Development of yield forecast model in bread wheat using regression analysis

Abstract

Background: Studies highlighted the possibilities of simultaneous crop failures in the world's "breadbaskets" (wheat) due to heatand 40% of the variability in inter-annual wheat production is already related to temperature extremes. The global yield numbers hide the degree of variability of wheat production,yet several environmental conditions pose a threat to wheat production.

Objective: The main objective of the study was to develop a regression model that fitted the dependent variable sufficiently well to account for the total variability.

Method: For this, sixty advance lines along with four standard checks were evaluated for fifteen yield-associated traits and eight quality traits during *Rabi* 2020-21 at the research area of Wheat and Barley section, Department of Genetics and Plant Breeding, CCS Haryana Agricultural University, Hisar. Multiple regression analysis revealed that 98.5% of the variability is explained by the studied morphological and quality traits.

Result: The stepwise regression analysis retained a total of seven traits (six morphological and one quality) *viz*. biological yield per plot, harvest index, grain weight per spike, flag leaf length, main spike weight, number of spikelets per spike and grain appearance score; explaining 97.8 % of the total variability.

Conclusion: The seventh model among all, indicated good yield predicting performance without modifying the traits.

Keywords: Regression model, Multiple regression, stepwise regression, andVariability

Introduction

Wheat is the leading cereal produced, consumed, and traded worldwide.Wheat is India's second most significant cereal crop and it plays an essential role in the country's food and nutritional security (Udhayan et al., 2023).With an unrivalled range of cultivation, it possesses the widest adaptability and is cultivated in around 100 nations globally. It grows in latitudes between 30° and 60° N and 27° and 40° S, having its origins in the Ethiopian Highlands and the Levant region to the east(Nuttonson, 1955; de Sousa et al., 2021).Moreover, half of the world's wheat is produced by the top five producers: China, India, the US, Russia, and the EU (FAOSTAT, 2020). India has been self-sufficient during the previous fifty years, rising to become the second-largest producer and a major exporter of wheat in the world.

Wheat has various end uses, each requiring specific conditions.Frequent extreme climate events, such as drought, heat, and frost have caused severe wheat yield losses during the last decades (Feng et al., 2020).The global yield numbers hide the degree of variability of wheat production,yet several environmental conditions pose a threat to wheat production. According to NOAA (2023), the last ten years (2013–2022) have been the warmest on record and the climate system has undergone several changes as a result of the temperature's consistent rise (IPCC, 2021). The average global temperature may increase by 2–5 °C by 2050, according to the Intergovernmental Panel on Climate Change's Fifth Assessment Report (IPCC, 2014). Food security has wide-reaching ramifications and is thought to be impacted by these changes, which include an increase in the frequency of extreme events, with "high confidence" (FAO, 2021). According to Braun *et al*. (2010) and Cossani and Reynolds (2012), heat stress affects around half of the world's wheat crop. The studies highlight the possibilities of simultaneous crop failures in the world's "breadbaskets" due to heat (Sarhadi *et al*. 2018; Gaupp *et al*. 2020; Kornhuber*et al*. 2020), and 40% of the variability in inter-annual wheat production is already related to temperature extremes (Zampieri *et al*. 2017).

These stresses are difficult to manage through agronomic approaches but there is good genetic variation for tolerance and recent research has been able to identify and characterize the traits associated with tolerance. As a result, many breeding programs include screening of such traitsin their selection processes. Breeding programmes can utilise a variety of approaches to improve genotype stress tolerance. These approaches are mostly centred on a wide range of field evaluations to help raise yield heritability or choose component traits that exhibit substantial heritability. Regression analysis can be used to quantify the extent to which grain yield is dependent on its constituent traits, or independent variables, as well as the proportionate contributions of each trait to the variance in grain yield as a whole. The cumulative contribution of the component traits and their order of importance in contributing to the total variance are found using more accurate techniques, such as multiple and stepwise regression analysis. A quantitative variable's value can be estimated *via* stepwise regression by examining how it relates to one or more quantitative variables. The relationship that allows changes in one variable to be used to anticipate changes in other variables is discussed in this paper.

Materials and Methods

The experimental material comprised sixty advanced lines of wheat along with four standard checks *viz*., WH 1021, WH 1124, DBW 90 and HD 3059 that were evaluated at the research area of the

Wheat and Barley Section, Department of Genetics and Plant Breeding, CCS Haryana Agricultural University, Hisar during *Rabi* 2020-21. The seeds of all the lines were sown with a hand plough in a Randomized Block Design (RBD) with three replications. Observations were recorded for fifteen morphological traits viz., days to heading, days to maturity, plant height, number of spikelets per spike, spike length, peduncle length, flag leaf length, main spike weight, number of grains per spike, grain weight per spike, 1000 grain weight, number of effective tillers per metre, grain yield per plot, biological yield per plot, harvest index from five plants chosen at random from each entry of the three replications. Eight quality traits *viz.*, grain appearance score, hectolitre weight, sedimentation value, wet gluten, dry gluten, total gluten, crude protein, and total soluble sugars were assessed for each replication and the average was taken for statistical analysis.

Statistical analysis

1. Multiple regression analysis: It defines the relative contribution of component traits to the grain yield (y) by applying the equation of Snedecor and Cochran (1981).
2. Stepwise regression analysis: Stepwise regression as suggested by Draper and Smith (1966) was used to determine the sequence of importance of variables in contribution to total yield.

Results and Discussion

Association through scatter diagram technique based on simple regression

Simple regression is the simplest statistical technique for examining the relationship between two variables. The line of best fit is represented as y = a + bx, where, 'a' represents the intercept and 'b' represents the slope. The coefficient of determination, R^2, indicated the contribution of individual variables to total variability, explained by the regression line. Each dot represents an advance line having two axis; x-axis containsthe mean value of independent traits, whereas, the y-axis corresponds to themean value of grain yield per plot under late sown conditions.

The line of best fit for some important yield contributing traits *viz*. biological yield per plot (y = 0.3348x + 37.509, R^2 = 0.7442), harvest index (y = 77.542x + 506.82, R^2 = 0.2419), 1000 grain weight (y = 38.998x + 1590.6, R^2 = 0.187), number of effective tillers per metre (y = 14.164x + 1783.7, R^2 = 0.138), number of grains per spike (y = 20.572x + 2070.3, R^2 = 0.1379), grain weight pe spike (y = 500.41x + 1869.8, R^2 = 0.3876), number of spikelets per spike (y = 82.794x + 1684.4, R^2 = 0.175), peduncle length (y = 57.414x + 1012.5, R^2 = 0.2182), flag leaf length (y = 100.55x + 795.6, R^2 = 0.4227) and hectolitre weight (y = 29.8x + 767.49, R^2 = 0.0424)are presented in Figure 1. The advance lines (dots) appearing closer to the trend line had a strong relationship between the variables.

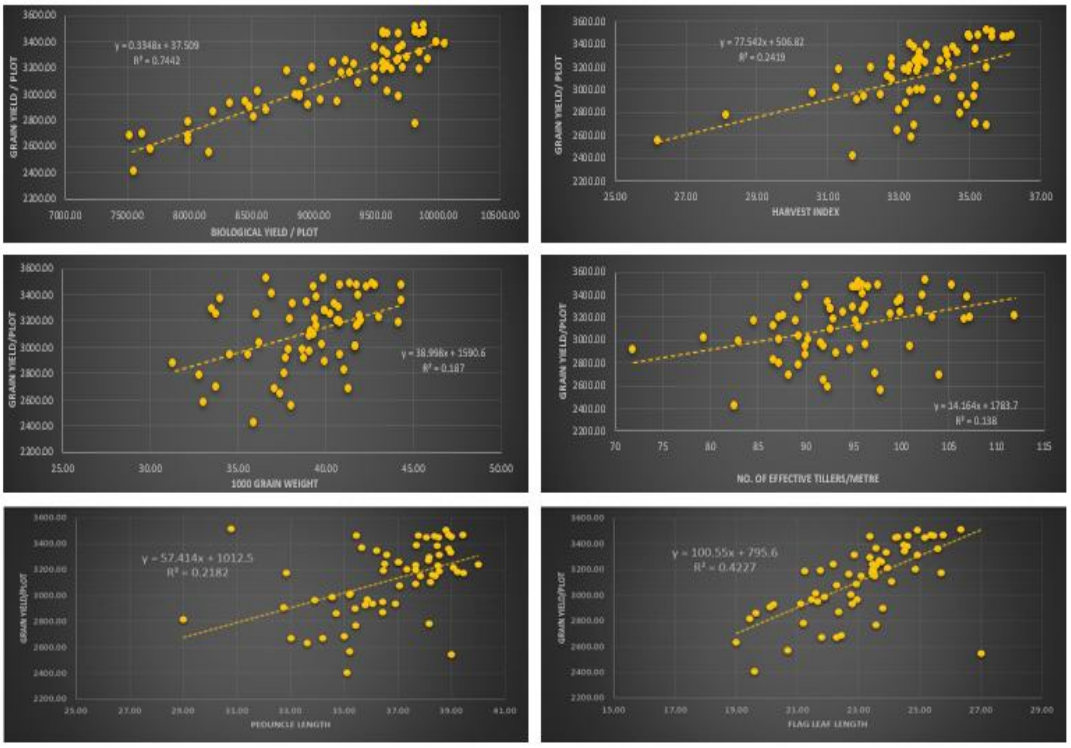

Figure 1: The line of best fit for different yield contributing traits

Multiple regression

The multiple regression model explained 98.5 per cent (R^2 = 0.985) of the total variability in grain yield per plot *via.* the studied morphological and quality traits. Similarly, Sobhanina*et al*. (2019) concluded that 94.5 per cent of the total variation in grain yield per m^2 was explained by biological yield per m^2, harvest index, relative water content and thousand grain weight.The coefficients of regression for different traits *viz*. days to heading (3.74), plant height (1.54), number of effective tillers per metre (-1.33), spike length (11.14), number of spikelets per spike (-12.35), flag leaf length (18.80), peduncle length (1.99), main spike weight (73.71), grain weight per spike (75.82), number of grains per spike (-1,54), 1000 grain weight (3.92), days to maturity (-4.90), biological yield per plot (0.27), harvest index (68.15), crude protein (0.04), sedimentation value (1.01), hectolitre weight (-1.39), grain appearance score (-37.19), wet gluten (3.50), dry gluten (-11.00), total gluten (-2.34) and total soluble sugars (-23.47) are presented in Table 1.Similar findings were also observed by Mansouri *et al.* (2018) and Khames*et al.* (2016).

Table 1: Multiple regression analysis of different morphological and quality traits in bread wheat under late sown conditions

Sr. No.	Traits	Coefficient ± S.E.
	Constant	**-1973.60 ± 490.02**
1	Days to heading	3.74 ± 3.71
2	Plant height (cm)	1.54 ± 1.33
3	No. of effective tillers/m	-1.33 ± 1.092
4	Spike length (cm)	11.14 ± 9.43
5	No. of spikelets/ spike	-12.35 ± 6.33
6	Flag leaf length (cm)	18.80 ± 5.67
7	Peduncle length (cm)	1.99 ± 3.527
8	Main spike weight (g)	73.71 ± 3.867
9	Grain weight per spike (g)	75.82 ± 0.01
10	No. of grains/spike	-1.54 ± 2.07
11	1000 grain weight (g)	3.92 ± 2.30
12	Days to maturity	-4.90 ± 2.57
13	Biological yield / plot (g)	0.27 ± 0.01
14	Harvest index (%)	68.15 ± 3.87
15	Crude protein (%)	0.04 ± 4.90
16	Sedimentation value (ml)	1.01 ± 1.15
17	Hectolitre weight (Kg/hl)	-1.39 ± 3.36
18	Grain appearance score	-37.19 ±18.71
19	Wet gluten (%)	3.50 ± 2.20
20	Dry gluten (%)	-11.00 ± 7.88
21	Total gluten (%)	-2.34 ± 1.33
22	Total soluble sugars (%)	-23.47 ± 14.64

where,

DH: Days to heading, **PH**: Plant height (cm), **NET/m**: Number of effective tillers per metre, **SL**: Spike length (cm), **NS/S**: Number of spikelets per spike, **FLL**: Flag leaf length (cm), **PL**: Peduncle length (cm), **MSW**: Main spike weight (g), **GW/S**: Grain weight per spike (g), **NG/S**: Number of grains per spike, **TGW**: 1000 grain weight (g), **DM**: Days to maturity, **BY/P**: Biological yield per plot (g), **HI**: Harvest Index (%), **CP**: Crude Protein (%), **SV**: Sedimentation Value (ml), **HW**: Hectolitre Weight (Kg/hl), **GAS**: Grain Appearance Score, **WG**: Wet Gluten (%), **DG**: Dry Gluten (%), **TG**: Total Gluten (%), **TSS**: Total Soluble Sugars (%)

Based on these results, the predicting model equation for the grain yield per plot (y) was formulated as follows:

y = -1973.60 + 3.74DH + 1.54PH – 1.33NET/m + 11.14SL – 12.35NS/S +18.80FLL + 1.99PL + 73.71MSW + 75.82GW/S – 1.54NG/S + 3.92TGW – 4.90DM + 0.27BY/P + 68.15HI + 0.04CP +1.01SV – 1.39HW -37.19GAS + 3.50WG -11.00DG – 2.34TG – 23.47TSS

Stepwise Regression

The stepwise analysis including entered or removed variables, partial R^2, model R^2 (cumulative R^2), P value for entered or removed variables, P value for model variables and standard error have been presented in Table 2. The stepwise regression analysis retained a total of seven traits (six morphological and one quality) *viz*. biological yield per plot (74.4 per cent), harvest index (20.3 per

cent), grain weight per spike (1.8 per cent), flag leaf length (0.6 per cent), main spike weight (0.3 per cent), number of spikelets per spike (0.3 per cent) and grain appearance score (0.1 per cent). These seven traits attributed 97.8 per cent of the total variation in grain yield per plot. The results are confirmed by the findings of Leilah and Al-Khateeb (2005), Mansouri *et al.* (2018) and Sobhanian*et al.* (2019).

Table 2: Relative contribution (partial and multiple R^2), parameter estimates along with their standard error and P values predicting wheat grain yield under late sown conditions

Step	Variables entered	Variables removed	Partial R^2	Model R^2	P value ER	Parameter estimate	Standard error	P value M
1	BY/P	-	0.744	0.744	0.000	0.274	0.010	0.000
2	HI	-	0.203	0.947	0.000	70.497	3.440	0.000
3	GW/S	-	0.018	0.965	0.001	76.842	22.157	0.016
4	FLL	-	0.006	0.971	0.000	17.705	4.710	0.002
5	MSW	-	0.003	0.974	0.001	79.137	21.573	0.004
6	NS/S	-	0.003	0.977	0.021	-12.058	5.064	0.058
7	GAS	-	0.001	0.978	0.044	-28.159	13.644	0.053

R^2: coefficient of determination, ***P* value ER**: *P* value for entered or removed variables, ***P* value *M***: *P* value for final model.

where,
BY/P: Biological yield per plot (g), **HI**: Harvest Index (%), **GW/S**: Grain weight per spike (g), **FLL**: Flag leaf length (cm), **MSW**: Main spike weight (g), **NS/S**: Number of spikelets per spike, **GAS**: Grain appearance score

The seven models could be described based on the results. Model 1 includes biological yield per plot, while model 2 has harvest index in addition. Model 3 can best be described using biological yield per plot, harvest index and grain weight per spike. The model 4, 5 and 6 have flag leaf length, main spike weight and number of spikelets per spike in addition.
The final predicting model 7 for grain yield per plot was formulated as:
y= -2238.50 + 0.27BY/P + 70.50HI +76.84GW/S + 17.70FLL +79.14MSW – 12.06NS/S – 28.16GAS

Conclusion

It can be concluded from the above findings that to remove the inefficient use of linear regression lines, multiple and stepwise regression models can be used. The wheat crop coefficient of multiple correlation (r = 0.985) indicates that the dependent and independent variables have a positive association, accounting for 98.5% of the explained variability.The coefficient of determination's significance (R^2 = 0.744) indicates that the regression model fitted the dependent variable sufficiently well to account for 74.4% of the variability, indicating the model's good predicting performance. In the same way, the adjusted R^2 = 0.744 value suggests that, even with the current correction, the regression model almost matches the variation in the data set without modifying the parameter.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc have been used during writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

References :

Braun, H.J., Atlin, G. & Payne, T. (2010). Multi-location testing as a tool to identify plant response to global climate change. In Reynolds MP (ed) *Climate change and crop production*, pp. 115. https://doi.org/ 10.1079/9781845936334.0115.

Cossani, C.M. & Reynolds, M.P. (2012). Physiological traits for improving heat tolerance in wheat. *Plant Physiology*. **160**(4),1710–1718. https:// doi.org/10.1104/pp.112.207753

Draper, N.R. & Smith, H. (1966). Allied Regression Analysis. Wiley, New York, 7407.

FAOSTAT (2020). Statistical Information. http://www.fao.stat.fao.org

Gaupp, F., Hall, J., Hochrainer-Stigler, S. & Dadson, S. (2020). Changing risks of simultaneous global breadbasket failure. *Natural Climate Change*. **10**, 54–57. https://doi.org/10.1038/s41558-019-0600-z

IPCC (2014). Climate change 2014, *Impact challenges and adaptation.* Cambridge University Press.

IPCC (2021). Global warming of 1.5 °C. An IPCC special report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change. In: Masson-Delmotte V, Zhai P, Pörtner HO, Roberts D, Skea J, Shukla PR, Pirani A, Moufouma-Okia W, Péan C, Pidcock R, Conners

S, Matthew JBR, Chen Y, Zhou X, Gomis MI, Lonnoy E, Maycock T, Tignor M, Waterfeld T (eds) Intergovernmental panel on climate change.

Khames, K. M., Abo-Elwafa, A., Mahmoud, A. M., & Hamada, A. (2016). Correlation, path-coefficient, normal and stepwise regression analyses *via* two cycles of pedigree selection in bread wheat (*Triticumaestivum* L). *Assiut Journal of Agricultural Sciences*, **47**(4), 84-108.

Kornhuber, K., Coumou, D., Vogel, E., Lesk, C., Donges, J.F., Lehmann, J. & Horton, R.M. (2020). Amplifed Rossby waves enhance risk of concurrent heatwaves in major breadbasket regions. *Natural Climate Change*. https://doi.org/10.1038/s41558-019-0637-z

Leilah, A.A. and Al-Khateeb, S.A. (2005). Statistical analysis of wheat yield under drought conditions. *Journal of Arid environments*, **61**(3), 483-496.

Mansouri, A., Oudjehih, B., Benbelkacem, A., Fellahi, Z. E. &Bouzerzour, H. (2018). Variation and relationships among agronomic traits in durum wheat [*Triticum turgidum* (L.)] under south Mediterranean growth conditions: Stepwise and path analyses. *International Journal of Agronomy*, **3**(1), 1-11.

NOAA National Centres for Environmental Information (2023). State of the climate: global climate report for annual.

Nuttonson, M.Y. (1955). Wheat-climatic relationships and the use of phenology in ascertaining the thermal and photo thermal requirements of wheat. American Institute of Crop Ecology, Washington, DC, USA.

Sarhadi, A., Ausín, M.C., Wiper, M.P., Touma, D. &Difenbaugh, N.S. (2018). Multidimensional risk in a nonstationary climate: joint probability of increasingly severe warm and dry conditions. *Science Advances.***4**(2),123-127. https://doi.org/10.1126/sciadv.aau3487

Snedecor, G.W. & Cochran, W.G. (1981). Statistical Methods, seventh ed. Iowa State University Press, Iowa, USA.

Sobhaninan, N., Heidari, B., Tahmasebi, S., Dadkhodaie, A. & McIntyre, C.L. (2019). Response of quantitative and physiological traits to drought stress in the SeriM82/Babax wheat population. *Euphytica*. **215**(2),32.

Zampieri, M., Ceglar, A., Dentener, F. &Toreti, A. (2017). Wheat yield loss attributable to heat waves, drought and water excess at the global, national and subnational scales. *Environmental Resources*, https://doi.org/ 10.1088/1748-9326/aa723b

de Sousa T, Ribeiro M, Sabença C, Igrejas G. The 10,000-year success story of wheat!. Foods. 2021 Sep 8;10(9):2124.

Feng P, Wang B, Li Liu D, Waters C, Xiao D, Shi L, Yu Q. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agricultural and Forest Meteorology. 2020 May 15;285:107922.

Udhayan N., Naik AD, Hiremath GM. An Economic Analysis of Wheat Cultivation in North-Karnataka, India. Int. J. Plant Soil Sci. 2023;35(20):939-45. Available from: https://journalijpss.com/index.php/IJPSS/article/view/3887