EVALUATION OF FULL FACTORIAL AND MULTIPLE REGRESSION METHODS IN 3^3 FACTORIAL DESIGN

A. ADAMU[1*], G. B. MELLER[1] AND BUHARI. S. A.[2]

[1]Department of mathematics, Gombe State University, Gombe
[2]Federal Polytechnic Kauran-Namoda, Zamfara State
Corresponding Author: E-Mail: abusiham@ymail.com

ABSTRACT

This research compared full factorial design model and multiple linear regression models in seed rates, row spacing and varieties of bread wheat yield. 3^3 full factorial design method and multiple linear regression were used for the analysis. The goodness of fit criteria used to evaluate the performance of structures was Akaike information criterion (AIC) and Bayesian information criterion (BIC). The data used composed of 270 observations for yield which were divided into three factors and three levels. The factors were A, B and C and the levels are 1, 2 and 3 for each factor. Analysis shows that the data have been tested and satisfied all the assumptions. Based on the results in this study, it was observed that coefficient of determination (R^2) has a highest percentage value in full factorial design model compare with multiple regression model. It was also found that full factorial design model is the most appropriate and accounted for most of the variability, According to AIC and BIC criteria.

Keywords: Full Factorial Design, Multiple Regression, Coefficient of Determination (R^2), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC).

INTRODUCTION

Factorial design is an important method or design to determine the effect of multiple variables on a response. Traditionally, experiments are designed to determine the effect of one variable upon one response. (Fisher, 1935), Showed that there are advantages of combining the study of multiple variables in the same factorial experiment. Factorial design can reduce the number of experiments one has to perform by studying multiple factors simultaneously. Additionally, it can be used to find both main effects (from each independent factor) and interaction effects (when both factors must be used to explain the outcome). Factorial design works well when interactions between variables are strong and important and where every variable contributes significantly (Trochim, 2006). Factorial designs can become cumbersome and have too many groups even with only a few factors (Williams, et al., 2006). Design of experiments is applicable to both physical processes and computer simulation models. Experimental design is an effective tool for maximizing the amount of information gained from a study while minimizing the amount of data to be collected. Factorial designs allow estimation of the sensitivity to each factor and the combined effects of two or more factors (Box, et al., 1978).

(Planta, 2006) Determined low-temperature tolerance and genetics potential in wheat (Tricitum aestivum) in 2^3 factorial design.

(Fareha, 2013) Determined the effects of process parameters on single fixation of reactive printing and crease resistance finishing of cotton fabric using 2^3 factorial designs.

Multiple regression analysis studies the simultaneous emotions that two or more independent variables may have over one dependent variable (Lefter, 2004). Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression, regression involving correlated responses such as time series and growth curves, regression in which the predictor or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression. The purpose of this paper is to find out the best model between the two-method stated earlier using some criteria like AIC, BIC and R^2.

This paper focuses on comparing full factorial design and multiple regression using seed rates, row spacing and varieties of bread wheat yield.

MATERIALS AND METHODS

The data used for this study is secondary data obtained from the Irrigation Scheme Maiduguri. The materials used composed of Varieties: V1-Local Variety, V2-R_{23}-BB-PCBWH-98 and V3-TOP'S NARO-CMB-PCBWH-1729. Seed Rate: 50kg/ha, 100kg/ha and 150kg/ha and Row Spacing: 15cm, 25cm and 35cm, the design was replicated 10 times in a 3^3 factorial design.

The trial was conducted at Lake Chad Research Institute Experimental Farm Maiduguri during the 2011 planting season. These consist of making plot sizes of 3m x 5m with 1m in-between. The experiment was completed same day to avoid introducing error due to planting same experiment on different days. The NPK and urea fertilizer were applied at split dosage, half at planting and the other half two weeks after germination. Weeding was carried out regularly as it was not part of the design.

Full Factorial Design

The three-level design is written as a 3^k full factorial design. It means that k factors are considered, each at three levels. These are (usually) referred to as low, intermediate and high levels. The levels are numerically expressed as 0, 1 and 2. We use the 0, 1, 2 schemes. Because, the three-level designs were proposed to model possible curvature in the response function and to handle the case of nominal factors at 3 levels. A third level for a continuous factor facilitates investigation of a quadratic relationship between the response and each of the factors.

The 3^3 Design Model

This is a design that consists of three factors, each at three levels. It can be expressed as 3x3x3=3^3 design. The model for such an experiment is given as follows

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk} + \varepsilon_{ijk} \qquad (1)$$

Where each factor is included as a nominal factor rather than as a continuous variable. Main effect has 2 degree of freedom, two-factor interactions have 4 degree of freedom and three-factor interactions have 8 degree of freedom and

Y_{ijk} = Is the yield of ith level of factor A, jth level of factor B and kth level of factor C

$\mu =$ Is the general mean independent of treatment effect or intercept (overall mean response of all observation).

A_i= Effect of ith level of factor A or variety.

B_j=Effect of jth level of factor B or seed rate.

AB_{ij}= Interaction effect of ith level of factor A or variety and jth level of factor B or seed rates.

C_k= Effect of kth level of factor C or Row spacing in between rows.

AC_{ik}= interaction effect of ith level of factor A or variety and kth level of factor C or row spacing (in between rows).

BC_{jk}= Interaction effect of jth level of factor B or seed rates and kth level of factor C or row spacing (in between rows).

ABC_{ijk}= Interaction effect of ith level of factor A or variety, jth level of factor B or seed rates and kth level of factor C or row spacing (in between rows).

ε_{ijk} = Is the random error associated with observing y_{ijk} and assumed iid~N$(0,\sigma^2)$.

We have 3^3 factorial design = 27 i.e ABC presented as follows

Layout/design

		A (LOW)			A (INTERMEDIATE)			A (HIGH)		
		0			1			2		
		C (LOW)			C (INTERMEDIATE)			C (HIGH)		
		0	1	2	0	1	2	0	1	2
B (LOW)	0	000	010	020	100	110	120	200	210	220
B (INTERMEDIATE)	1	001	011	021	101	111	121	201	211	221
B (HIGH)	2	002	012	022	102	112	122	202	212	222

2.3. Hypothesis Testing

H_o: $\mu = o$ versus H_a: $\mu \neq o$

H_o: $A_i = o$ versus H_a: $A_i \neq o$

H_o: $B_j = o$ versus H_a: $B_j \neq o$

H_o: $AB_{i\,j} = o$ versus H_a: $AB_{i\,j} \neq o$

H_o: $C_k = o$ versus H_a: $C_k \neq o$

H_o: $AC_{i\,K} = o$ versus H_a: $AC_{i\,k} \neq o$

H_o: $BC_{j\,K} = o$ versus H_a: $BC_{j\,k} \neq o$

H_o: $ABC_{i\,j_K} = o$ versus H_a: $ABC_{i\,j_k} \neq o$

Parameters Estimation

Inferences on specific factor effects requires the estimation of the parameters of ANOVA models such as blocks, treatments, interactions, error and total are given below (Robert, et al., 2003);

Correction factor (CF) = $\frac{y_{...}^{2}}{n} = n\bar{y}_{...}^{2}$ (2)

Where n is the total number of observation

Total sum of square (TSS) = $\sum_{i}\sum_{j}\sum_{k}\left(y_{ijk} - \bar{y}...\right)^{2}$ (3)

Sum of square A (SSA) = $\frac{\sum_{i} y_{i..}^{2}}{bcr} - CF$ (4)

Sum of square B (SSB) = $\frac{\sum_{j} y_{.j.}^{2}}{acr} - CF$ (5)

Sum of square C (SSC) = $\frac{\sum_{k} y_{..k}^{2}}{abr} - CF$ (6)

Sum of square of AB (SSAB) = $\frac{\sum_{i}\sum_{j} y_{ij.}^{2}}{cr} - CF - SSA - SSB$ (7)

Sum of square AC (SSAC) = $\frac{\sum_{i}\sum_{k} y_{i.k}^{2}}{br} - CF - SSA - SSC$ (8)

Sum of square BC (SSBC) = $\frac{\sum_{j}\sum_{k} y_{.jk}^{2}}{ar} - CF - SSB - SSC$ (9)

Sum of square ABC (SSABC) =

$$\frac{\sum_{i}\sum_{j}\sum_{k} y_{ijk}^{2}}{r} - CF - SSA - SSB - SSC - SSAB - SSAC - SSBC$$ (10)

Sum of square Error (SSE) $= TSS - SSA - SSB - SSC - SSAB - SSAC - SSBC - SSABC$ (11)

Analysis of Variance for 3^3 Factorial Experiment in RCB Design

Table 1: Summary of ANOVA for 3^3 in RCB Design

Source of Variation	Degree of Freedom	Sum of Square
Replication	r-1	
Treatment	abc-1	$27\sum_{i=1}^{r}\left(\bar{y}_{i...}-\bar{y}_{....}\right)^2$
Variety (A)	a-1	9r $\sum_{j=o}^{2}\left(\bar{y}_{.j..}-\bar{y}_{....}\right)^2$
Seed Rate (B)	b-1	
Row Spacing(C)	c-1	9r $\sum_{k=o}^{2}\left(\bar{y}_{..k.}-\bar{y}_{....}\right)^2$
A X B	(a-1)(b-1)	9r $\sum_{l=o}^{2}\left(\bar{y}_{...l}-\bar{y}_{....}\right)^2$
A X C	(a-1)(c-1)	3r $\sum_{jk}\left(\bar{y}_{.jk.}-\bar{y}_{.j..}-\bar{y}_{..k.}+\bar{y}_{....}\right)^2$
B X C	(b-1)(c-1)	3r $\sum_{jl}\left(\bar{y}_{.j.l}-\bar{y}_{.j..}-\bar{y}_{...l}+\bar{y}_{....}\right)^2$
A X B X C	(a-1)(b-1)(c-1)	3r $\sum_{kl}\left(\bar{y}_{..kl}-\bar{y}_{..k.}-\bar{y}_{...l}+\bar{y}_{....}\right)^2$
Error	(r-1)(abc-1)	$r\sum_{jkl}\left(\bar{y}_{.jkl}-\bar{y}_{.jk.}-\bar{y}_{.j.l}-\bar{y}_{..kl}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-\bar{y}_{....}\right)^2$
Total	r(abc-1)	$\sum_{ijkl}\left(\bar{y}_{ijkl}-\bar{y}_{i...}-\bar{y}_{.jkl}+\bar{y}_{....}\right)^2$

Where r is the number of replications and a, b, c are levels of the three factors A, B, C respectively.

Multiple Regression Co-efficient

Regression analysis is a useful technique in modelling and analyzing several variables, when the focus involves identifying the relationship between a dependent variable and one or more independent variables. It is widely used in different fields of study. For instant, in reliability and life-testing experiments, often one of the primary purposes is to study the effect of covariates on the failure time distribution and to develop inference on the survival probability or some other reliability characteristics of equipment. A model of the relationship is hypothesized and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given value for the independent variables. The variability of a dependent variable y can be explained by a function of several independent variables, $x_1, x_2,.., x_n$.

The multiple regressions for three independent variables is given as;

$$Y=\beta X+\varepsilon \qquad (12)$$

Where

Y = is the vector of observations of a dependent variable

β = is the vector of parameters

X = is the matrix of observations of independent variables

ε = is the vector of random errors

The vector of estimated values of a dependent variable can be expressed by using X and Y

$$\hat{Y}=Xb=X'\left(X'X\right)^{-1}XY \qquad (13)$$

The variance σ^2 is estimated by

$$S^2=\frac{SS_{RES}}{n-p}=MS_{RES} \qquad (14)$$

Where $SS_{RES}=e'e$ is the residual sum of square.

(n-p) = is the degrees of freedom

P = is the number of parameters in the model

MS_{RES} = is the residual mean square

Akaike Information Criterion (AIC)

Akaike information criterion is a measure of the goodness of fit or a test for goodness of fit of an estimated statistical model (Litell, et al., 1996). The formula for the criterion is given as follow:

AIC = 2l + 2d (15)

Where

l = is the log likelihood evaluated at the parameter estimates or restricted log-likelihood maximum value and **d** is parameter number.

Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) or Schwarz bayes information criterion (SBC) is a criterion for model selection among a finite set of models. It is based in point of the likelihood function and it is closely related to Akaike information criterion (AIC). In facts, Akaike was so impressed with Schwarz’s Bayesian formalism that he developed his own Bayesian formalism, now often referred to as the ABIC for “Akaike Bayesian Information Criterion” (Litell, et al., 1996). The formula for the BIC is given as

BIC = -2lnL + kln(n) (16)

Where L is the restricted log-likelihood maximum value k = is the parameter number and n is the observation number

Coefficient of Determination (R^2)

R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly fits the data.

Values of R^2 outside the range 0 to 1 can occur where it is used to measure the agreement between observed and modeled values and where the "modeled" values are not obtained by linear regression and depending on which formulation of R^2 is used. If the first formula above is used, values can be greater than one. If the second expression is used, there are no constraints on the values obtainable. The formula for R^2 is given below

$$R^2 = \frac{SS_{res}}{SS_{tot}}$$

Where

$$SS_{res} = \sum_{i=1}^{n} \left(y_i - \hat{y}\right)^2 \qquad (17)$$

$$SS_{reg} = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 \qquad (18)$$

$$SS_{tot} = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \qquad (19)$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (20)$$

n = number of observations

The notations SS_R and SS_E should be avoided, since in some texts their meaning is reversed to Residual sum of squares and Explained sum of squares, respectively.

Adjusted R^2

The use of an adjusted R^2 (often written as $\bar{R}^2$ and pronounced "R bar squared") is an attempt to take account of the phenomenon of the R^2 automatically and spuriously increasing when extra explanatory variables are added to the model. It is a modification due to (Theil, 1961), of R^2 that adjusts for the number of explanatory terms in a model relative to the number of data points. The adjusted R^2 can be negative, and its value will always be less than or equal to that of R^2. Unlike R^2, the adjusted R^2 increases when a new explanator is included only if the new explanator improves the R^2 more than would be expected by chance. The adjusted R^2 is defined as: $\bar{R}^2 = 1 - \frac{SS_{res}/df_e}{SS_{tot}/df_t}$ (21)

RESULTS

Analysis of Variance for YIELD, Using Adjusted SS for Tests

Table **2** shows that two factors are significant. That is factor A and factor C are significant, F_{cal} at df (2, 243) is greater than F_{tab} at df (2, 243), $\alpha = 0.05$ and P<0.05. but factot B is not significant, at df (2, 243), $\alpha = 0.05$, also all the interactions are significant A*B, A*C, B*C at df (4, 243), $\alpha = 0.05$ and A*B*C at df (8, 243), $\alpha = 0.05$. Hence, it can be concluded that the yield of seed rates, row spacing, and varieties have different effects at different levels which indicates that there is significant different in yield between different factors. In addition, R^2 is moderately ok which explained the variability of the data.

Table 3 ANOVA for regression also shows that only factor A is significant while the rest are not significant. Thus, it can be concluded that there is no significant difference on the yield of bread wheat in the effects of seed rates, row spacing and varieties because all the F_{cal} at different df are less than the F_{tab} in order word all the P>0.05. Hence it would be concluded that there is no different between the yields of bread wheat. In addition, R^2 value is small which shows that the variability of the data explained is not much compare to that of full factorial.

Table 2: ANOVA for Yield, Using Adjusted SS for Test

SOURCE	**DF**	**Seq SS**	**Adj SS**	**Adj MS**	**F**	**P**
SEED RATES	2	377931	377931	188966	**5.76**	0.004
VARIETIES	2	31517	31517	15759	0.48	0.619
ROW SPACING	2	204013	204013	102006	**3.11**	0.047
SEED RATES *VARIETIES	4	644273	644273	161068	**4.91**	0.001
SEED RATES*ROW SPACING	4	362224	362224	90556	**2.76**	0.029
VARIETIES*ROW SPACING	4	685262	685262	171315	**5.22**	0.000
SEED RATES*VARIETIES*ROW SPACING	8	758717	758717	94840	**2.89**	0.004
ERROR	243	7978319	7978319	32833		

TOTAL	269	1.1E+07	1.1E+07

NB: Bolded F_{cal} Values indicates that the factors and the interactions are significant.

S = 181.198   R-Sq = 27.75%   R-Sq(adj) = 20.02%

AIC = 1215.05   BIC = 1207.05

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk} + \varepsilon_{ijk} \quad (22)$$

10121817 = 0 + 37793 + 31517 +204013 + 64427 + 362224 + 685262 + 758717 + 7978319

Table 3: ANOVA for Multiple Regression

SOURCE	DF	SEQ SS	ADJ SS	ADJ MS	F	P
REGRESSION	3	231045	231045	77015	1.88949	0.13077
SEED RATES	1	176075	176075	176075	**4.33217**	0.03834
VARIETIES	1	14031	14031	14031	0.34522	0.55733
ROW SPACING	1	40939	40939	40939	1.00726	0.31647
ERROR	266	10811212	10811212	40644		
LACK-OF-FIT	23	2832893	2832893	123169	3.75143	0.00000
PURE ERROR	243	7978319	7978319	32833		
TOTAL	269					

NB: Bolded F_{cal} Value indicate that the factor is significant. Also we used Mean Square Error to obtain the F and P values of Regression, Seed rates, Varieties and Row spacing, then Mean Square Pure Error to obtain F and P values for Lack of fit.

S = 201.603

R-Sq = 2.1%

R-Sq(adj) = 1.0%

AIC =1250.68

BIC = 1242.68

From Table 4, all the criteria shows that full factorial design model is better than multiple regression model using such type of data because, if you are comparing R^2 with different models the one that has the largest percentage is the best while AIC and BIC indicates the model that has lowest values of AIC and BIC is the best. Also MSE of full factorial design model is the least compare to that of multiple regression hence full factorial design is the best.

Table 4: Summary for Full Factorial and Multiple Regression Model

Model Selection	Full Factorial	Multiple Regression
R^2	27.75%	2.10%
AIC	1215.05	1250.68
BIC	1207.05	1242.68
MSE	32833	40644

Figure **1** indicates that the distribution of residuals is normal. Because, from the figure you can see that the residuals (dots in red colour) resemble a straight line therefore the normality assumption hold. Because the data is said to be normal when the probability plot is linear in shape.

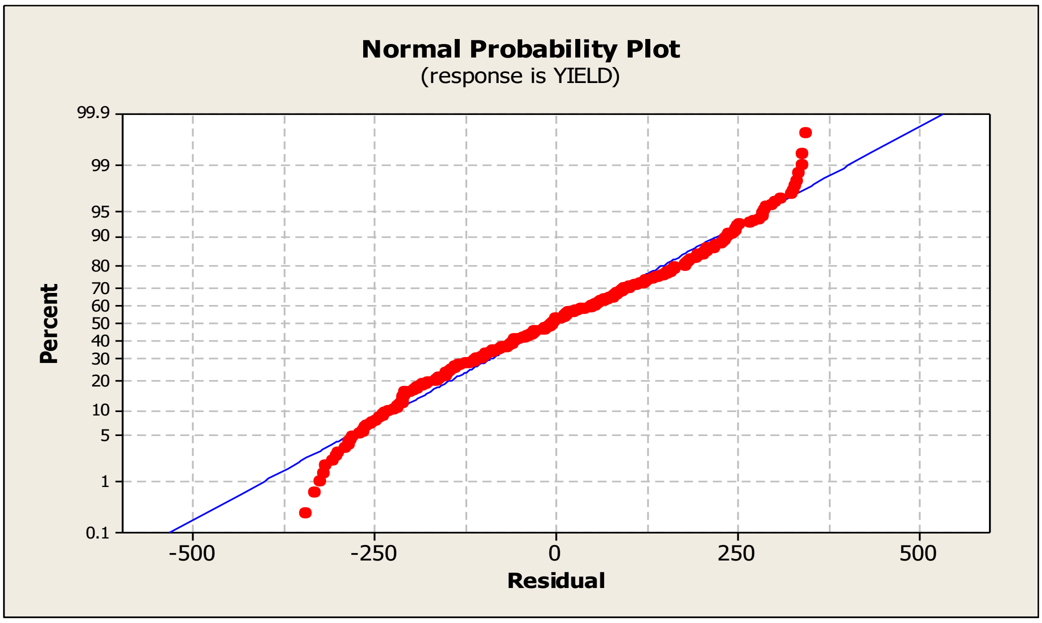

Figure 1: Normal Probability

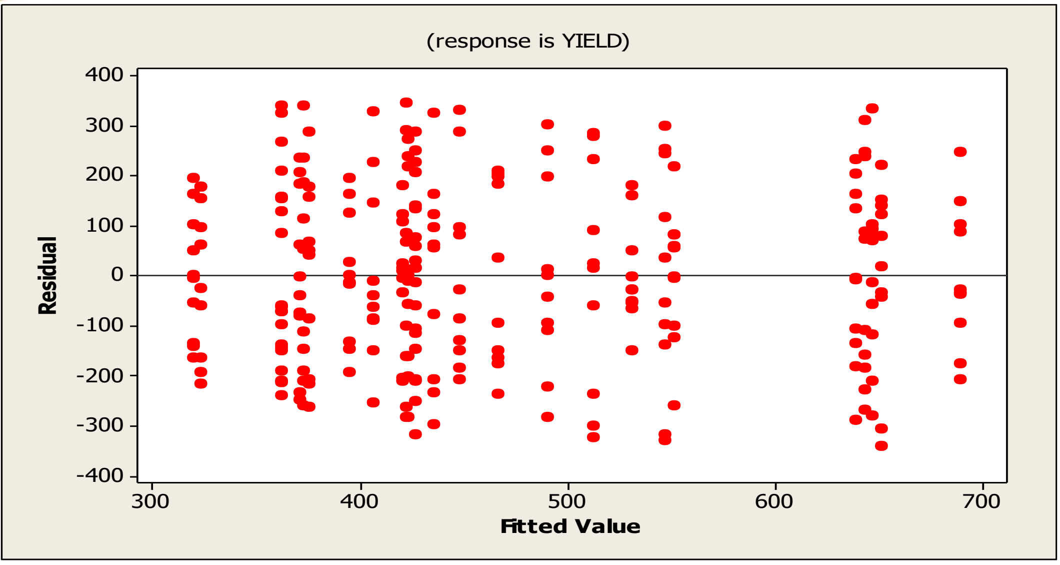

Figure 2: Constant Variance

Figure 2 can be checked with residuals versus fits plot. This shows that there is constant variance since it didn't show any recognizable patterns, the residuals (dots in red colours) scattered all over the plot. Residuals increases as the fitted values increases.

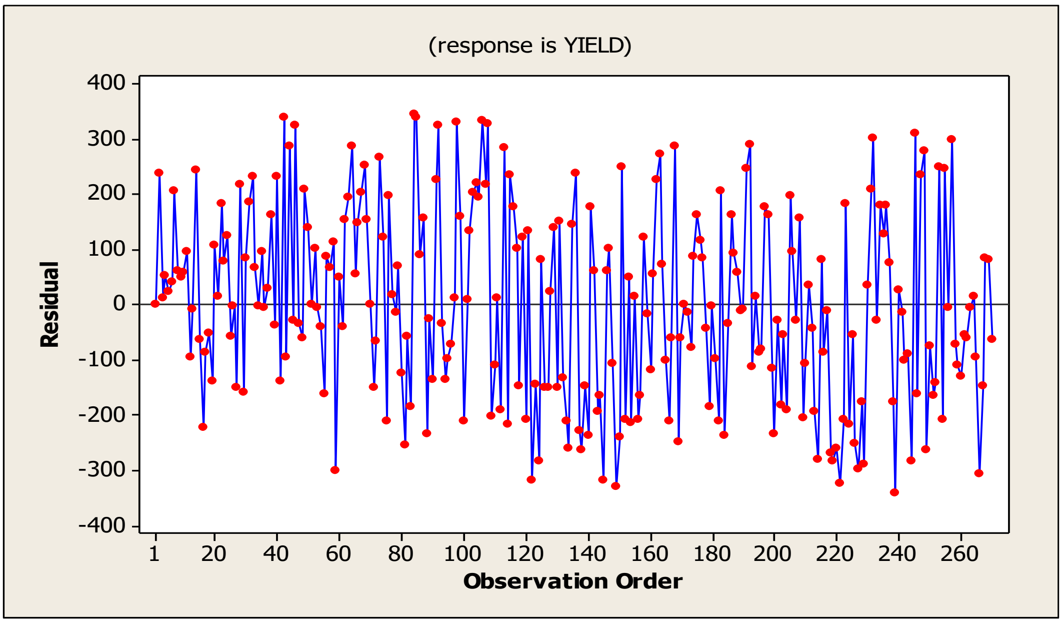

Figure 3: Independence

Figure 3 can also be checked with residuals versus order. A positive correlation or a negative correlation indicates that the assumption is violated. The above figure shows that the plots (dots in red colour attached with blue lines) are independent since both do not reveal any pattern. Therefore, the independence assumption is satisfied. In summary the figures satisfied almost all the assumptions which make the data to be good.

DISCUSSION

3^3 multi-level factorial design ANOVA methods and multiple regression ANOVA methods were used in this research to find out the best model among them using R^2, AIC and BIC model selection in seed rates, row spacing and varieties of bread wheat yield. The data satisfied the assumption of normality, independence and constant variance based on the outcome of the result. In this study, the set of data used was tested for adequacy and found to satisfy the assumption of normality, independence and homogeneity. Multiple regressions analysis between the dependent variable (yield) and the three factors were mild positive. 3^3 multilevel full factorial design analysis shows that two factors A and C are significance while the other factor B is not significance. Null hypothesis indicates that the hypothesis should be rejected and conclude that there is significance difference between the yields. Coefficient of determination (R^2) also indicates that the analysis is satisfied. Full factorial design is the best model compare to multiple regression model since; it gives the highest percentage value of coefficient of determination (R^2). Hence all the analysis uses $\alpha = 0.05$ significance level and SPSS software version 24.0 was used for the analysis.

CONCLUSION

This research is concerned with comparing full factorial design model and multiple regression model to determine best model in seed rates, row spacing and varieties of the bread wheat yield. Based on R^2, AIC, BIC and MSE full factorial design was the most appropriate model. Full factorial design method allow a large number of variables to be investigated in a compact trial, enable outliers in the data to be identified and provide detailed process knowledge. R^2(adj) penalizes the statistic as extra variables in the model.

REFERENCES

Box, G. E. P., Hunter, W. G. & Hunter, J. S., 1978. *Statistics for Experiments.* New York: Wiley.

Fareha, A., 2013. Effects of process Parameters on a Single Step Fixation of Reactive printing and Crease Resistance Finishing of Cotton Fabrics using 2 by 3 Factorial Design. *International Journal of Textile Science,* pp. 7-11.

Fisher, R. A., 1935. *The Design of Experiments.* Edinburgh, Scotland: Oliver and Boyd.

Lefter, C., 2004. *Marketing Researches.* Brasov: Published by Infomarket.

Litell, R. C., Milleken, G. A., Stroup, W. W. & Wolfinger, R. D., 1996. *SAS System for Mixed Models.* Cary, N. C: SAS Institute.

Planta, E., 2006. *Evidence for active Electron flow in twig Chlorenchyma in the Presence of an Extremely deficient Linear Electron Transport Activity.* s.l.:s.n.

Robert, L. M., Richard, F. G. & James, L. H., 2003. *Statiscal and Analysis of Experiments.* New York: John Wiley and Sons Publication.

Theil, H., 1961. *Economic Forecast.* Holland: Armsterdam. North.

Trochim, W. M. K., 2006. *Factorial Design.* s.l.:Research Methods Knowledge Base.

Williams, K. D. et al., 2006. *Evaluation of a Component of the Cloud response to climate in an Inter-comparison of Climate Models.* s.l.:s.n.