Sindh Univ. Res. Jour. (Sci. Ser.) Vol.44 (4) 627-634 (2012)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Yield Prediction Models For Brassica Juncea Based On Various Trait Selection Algorithms

A. GHAFOOR, I. A. ARSHAD* A.W. SHAIKH**

Department of Statistics, Govt. College 75 S.B., Sargodha

Received 12th April 2012 and Revised 17th September 2012

Abstract: Residual analysis and various traits selection procedures are performed and compared on brassica juncea(mustard) data which consists of 33 genotypes and nine different morphological and yield related traits. In Residual analysis it was observed that genotype UCD-6/21and P53-126/40 have outlying morphological traits because of high values of DFI, DFC and PH, so it is suggested to brassica juncea breeders to control DFI and PH traits because these traits have negative impact and have insignificant effect towards oil content (OC%). Also genotypes UCD-10/15 and P118-R8 have outlying values of OC% and needed to genetically improvement and cannot be recommended for adoption at wide scale but in influential analysis it can be concluded that although genotype UCD-6/21and P53-126/40 are outlying with respect to morphological traits taken as fixed and genotypes UCD-10/15 and P118-R8 are outlying w. r .t. OC% but none of them are influential observation in fitting of prediction models. Before developing prediction models based on various traits selection procedures, assumptions of linearity, normality of residuals, homoscedasticity of error variances, autocorrelation of residuals and multicollinearity were tested and found appropriate accept the multicollinearity. Due to the problem of multicollinearity in the data, ridge regression was applied as a remedy. Besides the development of prediction models, one of the major objectives was to point out morphological traits which effect negatively towards yield and it was observed that days to flowering initiation (DFI) has negative significant effect on OC% in all traits selection methods and it requires controlling it genetically by the breeders. Also days to flowering completion (DFC), YIELD and 100-SW have positive significant effect on OC%. In context of prediction models three models, based on Stepwise, PRESS combined with subset regression and ridge regression were developed, compared and finally one model was suggested on the basis of several measures of goodness of fit for prediction of OC % of brassica juncea genotypes. The effects of traits and the fitted models were verified by simulation (through program written in Minitab), in which 1000 random samples each of size 33 were generated using Normal, lognormal and exponential distribution and it was concluded that random samples generated through normal distribution gives close results to the original results as compared to lognormal and exponential distribution.

Keywords: Residual analysis; PRESS Statistic; Ridge regression; Lognormal; Exponential distribution.

1 . INTRODUCTION

Achievement of self-sufficiency in edible oil production and agricultural production is very important step for economic growth of any country. Brassica juncea is the oldest oilseed crop among the four most important crops grown for edible oil. During the last two decades, local production of edible oil has not increased at the rate needed for domestic consumption. Its, per hectare yield in Pakistan is far lower than the other countries. The consequent gap between production and consumption of edible oil is met through import. Among all oilseed crops, it contains high percentage of oil in seeds and being a short duration crop, it can be easily adjusted, without disturbing major cropping pattern in Pakistan.

Increased oil productivity of brassica juncea can be obtained by genetic improvement of genotypes and using efficient production practices. Breeders are interested to develop such genotypes which are high in yield and in knowing which traits of brassica juncea are contributing positive or negative towards oil yield.

Cerioli and Riani (1999), suggested a unified approach to the exploratory analysis of spatial data. This leads to the identification of spatial outliers that is extreme observations with respect to their neighboring values. The effectiveness of the suggested method in detecting masked multiple outliers. Depczynski and Frost (2000), described that principal components (PCs) gives the best calibration model. Normally factor selection is based on deterministic methods like top-down procedures, forward, backward or stepwise variable selection or correlated principal component regression (CPCR). The performance of Eigen-value and correlation ranking was compared. Whitley (2000), proposed a method for variable selection and its performance using three typical QSAR data sets. The aims of this procedure were to generate a subset of descriptors from any given data set in which the resultant

++Corresponding author: A.W. SHAIKH awshaikh786@yahoo.com Cell No. +92-3322604661
*Department of Mathematics and Statistics, International Islamic University, Islamabad
**Institute of Mathematics and Computer Science, University of Sindh, Jamshoro

variables are relevant, redundancy is eliminated, and multicollinearity is reduced. Faqir *et al.* (2005), discussed various selection methods such as best subset regression, ridge regression and PCR to developed different models and diagnosed the multicollinearity problem in multivariate data set using ridge trace and other methods.

In the present study, an attempt has been made to ascertain different quantitative dimensions of Brassica juncea crop by applying various traits selection techniques. This study will helpful for breeders to control such brassica juncea traits which are contributing negatively towards brassica juncea oil yield and in predicting oil yield of specific genotype on the basis of its morphological traits. It will also provide good comparison of traits selection methods such as stepwise regression, PRESS combined with subset regression, and ridge regression for the selection of best prediction model for predicting oil yield. The main objectives of the study are to:

- Detect the outlying traits of genotypes to avoid its effects on prediction.
- Classify morphological traits contributing positively and negatively towards oil yield.
- Compare and develop valid and stable prediction models based on various traits selection statistical procedures.
- Verify the results through simulation using various probability models.

2. MATERIALS AND METHODS

Brassica juncea data consists of 33 genotypes and nine different traits such as Days to Flower Initiation (DFI), Days to Flower Completion (DFC), Days to Maturity (DM), Plant Height (PH cm), yield (g/4rows) 100-Seed Weight (100-SW gm), Oil Content (OC%), Erucic Acid (EA%), and Glucose Inolates (GI µmol/g) are considered.

In selection of regression model on particular data, one is not certain in advance, whether or not, selected model is appropriate for data in hand. Several assumptions such as linearity of regression function, normality of residuals, homoscedasticity and autocorrelation among errors and multicollinearity among predictors must be appropriate for given data. Normality of residuals can be tested using normal probability plot of residuals and both linearity of model and homoscedasticity of error variances can be checked by residual plot. Also autocorrelation and multicollinearity can be tested using Durbin-Watson (1950) test and variance-inflation factors (VIF) respectively.

2.1 Outlying observation of traits

Frequently in regression analysis, data contains outliers i.e. observations which are well separated from main cloud of data. These outlying observations may involve large residuals and have dramatic effects on the fitted regression function and ultimately on prediction. It is therefore necessary to decide whether these should be retained or eliminated from data.

In regression, an observation may be outlying w. r. t. its response variable Y (OC %), its predictor traits values (morphological traits) or both. In case of regression with one or two predictor variables, it is relatively simple to identify outlying and influential observations. In case of multiple predictor variables, more powerful methods for outlying detection are needed which are discussed in the following sections.

2.1.1 Identification of outlying fixed traits

It is useful to recognize the hat matrix (*H*) of morphological traits (considered as fixed for regression purpose) of Brassica juncea and in regression residuals e_i can be expressed as a linear combination of the observations Y (OC %) by means of the hat matrix, H given as:

$$H = X(X^TX)^{-1}X^T$$

and ith diagonal value of hat matrix is given as:

$$h_{ii} = X^T{}_i(X^TX)^{-1}X_i \qquad (2.1)$$

where $X^T{}_i = [1 \quad X_{i.1} \ldots \quad \ldots X_{i.p-1}]$ is simply ith column of data matrix, pertaining to the ith sample observation and p is the number of parameters in the regression model including intercept term. The diagonal element h_{ii} of hat matrix is called the leverage of the ith observation. It indicates whether or not the morphological traits for ith observation (genotype) is outlying, because h_{ii} is a measure of the distance between the morphological traits values for the ith observation and the means of the traits values for all *n* observations (genotypes). Thus, a large leverage value h_{ii} indicates that the ith observation is distant from the center of the traits observations. A leverage h_{ii} is usually considered to be large if it is more than twice as large as the mean leverage value, denoted by, $\bar{h}$ and is defined as:

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{P}{n}$$

2.1.2 Identification of outlying Y observations

Detection of outlying oil content (OC% taken as response variable) is based on examination of studentized deleted residuals explained in John *et al.* (1987). Studentized residual can be defined as the ratio of residual to the square root of mean square error (MSE) i.e. $e_i/\sqrt{MSE}$ and adding two refinements which are given as:

- Incorporating different variance of e_i i.e. σ^2 (e_i), instead of σ^2 (e) to give recognition of difference in their sampling error and can be defined as:

$$\sigma^2(ei) = \sigma^2(1 - h_{ii})$$

and its estimated unbiased variance is:

$$S^2(e_i) = \text{MSE}(1 - h_{ii})$$

so then studentized residual is $e_i^* = e_i / s_{e_i}$.

- Second refinement is to measure i^{th} residual when fitted regression is based on the observations excluding the i[th] one, so in this way fitted value $\hat{y}_{(i)}$ (fitted value obtained without using i^{th} observation) will not be influenced by the i[th] observation, then deleted residual can be calculated as: $d_i = Y_i - \hat{Y}_{(i)}$ and estimated variance of d_i can be defined as:

$$S^2_{(d_i)} = MSE_{(i)}\{1 + X_i^T (X_{(i)}^T X_{(i)})^{-1} X_i\}$$

Where X_i is the X-observations vector (vector of morphological traits) for the i[th] observation, $MSE_{(i)}$ is the mean square error when the i[th] observation is omitted in fitting regression function and $X_{(i)}$ is the X matrix with the i[th] observation deleted. Studentized deleted residual d_i^* can be obtained by dividing deleted residual d_i by its estimated standard deviation and given as:

$$d_i^* = \frac{d_i}{s_{d_{(i)}}} \sim t(n - p - 1)$$

Sampling distribution of studentized deleted residuals follows student "t" distribution with "n-p-1" degrees of freedom. Fortunately studentized deleted residual d_i* can be calculated without fitting regression functions in which i[th] observation omitted. It can be shown that an algebraically equivalent expression for di* is:

$$d_i^* = e_i \sqrt{\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2}} \qquad (2.2)$$

Thus studentized deleted residual d_i* can be calculated from the residual e_i, the error sum of squares SSE and the leverage value h_{ii}, form fitted regression based on all "n" observations.

2.1.3 Identification of influential observations

The next step is to ascertain whether or not these observations are influential in effecting the fit of the regression model, or possibly leading to serious distortion in model. One measures of influence for the i[th] observation on the fit is to find. $b - b_{(i)}$, i.e. difference between the vector b of the estimated regression coefficients based on all "n" observations and the vector $b_{(i)}$ based on n-1 observations when i[th] observation is deleted. Another possible measure of influence of the i[th] observation is the difference between the fitted value $\hat{Y}_i$ based on the regression with all n observations and the fitted value $\hat{Y}_{(i)}$ obtained when the i[th] observation is deleted i.e. $\hat{Y}_i - \hat{Y}_{(i)}$. But most widely used measure of identification of influential observation is *Cook's distance* denoted by D_i which is the combined impact of the differences in the estimated regression coefficients when the i[th] observation is deleted and is defined as:

$$D_i = \frac{(b - b_{(i)})^t X^t X (b - b_{(i)})}{pMSE} \sim F_{(1-\alpha;\, p,\ n-p)}$$

It is useful to relate D_i to F-distribution i.e. if the percentile value is less than about 10% or 20%, the i[th] value has little influence on the fitted regression model and if on the other hand, the percentile value is near 50% or more, the distance between the vector b and $b_{(i)}$ is then considered large, implying that the i[th] observation has a substantial influence on the fit of the regression model. Fortunately, Cook's distance D_i can be calculated without fitting new regression function (when the i[th] observation is deleted). An algebraically equivalent expression is defined as:

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1 - h_{ii})^2}\right] \qquad (2.3)$$

Equation (2.3) shows that D_i depends on two factors, firstly the size of the residual e_i and the secondly the leverage value h_{ii}. The larger is either e_i or h_{ii}, the larger would be the D_i. Thus, i[th] observation would be influential if one of the following conditions fulfill.

- If residual e_i is larger and leverage h_{ii}, is moderate.
- If leverage h_{ii} is large and residual e_i, is moderate.
- If both the leverage value h_{ii} and residual e_i are larger.

Influential observations can be separated using test of significance criteria that any observation having Cook's distance larger than $F_{(\alpha;\ p,\ n-p)}$ should be considered as influential observation.

2.2 Traits selection procedures

Brassica juncea breeders usually interested in selection of those morphological and yield related traits which are contributing significantly and are more effective in predicting OC % of a specific genotype on the basis of its morphological traits. In this context following three traits selection procedures have been discussed with confidence intervals for various parameters of traits and are given as:

- Stepwise model selection.
- PRESS model selection.
- Ridge regression.

2.2.1 Ridge regression

One of the major problem that one encounters in multivariate data is multicollinearity i.e. presence of high correlation among predictors. In this situation ridge regression provides prediction model free from multicollinearity discussed in John *et.al*. (1987). Ridge regression as remedy of multicollinearity was proposed by Hoerl *et al*. (1975), is the modification of ordinary least square regression which allows biased estimates of parameters called ridge estimates b^R but even then are more precise (having less standard errors of estimates) than ordinary regression estimates which are defined as:

$$(r_{XX} + C * I) b^R = r_{YX}$$

Where "C" is the biasing constant described the amount of bias in estimates and its value is decided by observing ridge trace (diagrammatic representation of ridge estimates, for various values of C in the range $0 \leq c \leq 1$). Also r_{XX}, r_{YX} and I are correlation and identity matrices respectively. Horel *et al*. (1975), also suggested a mathematical way of calculating value of "C" at which ridge trace becomes stable instead of deciding "C" by personal judgment in ridge trace which is given as:

$$c = \frac{(p-1)S^2}{\{b_z(0)\}^T\{b_z(0)\}}$$

Where (p-1) and S^2 are the number of parameters in model (excluding β_0) and error variance respectively and

$$\{b_z(0)\}^T = \{b_{1z}(0)\ b_{2z}(0) \ldots b_{(p-1)z}(0)\}$$
$$= \{\sqrt{S_{11}}b_1(0)\ \sqrt{S_{22}}b_2(0) \ldots \sqrt{S_{(p-1)(p-1)}}b_{(p-1)}(0)\}$$

2.2.2 Stepwise model selection

Stepwise selection procedure is a two way computing algorithm one is backward elimination procedure which begins with the largest regression and subsequently reduces to the best model and other is forward selection method which is successive insertion of traits in turn until one reaches to the desired best prediction model. A detailed discussion of the method along with computing algorithm is given in Draper and Smith (1980).

2.2.3 PRESS model selection

Prediction error sum of squares is a criterion for selection of predictor variables was proposed by Allen (1971). It is used here combined with subset regressions, mean square error and C_p statistics, to arrive at final best model. PRESS computing algorithm is given as:

- Delete first observation in OC% and in predictors (morphological, taken as fixed) variables.
- Fit all possible regression models to remaining (n-1) observations.
- Calculate predictive discrepancies $Y_i - \hat{Y}_{ip}$ for each observation.
- Similarly obtain $Y_1 - \hat{Y}_{1p}$, $Y_2 - \hat{Y}_{2p}$, $Y_3 - \hat{Y}_{3p}$ and so on up till $Y_n - \hat{Y}_{np}$ for all possible regression models.

For each regression model obtain predictive discrepancies sum of squares $\sum_{i-1}^{n}(Y_i - \hat{Y}_{ip})^2$ and decide about prediction model having smallest predictive discrepancy sum of squares.

2.3 Stability of prediction models

In prediction model, it is a good practice to test the stability of suggested models because any selected model may be subject to the prediction bias because the predictive ability of the model for the data on which the model is based may be greater than the model's predictive ability for new data. In practice, the stability of regression model can be tested using following steps:

- Fit model to some portion of the data and consider the remaining data as future data.
- Make prediction of observed response of whole data using model constructed in first step.
- Compare the predicted values in step 2 with original response variable values.

The prediction model is considered the most valid or stable model in which there exists a close agreement between the original and predicted values of the response.

2.4 Monte-Carlo simulation

A Monte-Carlo simulation described in Demodar, (2003) proceeds as follows:

- Generate random sample from any probability model say from normal, lognormal or exponential distribution and call them as U_i.
- Using estimate of developed prediction model as true regression parameters and with predictors, values of the random response Y_i can be thus generated.
- Repeating steps many times and regressing each time for estimates then averaging all estimates one can get estimates of simulated model and their standard errors.

3. RESULTS AND DISCUSSIONS

Testing of various regression assumption and residual analysis are important steps before developing prediction models based on various traits selection procedures. Regression analysis is performed by considering OC% as response variable and all other morphological and yield related traits discussed earlier are considered as predictors.

3.1 Verification of assumptions

In residual analysis, assumptions of normality of residuals, linearity of model, homoscedasticity of error variance, autocorrelation and multicollinearity are verified. Normality of residuals is one of the basic assumptions, necessary to make test of significance about regression parameters, and is verified using normal probability plot given in Fig 3.1. All points in Fig. 3.1 fall reasonably along the diagonal line, which suggests that errors are approximately normally distributed. Similarly plot of studentized residuals against fitted values $\hat{Y}_i$'s is given in Fig 3.2, verifies the assumption of appropriateness of linear regression model and as well as the constancy (Homoscedasticity) of error variance because of random scatterness of residuals along horizontal line.

In further analysis, Minitab provided no evidence of lack of fit as P-value > 0.1. Similarly

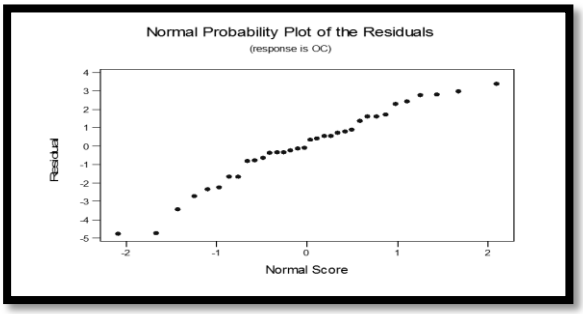


Fig 3.1 Normal Probability Plot of Residuals.

autocorrelation is tested using method proposed by Durbin and Watson (1950), to test the hypothesis of "No autocorrelation", the Durban-Watson test statistics is thus obtained is $d^*_i = 1.87$, which is much close to the "2" and it leads to "not rejection of null hypothesis of No autocorrelation". So there is no autocorrelation problem

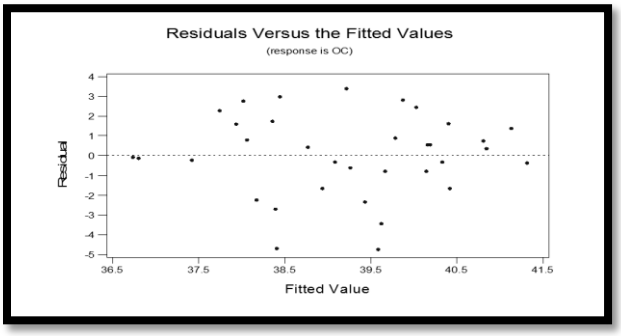


Fig.3.2 Fitted values versus Residuals.

in the data. Finally, Variance Inflation Factor (VIF) is used as a device of detection for multicollinearity among the predictors (traits) and using rule of thumb only two traits such as DFI and DFC have VIF as 38 and 40 respectively which both are greater than 10 (rule of thumb). VIF for remaining traits are found less than ten, which means at some extent multicollinearity problem exists in the data. The effect of this linear dependency upon estimates is removed through ridge regression in next section while choosing best prediction models for predicting oil content (OC%).

3.2 Residual analysis

In **(Table 3.1),** Brassica juncea genotypes, leverage (h_{ii}), studentized deleted residuals (d^*_i) and Cook's distances are listed in column 1, column 2, column 3 and column 4 respectively. Note that two largest leverage values $h_{3,\ 3} = 0.91781$ and $h_{28,\ 28} = 0.77531$ corresponding to the genotypes UCD-6/21and P53-126/40 respectively are greater than $2^* \bar{h} = 0.5294$, So it is concluded that genotype UCD-6/21and P53-126/40 have outlying fixed traits. To identify outlying oil content (OC%) of genotypes, studentized deleted residuals are examined and Student t-distribution is used to ascertain how far in tails, the absolute values fall. The studentized deleted residuals (d^*_i) are calculated by using relation (2.2) and it is clear that genotypes UCD-10/15 and P118-R8 both have the largest absolute studentized deleted residuals and are greater than the critical value $t_{(0.95,\ 26)} = 2.056$. So it can be concluded that genotypes UCD-10/15 and P118-R8 have outlying values of response variable OC%.

Cook's distances for genotypes are obtained using (2.3), to decide whether these observations are influential in fitting of models or not. According to test of significance criterion that any genotype having Cook's distance greater than $F_{(0.05,\ 9,\ 26)} = 2.30$ should be considered as influential observation. So genotype "UCD-10/15" and "P118-R8" which are concluded as outlying w.r.t fixed traits, has Cook_D = 0.29854 and 0.03452 much lesser than 2.30. Hence, it appears that genotypes UCD-10/15 and P118-R8 do not influence the fit of prediction model. Similarly the genotypes UCD-10/15 and "P118-R8" which are concluded as outlying w.r.t. genotypes OC%, also have Cook's distances 0.12312 and 0.06349 respectively much lesser than 2.30. So finally it can be concluded that although genotype UCD-6/21and P53-126/40 are outlying with respect to fixed traits and genotypes UCD-10/15 and P118-R8 are outlying values w.r.t OC% but non of them are influential observation so their influence in further analysis and in development of prediction models is not strong enough to call for remedial measures.

Table 3.1 Statistics obtained from Residual analysis

Genotypes	h_{ii}	d^*_i	Cook_D
UCD-6/4	0.13609	-0.34780	0.00219
UCD-6/18	0.26297	0.20162	0.00168
UCD-6/21	0.91781	-0.48295	0.29854
UCD-8/4	0.25261	0.66167	0.01682
UCD-10/8	0.25775	-0.38013	0.00577
UCD-10/15	0.18479	-2.41518	0.12312
UCD-44/4	0.27038	-1.15728	0.05441
DATA-SUR	0.34571	-0.03922	0.00009
PIRI	0.19361	1.63279	0.06668
KM-156	0.20101	0.74494	0.01579
P11-R1	0.29855	-1.12890	0.05961
P12-R1	0.16804	-1.63189	0.05604
P23-R2	0.12337	-0.27790	0.00125
P25-R2	0.21675	-1.30302	0.05079
P27-R2	0.27012	-0.80904	0.02729
P34-R3	0.25751	-0.17537	0.00123
P37-R3	0.31742	-0.11915	0.00076
P41-R3	0.21293	1.15837	0.03979
P56-R4	0.26356	1.37100	0.07220
P79-R6	0.20586	0.75089	0.01653
P80-R6	0.20253	1.07563	0.03244
P83-R6	0.13506	1.28683	0.02800
P104-R7	0.10594	-0.73351	0.00722
P105-R7	0.16049	0.33046	0.00241
P118-R8	0.11431	-2.27261	0.06349
P129-R9	0.24897	0.16927	0.00110
P-20	0.25503	0.83449	0.02681
P53-126/40	0.77531	-0.29455	0.03452
PR-25	0.21263	0.36589	0.00416
RL-15	0.37473	0.28726	0.00570
RL-16	0.34661	0.27686	0.00469
B-1	0.18222	0.40496	0.00420
S-9	0.18215	1.40587	0.04707

3.3 Models for yield prediction

In this context prediction models based on three traits selection procedures, discussed in section 2.2, are developed and compared for predicting oil content OC% on the basis of morphological and yield related traits.

3.3.1 Stepwise model selection

In stepwise regression using backward traits selection procedure, values of mean square error varies from 6.667 to 5.238 with brassica juncea traits included from all traits to only six traits and finally the model chosen is:

$$OC\% = 30.67 - 0.382^{*}\ DFI + 0.35^{*}\ DFC + +0.0039^{*}\ YIELD + 0.32^{*}\ (100SW) + 0.0075^{NS}\ EA - 0.035^{NS}\ GI \quad (1)$$

Regression estimate for DFI has significant negative effect towards brassica juncea oil content% i.e. with the increase in each unit in DFI, the oil content % on the average decreases but on the other hand DFC and 100-SW both have significant positive effect toward oil content% i.e. with the increase in each unit of these plant traits, oil content% on the average increases. Moreover contribution of YIELD, EA and GI are found insignificant towards oil content% and insignificant of YIELD is quite surprising in this case.

3.3.2 PRESS model selection

The final PRESS model selected using MSE, PRESS and Cp criteria is given as under:

$$OC(\%) = 32.06 - 0.386^{*}\ DFI + 0.361^{*}\ DFC - \ 0.0109^{NS}\ DM + 0.00408^{*}\ YIELD + 0.277^{*}\ 100SW + 0.0752^{NS} EA - 0.0315^{NS}\ GI \quad (2)$$

Model (2) having P = Cp = 7.0 (Ideal situation) and MSE = 5.46858 along with predictive error sum of square (PRESS) = 5.236 which is smallest among all possible best subsets models. In model (2) DM has insignificant effect on OC% but similar to model in (1) DFI is found to be significant negatively towards OC% and DFC, YIELD, and 100-SW have significant positive effect toward OC% and these traits have same interpretation as given in previous section.

3.3.3 Ridge regression

In analysis severe pair wise correlation and high values of VIF among morphological traits such as DFI and DFC are observed and to get rid of such linear dependency and its effects on estimated parameters, ridge regression is used.

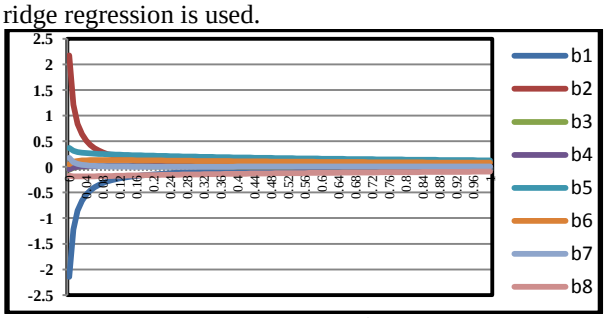


Fig. 3.3 Ridge trace for 0< C<1

Due to severe non-orthogonality in the data it is clear from **(Fig 3.3)** that the values of the least square estimates of the prediction models are initially large and unstable and then with increase of c, estimated parameters gradually becomes stable first at c=0.48. So final ridge regression model which is selected from series of C (c=0.48, point at which ridge trace becomes stable also calculated using relation given in section 2.2.3) values is given in equation (3).

$$OC\% = 35.199 - 0.02968^{*} DFI + 0.0179^{*} DFC + 0.000548^{NS} DM + 0.001636^{NS} PH + 0.00487^{*} YIELD + 0.9339^{*}\ 100SW + 0.014264^{NS} EA - 0.05462^{NS} GI \quad (3)$$

Similarly in Ridge model (3), estimates for DFC, YIELD and 100-SW are found positively significant towards OC% which means that with the increase of these traits, OC% also increases but the effect of DFI is found negatively significant towards brassica juncea OC%.

3.4 Stability of prediction models

Prior to recommending a model for prediction, it is important to test the stability of the model w.r.t. future prediction. For this purpose complete data set is divided into two parts, one part of data set contains first 23 observations and other contains last 10 observations (which are considered as future observations). Residual sum of squares for futures observations is obtained from models based on the first portion of data using all traits selection methods. It was observed that residual sum of squares of future observation in PRESS model (3) is 13.607 which is smallest as compared to Stepwise model (17.69), and ridge regression (21.95). So it can be concluded that PRESS model given in (2) is more stable model for predicting future OC% as compared to the stepwise and ridge models.

3.5 Simulation

Monte Carlo simulation experiment described in Demodar (2003) was performed and for this purpose 1000 random samples each of size 33 were generated from normal (0, 1), lognormal (0, 1) and exponential (1) distribution using an algorithm generated in Minitab to generate artificial samples of OC%. Simulated models based on different probability models are given as:

3.5.1 Models generated in Normal distribution

- Stepwise regression model.

$$OC = 31.67 - 0.347^{*}\ DFI + 0.317^{*}\ DFC + 0.0036^{*}\ YIELD + 0.189^{*}\ (100SW) + 0.0078^{NS}\ EA - 0.034^{NS}\ GI$$

- Press regression model.

$$OC(\%) = 32.06 - 0.388^{*}\ DFI + 0.371^{*}\ DFC - \ 0.0112^{NS}\ DM - 0.00423^{NS}\ YIELD + 0.277^{*}\ 100SW + 0.0761^{NS} EA - 0.0313^{NS}\ GI$$

- Ridge regression model.

$$OC\% = 46.08 - 0.020\ DFI + 0.002^{*} DFC + 0.034^{NS} DM + 0.013^{NS} PH + 0.003^{NS} YIELD + 0.53^{*}\ 100SW + 0.014^{NS} EA - 0.056^{NS} GI$$

Table 3.2 Confidence intervals of regression parameters based on various algorithms using various probability models.

b_i	Traits	Confidence Intervals(Stepwise Regression)						Confidence Intervals (PRESS Regression)						Confidence Intervals(Ridge Regression)					
		Normal (0,1)		Lognormal (0, 1)		Exp(1)		Normal (0,1)		Lognormal (0, 1)		Exp(1)		Normal (0,1)		Lognormal (0, 1)		Exp(1)	
b_0	Intercept	-4.13	67.46	-8.46	68.98	1.94	61.35	-14.83	72.45	-14.9	85.23	-12.69	80.87	-12.94	105.11	-33.3	103.16	-28.42	87.02
b_1	DFI	-1.28	-0.58	-0.30	0.69	-1.32	-0.50	-1.38	-0.51	-1.28	-0.65	-1.52	-0.68	-1.95	-0.99	-1.89	-0.95	-1.97	-1.09
b_2	DFC	0.57	1.28	0.67	1.32	0.51	1.23	0.53	1.31	-0.63	0.195	0.66	1.42	0.93	1.93	0.84	1.21	0.07	1.19
b_3	DM	****		****		****		-0.26	0.29	-0.25	0.25	-0.29	0.26	-0.28	0.22	-0.27	0.27	-0.25	0.28
b_4	PH	****		****		****		******		******		******		-0.18	0.16	-0.17	0.18	-0.16	0.17
b_5	YIELD	0.009	0.01	0.009	0.014	0.009	0.015	0.0084	0.02	0.009	0.01	0.008	0.014	0.008	0.017	0.007	0.015	0.009	0.02
b_6	100-SW	0.20	4.12	3.93	5.04	0.97	3.77	3.76	4.10	2.12	4.00	1.33	3.21	1.27	4.33	3.00	5.67	2.80	4.06
b_7	EA	-0.47	0.48	-0.40	0.42	-0.38	0.44	-0.37	0.60	-0.43	0.50	-0.33	0.53	-0.39	0.38	-0.46	0.43	-0.43	0.49
b_8	GI	-0.19	0.123	-0.20	0.13	-0.21	0.16	-0.22	0.12	-0.19	0.12	-0.20	0.15	-0.22	0.104	-0.22	0.13	-0.20	0.11

3.5.2 Models generated in Exponential distribution

- Stepwise regression model.

$OC\% = 31.65 - 0.41^{*}\ DFI + 0.3587^{*}\ DFC + 0.0043^{*}\ YIELD - 0.0201^{*}\ (100SW) + 0.027^{NS}\ EA - 0.025^{NS}\ GI$

- Press regression model.

$OC(\%) = 33.89 - 0.41^{*}\ DFI + 0.368^{*}\ DFC - \ 0.017^{NS}\ DM - 0.0043^{NS}\ YIELD + 0.316^{*}\ 100SW + 0.095^{NS} EA - 0.026^{NS}\ GI$

- Ridge regression model.

$OC\% = 29.30 - 0.087^{*} DFI + 0.073^{*} DFC + 0.0128^{NS} DM + 0.0061^{NS} PH + 0.0052^{NS} YIELD + 1.128^{*}\ 100SW + 0.029^{NS} EA - 0.0474^{NS} GI$

3.5.3 Models generated in lognormal distribution

- Stepwise regression model.

$OC = 30.26 - 0.305^{*}\ DFI + 0.28^{*}\ DFC + 0.004^{*}\ YIELD + 0.541^{*}(100SW) + 0.0081^{NS}\ EA - 0.037^{NS}\ GI$

- Press regression model.

$OC(\%) = 33.16 - 0.315^{*}\ DFI + 0.381^{*}\ DFC - \ 0.0002^{NS}\ DM - 0.0043^{NS}\ YIELD - 0.058^{*}\ 100SW + 0.035^{NS} EA - 0.038^{NS}\ GI$

- Ridge regression model.

$OC\% = 34.93 + 0.029^{*} DFI - 0.024^{*} DFC + 0.000548^{NS} DM + 0.001636^{NS} PH - 0.004^{NS} YIELD + 1.33^{*}\ 100SW - 0.019^{NS} EA - 0.04^{NS} GI$

Models obtained from simulation using randomly generated data and confidence interval of true parameters are listed in **(Table 3.2)**, it is clear that estimated regression coefficients (for stepwise, PRESS, ridge regression) of DFC (b_2), YIELD (b_5) and 100-SW (b_6) are found positively significant in all probability models accept the estimated coefficient of YIELD in lognormal case is found negatively significant which is quite surprising but the effects of EA and GI are found insignificant towards OC% in all traits selection models. The only trait, DFI has negatively significant effect on OC% in all models and interpreted as, with the increase in DFI, the OC% decreases, so it is necessary for breeders to control the DFI genetically.

Also it is clear from simulated models and confidence intervals presented in **(Table 3.2)** that the estimated parameters which are found positively significant are also observed positively significant in simulated models and those that are observed negatively significant are also observed same in simulated models. It is also observed that the ranges of estimated parameter of simulated models which are obtained using random samples from normal probability model are not wider as compared to the lognormal and exponential probability models and some unexpected results are observed using lognormal probability model, so it means that generation of random samples from normal distribution in simulation is more appropriate as compared to other probability models.

4. CONCLUSIONS

In residual analysis genotype UCD-6/21and P53-126/40 have outlying morphological traits because of high value of DFI, DFC and PH, so it is suggested to Brassica juncea breeders to reduce DFI and PH trait because these contribute negatively towards OC%. Also genotypes "UCD-10/15" and "P118-R8" both have outlying values of response variable OC % (in minimum sense) and needed to genetically improvement and cannot be suggested for adoption at wide scale.

In prediction models such as stepwise model (1), PRESS model (2) and ridge model (3), it is observed that increase in DFC, YIELD and 100-SW, OC% significantly increases but on the other hand increase in DFI, OC% significantly decreases. So it is important to control increase in plant trait such as DFI that causes decrease in oil content%. On the basis of various measures of goodness of fit such as MSE, C_p and predictive error sum of squares (PRESS), PRESS combined with subset regression model (2) can be suggested as best prediction model for predicting OC%. Also generation of random samples from normal probability model for simulation is more appropriate as compared to other probability models.

REFERENCES:

Allen, D.M. (1971) "Mean square error of prediction as a criterion for selecting variables", Technometrics, **(13):** 469-475.

Cerioli, A. and M. Riani, (1999) "The ordering of spatial data and the detection of multiple outliers", Journal of Computational and Graphical Statistics. 8 **(2):** 239-258.

Demodar, N. G. (2003) "Basic Econometrics", 5th Edition McGraw Hill Companies Inc., New York (U.S.A.).

Depczynski, U. and V.J. Frost, (2000) "Genetic algorithms applied to the selection of factors in principal component regression", Analytica Chimica Acta. 420 **(2)**: 217-227.

Draper, N. R. and H.Smith, (1980) "Applied Regression Analysis", J. Wiley and Sons.2nd edit, USA.

Durbin, J. and G. S. Watson (1950) "Testing for serial correlation in least square regression", Biometrika. *(***37):** 409-428.

Faqir, M. I. A. Arshad, and A. Ghafoor, (2005), "Development of various prediction models for Mash yield with comparison", Journal of Science and Technology Peshawar. 29 **(1)**:11-18.

John, N., W. William, and H. K. Michal, (1987) "Applied linear statistical model", 2nd Edition Toppan Comp., Tokyo (Japan).

Horel, A.E., R.W. Kennared, and K.F. Baldwin, (1975) "Ridge regression: some simulations," communication in statistics. **(4**): 105-123.

Whitley, D.C. (2000) "Unsupervised forward selection: a method for eliminating redundant variables", Journal of Chemical Information and Computer Sciences. 40 **(5)**: 1160-1168.