustful. From models with predictors selected by shrinkage methods, for the Days to Heading showed that there seems sensible prediction as their predicted value increase continuously as a function of the actual values, but we should also noted that there is sounding variability which may make the prediction uncertain.

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 Days to Heading, 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 method

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.

[4] Del Moral, L. F., Rharrabti, Y., Villegas, D., & Royo, C. (2003). Evaluation of Grain Yield and Its Components in Durum Wheat under Mediterranean Conditions Funding for this study was provided by the Spanish government throughout INIA Project SC97-039-C2 and CICYT Project AGF99-0611-CO3. 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

[7] Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, pp. 337-387). New York: Springer series in statistics.

[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 R. In Proceedings of the SAS Global Forum 2015 Conference. Cary, NC: SAS Institute Inc. http://support. sas. com/rnd/app/stat/papers/2015/PenalizedRegression_LinearModels. pdf.

[10] Kabbaj, H., Sall, A. T., Al-Abdallat, A., Geleta, M., Amri, A., Filali-Maltouf, A., ... & Bassi, F. M. (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 J. Plant Sci, 2(3), 254-256.

[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., Maalouf, F., ... & Royo, C. (2010). 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(335.344).

[16] Nishida, K. (2017). Skewing Methods for Variance-Stabilizing Local Linear Regression Estimation. arXiv preprint arXiv:1704.04356.

[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., Hanson, J., Dulloo, M. E., Ghosh, K., & Nowell, A. (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 (Methodological), 267-288.

[21] Tibshirani, R., Wainwright, M., & Hastie, T. (2015). Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC.

[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., & Hodkin, T. (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 (Doctoral dissertation, International Environmental Modelling and Software Society (iEMSs)).

[25] Yaffee, R. A. (2002). Robust regression analysis: some popular statistical package options. ITS Statistics, Social Science and Mapping Group, New York State University, downloaded on Dec, 23, 2009.

[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.