xIlkogretim Online - Elementary Education Online,2021; Vol 20 (Issue 2): pp.1184-1195 http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.02.134

Modeling Wheat Productivity using Hierarchical Regression: A way to Address Food Security Concerns

Muhammad Islam, PhD Scholar, Department of Statistics, The Islamia University of Bahawalpur, Pakistan mislam6667@gmail.com,
Farrukh Shehzad, Assistant Professor, Department of Statistics, The Islamia University of Bahawalpur, Pakistan
Muhammad Omar, Assistant Professor, Department of Computer Science, The Islamia University of Bahawalpur, Pakistan

Abstract- Food security has been considered as a major concern for many countries since several years. Food is basic necessity of human beings all over the world. World's Population and demand of food is increasing across the years while its production is not enough to meet this challenge of food security. Wheat is paramount food crop all over the world. Main objectives of this study is to identify data set (data units form) that could provide a model with better prediction capability and to investigate the significant factors for wheat yield enhancement. Hierarchical regression analysis is applied on the data taken from Crop Reporting Service, Agriculture Department of Punjab, Pakistan. Three more data sets (clusters) generated from the original data set. Model selection criteria, adjusted R^2, ΔR^2, MSE AIC, SIC, Wi (AIC) and ER (AIC) have been exercised on these models. The result indicates that clustering improved the R^2, Cronbach's alpha and reduced the variance, MSE, AIC, SIC. The best model is selected on the basis of prediction capability and it can be helpful for precise estimation of food to cope with the coming challenge of food security.

Keywords: Hierarchical regression, multiple regressions, weighted least squares, wheat productivity

I. INTRODUCTION

Food is a fundamental need of life and agriculture is the main stream for the food security and for food availability. The world's population is expected to reach 9.1 billion up to 2050 and the major contribution for this increase in the world population will come from developing countries (FAO, 2009; Nelson et al., 2010). Food production must increase about 70% to meet this challenge and it will be double for the developing countries (Kagan, 2016). It is foremost for us to increase global food production by (70–100)% to meet the feed requirement of the world in 2050 (McKenzie & Williams, 2015). According to the international food policy research institute (IFPRI) each day our world witnesses 800 (million) people go hunger. Despite of that world population is increasing, growth rates of yields for major cereals is decreasing in the world (FAO, 2009). According to economic survey of Pakistan, agriculture is the biggest sector of Pakistan and contributing about 21% of GDP and providing employment to 45% of Population (ESP, 2013; Raza, Ali, & Mehboob, 2012). Economic survey of Pakistan 2018-19 reported that share of agriculture is now reduced from 21% to 18.5% and employment from 45% to 38.5%. Wheat is a staple food crop of Pakistan and it ranks first in acreage and production among all food crops. According to FAO Pakistan is 7th largest producer of wheat in the world (UAF, 2014). According to economic survey of Pakistan 2010-11, population growth rate in Pakistan is very high about 2.05% while it is low in neighboring countries like Bangladesh, Bhutan, India, Maldives, Nepal, Srilanka about 1.7%, 1.4%, 1.5%, 1.9, 2.0% and 0.5% respectively. It is alarming that population growth rate in Pakistan is now increases to 2.40% and total population is 212.8 (Millions) from 155.4 (Millions) in 2005-06. The agriculture growth rate of Pakistan is reached low at 0.85% in 2018-19 from 2.7% in 2012-13, 2.5% in 2013-14, 2.1% in 2014-15, 0.2% in 2015-16, 2.2% in 2016-17 and 3.9% in 2017-18 (ESP, 2019). The wheat crop production in 2013-14 was 25979(000) tonnes at growth rate 7.3% while in 2018-19, it was 25195(000) tonnes at growth rate 0.5% and 25076(000) tonnes at growth rate -6.0% in 2017-18. The average yield in kg/hec wheat (2002-03 to 2006-07) in Pakistan was 2496 and in UK, Germany, France, China, Poland, Italy, India, USA, Argentina were 7779, 7289, 6760, 4345, 3765, 3470, 2643, 2783, 2578 which shows in Pakistan yield loss is -67.91%, -65.76%, -63.08%, -42.55%, -33.71%, -28.07%, -5.56%, -10.31%, -3.18% as compared to these countries (CRS, 2008; Islam, 2015). Population growth rate in Pakistan is still high and production of wheat crop is still low as compared to others countries. With the current rate of population growth, it is estimated that in 2050 Pakistan will attain the 4th position in term of population in the world instead of 6th (Ahmad & Farooq, 2010). To meet the food security, with the reduction of population, it is very necessary to increase the yield of wheat crop (Islam, 2015)

Qayyum and Pervaiz (2013) presented descriptive study for the factors affecting wheat yield and they studied the variables DAP, urea , plough, level, water, source of seed, variety, spray, sowing time, harvesting time, rainfall and humidity and they presented the multiple regression model for the projection of wheat crop in Punjab based on 34 regressors (Qayyum, 2011; Qayyum & Pervaiz, 2013). Bajkani et al. (2014) reported that traditional practice resulting the low production of wheat crop. Hussain (2010) reported that by giving the better inputs food grain crop characterized increasing returned to scale. Tariq et al. (2014) reported that per capita availability of wheat was 198 kg per annum in 2014, it would be 105 kg per annum in 2031 and 84 kg per annum in 2050 due to rising trend of population and adverse climatic effect. In Pakistan yield of the wheat crop is low while its population growth rate is still high as compared to competitive countries. In recent era, the focus of researchers and policy makers has been diverted towards the physical availability of food facilitated by sufficient agricultural production. Geographically Pakistan is divided in four major provinces named Punjab, Sindh, Balochistan and Khyber Pakhtunkhwa. Pakistan is at standing 6th populous state in the world and about 53% of total population of Pakistan is situated in Punjab. The 75% of total wheat area is sown in Punjab. Crop reporting service agriculture department, Punjab is very large and only statistical organization having sound statistical mechanism and responsible for the estimation of all major and minor crop acreage, yield and production. The data of CRS is valid and further published by Pakistan bureau of statistics and Punjab bureau of statistics nationally and internationally for better understanding for the researchers and policy makers. A detail interview of input used is carried out after the yield estimation procedure by CRS. The CRS is reporting the production and acreage estimate to the government of Pakistan without opted proper statistical modeling techniques for the forecast estimation of wheat crop production. Qayyum (2011) presented multiple linear regression model for the projection of wheat crop production in the Punjab by using the 34 regressors with MSE 4.199 and R^2 0.449. According to Neter et al. (1989) there should be as possible as to choose the minimum number of regressors for the projection of response variable for best regression modeling and it is very difficult and laborious to handle the many parameters in prediction and such models may generalize poor. The current study reduced the regressors from 34 to 15 and introduced new datasets generators/clusters which improved the R^2, Cronbach's alpha, reduced the variance, MSE, AIC, SIC and produced the better Wi(AIC) and ER (AIC) to select the best model. For better understanding the food security and its availability the accurate and precise statistical model is very essential for each level of estimation and forecasting for the future prediction of food to prevent the people of universe from hunger. The present study focuses to develop the statistical hierarchical regression model for various levels for the accurate and precise yield estimation modeling in the Pakistan in context of food availability and food insecurity situation.

II. MATERIAL ANDMETHODS

2.1 Data Collection and Preparation of Data Sets

For the present study large amount of secondary data for the years 2016-17 to 2019-20 is taken from crop reporting service (CRS) agriculture department of Punjab (Pakistan). The CRS is only agricultural statistical data based organization working independently since 1978 and responsible for estimation of yield, production and area of all crops, fruits and vegetables (Qayyum & Shera, 2019). These estimates are published by Pakistan bureau of statistics (PBS), Punjab bureau of statistics (PBOS) and many other government agencies for the researchers and policy decision. List frame sampling technique (LFS) was applied by CRS up to 2018-19 to selected the samples (villages) with systematic random sampling (S_YRS) in which complete village was taken as basic unit but after 2018-19, CRS shifted the sampling technique to two stage area frame sampling (AFS). At stage-I probability proportional to size (PPS) is applied to select the sample village first and then at stage-II simple random sampling (SRS) is applied to select the segment area from village (Qayyum, 2011; Qayyum & Shera, 2019). CRS selected the 1240 villages by LFR and then shifted to 5500 segments selected by AFS covering all the districts of Punjab. The crop cut experiment C.C.E'S is carried out in selected samples villages (segments) by applying SRS (Qayyum, 2011; Qayyum & Pervaiz, 2013). This data sets is consists of 26430 crop cut experiments (C.C.E'S). Three more data sets generated from 26430 C.C.E'S (tab:2.1) to get the better model performance measures.

Table 2.1 Data sets/Clusters

Data preprocessing data sets and clusters			
Dataset1	Dataset2 (Cluster1)	Data set 3 (Cluster 2)	Data set 4 (Cluster 3)

26430 crop cut experiments (CCE) plots in Punjab, Pakistan	6034 sample points of Punjab (2016-17 to 2019-20)	145 Tehsils of Punjab	36 Districts of Punjab

Author's contribution

Datasets2-4 (clusters1-3) comprise of the centroid base point and proportion of data set at sample point villages, tehsils and district level.

$$\bar{y}_{i_{c_k}} = \frac{\sum_{j=1}^{N_{j_{n_k}}} y_{i_{c_k}}}{N_{j_{n_k}}} \quad \ldots\ldots(1) \qquad \bar{P}_{i_{c_k}} = \frac{\sum_{j=1}^{N_{j_{n_k}}} P_{i_{c_k}}}{N_{j_{n_k}}} \ldots\ldots(2)$$

where i= 1, 2…8 (quantitative variable), j = 1, 2… $N_{j_{n_k}}$ (jth observation of ith predictors in k^{th} cluster , k= 1, 2, 3 clusters, $N_{j_{n_k}}$ = total no. of jth observation of ith predictor in k^{th} cluster, $\bar{y}_{i_{c_k}}$ = mean of the ith quantitative predictor (response variable) in k^{th} cluster.

where i= 1,2,…8 categorical binary variables, j= 1, 2... $N_{j_{n_k}}$ (jth observation of the presence of ith binary variable in k^{th} cluster, k= 1, 2, 3 clusters, $N_{j_{n_k}}$ = total no. of jth observation of ith binary variable in k^{th} cluster, $\bar{P}_{i_{c_k}}$ = proportion of the ith predictor (binary) in k^{th} cluster.

2.2 Regression Analysis

The hierarchical regression analysis is applied on four data sets with two estimation methods known as least square or weighted least square (if required). Hierarchical regression analysis is based on theoretical decisions and measures the influence of the several predictors in sequential way such that predictors are set in blocks to determine the change in R^2 and change in F-statistic along to select the best regression model (Cohen, 2008; Pedzahur, 1997; Petrocelli, 2003; Rosenthal, 2017; Wampold & Freund, 1987). For the current study predictors are set in two blocks, block 1 contained seven quantitative variables while in block 2, eight categorical (binary) variables used to build the hierarchical regression model for the projection of wheat crop production.

Model 1 $\quad Y=X\beta+\epsilon, \quad \beta = (X'X)^{-1}X'Y \ldots\ldots(3)$

Y = response variable, β = coefficients of 07 predictors block 1 ε_i = error term

Model 2 $\quad Y=X\beta+\epsilon, \quad \beta = (X'X)^{-1}X'Y \ldots\ldots(4)$

Y = response variable, β = coefficients of all 15 predictors in the model 2, ε_i = error term.

Gujarati (2003) defined heterogeneity is widespread problem especially in case of large amount of cross-sectional data. Neter et al. (1996) and Rawlings, Pantula, and Dickey (2001) presented the weighted least square model using the weight $=1/\sigma^2$ as

$$WY = WX\beta + W\varepsilon, \quad Y^* = X^*\beta + \varepsilon^* \ldots(5) \qquad \hat{\beta}_W=(X^{*\prime}X^*)^{-1}X^{*\prime}Y^* \ldots(6)$$

The feature variables studies are

Table 2.2 Features of variables used in this study

1	Average yield of wheat crop	7	No. of pest spry	13	Planting period November (yes or no)
2	Fertilizer urea in kg/acre	8	No. of weeds spry	14	Land is irrigated (yes or no)
3	Fertilizer DAP in kg/acre	9	Seed treatment (yes or no)	15	Large farmers area >25 acres (yes or no)
4	Other fertilizers in kg/acre	10	Soil type chikny loom (yes or no)	16	Seed type (un-certified or certified)
5	No. of water	11	Varieties Galaxy, Faisalabad, Sahr (GFS) yes or no		
6	Seed quantity used in kg/acre	12	Harvesting period April, 1-20 (yes or no)		

2.3 Goodness of fit and model selection procedures

Regression modeling is itself a very crucial task in the applied statistical analysis. Gujarati and Porter (2004) elaborated as crop yield have diverse phenomenon and depending on various predictors. According to Gujarati (2003) following are the basic consideration for the best regression model selection criterion.

Lower value of mean square error (MSE) and higher value of coefficient of determination R^2 should be preferred

$$MSE = \frac{\sum_i^n (y_i - \bar{y}_i)^2}{n} \quad \ldots\ldots (7) \qquad R^2 = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2} \quad \ldots\ldots (8) \qquad R^2_{adj} = 1 - (1 - R^2)\left(\frac{n-1}{n-k}\right) \quad \ldots\ldots (9)$$

Significance of the regression coefficient is determine by P-value or t-statistic and the significance of the overall model is determined by the F-statistic

$$t = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \quad \ldots\ldots (10) \qquad F = \frac{SSR_{eg}/k - 1}{SSR_{es}/n - k} \quad \ldots\ldots (11)$$

It is necessary to take into account the logically or theoretically relevance of the predictors to the response and their statistical significance in the model building (Gujarati, 2003). The Akaike Information Criteria (Akaike, 1973) introduced the model selection criteria with unified way as log-likelihood functions with simple penalties and lower value of AIC should be preferred for best regression model(Banks & Joyner, 2017; Dziak, Coffman, Lanza, & Li, 2012; Gujarati, 2003).

$$\text{AIC} = e^{2k/n}\frac{\sum \hat{u}_i^2}{n} = e^{2k/n}\frac{RSS}{n} = \qquad \ln \text{AIC} = \left(\frac{2k}{n}\right) + \ln\left(\frac{RSS}{n}\right) \ldots (12)$$

"k" is the no. of regressors including the intercept and "n" is the number of observations. The term "2k/n" is defined as penalty factor for AIC (Gujarati, 2003). Akaike weight "W_i" is a value lies between 0 and 1, and the sum of Akaike weights is unity which can be considered as the probability that a given model is the best regression model. The higher value of AIC weights indicates that model is good fit and weights of all models sum to unity which means weight gives a probability about each model's selection as a best model (Banks & Joyner, 2017). The evidence ratio "ER" is used to compare the efficiency of any two models

$$W_i(AIC) = \frac{exp\left\{-\frac{1}{2}\Delta_i(AIC)\right\}}{\sum_{k=1}^n exp\left\{-\frac{1}{2}\Delta_i(AIC)\right\}} \quad \ldots\ldots (13) \qquad ER = \frac{exp\left(-\frac{1}{2}\Delta_{best}\right)}{exp\left(-\frac{1}{2}\Delta_i\right)} = \frac{Weight_{best}(AIC)}{W_j(AIC)} \quad \ldots\ldots (14)$$

The Schwarz (1978) information criterion (SIC) presented that lower value of (SIC) should be preferred (Gujarati, 2003; Neath & Cavanaugh, 1997)

$$\text{SIC} = n^{k/n}\frac{\sum \hat{u}_i^2}{n} = n^{k/n}\frac{RSS}{n} = \qquad \ln \text{SIC} = \left(\frac{k}{n}\right) lnn + \ln\left(\frac{RSS}{n}\right) \ldots (15)$$

The term "((k/n) ln n)" is defined as the penalty factor. "k" is the no. of regressors including the intercept and "n" is the number of observations under study.

Normality, linearity and constant error variance is checked by P-P plots and graphical presentation. The multicollinearity is checked by VIF and non-constant error variance is also checked by Breusch–Pagan test which is developed in 1979 by Trevor Breusch and Adrian Pagan (Breusch & Pagan, 1979) and Koenker test developed by Koenker and Bassett (Koenker, 1981). The autocorrelation is checked by Durbin–Watson "d" statistic and reliability analysis for degree of relevance and consistency of predictors with reference to the measure of response is checked by Cronbach's Alpha "α" (Bland & Altman, 1997; Qayyum, 2011; Sekaran & Bougie, 2016)

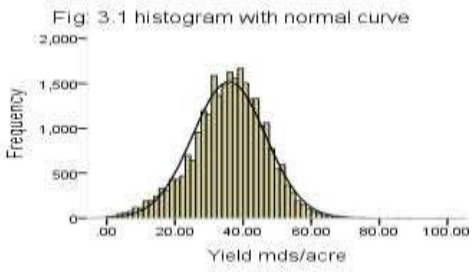
$$\text{Durbin–Watson statistic} = d = \frac{\sum_{t=2}^{t=n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n}\hat{u}_t^2} \ldots (16) \qquad \text{Cron bach's Alpha} = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right) \ldots (17)$$

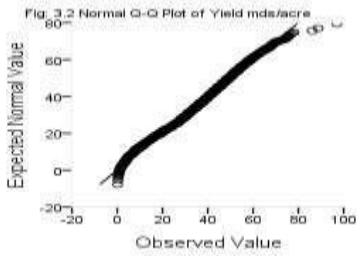
"K" is the no. of items, s_i^2 is the variance of ith item and s^2_T is the variance aggregates items

III. Results and Discussion

3.1 Normality of Data

In most of the applied statistical analysis, check for normality is a pre-requisite (Gujarati, 2003). For large data set central limit theorem supports normality. The graphical presentation (fig: 3.1) check the normality of response variable in data set by histogram with normal plots and Q-Q plots (fig:3.2) predicts to follows the normal distribution.





3.2 Hierarchical regression analysis

For the hierarchical regression analysis partition of predictors is carried out in two blocks (tab: 3.2). Block 1 contains seven quantitative and block 2 contains eight binary categorical predictors use to seek out best regression model. Model 1 based on seven quantitative predictors and model 2 is based on all 15 predictors (quantitative and categorical). Table 3.5 revealed, for data set 1 (collected data) the R^2 for model 1 is 0.197 and for model 2 is 0.266 with change in R^2 (ΔR^2) is 0.068. The value of F-statistic is found significant for both models. The MSE are 84.01 in model 2 and 91.82 model 1, shows significant decrease in model 2, elaborate for the projection of wheat yield, the model 2 found good. To enhance the reliability and precision of results for model 2, data preprocessing technique is opted, divide the data set in four sets and three clusters which rise R^2, fall the MSE (tab:3.5), avoid the wide dispersion (tab:3.3) and increases the reliability (tab:3.4). It is evident (tab: 3.3, fig:3.3), the variance in the data set is gradually decreases from data sets 1 to data set 4, separately for all the variables indicate the upturns of precision for data sets. Cronbach's alpha investigates the internal consistency or average correlation of items in a survey instrument to examine its reliability (Cronbach, 1951; Schmitt, 1996). The value of Cronbach's alpha increases from 0.347 to 0.642 varying from data set 1 to cluster 1-3 (table 3.4) predicts reliability of data sets increases. Tab: 3.5, for cluster 1 (data set 2), R^2_{adj} in model 2 is 0.287 which is larger than from model 1 R^2_{adj} 0.220, ΔR^2 is 0.068 with the significance of F-statistic and MSE is lower in model 2 is 60.39 from model 1 is 66.06, for cluster 2 (data set 3) model 2, R^2_{adj}= 0.823 and MSE= 11.24 against model 1 R^2_{adj} =0.718 and MSE= 17.93 and for cluster 3 (data set 4) model 2, R^2_{adj}= 0.862 MSE= 7.10 against model 1 R^2_{adj}=0.743 and MSE= 13.25. It shows for all clusters high value of R^2_{adj} obtained model 2 ($R^2_{model2i} > R^2_{model1i}$) and low value of MSE is obtained from model 2 ($MSE_{model2i} < MSE_{model1i}$). It is elaborating that for all clusters model 2 is better than from model 1 and will be applied for the projection of wheat crop yield at their various levels varying from data set 1 to 4. The data preprocessing improved the coefficient of determination from to 0.265 to 0.862 decreases the MSE from 84.01 to 7.10, upturns the reliability from 0.347 to 0.642 and raises the precision of results with the significance of the F statistic.

Table 3.2 Application of hierarchical regression analysis of block 1 and block 2

Hierarchical regression	Response variable	Name of Regressors	Blocks
Model 1	Average yield of wheat crop in mds/acre Punjab, Pakistan	1. Fertilizer urea in kg/acre 2. Fertilizer DAP in kg/acre 3. Other fertilizers in kg/acre 4. No. of water 5. Seed quantity used in kg/acre 6. No. of pest spry 7. No. of weeds spry	Block1=07 (predictors)
		1. Seed treatment (yes or no) 2. Soil type chikny loom (yes or no) 3. Application of advanced varieties(yes or no) 4. Harvesting period April (1-20)(yes or no) 5. Planting period Nov (yes or no) 6. Land is irrigated(yes or no) 7. Large farmers area >25 acres (yes or no) 8. Seed type (home based or certified)	Block 2 = 08 (predictors)
Model 2		1. Fertilize urea in kg/acre 2. Fertilizer DAP in kg/acre 3. Other fertilizers in kg/acre 4. No. of water 5. Seed quantity used in kg/acre 6. No. of pest spry 7. No. of weeds spry: 8. Seed treatment (yes or no) 9. Soil type chikny loom (yes or no) 10. Application of advanced varieties (yes or no) 11. Harvesting period April (1-20) (yes or no) 12. Planting period November (yes or no) 13. Land is irrigated(yes or no) 14. Large farmers area >25 acres(yes or no) 15. Seed type (home based or certified)	Total =15 (predictors)

Table 3.3 Variances of the data sets

Parameters	Variances				Parameters	Variances			
	Data set 1	cluster 1	Cluster 2	cluster 3		Data set1	cluster 1	cluster 2	cluster 3
Average yield of wheat	114.35	84.70	63.65	51.64	Seed treatment (yes or no)	0.055	0.045	0.007	0.003
Fertilizer Urea in kg/acre	1023.4	808.8	721.0	597.1	Soil type	0.223	0.207	0.052	0.041
Fertilizer DAP in kg/acre	291.26	188.5	86.45	50.45	Varieties (GFS) (yes or no)	0.184	0.121	0.033	0.028
Other fertilizers in kg/acre	272.60	159.2	23.51	15.20	Harvest April (1-20) (yes or no)	0.247	0.222	0.084	0.068
No. of water	2.408	2.015	2.072	1.849	Planting November (yes or no)	0.113	0.078	0.035	0.020
Seed quantity in kg/acre	33.51	27.14	14.97	13.63	Land is irrigated (yes or no)	0.044	0.04	0.089	0.062
N. of pest spry	0.198	0.196	0.044	0.040	farmers area >25 acres (yes or no)	0.227	0.137	0.035	0.019
No. of weed spry	0.178	0.127	0.076	0.055	Seed type (un-certified or certified)	0.193	0.143	0.037	0.010

fig: 3.3, Graph for the variances in data sets

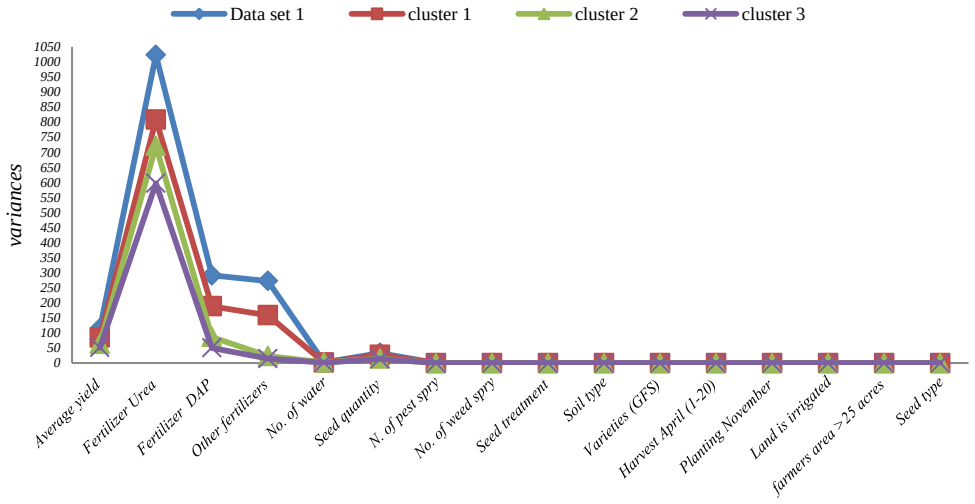


Table 3.4 Reliability Statistics for data sets

Cronbach's Alpha	Data set 1	cluster 1	cluster 2	cluster 3
	0.347	0.387	0.635	0.642

Table 3.5 Hierarchical regression analysis for data sets

Data sets	Models	R	R^2	Adj R^2	S.E of estimate	Change Statistics R^2	Change Statistics Sig. F	MSE
Data set 1	1	0.444	0.197	0.197	9.58	0.197	0.00	91.82
	2	0.515	0.266	0.265	9.16	0.068	0.00	84.01
Data sets 2 (Cluster 1)	1	0.470	0.221	0.220	8.12	0.221	0.00	66.06
	2	0.537	0.289	0.287	7.77	0.068	0.00	60.39
Data sets 3 (Cluster 2)	1	0.856	0.732	0.718	4.23	0.732	0.00	17.93
	2	0.917	0.842	0.823	3.35	0.110	0.00	11.24
Data sets 4 (Cluster 3)	1	0.891	0.795	0.743	3.64	0.795	0.00	13.25
	2	0.960	0.921	0.862	2.66	0.127	0.005	7.10

For the large data set, the graphical presentation is better to predicts or detect the normality (Gujarati, 2003). Figure 3.1.1 to 3.1.12 shows histogram with normal curve and P-P plots for residuals to predicts the error term fallows normality and scatter plots for the predicted and residual to shows to exhibits the pattern of heterogeneity in data for data set 1-4. The error term for the original data set of 26430 crop cut experiments (C.C.E'S), data set 2 (cluster 1) having 6034 sample points, data set 3 (cluster 2) comprising 145 sample points and data set 4 (cluster 3) based on 36 sample points exhibits to fallow the normality but fig: 3.1.3 and 3.1.6 agreements of predicted with residual and significant values of Breusch-Pagan and Koenker test (tab:3.7), shows that the error term did not fallows the homogeneity and heterogeneity exist for data set 1-2 while In cluster 2 and 3, the heterogeneity disappeared in the data set and fulfills the assumption of homogeneity.

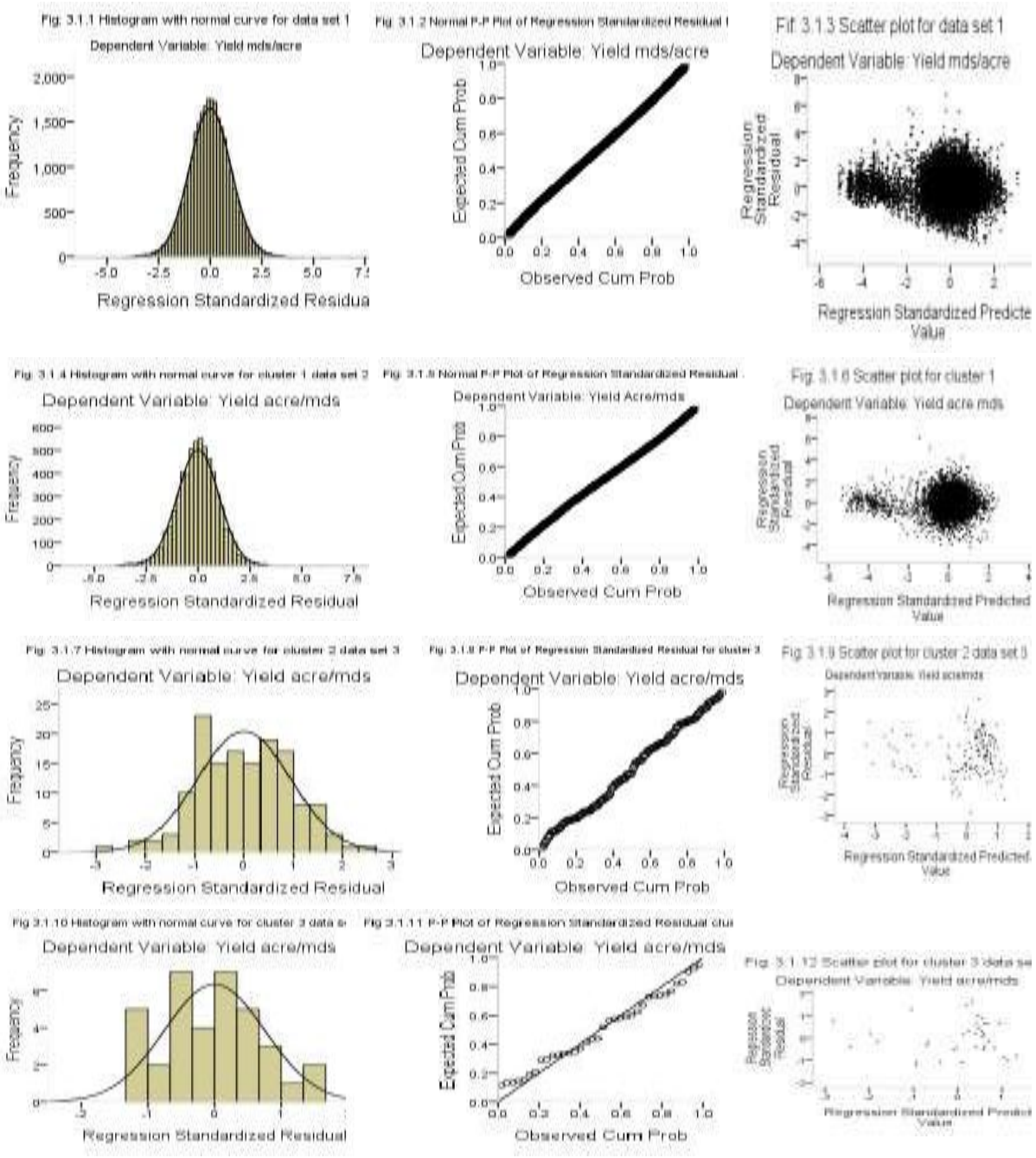


Table 3.6 indicates that value of adjusted R^2 is gradually increases from data sets 1 to cluster 1-3 from 0.265 to 0.862 ($R_i^2 < R_{i+1}^2$) and MSE is decreasing from 84.01 to 7.10 ($MSE_i > MSE_{i+1}$). The values of Durban-Watson test for the detection of autocorrelation indicates that data are free from autocorrelation. According to Gujarati 2004, there should be essential the presence of non-constant error variance in the data in case of large amount of cross-sectional data, so before selection of best

model, firstly proceed to check the heterogeneity in the data set. According to Breusch-Pagan test and Koenker test (tab: 3.7) for the detection of heterogeneity revealed that for data set 1 and data set 2 (cluster 1), the significance value are less than 0.05, the rejection of null hypothesis indicates that there are heterogeneity in data set and for cluster 2 and cluster 3, the significance values are greater than 0.05, acceptance of null hypothesis, indicates homoscedasticity in data set. Because of the presence of heterogeneity in the data set 1 and cluster 1, weighted least square are also applied to the data with the aim to select the best regression model and from table 3.6, the coefficient of determination is gradually increases from data set 1 to cluster 1-3 from 0.265 to 0.862 in multiple regression analysis and its values are increases by changing the estimation procedure by weighted least square (WLS) from 0.265 to 0.315 for data set 1, from 0.287 to 0.376 for cluster 1, from 0.823 to 0.880 for cluster 2, from 0.862 to 0.954 for cluster 3. The values of mean square error (MSE) are also very low in WLS as compared to multiple regression analysis from 84.04 to 1.638 for data set 1, from 60.39 to 1.67 for cluster 1, from 11.24 to 1.549 for cluster 2 and from 7.10 to 1.75 for cluster 3. It is cleared here that after data preprocessing the R^2 increasing and MSE are decreasing. For hierarchical multiple regression the no. of significant regressors are 13, 12, 07 and 04 for cluster 1, 2, 3, 4 respectively and for weighted least square by taking the weights as $w_{i=} 1/\sigma^2$ they are 13 both for data set 1 and 2 at 5% level of significance.

Table 3.6 Hierarchical regression models for un-weighted and weighted least square

Data sets	Hierarchical MLR					Hierarchical WLS Regression				
	R^2	Adj R^2	D.W	MSE	F(Sig)	R^2	Adj R^2	D.W	MSE	F(Sig)
Data set 1	0.266	0.265	1.14	84.01	637.2**	0.316	0.315	1.13	1.638	811.6**
Data set 2 (cluster 1)	0.289	0.287	1.37	60.39	162.9**	0.378	0.376	1.37	1.670	243.4**
Data set 3 (cluster 2)	0.842	0.823	1.27	11.24	45.7**	0.893	0.880	1.28	1.549	71.5**
Data set 4 (cluster 3)	0.921	0.862	1.73	7.10**	15.6**	0.974	0.954	1.68	1.750	48.99**

** shows significant results

Table 3.7 Breusch-Pagan and koenker test for heterogeneity in error term

	Data set 1		Data set 2 (Cluster 1)		Data set 3 (Cluster 2)		Data set 4 (Cluster 3)	
	LM	Sig	LM	Sig	LM	Sig	LM	Sig
Breusch–Pagan test	285.09	0.000	133.5	0.000	22.3	0.099	10.5	0.78
Koenker test	211.68	0.000	90.5	0.000	23.6	0.072	16.4	0.36

Null hypothesis (H_0): Heteroscedasticity not present (homoscedasticity) in data
If sig-value less than 0.05, reject the null hypothesis.
Note: Breusch-Pagan test is a large sample test and assumes the residuals to be normally distributed

3.3 Akiake and Schwarz information criterion

Burnham and Anderson (2002) introduced the model selection method called Akiake and schwarz information criteria. Table 3.8 indicates that values of AIC are 4.43, 4.10, 2.52, and 2.26 and SIC are 4.44, 4.12, 2.85 and 2.97 for data set 1 to 4, respectively. These values shows that AIC are decreasing ($AIC_i > AIC_{i+1}$). Moreover it has been observed that the highest value of the AIC weights is obtained by cluster 3 with probability = 0.38 against cluster 2, 1, dataset 1which is 0.34, 0.15 and 0.13 ($AIC_{weights} < AIC_{weights\ i+1}$) which support that cluster 3 is best model as highest value of AIC weights is obtained from cluster 3 while the sum of the weights probability is unity. The evidence ratio of AIC indicates for cluster 3 against data set1 (w_3/w_d) is 2.96 which means cluster 3, is 2.96 more likely to data set 1, evidence ratio for cluster 3 against cluster 1 (w_3/w_1) is 2.51 indicated that model cluster 3 is 2.51 more likely to model cluster 1 and likeliness of cluster 3 on cluster 2 (w_3/w_2) is 1.14, which mean cluster 3 predictive model is 1.14 more likely to clusters 2.

Table 3.8 AIC, SIC, Weights (AIC) and ER (AIC) for data sets

Data sets	N	K	AIC	ΔAIC	Wi	ER	SIC
Data set 1	26430	16	4.43	2.17	0.13	2.96	4.44
Dataset2(cluster1)	6034	16	4.10	1.84	0.15	2.51	4.12
Dataset3(cluster2)	145	16	2.52	0.26	0.34	1.14	2.85
Data set 4 (cluster 3	36	16	2.26	----	0.38	----	2.97

Cluster 3 shows high R^2, $AIC_{weights}$ and less value of MSE, AIC, SIC, with better evidence ratios against others. The model based on cluster 3 is best regression model for the projection of wheat crop production and clustering improved the results of the estimates. Due to existence of heterogeneity the WLS approach gives the better results for data set 1 and cluster 1 while WLS approach is not being applied on cluster 2 and 3 because their exist homoscedasticity in data (table 3.7). Table 3.9 to 3.11 shows the significance of regressors and its slope coefficients for various levels to shows the effects of the regressosrs on the response variable. Hierarchical multiple and WLS regression, data set 1, there are 13 regressors are significant i.e. Fertilizer application urea, DAP, other fertilizers, water, pest spry, weeds spry, seed treatment, soil type, harvesting period April (1-20), planting period November, irrigated land, large farmers and seed type. The cluster 1, the regressors are significant for MLR i.e Fertilizer application urea, DAP, other fertilizers, seed quantity, pest spry, weeds spry, soil type, harvesting period April (1- 20), planting period November, irrigated land, large farmers and seed type while for WLS all regressors are significant except seed treatment and varieties. At cluster 2, the significant regressors are fertilizer application urea, DAP, other fertilizers, pest spry, soil type, harvesting April 1-20 and irrigated land. For cluster 3 the significant regressors are fertilizer application DAP, other fertilizers, soil type and harvesting April 1-20. These analysis shows that all fertilizers, soil type, harvesting period advanced varieties, spry, seed treatment planting time, and irrigated land are effective factors for better response of wheat crop forecast to meet the food security condition and precise estimate of forecasting for food availability for the growing population explosion need in future.

Table 3.9 Hierarchical regression for data set 1

Data set 1 Coefficients	Hierarchical regression MLR, data set 1				Hierarchical weighted regression data set 1			
	B	t.	Sig.	VIF	B	t.	Sig.	VIF
(Constant)	8.307	14.03	0.000		8.099	14.35	0.000	
Urea	0.053	24.49	0.000	1.48	0.051	24.14	0.000	1.58
DAP	0.115	32.21	0.000	1.17	0.111	31.75	0.000	1.19
Other fertilizer Qty	0.055	15.70	0.000	1.06	0.053	14.52	0.000	1.05
No of water:	0.130	2.74	0.006	1.70	0.142	3.00	0.003	1.86
Seed quantity	-0.010	-0.89	0.372	1.25	-0.004	-0.35	0.726	1.28
Pest spry	1.106	8.51	0.000	1.05	1.084	8.21	0.000	1.06
Weeds spry	1.290	8.94	0.000	1.17	1.272	8.92	0.000	1.22
Seed treatment	0.818	3.36	0.001	1.03	0.935	3.89	0.000	1.03
Soil type (chikni loom)	2.836	23.16	0.000	1.05	2.886	23.39	0.000	1.06
Varieties (GFS)	0.064	0.48	0.629	1.03	0.142	1.07	0.285	1.03
Harvesting April, 1-20	1.831	14.83	0.000	1.18	1.866	15.19	0.000	1.19
Sowing November	2.406	13.87	0.000	1.07	2.352	13.97	0.000	1.10
Irrigated land	11.099	33.41	0.000	1.52	11.289	39.34	0.000	1.73
Large farmers	0.258	2.14	0.032	1.04	0.259	2.19	0.029	1.04
Seed type	0.891	6.66	0.000	1.09	0.829	6.44	0.000	1.09

Table 3.10 Hierarchical regression for data set 2 (cluster 1)

Data set 2 (cluster 1) Coefficients	Hierarchical regression MLR, cluster1				Hierarchical weighted regression cluster 1			
	B	t.	Sig.	VIF	B	t.	Sig.	VIF
(Constant)	12.971	11.8	0.000		12.07	11.4	0.000	
Urea	0.052	11.7	0.000	1.619	0.048	11.3	0.000	1.735
DAP	0.127	15.8	0.000	1.208	0.121	16.2	0.000	1.221
Other fertilizer Qty	0.061	7.5	0.000	1.055	0.055	7.1	0.000	1.053
No of water:	0.113	1.2	0.240	1.848	0.329	3.4	0.001	2.108
Seed quantity	-0.09	-4.1	0.000	1.299	-0.078	-3.6	0.000	1.352
Pest spry	0.68	2.9	0.004	1.077	0.691	2.9	0.004	1.075
Weeds spry	0.903	2.9	0.003	1.199	1.262	4.2	0.000	1.289
Seed treatment	-0.023	0.0	0.963	1.048	0.452	1.0	0.338	1.049

Soil type (chikni loom)	1.691	7.5	0.000	1.057	1.785	7.8	0.000	1.05
Varieties (GFS)	-0.306	-1.0	0.301	1.059	0.003	0.0	0.992	1.075
Harvesting April, 1-20	1.578	6.5	0.000	1.315	1.514	6.3	0.000	1.342
Sowing November	2.829	7.6	0.000	1.084	2.657	7.6	0.000	1.145
Irrigated land	11.26	17.8	0.000	1.616	11.15	21.6	0.000	1.925
Large farmers	0.555	2.0	0.046	1.062	0.618	2.3	0.023	1.082
Seed type	1.04	3.7	0.000	1.113	0.749	2.9	0.003	1.116

Table 3.11 Hierarchical regression analysis for data set 3 (cluster 2) and data set 4 (cluster 3)

Coefficients	Hierarchical regression MLR cluster 2				Hierarchical regression MLR cluster 3			
	B	t.	Sig.	VIF	B	t.	Sig.	VIF
(Constant)	6.301	1.00	0.317		15.939	0.97	0.44	
Urea	0.047	2.02	0.046	5.08	0.019	0.39	0.69	6.93
DAP	0.197	3.52	0.001	3.49	0.440	2.38	0.027	8.49
Other fertilizer Qty	0.167	2.362	0.02	1.51	0.366	2.30	0.032	1.89
No of water:	0.238	0.512	0.608	5.67	0.325	0.34	0.739	8.44
Seed quantity	-0.137	-1.065	0.29	3.19	-0.527	-1.885	0.074	5.26
Pest spry	3.261	1.983	0.049	1.54	5.194	1.61	0.122	2.06
Weeds spry	-1.985	-0.90	0.37	4.71	-1.218	-0.237	0.815	7.35
Seed treatment	-4.941	-1.332	0.18	1.24	-7.13	-0.648	0.524	2.01
Soil type (chikni loom)	7.526	5.03	0.00	1.49	10.42	3.05	0.006	2.37
Varieties (GFS)	1.211	0.594	0.554	1.74	0.40	0.096	0.925	2.34
Harvesting April, 1-20	5.403	3.88	0.00	2.08	8.23	2.86	0.01	2.79
Sowing November	4.272	1.806	0.073	2.52	1.64	0.143	0.888	13.15
Irrigated land	9.838	3.813	0.00	7.63	7.4	0.964	0.347	18.17
Large farmers	1.42	0.672	0.503	2.01	2.83	0.457	0.653	3.56
Seed type	0.53	0.328	0.74	1.25	1.916	0.359	0.723	1.43

IV. CONCLUSIONS

In current scenario, food production does not meet with the demand of food particularly for a country like Pakistan. Better prediction and investigation of significant factors for food may lead to overcome the issue of food security to some extent. In this study, effort is made to model the agriculture data on wheat crop. In hierarchical regression analysis block 2 (model 2 with 15 predictors) gives the better results. Amongst models of four data sets, data on districts (cluster 3) shows highest value of R^2, AIC weights and less value of MSE, AIC and SIC with better ER as compare to others data sets. These analysis shows that all fertilizers, soil type, harvesting period advanced varieties, spry, seed treatment planting time and irrigated land are significant factors to get the better yield of wheat crop to tackle food security. Data preprocessing (data sets clusters) is power full source, improve the precision of estimates to reduce the variation in variable and enhanced the reliability in the data set at various levels. The hierarchical regression analysis on four data sets by two estimation methods known as least square or weighted least square with the checks of R^2, MSE and information criterion metrics gives better understanding for precise estimate and best model selection for the projection of wheat crop production to cope with the food security in the region and efficacies of the predictors at their different level.

REFERENCES

1. Ahmad, M., & Farooq, U. (2010). The state of food security in Pakistan: Future challenges and coping strategies. *The Pakistan Development Review*, 903-923.

2. Bajkani, J. K., Ahmed, A., Afzal, M., Jamali, A. R., Bhatti, I. B., & Iqbal, S. (2014). Factors Affecting Wheat Production in Balochistan Province of Pakistan". *Journal of Agriculture and Veterinary Science, 7*(12), P73-80.

3. Banks, H. T., & Joyner, M. L. (2017). AIC under the framework of least squares estimation. *Applied Mathematics Letters, 74*, 33-45.

4. Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *Bmj, 314*(7080), 572.

5. Breusch, T., & Pagan, A. (1979). A simple test for heteroscedasticity and random. *Econometrica, 47*(5).

6. Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference A practical information theoretic approach second edition.

7. Cohen, B. H. (2008). *Explaining psychological statistics*: John Wiley & Sons.

8. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika, 16*(3), 297-334.

9. CRS. (2008). *Punjab agriculture statistics 2003-04 to 2007-08*.

10. Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). Sensitivity and specificity of information criteria.

11. ESP. (2013). Economic survey of Pakistan(ESP).

12. ESP.(2019). *Economic Survey of Pakistan*.

13. FAO. (2009). *17- How to Feed the World in 2050*.

14. Farooq, A., Ishaq, M., Yaqoob, S., & Sadozai, K. N. (2007). Varietal adoption effect on wheat crop production in irrigated areas of NWFP. *Sarhad Journal of Agriculture, 23*(3), 807.

15. Gujarati, D. N. (2003). Basic Econometrics" fourth edition McGraw-Hill. *New York*.

16. Gujarati, D. N., & Porter, D. C. (2004). Basic econometrics (ed.) McGraw-Hill. *Irwin, a business*.

17. Hussain, A. (2010). *Economic analysis of staple food-grain crops: varieties'input-output comparison, economic practices and significance in the economy of district Swat.* University of Peshawar.

18. Islam, M. (2015). *Factors affecting major food crops production, a case study of district Bahawalpur.* (M.Phil Applied Statistics analysis on Agriculture Data), The islamia university of Bahawalpur.

19. Kagan, C. R. (2016). At the nexus of food security and safety: opportunities for nanoscience and nanotechnology: ACS Publications.

20. Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of econometrics, 17*(1), 107-112.

21. McKenzie, F. C., & Williams, J. (2015). Sustainable food production: constraints, challenges and choices by 2050. *Food Security, 7*(2), 221-233.

22. Neath, A. A., & Cavanaugh, J. E. (1997). Regression and time series model selection using variants of the Schwarz information criterion. *Communications in Statistics-Theory and Methods, 26*(3), 559- 580.

23. Nelson, G. C., Rosegrant, M. W., Palazzo, A., Gray, I., Ingersoll, C., Robertson, R., . . . Ringler, C. (2010). *Food security, farming, and climate change to 2050: scenarios, results, policy options* (Vol. 172): Intl Food Policy Res Inst.

24. Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (Vol. 4): Irwin Chicago.

25. Neter, J., Wasserman, W., & Kutner, M. H. (1989). Applied linear regression models.

26. Pedzahur, E. (1997). Multiple regression in behavioral research: Explanation and prediction. *London, UK: Wadsworth, Thompson Learning*.

27. Petrocelli, J. V. (2003). Hierarchical multiple regression in counseling research: Common problems andpossibleremedies. *Measurement and evaluation in counseling and development, 36*(1),9-22.

28. Qayyum, A. (2011). *Model based wheat yield estimation in the Punjab, Pakistan.* GCUNIVERSITY LAHORE.

29. Qayyum, A., & Pervaiz, M. K. (2013). A detailed descriptive study of all the wheat production parameters in Punjab, Pakistan. *African Journal of Agricultural Research, 8*(31), 4209-4230.

30. Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis: a research tool*: Springer Science & Business Media.

31. Raza, S. A., Ali, Y., & Mehboob, F. (2012). Role of agriculture in economic growth of Pakistan.

32. Rosenthal, S. (2017). Regression analysis, linear. *The International Encyclopedia of Communication Research Methods*, 1-15.

33. Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment, 8*(4), 350.

34. Sekaran, U., & Bougie, R. (2016). *Research methods for business: A skill building approach*: John Wiley & Sons.

35. Tariq, A., Tabasam, N., Bakhsh, K., Ashfaq, M., & Hassan, S. (2014). Food security in the context of climate change in Pakistan. *Pakistan Journal of Commerce and Social Sciences (PJCSS), 8*(2), 540- 550.

36. UAF. (2014). *Vision 2030. Sustainable Agriculture through learning discovery outreach*. Universityof agriculture Faisalabad.

37. Wampold, B. E., & Freund, R. D. (1987). Use of multiple regression in counseling psychology research: A flexible data-analytic strategy. *Journal of Counseling Psychology, 34*(4), 372.