UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics

Masterthesis

Quantifying the relationship between adaptive traits and agro-climatic conditions

Mehari Gebre Teklezgi

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :
Prof. dr. Olivier THAS

SUPERVISOR :
Mr. Zakaria KEHEL

2017
2018

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics

Masterthesis

Quantifying the relationship between adaptive traits and agro-climatic conditions

Mehari Gebre Teklezgi

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Olivier THAS

SUPERVISOR :

Mr. Zakaria KEHEL

Contents

List of Figures

List of Tables

Abstract

Durum wheat is an economically important and regularly eaten food for billions of people in the world. Consequently, wheat breeders over the past century have increased the productivity and adaptability via strong selection applied to genes controlling agronomical important traits. In the International Center for Agriculture Research in 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 as well as the association between these traits and those of the different agro-climatic characteristics.

Ordinary multiple linear regression with stepwise variable selection method, 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 which were chosen from the International Center for Agriculture Research in the Dry Areas (ICARDA) genebanks. Each of the models are fitted on Days to Heading, Days to Maturity, Plant Height, Grain Weight and Thousand Kernel Weight response variables with 57 predictor variables, independently. The penalized based models used data splitting into training on which the model is fitted and test data set on which the fitted model is validated. Ordinary least square and weighted least square estimation methods are also used for parameter estimation and prediction of post model selection.

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 Days to Heading, Days to Maturity, Plant Height and Grain Weight using both ordinary and shrinkage based models. Longitude affects significantly the Thousand Kernel Weight using the ordinary MLR model, However, there is no significant predictor which affects the Thousand Kernel Weight from the shrinkage based MLR models. But longitude affects it significantly using the ordinary MLR model. It is revealed that 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. Furthermore, for the Days to Heading and Grain Weight showed that there seems better prediction as their predicted values increase continuously as a function of the actual values though there is considerable variability. However, the Lasso based model used for Thousand Kernel Weight is not predicting well.

In conclusion, inferences and predictions by the ordinary MLR models are not trusted due to the presence of multicollinearity in the model fitting, and violation of some model assumptions after model fitting. 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 Plant Height, Days to Heading and grain Weight.

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

Acknowledgements

First and foremost, I would like to thank to the Almighty God for giving me the strength, knowledge, ability and opportunity to undertake this research study, and to persevere and complete it. Without his care and blessings, this achievement would not have been possible. I will always praise you.

I would like to express my deepest gratitude to my advisor **Prof. dr. Olivier THAS** for his great supervision, suggestions and understanding throughout this process. I am grateful to have had you as a supervisor, as you were always very supportive and tried to create time to discuss my numerous questions.

I would also like to give my warmest thank to my advisor **Mr. Zakaria KEHEL** for his great supervision and suggestions throughout this process. I really apperciate you for all your clear explanations of my questions.

I would like to thank all the professors at CenStat as this thesis would not be accomplished without their teaching and guidance during the past two years.

My acknowledgement also goes to the VLIR-UOS scholarship that gave me the chance to join to Hasselt University, and having magnificent experiences.

1 Introduction

1.1 Background

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]. Durum wheat (Triticum durum) 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 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 abiotic 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].

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 require. Thus, increasingly breeders will be forced to seek the variation they require from genetic resource collections conserved in geenbanks. 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].

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 genbank 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 have 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]. 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, [5] 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.

A variance study for Grain Yield and yield components held by [8] 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. Another study was carried out from 13 Mar, 2007 through 12 May, 2009 at the University of Arizona Maricopa Agricultural Center, Maricopa by [17], and suggested that promising increases in overall temperature have a negative effect on spring durum wheat yield. Moreover, a field study was carried out on the tolerance of durum wheat to high temperatures using analysis of variance at Elvas, Portuguese by [13], and stated that Grain yield and individual grain weight were considerably affected by temperature increase.

A study by [14] evaluated phonological traits of durum wheat such as Days to Heading and Plant Height in highly different rainfall conditions in Mediterranean countries (Italy, Morocco, Spain, Syria, and Tunisia), and others. 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 Days to Heading, Plant Height and Thousand Kernel Weight.

1.2 Objectives of the study

- Main objective of the study

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 different traits as well as the association between these traits and those of the different agro-climatic characteristics.

- Specific objectives

Five different specific objectives will be addressed in this study:-

1. 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.
2. To investigate the predictive value of the agro-climatic variables on the future observations of Day to Maturity of durum wheat landraces, and to study their association.
3. To assess the predictive value of the agro-climatic variables on the future observation of Plant Height of the durum wheat landraces, and to study their association.
4. To examine the predictive value of the agro-climatic variables on the future observation of Thousand Kernel Weight of thedurum wheat landraces, and to study their association.
5. To examine 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.

1.3 Data description

238 durum wheat landraces were chosen from the International Center for Agricultural Research in the Dry Areas (ICARDA) genebanks. The landraces were 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 5 different response variables (Table 1). 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. 3. Plant Height (PHT): is the height of the plant from ground to top of spike measured in centimeter, excluding awns. 4. Thousand Kernel Weight (TKW): which is the weight in grams of 1000 well-developed whole grains, dried to 13% moisture content. 5. Grain Weight (GRY): is weight of grains that was harvested, and registered on a scale of kilogram per hectar.

In this study, 57 environmental variables including geographic coordinates: longitude and latitude were used (appendix-I, Table 10). 36 out of the 55 are monthly long term averages

Table 1: Summary of the dependent variables used in this study.

Existed Variable Name	Description	Data type	Unit
DHE	Days to heading	continuous	days
DME	Days to Maturity	continuous	days
PHT	Plant Height	continuous	cm
TKW	Thousand kernel weight	continuous	gram
GRY	Grain yield	continuous	Kg per hectar

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).

2 Methodology

In this section, the different methodologies used to adress our objectives are presented. In order to achieve our study goals, different statistical methods are used. Firstly, exploratory data analysis was used in order to get insight in to and explore the data. Secondly, several statistical methods were used to fit and select parsimonious models which adequately describe the data. Ordinary multiple linear regression models were first used to quantify the relationship between the adaptive traits and the agro-climatic variables. Penalized regression methods, such as Lasso method and Elastic net methods, are used as variable selection and estimation methods, aiming at finding good prediction models.

2.1 Exploratory Data Analysis

This section describes the statistical techniques which are used in data exploration. Summary statistics of the five response variables are reported. Several exploratory plots were created to explore the correlations between the predictor variables, and Variance Inflation Factors(VIF) were computed in order to better understand issues related to multicollinearity.

2.2 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:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i \tag{1}$$

Where, n and p are the number of observations and the number of predictors, respectively. Y_i is the response for the i^{th} observation(i=1 , 2, 3, ...238). X_{ij} is the j^{th} predictor for the i^{th} observation, β_0 is the intercept. β_j is the effect parameter of the j^{th} predictor. ϵ_i are independent and identically normally distributed with mean 0 and constant variance σ^2. This model is applied for the five response variables (Days to Healing, Days to Maturity, Plant Height, Thousand Kernel Weight and Grain Weight), independently.

In a multiple linear regression model, as with all statistical models, 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 (as we will see in the result section), 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 regression [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 later).

2.3 Penalized Regression Methods:

Statistical model selection process based on shrinkage methods works 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,

$$MSE_j(\lambda) = 1/n_k \sum_{i\varepsilon jthpart} (Y_i - \hat{Y}_i^{-k}(\lambda))^2 \tag{2}$$

where $\hat{y_i}^{-k}$ is the predicted value from the fitted model without the observations in the k^{th} left out part, and n_k is the number of observations in the k^{th} group. Finally, the CV estimate of the MSE is computed as

$$MSECV(\lambda) = 1/k \sum_{i=1}^{k} MSE_j(\lambda) \tag{3}$$

This is done for many values of λ and choose 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 are 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: Understanding the bias-variance trade-off is very important in understanding the added value of penalized regression for prediction purposes. The bias-variance trade-off indicates the exchange of bias and variance, i.e by introducing bias in to the OLS estimators, the variance may reduce substantially. The MSE of a model is the sum of the variance of the predictions and the squared bias [7], and it is given by:

$$MSE = E[(y - \hat{y})2] = Var(y) + Var(\hat{y}) + (E(\hat{y}) - E(y))2 = \sigma^2 + Var(\hat{y}) + bias(\hat{y})^2 \tag{4}$$

Where y and $\hat{y}$ are the actual and predicted responses. However, as σ^2 is an uncontrollable error, that does not depend on the models to be evaluated with the MSE. Hence, it can be ignored in understanding the importance of the bias-variance trade-off for prediction models.

2.3.1 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

$$\hat{\beta}^{lasso} = argmin_\beta \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5}$$

Where $\lambda \geq 0$.
In this penalized 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 $\lambda = 0$, 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.

2.3.2 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

$$\hat{\beta} = argmin_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \tag{6}$$

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].

2.4 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.

2.4.1 Ordinary Least Square (OLS)

Ordinary least squares is 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:

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2 \tag{7}$$

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

2.4.2 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

$$\sum_{i=1}^{n} W_i(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2 \tag{8}$$

where W_i is the weight of the i^{th} observation. WLS residuals are given by $\sqrt{W_i}(Y_i - \hat{Y}_i)$ where $W_i = 1/\sigma_i^2$ with σ_i^2 is the error variance for observation i.

The error variance is calculated as follow. Firstly, residuals (e_i) are calculated, and then a model with the response variable squared residual(e_i^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.

2.5 Software

Some exploratory plots were done using R version 3.4.3. However, all model fittings and post model selection data analysis were performed using SAS version 9.4.

3 Results and Discussion

In this section, results from the exploratory data analysis, model fittings and findings from the final selected models, are presented. For the exploration, tables with summary statistics of the five response variables were given so as to explore the nature of the data. A heat map of the correlation among all predictors was constructed in order to reveal the co-linearity among them. Additionally, tabular as well as histogram representations were used to assess the variance inflation factor in order to understand the multicollinearity. Different models fitted with ordinary and shrinkage regression methods are constructed and compared so as to select the best fitted model in terms of predictive power. This is followed by checking the model assumptions. Finally, the best selected models are estimated using ordinary least square estimation method. Besides, weighted least square estimation method is used in order to find more efficient estimates.

3.1 Exploratory Data Analysis

Summary statistics of the five response variables are presented (Table 2). It is revealed that the total number of observation is 238 for all the variables, with no missing data. The variability among the measurements of Days to Maturity (DMA) is smaller as compared to the other variables, whereas that of the Grain Weight (GRY) is higher. For Days to Heading (DHE) and Days to Maturity(DMA), there is 21 and 25 days respectively between the earliest and latest accession. For Plant Height (PHT), the tallest accession has almost double height of the shortest one. Similar pattern can be observed for Thousand Kernel Weight (TKW). The Grain Weight (GRY) 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.

Table 2: Summary statistics for the five response variables.

Variable	N	Sum	Mean	Std Dev	Minimum	Maximum
DHE	238	34635.00	145.53	4.17	134.00	155.00
DMA	238	42490.00	178.53	3.06	172.00	197.00
PHT	238	23075.00	96.95	11.52	61.00	118.00
GRY	238	8071.70	33.91	4.59	21.80	44.50
TKW	238	423241.55	1778.33	609.53	479.78	3285.11

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 with few predictors that are not assigned to any of these clusters. 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,

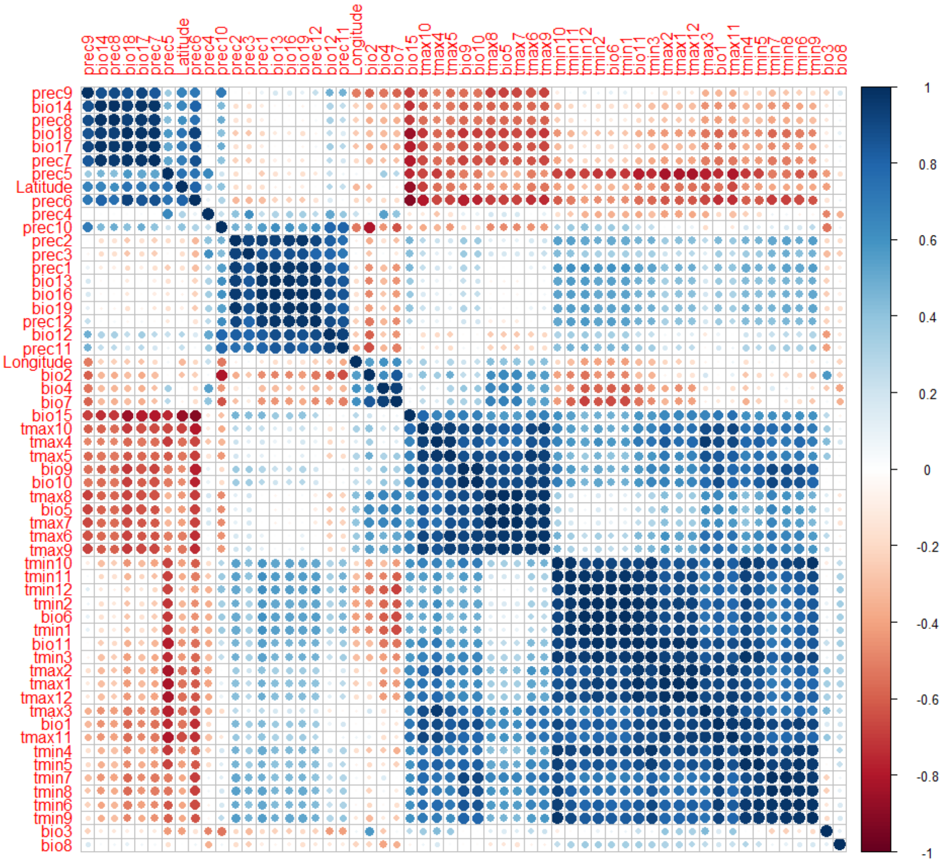


Figure 1: Heat map of the correlations between all the 57 predictors. The red color indicates the par-wise negative correlation whereas the blue color indicates pair-wise positive correlation. The white color is for no correlation.

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 have 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. See appendix-I, Table 10 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 used in this study should certainly be methods that work well in the presence of multicollinearity.

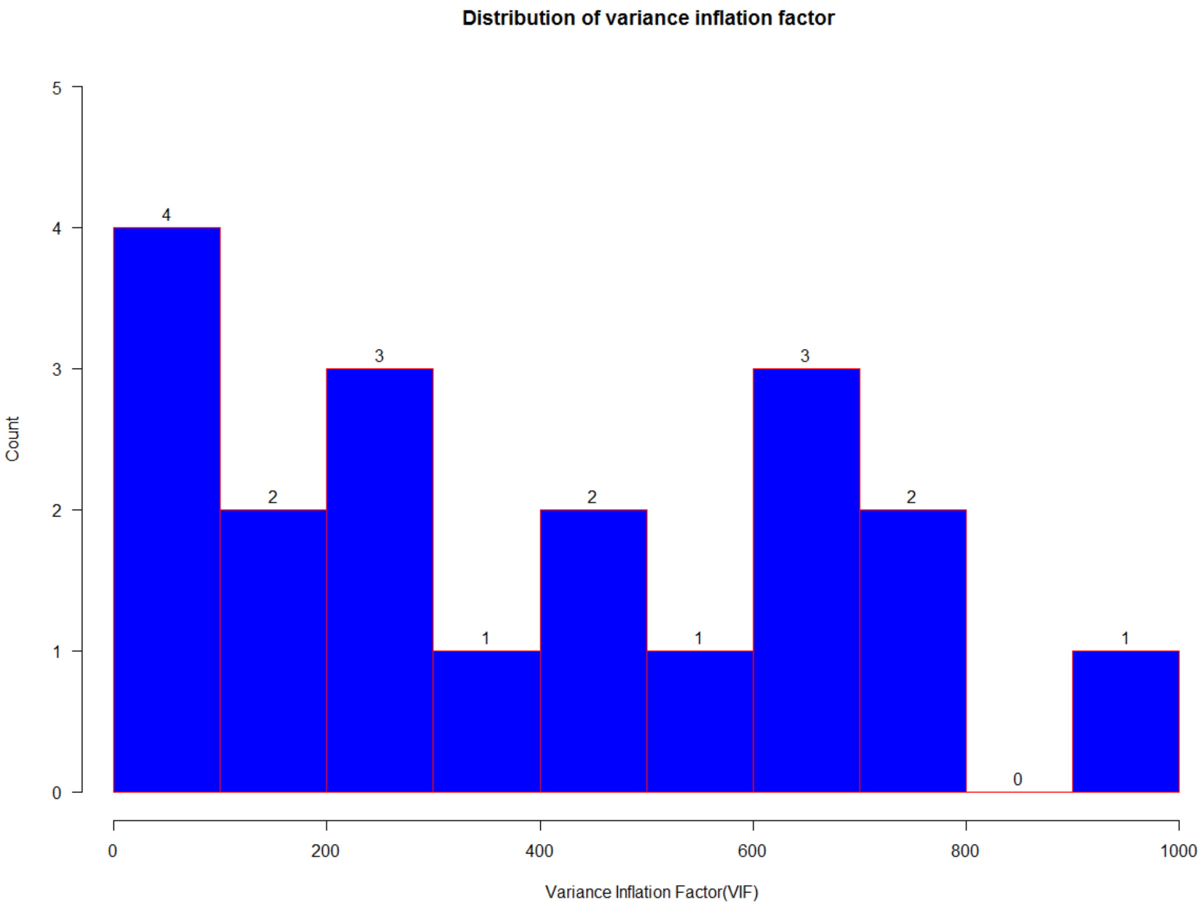


Figure 2: Histogram of Variance Inflation Factors (only the $VIF \leq 1000$ of 19 predictors are shown). The numbers on each bar are the number of predictors those their VIF are within the interval.

3.2 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 criteria. The respective fitted models are given in Table 3 with their respective RMSE, and Table 5 with all selected 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 3), four different partitions were used for each response, and Lasso and Elastic net methods were applied for each partition. Within each partition, root mean square errors (RMSEs) were presented for all the models. Based on this, the partitions in bold letter were selected for

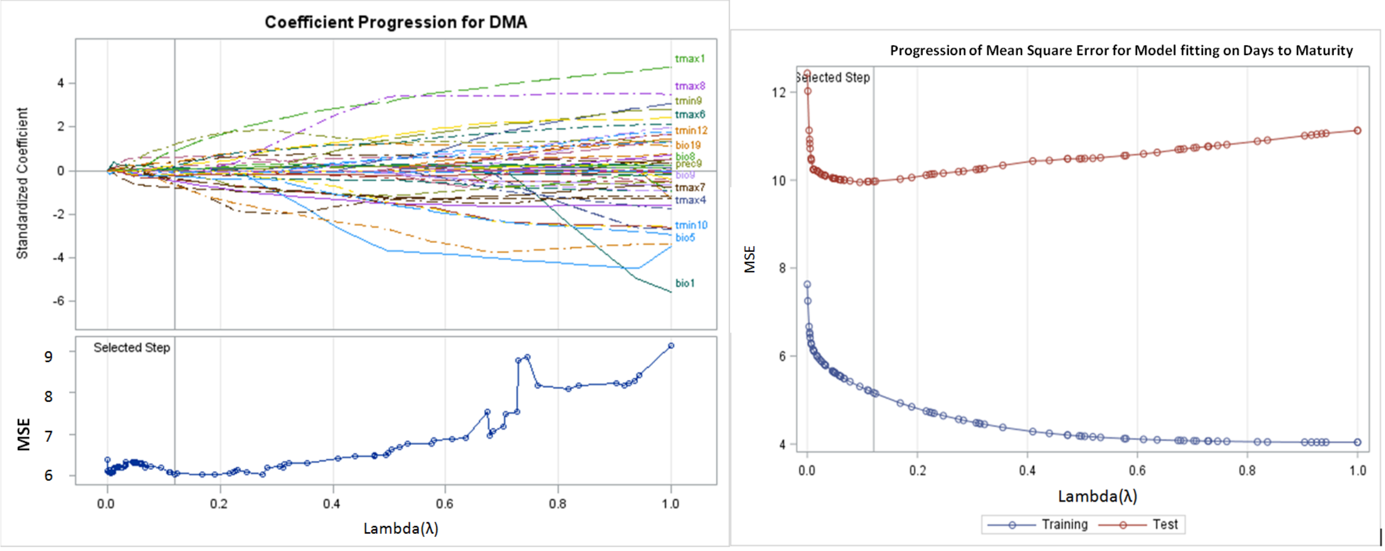
Table 3: Comparison of partitions for the shrinkage based MLR models in order to select the best partition which gives the optimal models, and comparison of predictive performance of all the three MLR models, based on RMSE.

Response	Partition	Model	RMSE
	20-80	Lasso	3.323
		Enet	3.323
DHE	**30-70**	**Lasso**	**3.159**
		Enet	**3.158**
	35-65	Lasso	3.538
		Enet	3.538
	40-60	Lasso	3.310
		Enet	3.310
	-	Ordinary MLR Model	3.290
	20-80	Lasso	2.781
		Enet	2.787
DMA	30-70	Lasso	3.010
		Enet	3.010
	35-65	**Lasso**	**2.506**
		Enet	**2.501**
	40-60	Lasso	2.904
		Enet	2.904
	-	Ordinary MLR Model	2.719
	20-80	**Lasso**	**10.369**
		Enet	**10.418**
PHT	30-70	Lasso	10.530
		Enet	10.530
	35-65	Lasso	10.654
		Enet	10.649
	40-60	Lasso	10.730
		Enet	10.730
	-	Ordinary MLR Model	10.897
	20-80	Lasso	624.326
		Enet	624.326
GRY	30-70	Lasso	579.170
		Enet	579.170
	35-65	Lasso	631.381
		Enet	631.381
	40-60	**Lasso**	**568.420**
		Enet	**568.420**
	-	Ordinary MLR Model	583.052
	20-80	**Lasso**	**4.456**
		Enet	**4.456**
TKW	30-70	Lasso	4.571
		Enet	4.571
	35-65	Lasso	4.605
		Enet	4.605
	40-60	Lasso	4.542
		Enet	4.542
	-	Ordinary MLR Model	4.516

MSE=MSECV= mean square error based on cross-validation
The selected partitions, and respective methods are in **bold** letters.

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 (Tables 6-8), for each response.

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.



(a) fig 3.1

(b) 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 the selection process, whereas Figure 3.2 is comparision between training and test sets.

Model assumptions were checked after model fitting. It is revealed from Table 4 of the normality test for the complete (original) data, and shown that the residuals found from the regression models fitted for DHE, GRY and TKW are normality distributed, whereas for DMA and PHT are not normally distributed, all at the 5% significance level. It should be noted that the normality assumption is needed only for the ordinary MLR models.

Table 4: 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.

Ordinary MLR Models using original data set						
Test(P-value)	**DHE**	**DMA**	**PHT**	**GRY**	**TKW**	
Normality(P-val)	0.098*	0.0001	0.0005	0.055*	0.078*	
White(P-val)	0.509*	0.628*	0.118*	0.214*	0.878*	
GOF(P-val)	0.405*	0.676*	0.297*	0.689*	0.194*	
Shrinkage methods based MLR Models using test data set						
	Lasso(DHE)	**Lasso(DMA)**	**Lasso(PHT)**	**Enet(PHT)**	**Lasso(GRY)**	**Lasso(TKW)**
White(P-val)	0.917*	0.641*	0.466*	0.953*	0.461*	0.144*
GOF(P-val)	0.352*	0.411*	0.576*	0.574*	0.135*	0.039

Tests with * showed the data is identically and independently normally distributed, have constant varinace and the model has no lack of fit at 5% level of significance.

In all the ordinary MLR models and MLR models whose predictors selected by shrinkage methods of all the data sets, the homogeneity of variance test showed that there is no evidence that shows heteroskedasticity at 5% level of significance. Results of the goodness of fit test for the ordinary MLR models based on the original data set indicated no evidence which shows model lack of fit. However, for the MLR models with predictors selected by the shrinkage methods, the model for TKW showed evidecne of lack of fit. This may happen due to the reason that the relationship between the response and predictor variables is not linear, see appendix-I, figure 11.2 as well as the effect of outliers.

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, is taken as a sample. For the other ordinary MLR models, results are given appendix-I (Figures 8 -9). Observations 26, 68 and 125 from Turkey; observations 165 (Turkey),177 and 179 (Greece); observations 88 and 168 (Turkey) and 192 (Jordan), and observations 103 (Turkey) and 192 (Jordan) are identified as an outliers for DMA, PHT, GRY and TKW, respectively. Unlike Figure 4, these Figures show no deviations from linearity.

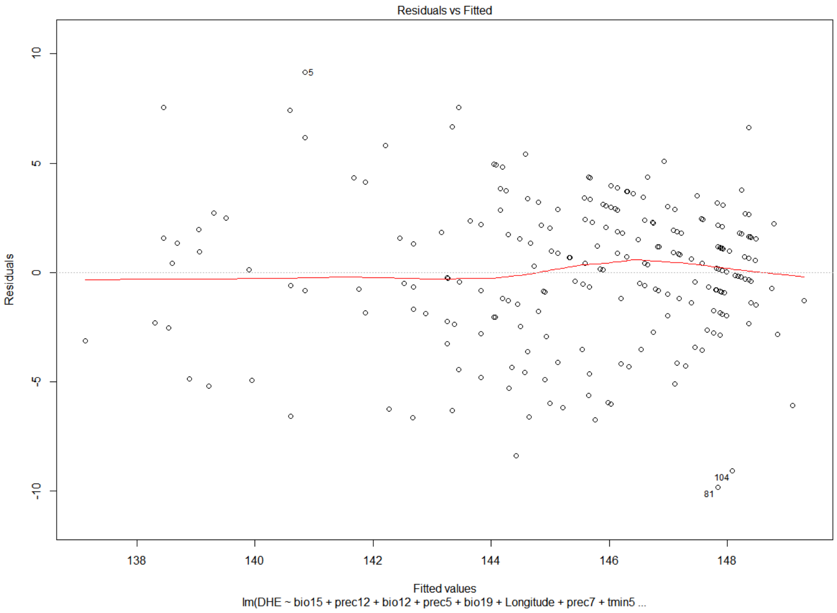


Figure 4: Plot of residuals versus predicted values for the complete data set of Days to Heading, for the ordinary MLR model. The most extreme observations are labeled with the row numbers of the data in the data set.

Moreover, it is observed from Figure 5 that 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. Figure 5 is given as a sample. For the other shrinkage based MLR models using the test data set, graphs are given in appendix-I (Figures 10-11). Observations 8, 19 and 31; observations 12, 19 and 24; observations 43, 45 and 49, and observations 8, 19 and 28 are identified as outliers for DHE, DMA, PHT and TKW, respectively. In many of these figures, it is observed that there seems some deviations from the linearity, especially for Days to Heading, Days to Maturity and Thousand Kernel Weight.

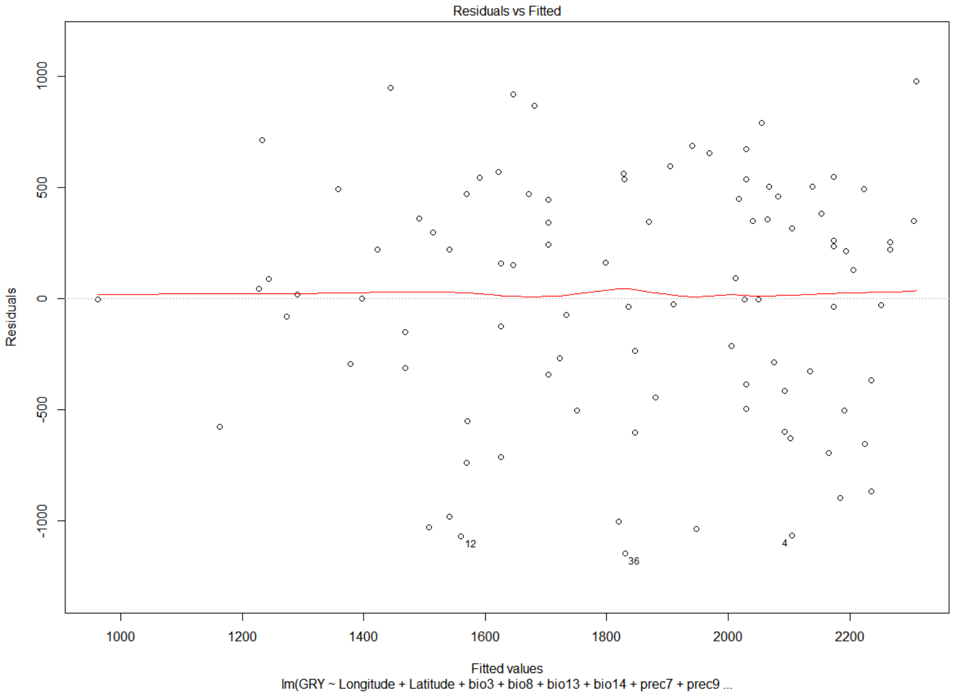


Figure 5: Plot of residual versus predicted values for test data set (40% of the data set) of Grain Weight, Lasso based MLR model. The red line is a smoothed high order polynomial curve to show the pattern of residual movement in order to assess linearity.

3.3 Inference Post Model Selection

For the shrinkage methods, for responses of DHE, DMA, 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 5. Based on the WLS estimation method, it is visible that prec12 and tmin6 have positive significant effect, while bio15 and tmin5 have negative significant effect on the Days to Heading of the plant. Making constant the other predictors in the model, as prec12 and tmin6 increase by a unit measure, the predicted 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, the predicted value of 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 the other hand, using the WLS 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 5: Parameter Estimates from OLS and WLS estimation methods of ordinary MLR models, for all the responses using complete data set.

Days to Heading(DHE)						
	OLS estimation method			WLS estimation method		
Effect	**Par.Es**	**St.Er**	**P-value**	**Par.Es**	**St.Er**	**P-value**
Intercept	151.839	3.310	< .0001*	153.081	3.132	< .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	0.0091	0.5239	-0.01029	0.00910	0.2591
prec5	0.01382	0.0375	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	-0.0311	0.03595	0.388	-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
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*

P-values indicated by * are significant at 5% level of significance.
Par.Es=Parameter Estimate , St.Er=Standard Error

Besides, to evaluate the predictability of these models, see Figure 6 for both WLS and OLS estimation methods for Days to Heading. Note that the Figures for the other response are given appendix-II.

It is noticed from Figure 6 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. Based on appendix-II, Figures 12 and 13 (fig 12.2 and 13.1), the MLR models for responses GRY, as well as PHT seem to have considerable variability. Their predictive values increase as a function of their actual values, but prediction may be questioning due to high variability. Moreover, Figures 12.1 and 13.2 on the Days to Maturity and Thousand Kernel Weight, respectively seem to have also noticeable variability. For Days to Maturity, most of the observations lie in the zone where the actual values between about 175 and 185 in both OLS and WLS methods. Besides, for Thousand Kernel Weight also seems to have high variability in which the data points are far from the diagonal line, and due to this

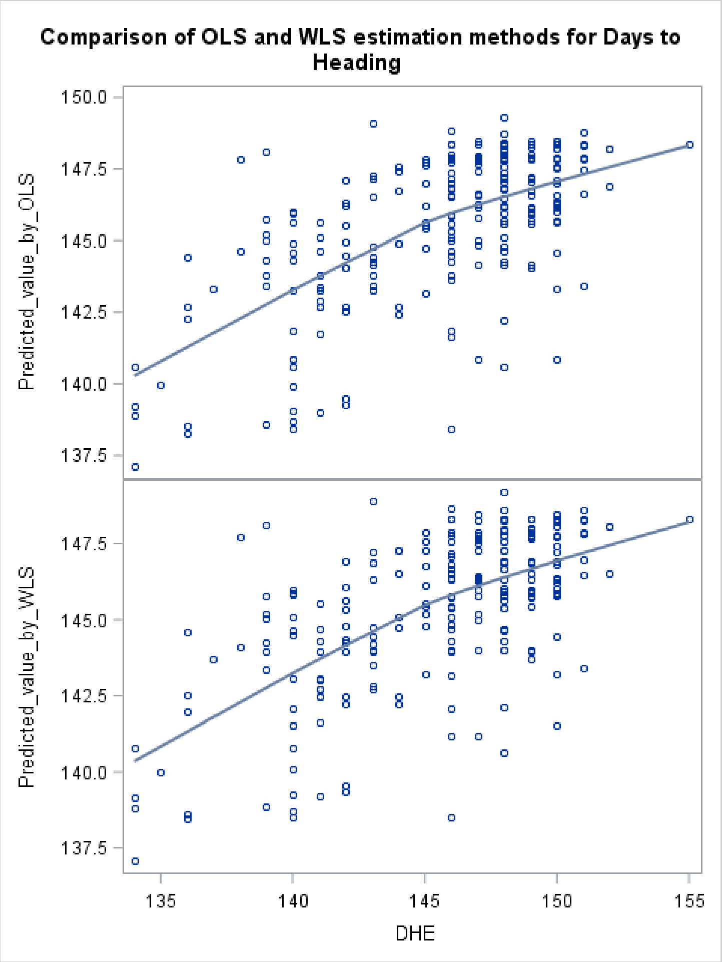


Figure 6: Predicted versus actual value for Days to Heading. Horizontal axis is actual value and vertical axis is predicted value from both WLS and OLS estimation methods for the complete data set using ordinary MLR model.

prediction may not be trustful.

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 like the RMSE given in Table 9. 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 and GRY Table 6. 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. In contrast, 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. 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 increase in bio3 and tmin11 results a decreasing for the grain Weight by 122.892 and 13.986 kg/hectar, 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/hectar as the tmin11 and bio3 increaesed by a unit measure, respectively.

Table 6: Parameter Estimates from OLS and WLS estimation methods in MLR models with the predictors selected by Lasso, for DHE and GRY using test data set.

Days to Heading(DHE)							
		OLS estimation method			WLS estimation method		
Effect	**Pen.Est**	**Par.Es**	**St.Er**	**P-value**	**Par.Es**	**St.Er**	**P-value**
Intercept	161.9171	171.326	23.80617	< .0001∗	158.53771	20.993	< .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	0.056417	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
Grain Weight(GRY)							
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	2.252377	0.1323
bio14	13.07108	60.72441	38.97756	0.12307	58.701857	37.695937	0.1232
prec7	3.415454	-51.4843	34.34114	0.1376	-67.40306	35.79828	0.0632
prec9	-14.2640	-11.3503	15.88892	0.47707	-3.17688	12.90839	0.8062
tmin9	-6.96310	-6.20974	8.50869	0.4675	-14.025477	8.16688	0.0896
tmin11	-0.66160	-19.20187	6.19670	0.002*	-13.98567	5.08390	0.0073*
tmax11	10.15632	19.44664	10.157987	0.0590	24.014147	9.61972	0.0145*

P-values indicated by * are significant at 5% level of significance.

Pen.Est=Penalized coefficient estimates

For Days to Maturity (Table 7) 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 7: 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)							
		OLS estimation method			WLS estimation method		
Effect	**Pen.Est**	**Par.Es**	**St.Er**	**P-value**	**Par.Es**	**St.Er**	**P-value**
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	0.113927	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

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 8. 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 some predictors on the Plant Height implied that the predictors might have no importance in growing the height of the durum wheat or the height of the plant might be shrunken (become short).

The penalized coefficinet estimates are presented in Tables 6, 7 and 8 for all the responses. In most of the parameters, these penalized estimates are somehow smaller in magnitude than the un-penalized coefficient estimates (estimates using test data set). However, in some parameters the penalized estimates are lagre in magnitude. This indicates that on the process of shrinking some of the parameters may be forced to have smaller magnitude whereas others to have larger values.

Table 8: Parameter Estimates from OLS and WLS estimation methods in MLR models with the predictors selected by Lasso, for PHT and TKW using test data set.

Lasso based MLR Model for plant Height(PHT)							
		OLS estimation method			WLS estimation method		
Effect	**Pen.Est**	**Par.Es**	**St.Er**	**P-value**	**Par.Es**	**St.Er**	**P-value**
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	0.765657	0.7448
bio12	0.004314	0.01090	0.01443	70.4544	0.02297	0.01506	0.1355
bio18	0.012706	0.168117	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	70.16765	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 Thousand Kernel Weight(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

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

To evaluate the predictability of the MLR models with predictors selected by shrinkage methods, see Figures 7, 14-15 (appendix-III) and their respective RMSEs (Table 9).

With the Elastic net based model (fig 7.2), 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 7.1) seems to have three subsections of predictive 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 a 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.

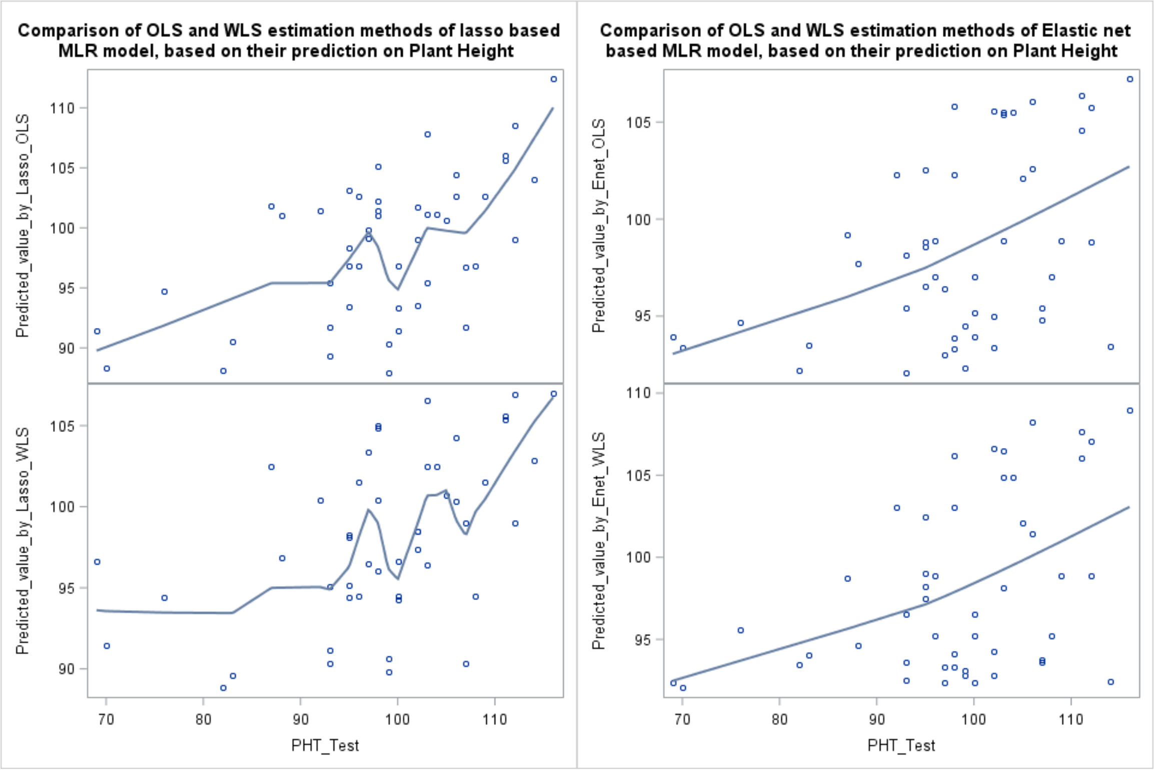
Table 9: Ordinary MLR and Shrinkage based Models based on OLS and WLS estimation methods and their predictability measured by RMSE, for each responses.

Ordinary MLR Models						
Methods(RMSE)	**DHE**	**DMA**	**PHT**	**GRY**	**TKW**	
OLS(RMSE)	3.301	2.724	10.897	583.052	4.516	
WLS(RMSE)	1.018	1.003	4.708	233.332	1.778	
Shrinkage methods based MLR Models						
	Lasso(DHE)	**Lasso(DMA)**	**Lasso(PHT)**	**Enet(PHT)**	**Lasso(GRY)**	**Lasso(TKW)**
OLS(RMSE)	3.139	2.996	9.207	9.695	566.176	4.147
WLS(RMSE)	1.14085	1.09548	1.13371	1.15434	1.06790	0.89481

RMSE=Root mean square error

Based on the Figures 14-15 appendix-III, it is observed that the model used for DHE (fig 14.1) seems to have good prediction with small variability. The Lasso based model for Days to Heading (fig 14.2) seems to have two zones of performance. The first one where actual values between around 170 and 177, within this zone the model might have better predictability with small variability. The second, zone is where actual values are above 177, from which it is observed that there seems high variability. This might result with low predictability. For the Grain Weight (fig 15.1), the predictability of the model seems better in the WLS comparing to the OLS though there seems high variability in both cases. In the other hand, for Thousand Kernel Weight the model doesn't show any clear relationship between the predicted and actual values. This might imply prediction is not sensible.

It is important to note that the prediction is more sensible for the WLS estimation method than that of the OLS as we can reveal the RMSEs (Table 9) in all the models are smaller in WLS estimation methods as compared to those of OLS estimation methods. Note that the RMSE of the models is reasonably smaller in the WLS than the OLS estimation method, but this is not visible in the predictive plots. This may happen due to the fact that the data is highly random dispersed.



(a) fig 7.1

(b) fig 7.2

Figure 7: Actual versus predicted values for Plant Height for both Lasso and Elastic net based MLR models, using both OLS and WLS estimation methods for test data set. Vertical axis is Predicted values, and horizontal axis is actual values.

In general, the parameter estimates from the ordinary MLR models are not trusted as the multicollinearity problem is not considered. Specially for prediction these models are not advisable. However, the estimates from the models with predictors selected by penalized methods are more reasonable as they 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 might 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 models with WLS estimation methods.

4 Conclusion

238 durum wheat landraces were chosen from the international center for agricultural research in the dry areas (ICARDA) genebanks aiming to investigate the predictive value of various types of long-term agro-climatic variables on future values of the different traits. Examining the association between these adaptive traits and the different agro-climatic variables is also another objective of this study. Five different traits of the durum wheat are used as response variables, and are assessed separately with 57 different agro-climatic predictor variables. From the results of exploratory data analysis, the heat map of the predictors depicted that there is high correlation among the predictor variables. It was also verified using the VIF for all the predictors, and is revealed that all the predictors have high VIF which indicated the existence of high multicollinearity.

In this study, different statistical models are employed in order to address the scientific questions. Various multiple linear regressions (MLR) models with different variable selection and estimation methods are used. Firstly, ordinary MLR models with stepwise selection criterion are used. Predictors selected by this selection method are included in the models. The ordinary MLR model used the complete (original) data set for fitting the models based on ordinary least square method. However, this model cannot consider and solve the problem of multicollinearity among the predictor variables. To solve this problem, penalized estimation methods, Lasso and Elastic net methods with cross-validation mean square error (MSECV) as a model selection criterion, are used. These methods used data partitioning to split the available data set into training and test data sets. The models are fitted on the training data set, and then validated (tested) on the test data set. The models with minimum MSE are selected and used for the analysis. Two models, one ordinary and the other one is Lasso based MLR models are fitted for each of the Days to Heading, Days to Maturity, Grain Weight and Thousand Kernel Weight. While for Plant Height, three models, ordinary MLR, Lasso based MLR and Elastic net based MLR models are fitted. Then after, from the model assumptions, some of them seem to be violated. Ordinary lease square (OLS) and weighted lease square (WLS) estimation methods are used for parameters estimation of post model selection, using the complete (original) data set for the ordinary MLR models, and test data set for the MLR models with predictors selected by shrinkage methods, for each response.

From the ordinary MLR models depicted that precipitation of month December (prec12), minimum temperature of month June (tmin6) and longitude have an increasing significant effect on the Days to Heading. While an increase in bio15 and minimum temperature of Month May (tmin5) implies to decrease in the number of Days to Heading. An increment of the maximum temperature of month August (tmax8) and bio18 increases Days to Maturity. But, increasing in bio14 tends to decrease Days to Maturity. An increment of bio3, bio7, bio13 and precipitations of months June and September (prec5 and prec9) tends to increase the Plant Height. Moreover, an increment of bio16 is inclined to decrease the Grain Weight. The only predictor that has an increasing effect on Thousand Kernel Weight is Longitude.

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. As the OLS method, bio15 and precipitation of October (prec10) have decreasing significant effect on Days to Heading. 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. 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. 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 Days to Heading, 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 Days to Maturity and Thousand Kernel Weight are 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 Days to Heading and 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 noted 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. Likwise, 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, Days to Heading 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.

[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®. 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-licensure-certification/technical-resources/calibration-training/11-weighted-least-squares-regression-calib.pdf. Accessed on May 5, 2018.

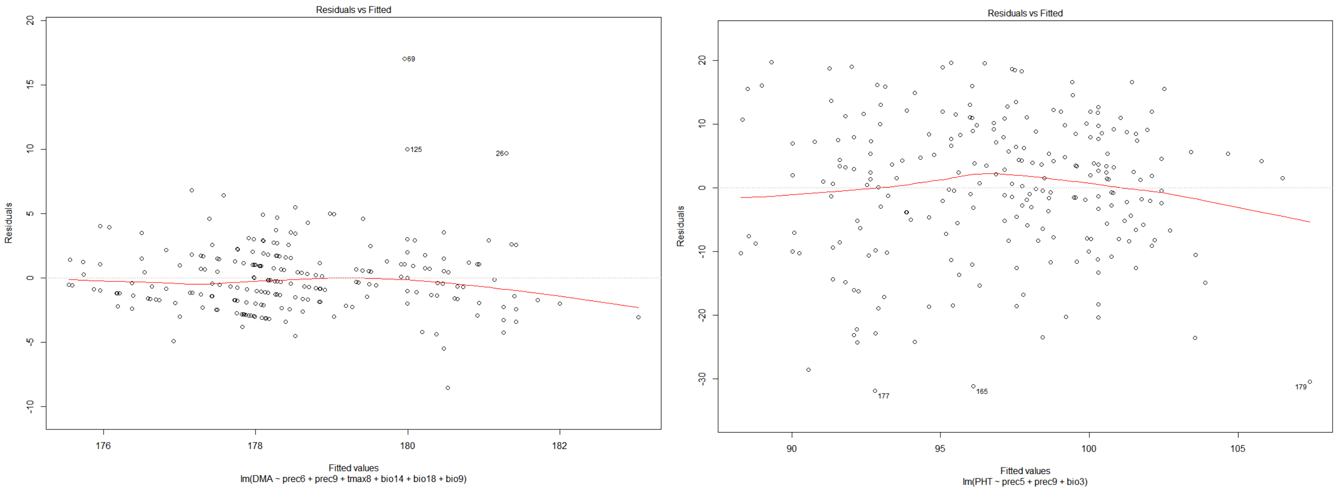
Appendix-I

Table 10: Description of all the predictors and their VIF obtained from MLR of OLS method.

Variable	**Description of variables**	**Variance Inflation**
Longitude	Longitude	29.986
Latitude	Latitude	39.614
bio1	Annual Mean Temperature	17027
bio2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	10853
bio3	Isothermality (bio2/bio7) (* 100)	141.424
bio4	Temperature Seasonality (standard deviation *100)	12262
bio5	Max Temperature of Warmest Month	2858.567
bio6	Min Temperature of Coldest Month	21815
bio7	Temperature Annual Range (bio5-bio6)	
bio8	Mean Temperature of Wettest Quarter	5.957
bio9	Mean Temperature of Driest Quarter	232.328
bio10	Mean Temperature of Warmest Quarter	7775.604
bio11	Mean Temperature of Coldest Quarter	12625
bio12	Annual Precipitation	4857.149
bio13	Precipitation of Wettest Month	310.976
bio14	Precipitation of Driest Month	342.137
bio15	Precipitation Seasonality (Coefficient of Variation)	79.946
bio16	Precipitation of Wettest Quarter	655.171
bio17	Precipitation of Driest Quarter	487.299
bio18	Precipitation of Warmest Quarter	162.717
bio19	Precipitation of Coldest Quarter	445.3
prec1	Precipitation of month 1	291.414
prec2	Precipitation of month 2	174.174
prec3	Precipitation of month 3	147.386
prec4	Precipitation of month 4	67.911
prec5	Precipitation of month 5	106.44
prec6	Precipitation of month 6	171.559
prec7	Precipitation of month 7	133.768
prec8	Precipitation of month 8	261.709
prec9	Precipitation of month 9	158.603
prec10	Precipitation of month 10	90.953
prec11	Precipitation of month 11	167.198
prec12	Precipitation of month 12	
tmin1	Minimum temperature of month 1	21897
tmin2	Minimum temperature of month 2	1374.762
tmin3	Minimum temperature of month 3	1179.093
tmin4	Minimum temperature of month 4	1447.048
tmin5	Minimum temperature of month 5	793.654
tmin6	Minimum temperature of month 6	1217.717
tmin7	Minimum temperature of month 7	1874.554
tmin8	Minimum temperature of month 8	1888.431

tmin9	Minimum temperature of month 9	1190.58
tmin10	Minimum temperature of month 10	1326.523
tmin11	Minimum temperature of month 11	951.6
tmin12	Minimum temperature of month 12	1300.925
tmax1	Maximum temperature of month 1	1673.961
tmax2	Maximum temperature of month 2	2122.59
tmax3	Maximum temperature of month 3	918.784
tmax4	Maximum temperature of month 4	830.08
tmax5	Maximum temperature of month 5	940.511
tmax6	Maximum temperature of month 6	1152.955
tmax7	Maximum temperature of month 7	1835.22
tmax8	Maximum temperature of month 8	1793.253
tmax9	Maximum temperature of month 9	1167.958
tmax10	Maximum temperature of month 10	1206.32
tmax11	Maximum temperature of month 11	795.273
tmax12	Maximum temperature of month 12	1419.534

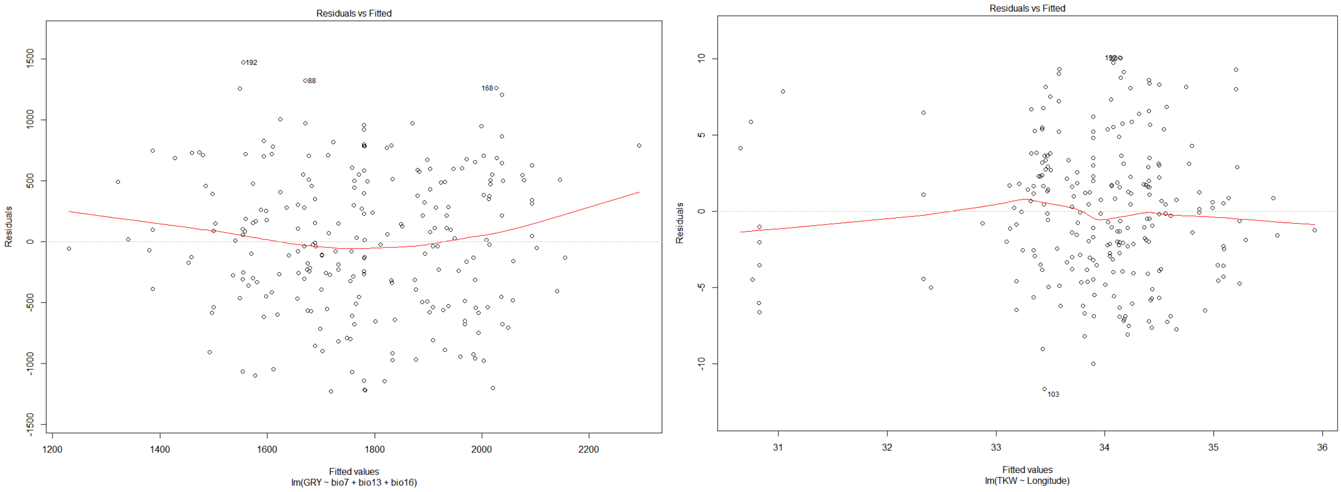
- Plots of residual versus fitted values for the complete (original) data set using the ordinary MLR models.



(a) fig 8.1

(b) fig 8.2

Figure 8: Plot of residual versus predicted values for the complete data sets of Days to Maturity and Plant Height of ordinary MLR models. Most extreme observations labeled with the row numbers of the data in the data set.

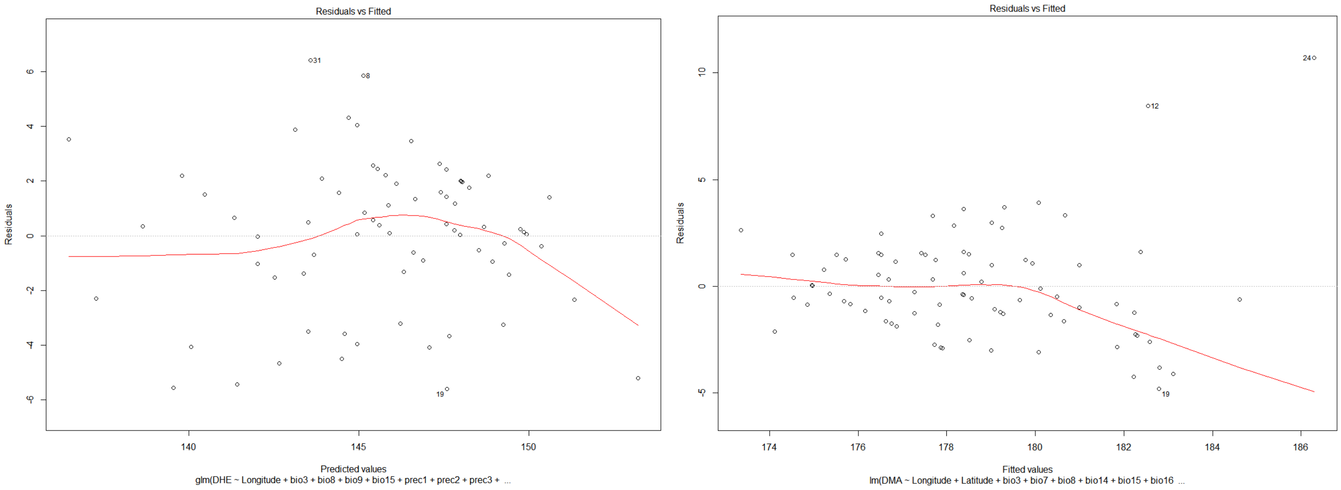


(a) fig 9.1

(b) fig 9.2

Figure 9: Plot of residual versus predicted values for the complete data sets of Grain Weight and Thousand Kernel Weight of ordinary MLR models. Most extreme observations labeled with the row numbers of the data in the data set.

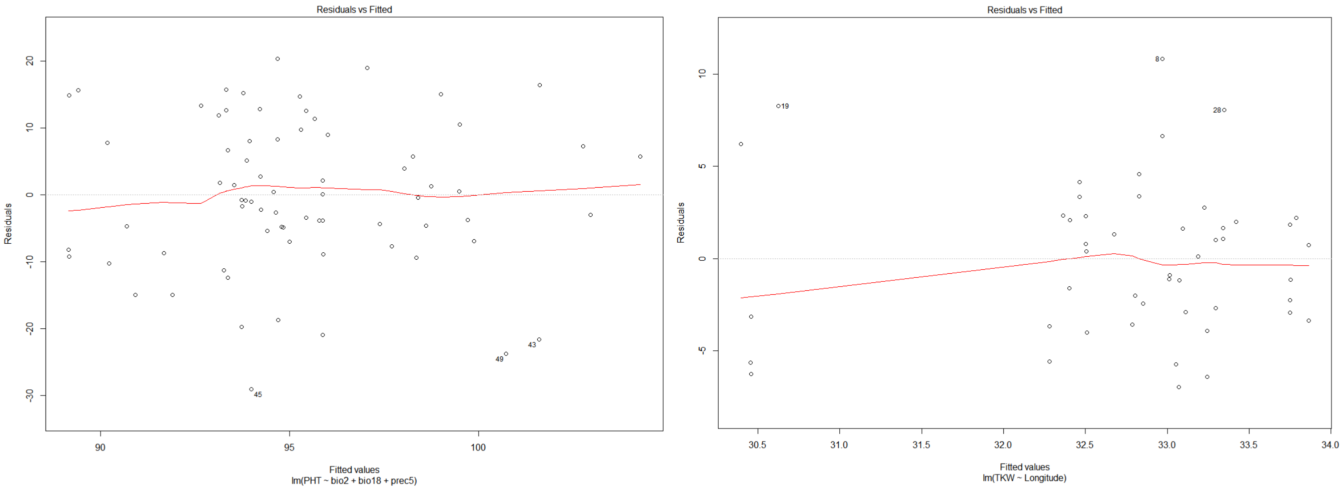
- Plots of residual versus fitted values for the test data set using lasso based MLR models.



(a) fig 10.1

(b) fig 10.2

Figure 10: Plot of residual versus predicted values for the test data sets of Days to Heading and Days to Maturity of lasso based MLR models. Most extreme observations labeled with the row numbers of the data in the data set.



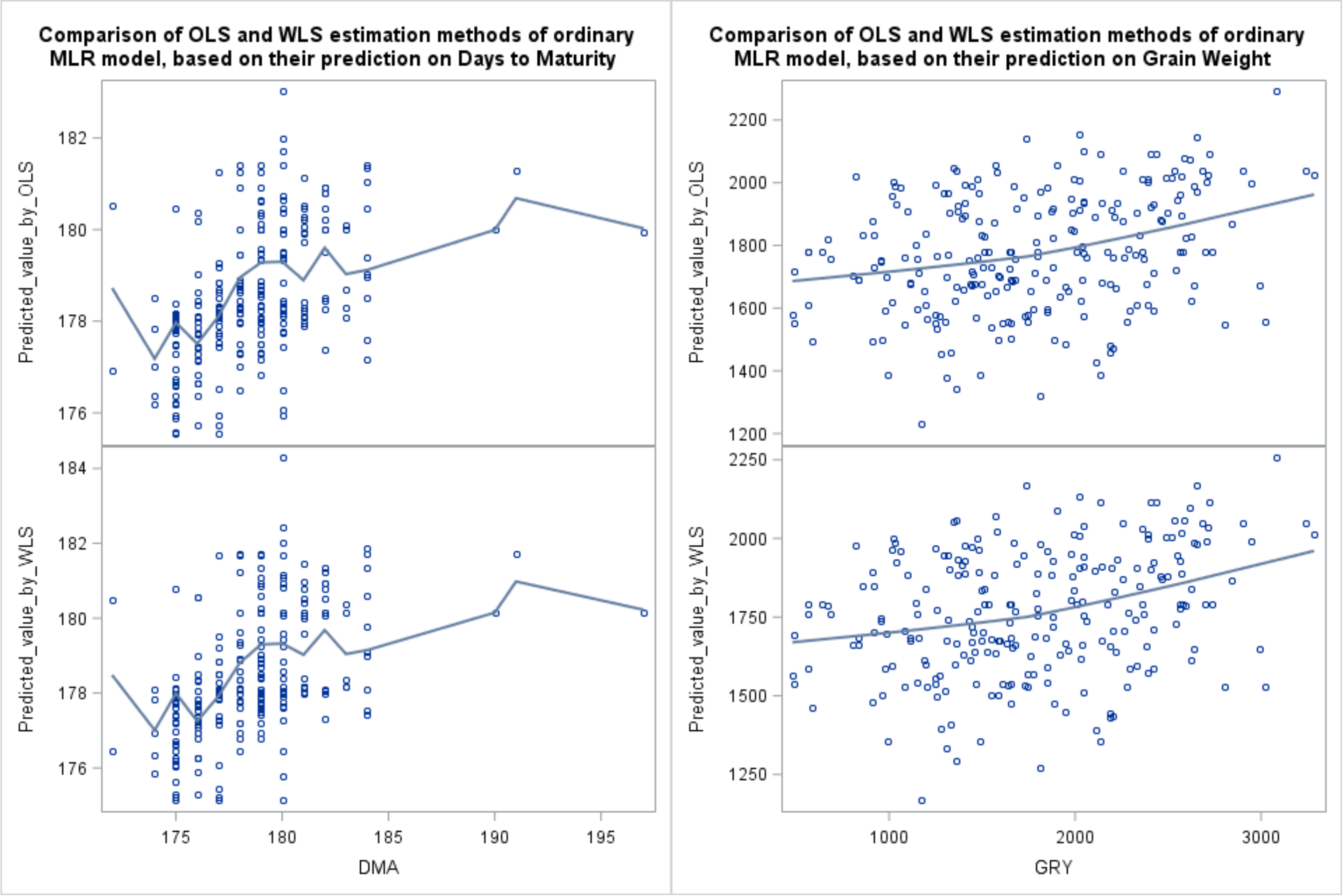
(a) fig 11.1

(b) fig 11.2

Figure 11: Plot of residual versus predicted values for the test data sets of Plant Height and Thousand Kernel Weight of lasso based MLR models. Most extreme observations labeled with the row numbers of the data in the data set.

Appendix-II

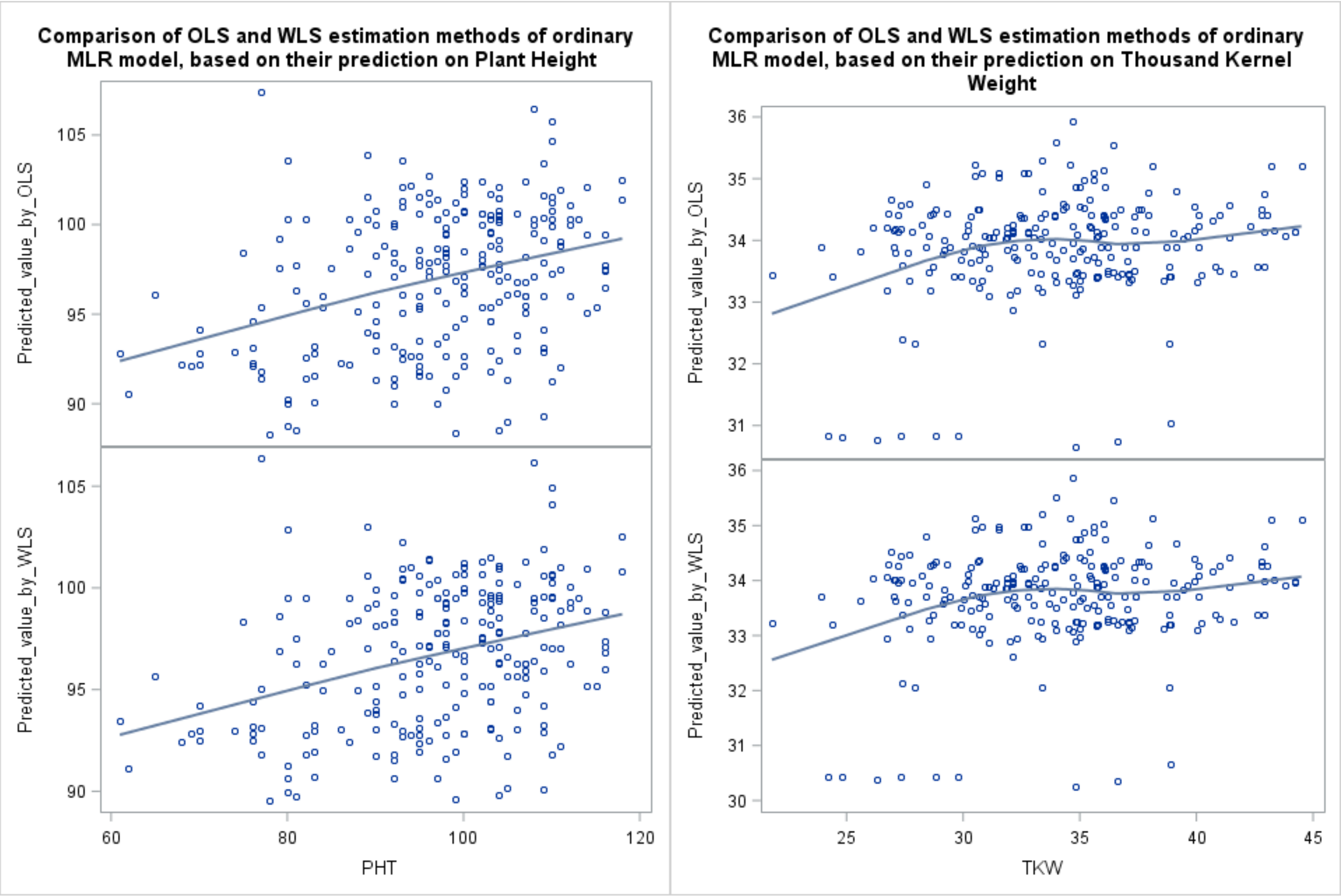
Predicted plots from the ordinary MLR models are given here.



(a) fig 12.1

(b) fig 12.2

Figure 12: Predicted versus actual values for Days to Maturity and Grain Weight using both OLS and WLS estimation methods by the ordinary MLR model. Horizontal axis actual values, vertical axis predicted values.



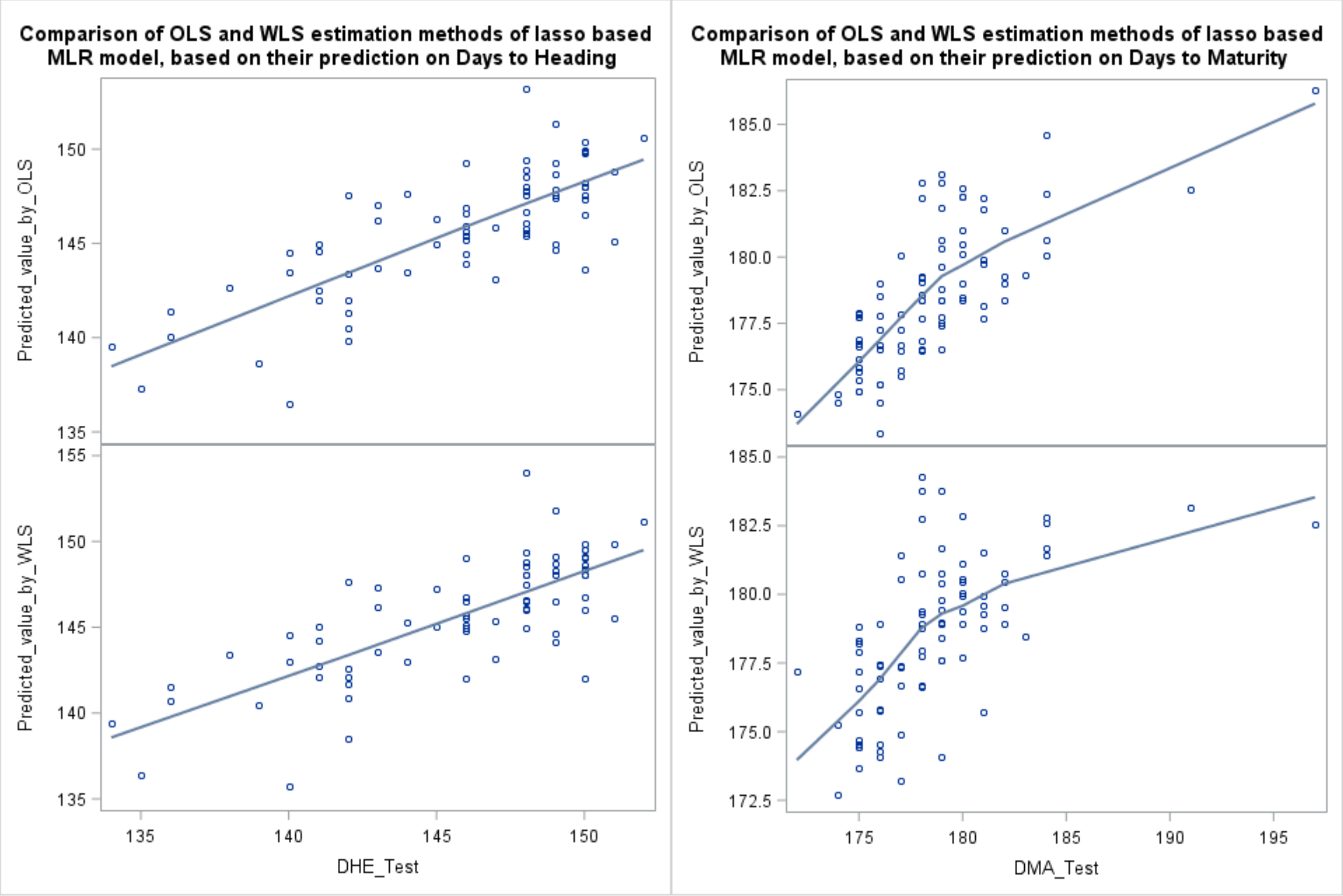
(a) fig 13.1

(b) fig 13.2

Figure 13: Predicted versus actual values for Plant Height and Thousand Kernel Weight using both OLS and WLS estimation methods by the ordinary MLR model. Horizontal axis actual values, vertical axis predicted values.

Appendix-III

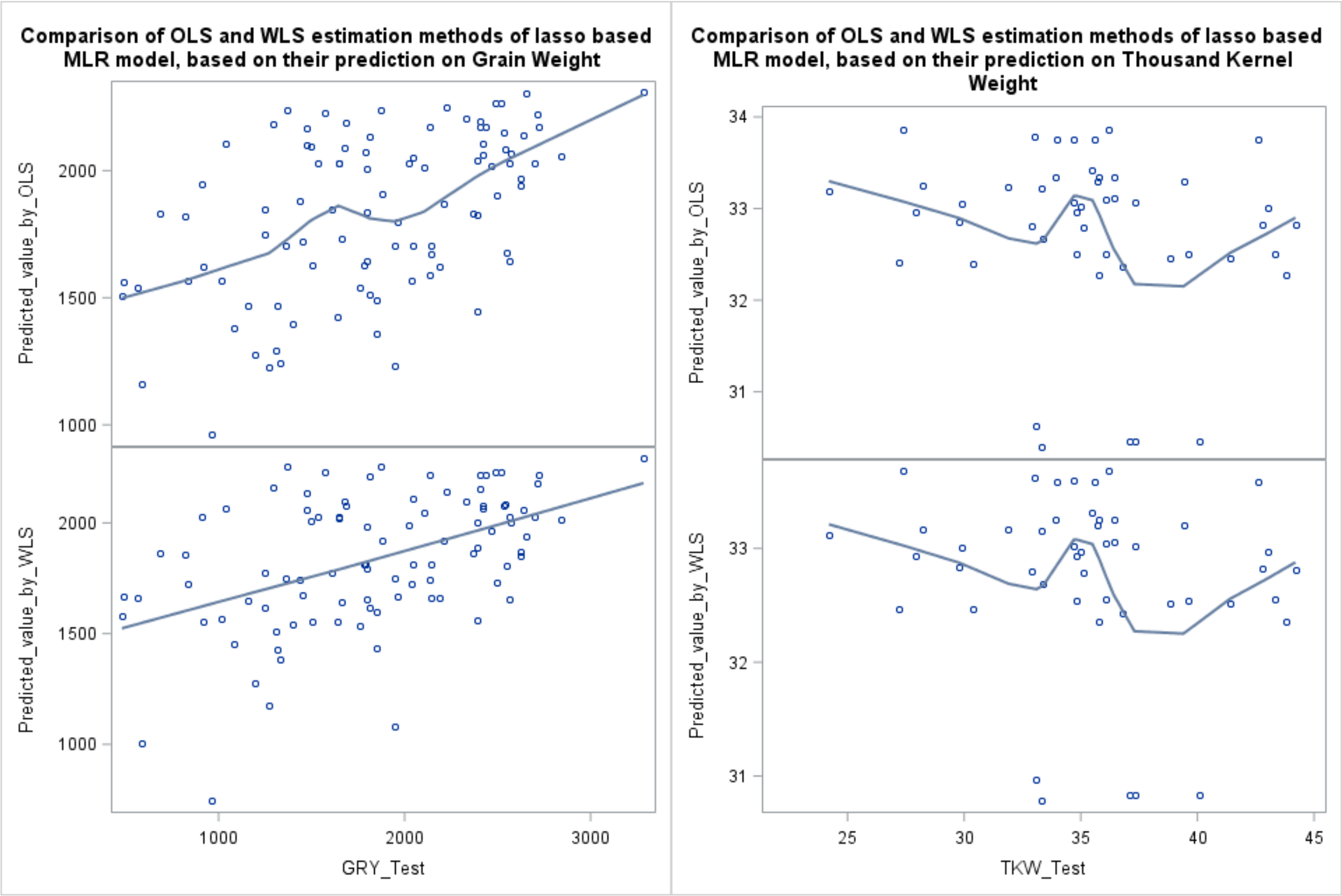
Predicted plots from the Penalized methods based MLR models, using test data set are given here.



(a) fig 14.1

(b) fig 14.2

Figure 14: Predicted versus actual values for Days to Heading and Days to Maturity using both OLS and WLS estimation methods by the shrinkage based MLR models. Horizontal axis actual values, vertical axis predicted values.



(a) fig 15.1

(b) fig 15.2

Figure 15: Predicted versus actual values for Grain Weight and Thousand Kernel Weight using both OLS and WLS estimation methods by the shrinkage based MLR models. Horizontal axis actual values, vertical axis predicted values.