Proceedings of The National Conference
On Undergraduate Research (NCUR) 2020
Montana State University, Bozeman MT
March 26-28, 2020

Analysis of Three Yield Components in Relation to Grain Yield of Wheat in Three Environments in Southeastern Idaho

Present Karmacharya
[1]Department of Mathematics & Statistics
Idaho State University
Pocatello, Idaho, USA

[2]Department of Plant Sciences
University of Idaho
Aberdeen Research & Extension Center
Aberdeen, Idaho, USA

Faculty Advisors: Dr. Xiaoxia Xie[1] and Dr. Jianli Chen [2]

Abstract

Increasing grain yield of wheat is an essential research goal to meet the demand of the increasing population. One important approach to increase grain yield is to manipulate major yield components, including productive tiler number per unit area (PTN), fertile spikelet number per spike (SNS), and thousand kernel weight (TKW). The present study aimed to elucidate the contributions of the three yield components to grain yield using the data collected in three field trials in Aberdeen in 2018 and 2019, and Ashton, Southeastern Idaho in 2018. Two software packages python and Minitab were compared to find the optimal analysis in finding the contribution of yield components with respect to grain yield of wheat (dependent variable). Statistical analysis methods like regression, standardization were used extensively on the wheat datasets which assisted in targeting the contribution effects of each yield component (components that help to result in yielding). The three datasets from the three environments were analyzed separately. The present study found that TKW contributed the most to grain yield in two Aberdeen trials, while contributed the least in Ashton trials in 2018. These novel findings explain why selecting high yield wheat cultivars should test breeding lines in multiple environments and multiple years, and the best yield components architecture may be different in different environments. Additional diverse trial data are needed to confirm these findings obtained from the present study.

Keywords: Grain yield (GY), Predictive Analytics, Yield components (PTN, TKW, SNS)

1. Introduction

The population of the world is continuously increasing, it is expected to grow specifically in developing countries that reside in Asia and Africa, and by 2050, the population could reach 9.3-10 billion (UN, 2017). Wheat is the most widely grown crop in the world and provides 20% of the daily protein and the food calories for 4.5 billion people. To support the increased population, wheat production will need to grow by at least 1.1 per year which would result in a 60% increase in yield by 2050 (Cordoba, 2011). Yield and agronomic improvements are crucial to ensure the demand for the future population.

Grain yield improvement can be achieved by the improvement of three major yield components, including productive tiller numbers per unit area (PTN), fertile spikelet number per spike (SNS), and thousand kernel weight (TKW). Productive tiller number (PTN) is defined by the number of tillers that produce spikes with seed set and is a very important component of grain yield (Li et al., 2011). Fertile spikelet number per spike (SNS) in wheat is defined by the number of spikelets that have produced seed. Thousand Kernel Weight (TKW) is phenotypically the most stable yield component(Sun et al., 2009) which has consistently higher heritability compared to kernel per spike.

There is a need for understanding how we can optimize yield components to increase grain yield as it is essential to ensure the demand for the future ever-growing population. There is limited research studying simultaneously for three major yield components concerning grain yield in diverse environments.

The purpose of this study is to investigate the contribution percentage of yield component for the grain yield of wheat by utilizing a fitted regression line, standardization of datasets, and another statistical analysis method. This study uses ANOVA multiple linear regression analysis to find the contribution percentage of each yield component by utilizing the coefficient table output that is supported by the residual plot and probability plot.

2. Methods

2.1 Data

Data used in the present study were acquired from the University of Idaho Aberdeen Research and Extension Center by Dr. Jianli Chen. The data was collected under the support of USDA-ARS, ….including grain yield and three major yield components(fertile spikelet's per spike, SNS; productive tiller number per unit area, PTN; and thousand kernel weight, TKW) in two field trials in Aberdeen and Ashton, Southeastern Idaho in 2018.

2.2 Data Mining

The datasets in excel were first imported to python, then analyzed by utilizing various python packages like Pandas, NumPy, Scikit-learn, and Matplot. The datasets were explored utilizing python pandas package to discern which part of datasets requires cleaning. Afterward, the datasets were cleaned i.e. removing outliers, null components, and filtering for required variables. The process of data cleaning helps us to focus on our main goal by filtering out unnecessary inputs for our paper. The cleaned data is then transformed by standardizing using python's Sklearn library for preprocessing data. The standardizing process is essential for the dataset since the variables are measured in different units as variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bias(Institute of Medicine (US), 2013). Since our main goal is to find the contribution percentage of each yield component with respect to the grain yield of wheat. The variables must contribute equally.

2.3 Statistical Analytics

The standardized data was used in the Analysis of Variance (ANOVA) for multiple linear regression analysis with the help of python's Statsmodels packages. ANOVA consists of calculations that provide information about levels of variability within a regression model and form a basis for significance tests. Multiple linear regression attempts to fit a regression line for a dependent variable (grain yield) using more than one independent variable (PTN, TKW, SNS). Also, the Minitab for data analytics was used for the paper where the standardized was imported into the Minitab spreadsheet. Then, the datasets were then fitted into the regression line. This provides us with the ANOVA and model summary of the dataset. The output is attached with residual plots, probability plots, and confidence intervals. These are necessary components to understand the model and develop appropriate intuition for the dataset to find the contribution of each yield component of the grain yield of wheat. Residual plots help to verify if the regression model is valid by checking constant variance. Probability plots help to verify if the normality condition of the model is satisfied which aids in checking the validity of prediction interval and confidence interval. The ANOVA table, probability plots, residual plots, confidence interval, R-square value, VIF, P-values, and F-values were carefully analyzed to find and verify the contribution percentage of yield components (PTN, TKW, SNS) of the grain yield (GY) of wheat.

3. Results and Discussion

The coefficient table that was obtained when the wheat dataset was fitted with the regression line provides the contribution percentage of each yield component (PTN, TKW, SNS) for the Aberdeen and Ashton locations. The contribution percentage was obtained by summing up three yield component coefficient values and dividing each of them by the summed values (i.e. percentage contribution of PTN = coefficient of PTN value/sum (of PTN, TKW, SNS) coefficient value). This gives us the contribution percentage for the yield components.

Table 1: Wheat yield coefficient for Aberdeen (AB) for year 2018 &19, and Ashton (ASH) location

Term	Coefficient	T-value	P-value	VIF
ABSNS-2018	0.2809	3.21	0.002	1.58
ABPTN-2018	0.2319	2.94	0.004	1.28
ABTKW-2018	0.4797	5.31	0.000	1.68
ABSNS-2019	0.3239	0.000	0.000	1.18
ABPTN-2019	0.1735	0.003	0.003	1.13
ABTKW-2019	0.8038	0.000	0.000	1.28
ASHSNS-2018	0.6041	9.56	0.000	1.15
ASHPTN-2018	0.4572	7.05	0.000	1.21
ASHTKW-2018	0.2864	4.41	0.000	1.22

The values obtained in Table 1 are exported from the Minitab where the standardized data values were fitted to the regression line providing us the regression equation. In table 1, we can notice that VIF for each component is low. The VIF is the variance inflation factor, which is the quotient of the variance in a model with multiple terms by the variance of the model with one term alone. This factor helps to quantify the severity of multicollinearity in ordinary least squares regression analysis. Since the VIF is low it tells us the respective component data are independent with each other which is one of the conditions that needs to be satisfied in OLS assumption.

The coefficient alone does not verify the contribution percentage of the yield components. Regression models need to satisfy the OLS assumptions for linear regression. Regression is a powerful analysis that helps us analyze multiple variables simultaneously to answer complex research questions. However, if we don't satisfy the OLS assumption, we might not be able to trust the results. The datasets obtained from the Aberdeen research center were independent and random. The error term should have constant variance i.e. no heteroscedasticity. This assumption is checked by the residual plot as shown in Figure1. To see if the variance of the error is constant, we must look in the residual plot for a random pattern i.e. no curve or line pattern in the residual. As seen in Figure1, the residual plot is random for AB 2018, AB2019, and ASH 2018. This satisfies the no heteroscedasticity assumptions.

.

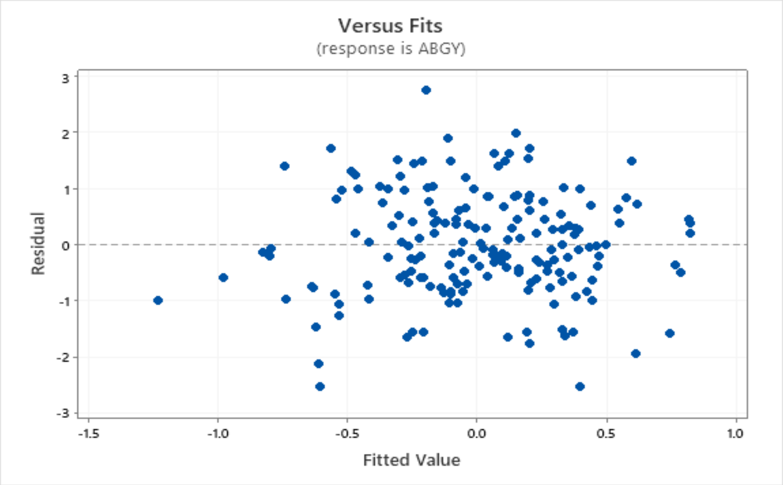

Fig: AB2018

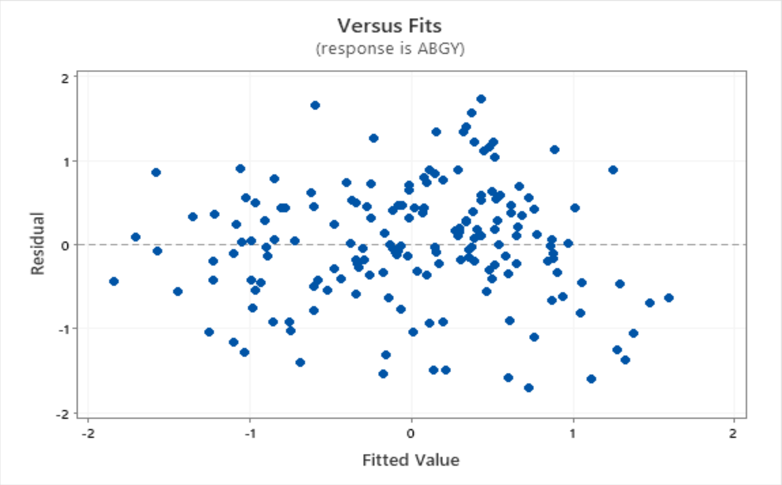

Fig: AB2019

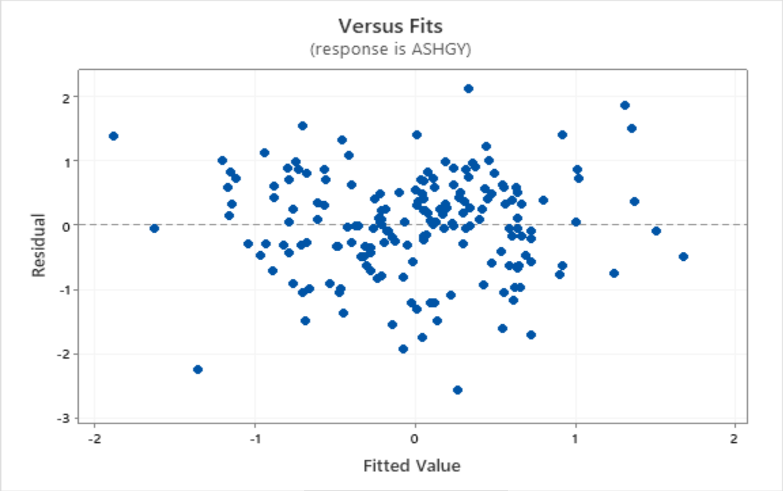

Fig: ASH2019

Figure 1: Residual plot for AB location for the year 2018 and 2019, for ASH location for the year 2018

Finally, the error term should be normally distributed and the condition must be satisfied. The probability plot is efficient in determining whether the residuals follow a normal distribution. To see if the dataset follows normal distribution one can easily interpret by following if the residual follows the straight line on the probability graph. The normal probability plots are shown in Figure 2 for the AB2018, AB2019, and ASH 2018. This shows the linear pattern which helps to satisfy the normality condition.

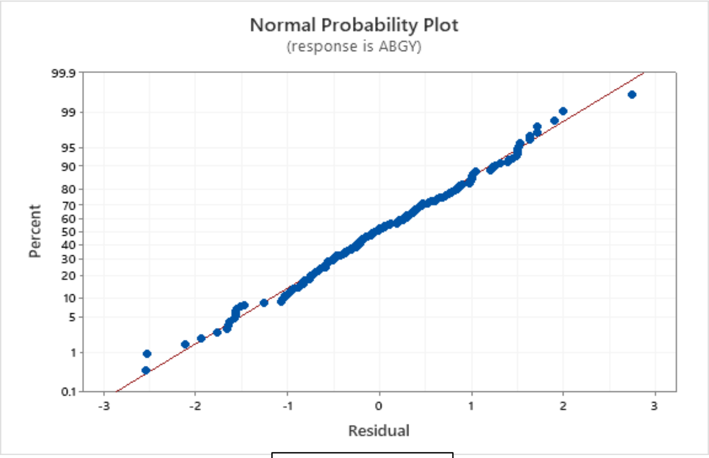

Fig: AB2018

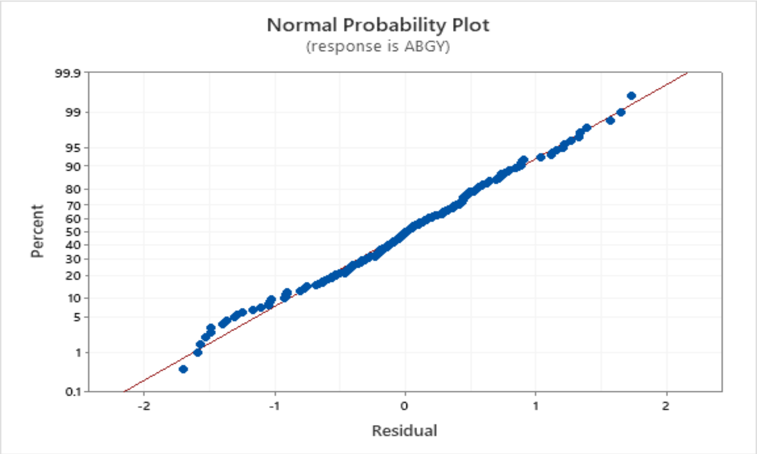

Fig: AB2019

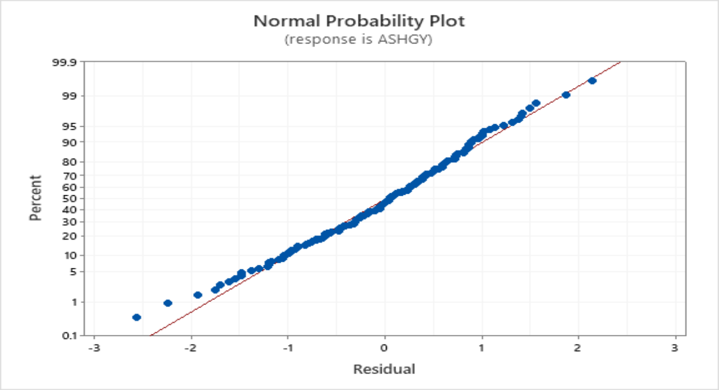

Fig: ASH2018

Figure 2: Normal probability plot of residuals for the AB location for the year 2018, 2019 and ASH location for the year 2018.

Since the wheat dataset satisfies the assumption for linear regression analysis. It is plausible to accept the contribution percentage obtained for the yield components (PTN, TKW, SNS). The contribution percentage obtained for the yield components can be seen in Table 2: Contribution percentage of yield components: PTN, TKW, SNS for AB location for the year 2018, Table 3: Contribution percentage of yield components: PTN, TKW, SNS for ASH location for the year 2018 and Table 4: Contribution percentage of yield components: PTN, TKW, SNS for AB location for the year 2019.

It is observed from Table 2 and Table 4 the yield component TKW has the highest contribution percentage in both the case and the yield component PTN has the lowest contribution percentage. This result shows us that there is some consistency in the result. In Table 2 and Table 3, the highest contribution percentage and lowest contribution both differ from each other. This result points out that there is some underlying correlation due to different environment locations

Table 2: Contribution percentage of yield components: PTN, TKW, SNS for AB location for the year 2018.

Yield Components	Contribution Percentage
PTN	23.7%
TKW	47.8%
SNS	28.30%

Table 3: Contribution percentage of yield components: PTN, TKW, SNS for ASH location for the year 2018.

Yield Components	Contribution Percentage
PTN	33.92%
TKW	21.25%
SNS	44.8%

Table 4: Contribution percentage of yield components: PTN, TKW, SNS for AB location for the year 2019.

Yield Components	Contribution Percentage
PTN	13.4%
TKW	61.8%
SNS	24.8%

4. Discussion

Among the three yield components studied, the TKW has the highest contribution to grain yield for the two Aberdeen trials, while the least contribution for the Ashton trial. This result corresponds to the different climate conditions in Aberdeen and Ashton. Ashton has favorable environments that allow wheat plants to produce more SNS and PTN, but smaller kernels, while Aberdeen has favorable environments that allow longer kernel development and grain filling. This suggests that selecting a high yield cultivar one should test breeding materials in multiple environments and multiple years. This also suggests that high yield cultivars have different yield component architecture in different environments. Additional field trials should be conduction in more diverse environments to confirm findings obtained from the present study.

5. Acknowledgment

I would like to express my special thanks to Dr. Jianli Chen who offered the data and guidance for this project. This project is also partially supported by the CPI (Career Path Internship) program at Idaho State University under mentor

Dr. Xiaoxia Xie. The data used in the present study was collected under the support by the National Research Initiative Competitive grant 2017-67007-25939 from the USDA National Institute of Food and Agriculture.

6. References

1. Liu, Dongli & Archer, Nicholas & Duesing, Konsta & Hannan, Garry & Keast, Russell. (2016). Liu et al 2016. https://www.researchgate.net/publication/303408121_Liu_et_al_2016

2. Velu, G., Singh, R. P., Huerta, J., & Guzmán, C. (2017). Genetic impact of *Rht* dwarfing genes on grain micronutrient concentration in wheat. *Field crops research*, *214*, 373–377. https://doi.org/10.1016/j.fcr.2017.09.030

3. GCARD (2012). *National Food security- The Wheat Initiative- an international Research Initiative for Wheat Improvement.* http://www.fao.org/docs/eims/upload/306175/Briefing%20Paper%20(3)-Wheat%20Initative%20-%20H%C3%A9l%C3%A8ne%20Lucas.pdf

4. United States Department of Agriculture (2020). *Grain: World Markets and Trade.* https://apps.fas.usda.gov/psdonline/circulars/grain.pdf

5. Institute of Medicine (US). Sharing Clinical Research Data: Workshop Summary. Washington (DC): National Academies Press (US); 2013 Mar 29. 5, Standardization to Enhance Data Sharing. Available from: https://www.ncbi.nlm.nih.gov/books/NBK137818/

6. United Nations (UN). World population project to reach 9.8 billion in 2050, https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html

7. Silos Cordoba. By 2050, a 60% increase in wheat production will be needed. https://siloscordoba.com/blog/grain-storage/by-2050-a-60-increase-in-wheat-production-will-be-needed/