

CORRELATION, PATH, AND LASSO REGRESSION ANALYSIS OF GRAIN YIELD AND SOME MORPHOLOGICAL CHARACTERISTICS IN BARLEY CULTIVARS

MAHMUT AÇIKGÖZ, SUNA AKKOL¹, EROL ORAL² AND TAMER ERYİĞİT^{3*}

Van Yuzuncu Yil University, Institute of Science and Technology, Van, Türkiye

Keywords: Barley, Correlation, Lasso regression, Path analysis

Abstract

This study evaluated 16 barley cultivars during the 2020-2021 winter season under irrigated conditions using correlation, path, and Lasso regression analyses. Correlation analysis revealed significant positive associations between grain yield and harvest index (HI), hectoliter weight (HLW), thousand-grain weight (TGW), grain filling rate per spike (GFRS), plant height (PH), number of sterile spikelets per spike (NSSS), and biological yield (BY). Path analysis demonstrated that HI and TGW exerted the strongest direct effects on grain yield. Lasso regression further identified HI, TGW, and BY as the most influential predictors, collectively explaining 79.28% of the yield variation. The consistency of HI and TGW across all statistical approaches highlights their reliability as selection criteria. These findings provide valuable insights for improving cultivar performance and optimizing barley breeding strategies under irrigated production systems.

Introduction

Barley (*Hordeum vulgare* L.) is serving as animal feed, a key raw material in malt and brewing industries, and a minor food crop (Zhou 2009, El-Hashash and El-Absy 2019, Huang *et al.* 2020). Türkiye ranks among the top 10 global producers, with barley second only to wheat in national cereal output (Benlioglu *et al.* 2025). The crop's adaptability to diverse climatic conditions sustains its agricultural and economic relevance worldwide (Sharafi *et al.* 2014, Sonmez 2021).

Grain yield in barley is shaped by complex interactions between vegetative and generative growth phases (Alqudah *et al.* 2016, Hu *et al.* 2021). Identifying the morphological determinants of yield is therefore critical for effective breeding. Correlation and path analyses provide a robust framework for partitioning these associations into direct and indirect components, while Lasso regression enables parsimonious model building through penalized variable selection (Tibshirani 1996). This study integrates all three methods to identify key yield predictors in 16 two-row barley cultivars grown under irrigated conditions in Van province, Eastern Anatolia, is essential for advancing barley breeding and sustainable production.

Materials and Methods

This study evaluated 16 two-row barley varieties: Akar, Anka-06, Anka-08, Anka-9, Anka-10, Anka-11, Asil, Avci-2002, Aydanhanım, Bozlak, Burakbey, Calabey, Larende, Olgun, Tosunpasa, and Tarm-92. Field experiments were conducted during the 2020 and 2021 winter seasons under irrigated conditions at Van Yuzuncu Yil University Field Crops Experiment Station (Van province, Eastern Anatolia). The site had an average annual temperature of 8.8°C and total precipitation of 376 mm; soils were loamy-clay with slightly alkaline pH and low organic matter content. A randomized complete block design with three replications (48 plots) was used; each

*Author for correspondence: <tamereryigit@yyu.edu.tr>. ¹Van Yuzuncu Yil University, Faculty of Agriculture, Department of Animal Science, Van, Türkiye. ²Van Yuzuncu Yil University, Faculty of Agriculture, Department of Field Crops, Van, Türkiye. ³Van Yuzuncu Yil University, Gevas Vocational School, Department of Plant and Animal Production, Van, Türkiye.

plot was 2 m x 0.68 m with 0.17 m row spacing and five rows. Seeding density was 450 seeds m⁻². Fertilization comprised 80 kg ha⁻¹ P₂O₅ and 100 kg ha⁻¹ N (split application: 50% pre-planting, 50% at rooting). Supplemental irrigation totaling 114 mm was applied twice (stem elongation and spike formation stages). Weed control was manual within plots and mechanical between blocks (Yüksel and Incekarakaya 2022). At harvest, 20 plants per plot were randomly selected from a 1.02 m² central area; measured traits included harvest index (HI), hectoliter weight (HLW), thousand-grain weight (TGW), number of days to spike (NDS), number of fertile spikes per square meter (NFSSM), grain filling rate per spike (GFRS), plant height (PH), spike length (SL), number of fertile spikelets per spike (NFSS), number of sterile spikelets per spike (NSSS), number of grains per spike (NGS), and biological yield (BY).

Pearson correlation coefficients were calculated for all trait pairs using:

$$r_{yx} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (1)$$

Path analysis partitioned correlations of 12 independent variables on grain yield into direct (P_{yxi}) and indirect effects. The system of equations is:

$$\begin{aligned} r_{x_1y} &= P_{yx_1} + r_{x_1x_2} P_{yx_2} + \dots + r_{x_1x_n} P_{yx_n} \\ r_{x_2y} &= r_{x_2x_1} P_{yx_1} + P_{yx_2} + \dots + r_{x_2x_n} P_{yx_n} \\ r_{x_ny} &= r_{x_nx_1} P_{yx_1} + r_{x_nx_2} P_{yx_2} + \dots + P_{yx_n} \end{aligned} \quad (2)$$

These equations can be expressed as $B = A \times P$, where A is the correlation matrix of independent variables, B is the correlation vector between grain yield and each independent variable, and P represents the path coefficients (Yilkan *et al.* 2020). Path coefficients were estimated as $P = A^{-1}B$. Lasso regression (Tibshirani 1996) was applied to select the most influential predictors by minimizing:

$$\hat{\beta}_{Lasso} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n (y_i - \sum_{j=0}^k x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^k |\beta_j| \right\} \quad (3)$$

Here y_i is grain yield, x_{ij} are yield components, β_j are regression coefficients, and λ is the regularization parameter (Ciftsuren and Akkol 2018, Das *et al.* 2018). The optimal λ was selected by cross-validation.

Results and Discussion

The correlation coefficients are presented in Table 1. Grain yield showed strong positive and significant correlations with HI ($r = 0.876^{**}$), TGW ($r = 0.659^{**}$), GFRS ($r = 0.655^{**}$), HLW ($r = 0.602^{**}$), BY ($r = 0.414^{**}$), NSSS ($r = 0.330^{*}$), and PH ($r = 0.320^{*}$). Among these, HI had the strongest association, consistent with studies conducted under similar ecological conditions (Kaydan and Yağmur 2007, Akkol *et al.* 2018, Yüksel and Incekarakaya 2022). Strong inter-trait correlations were also observed: HI with GFRS ($r = 0.763^{**}$), HLW ($r = 0.663^{**}$), and TGW ($r = 0.623^{**}$), confirming their interconnected contributions to yield. The NDS showed significant negative correlations with NFSSM ($r = -0.707^{**}$), NFSS ($r = -0.713^{**}$), and SL ($r = -0.620^{**}$), indicating that delayed spiking adversely affects spike development. GFRS was negatively correlated with NSSS ($r = -0.350^{*}$) and BY ($r = -0.337^{*}$), and NFSS showed a strong positive correlation with SL ($r = 0.840^{**}$) and BY ($r = 0.460^{**}$).

Path analysis (Table 2) revealed that HI exerted the largest direct effect on grain yield ($p = 0.734^{**}$; 66.38%), followed by NFSS ($p = 0.282$; 34.29%), NDS ($p = 0.243$; 27.35%), and TGW ($p = 0.229^{*}$; 25.03%). The dominance of HI as a direct yield driver aligns with earlier reports

Table 1. Correlation coefficients and significance levels among the traits.

	GY	HI	HLW	TGW	NDS	NFSSM	GFRS	PH	SL	NFSS	NSSS	NGS	BY
GY	1												
HI	0.876**	1											
HLW	0.602**	0.663**	1										
TGW	0.659**	0.623**	0.615**	1									
NDS	0.093ns	0.119ns	0.280ns	0.219ns	1								
NFSSM	0.106ns	0.063ns	-0.221ns	-0.178ns	-0.707**	1							
GFRS	0.655**	0.763**	0.453**	0.424**	0.066ns	0.095ns	1						
PH	0.320*	0.342*	0.345*	0.285*	-0.285*	0.051ns	0.185ns	1					
SL	0.098ns	0.069ns	-0.053ns	0.088ns	-0.620**	0.478**	0.141ns	0.493**	1				
NFSS	0.157ns	0.056ns	-0.025ns	0.025ns	-0.713**	0.532**	0.146ns	0.411**	0.840**	1			
NSSS	0.330*	0.349*	0.048ns	0.213ns	0.095ns	-0.022ns	-0.350*	-0.046ns	-0.173ns	-0.153ns	1		
NGS	0.227ns	0.234ns	0.378**	0.366*	0.384**	-0.221ns	0.214ns	0.069ns	-0.055ns	0.144ns	0.100ns	1	
BY	0.414**	0.333*	0.373**	0.217ns	-0.277ns	0.130ns	-0.337*	0.346*	0.211ns	0.460**	0.171ns	0.093ns	1

ns: non-significant; significant levels: * $p < 0.05$; ** $p < 0.01$. GY: grain yield; HI: harvest index; HLW: hectoliter weight; TGW: thousand-grain weight; NDS: no. of days to spike; NFSSM: no. of fertile spikes per m^2 ; GFRS: grain filling rate per spike; PH: plant height; SL: spike length; NFSS: No. of fertile spikelets per spike; NSSS: no. of sterile spikelets per spike; NGS: no. of grains per spike and BY: biological yield.

Table 2. Path coefficients and contribution margins (%) of traits affecting grain yield in barley.

Traits	Directly effect	Indirectly Effect												Correlation
		HI	HLW	TGW	NDS	NFSSM	GFRS	PH	SL	NFSS	NSSS	NGS	BY	
HI	0.734**		-0.046	0.142	0.029	0.011	-0.050	0.013	-0.012	0.016	0.011	-0.006	0.034	0.876**
	66.381		4.204	12.891	2.617	0.990	4.546	1.147	1.070	1.439	1.039	0.568	3.107	100
HLW	-0.070	0.486		0.141	0.068	-0.039	-0.030	0.013	0.009	-0.007	0.002	-0.010	0.038	0.601**
	7.681	53.297		15.401	7.448	4.231	3.275	1.402	1.012	0.760	0.174	1.111	4.208	100
TGW	0.229*	0.457	-0.043		0.053	-0.031	-0.028	0.011	-0.015	0.007	0.007	-0.010	0.022	0.659**
	25.031	50.096	4.722		5.819	3.393	3.057	1.157	1.665	0.772	0.769	1.074	2.445	100
NDS	0.243	0.087	-0.020	0.050		-0.123	-0.004	-0.011	0.107	-0.201	0.003	-0.010	-0.028	0.093 ^{ns}
	27.348	9.828	2.206	5.622		13.872	0.490	1.187	12.065	22.666	0.350	1.158	3.206	100
NFSSM	0.174	0.046	0.016	-0.041	-0.172		-0.006	0.002	-0.083	0.150	-0.001	0.006	0.013	0.105 ^{ns}
	24.595	6.487	2.187	5.720	24.208		0.885	0.268	11.651	21.165	0.104	0.835	1.894	100
GFRS	-0.066	0.559	-0.032	0.097	0.016	0.017		0.007	-0.024	0.041	0.012	-0.006	0.035	0.655**
	7.234	61.408	3.489	10.623	1.763	1.824		0.753	2.682	4.525	1.265	0.630	3.805	100
PH	0.037	0.251	-0.024	0.065	-0.069	0.009	-0.012		-0.085	0.116	-0.002	-0.002	0.036	0.320*
	5.231	35.448	3.418	9.203	9.772	1.263	1.723	12.037	16.400	0.214	0.259	5.032	100	
SL	-0.173	0.050	0.004	0.020	-0.151	0.083	-0.009	0.018		0.237	-0.006	0.001	0.022	0.098 ^{ns}
	22.302	6.480	0.483	2.594	19.457	10.768	1.202	2.359	30.624	0.734	0.191	2.807	100	
NFSS	0.282	0.041	0.002	0.006	-0.173	0.093	-0.010	0.015	-0.145		-0.005	0.004	0.047	0.157 ^{ns}
	34.290	5.019	0.209	0.692	21.048	11.263	1.168	1.850	17.633	0.612	0.470	5.745	100	
NSSS	0.033	0.256	-0.003	0.049	0.023	-0.004	-0.023	-0.002	0.030	-0.043		-0.003	0.018	0.330*
	6.772	52.664	0.694	10.034	4.727	0.804	4.749	0.351	6.143	8.899	0.549	3.613	100	
NGS	0.027	0.172	-0.027	0.084	0.093	-0.039	-0.014	0.003	0.010	-0.041	0.003		0.010	0.227 ^{ns}
	5.153	33.029	5.092	16.064	17.928	7.406	2.710	0.487	1.830	7.833	0.630	1.837	100	
BY	0.103	0.245	-0.026	0.050	-0.067	0.023	-0.022	0.013	-0.036	0.130	0.006	-0.002		0.414**
	14.257	33.851	3.617	6.855	9.303	3.150	3.069	1.774	5.050	17.952	0.777	0.345	100	

The second line in each cell shows the contribution margin (%). The sum of direct and indirect effects equals the correlation coefficient shown in the last column. ns: non-significant; significant levels: * $p < 0.05$; ** $p < 0.01$.

(Shalaby *et al.* 1988, Monouchi 2006, Yılkan *et al.* 2020). HI also exerted large indirect positive effects through TGW (12.89%) and BY (3.11%), amplifying its overall influence. TGW showed a major indirect positive effect via HI (50.10%), reinforcing its role as a secondary yield determinant (Kıran 1999). GFRS had a negative direct effect ($p = -0.066$; 7.23%) but large positive indirect effects through HI (61.41%) and TGW (10.62%), indicating that its net contribution to yield is primarily mediated through these traits (Garcia del Moral *et al.* 1991). SL exerted a negative direct effect (-22.30%) but positive indirect effects through NFSS (30.62%) and NFSSM (10.77%), demonstrating that its influence on yield is nuanced (Kara and Akman 2007). NSSS reduced yield directly (6.77%), consistent with drought-induced sterility reported by Kumar *et al.* (2017). The biological yield also contributed positively through indirect pathways mediated by HI (33.85%) and NFSS (17.95%), supporting its selection utility in combination with HI (Ereku and Kohn 2006, Kaydan and Yagmur 2007).

Lasso regression identified three significant predictors of grain yield, yielding: $GY = 0.686HI + 0.134TGW + 0.075BY$ ($R^2 = 0.7928$). These three variables collectively explained 79.28% of yield variation (Fig. 1), with HI as the dominant predictor. A unit increase in HI, TGW, and BY resulted in 0.686-, 0.134-, and 0.075-unit increases in grain yield, respectively. Lasso regression has been successfully applied for variable selection in crop studies (Akkol *et al.* 2018, Çiftsüren and Akkol 2018, Das *et al.* 2018), and the present application confirms its utility for identifying parsimonious yield prediction models in barley.

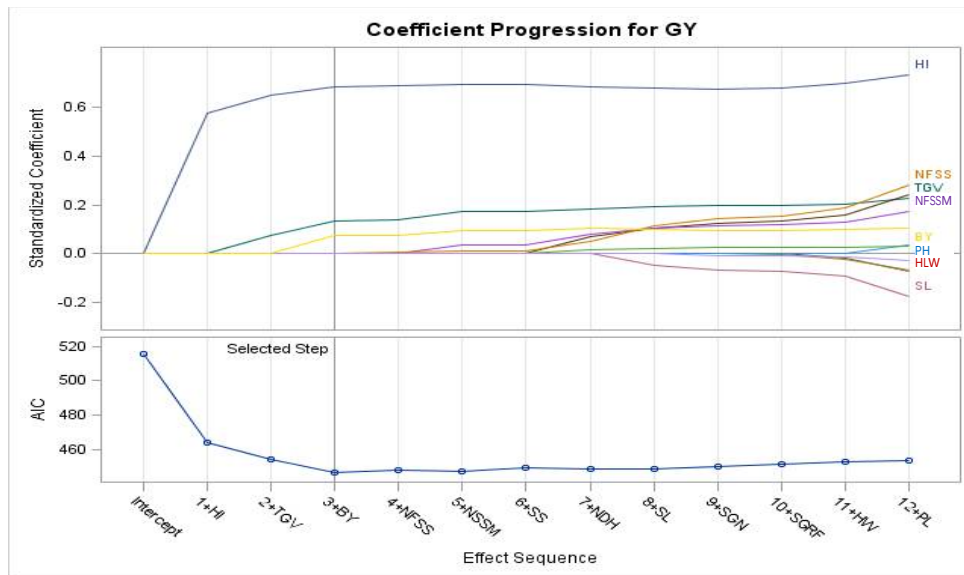


Fig. 1. Variable selection procedure and monitoring of regression coefficients during Lasso regression analysis. The vertical dotted line indicates the optimal λ value selected by cross-validation corresponding to the model: $GY = 0.686HI + 0.134TGW + 0.075BY$ ($R^2 = 0.7928$).

Table 3 summarizes the significance of yield components across all three methods. HI and TGW were consistently significant in correlation, path, and Lasso analyses, confirming their primacy as selection criteria for barley breeding programs. BY was significant in both correlation and Lasso regression. The convergence of results across three independent statistical approaches strengthens the reliability of these findings (Dokuyucu *et al.* 2001, Yuksel and Incekarakaya 2022).

Table 3. Yield components identified as significant through correlation, path, and Lasso regression analyses.

Variables	Correlation	Path	Lasso
Harvest index (HI)	*	*	*
Hectoliter weight (HLW)	*		
Thousand grain weight (TGW)	*	*	*
Number of days to spike (NDS)			
Number of fertile spikes per m ² (NFSSM)			
Grain filling rate per spike (GFRS)	*		
Plant height (PH)	*		
Spike length (SL)			
Number of fertile spikelets per spike (NFSS)			
Number of sterile spikelets per spike (NSSS)	*		
Number of grains per spike (SGN)			
Biological yield (BY)	*		*

*: significant; blank: non-significant.

In conclusion, this study demonstrated that HI, TGW, and BY were the most reliable determinants of grain yield in two-row barley under irrigated conditions. HI consistently dominated all three analytical methods, while TGW contributed substantially through both direct and indirect pathways. Incorporating these traits as primary selection criteria in barley breeding programs would support the development of higher-yielding cultivars adapted to irrigated production systems in Eastern Anatolia and analogous environments.

Acknowledgements

The authors thank Van Yuzuncu Yil University for providing the experimental facilities.

References

- Akkol S, Arpali D and Yagmur M 2018. Adaptive lasso analysis for grain yield and yield components in two-rowed barley under rainfed conditions. *CR. Acad. Bulg. Sci.* **71**(9): 1279-1287. doi:<https://doi.org/10.7546/CRABS.2018.09.17>
- Alqudah AM, Koppolu R, Wolde GM, Graner A and Schnurbusch T 2016. The genetic architecture of barley plant stature. *Front. Genet.* **7**(117): 1-13. doi: <https://doi.org/10.3389/fgene.2016.00117>
- Benlioglu B, Bilir M, Akdogan G, Ahmed HaA, Ergun N, Aydogan S and Emrebas T 2025. Phenotypic characterization of two-row barley (*Hordeum vulgare* L.) germplasm conserved in Osman Tosun Genebank of Türkiye by multivariate analysis model. *Genet. Res. Crop Evol.* **72**(2): 1567-1584. doi: <https://doi.org/10.1007/s10722-024-02032-0>
- Çiftsuren MN and Akkol S 2018. Prediction of internal egg quality characteristics and variable selection using regularization methods: ridge, LASSO and elastic net. *Arch. Anim. Breed.* **61**(3): 279-284. doi: <https://doi.org/10.5194/aab-61-279-2018>
- Das B, Nair B, Reddy VK and Venkatesh P 2018. Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. *Int. J. Biometeorol.* **62**(10): 1809-1822. doi: <https://doi.org/10.1007/s00484-018-1583-6>
- Dokuyucu T, Cesurer L and Akkaya A 2001. Determination of yield and yield components in some bread wheat (*Triticum aestivum* L.) varieties. *J. Sci. Engin.* **4**(1): 109-117.
- El-Hashash E and El-Absy K 2019. Barley (*Hordeum vulgare* L.) Breeding. In: Al-Khayri J, Jain S and Johnson D (Eds), *Advances in Plant Breeding Strategies: Cereals*. Springer Cham. pp. 1-45. doi: https://doi.org/10.1007/978-3-030-23108-8_1.

- Ereku O and Kohn W 2006. Effect of weather and soil conditions on yield components and bread-making quality of winter wheat and winter triticales varieties in North-East Germany. *J. Agron. Crop Sci.* **192**(6): 452-464. doi: <https://doi.org/10.1111/j.1439-037X.2006.00234.x>
- Garcia del Moral L, Ramos J, Garcia del Moral M and Jimenez-Tejada M 1991. Ontogenetic approach to grain production in spring barley based on path-coefficient analysis. *Crop Sci.* **31**(5): 1179-1185. doi: <https://doi.org/10.2135/CROPSCI1991.0011183X003100050021X>
- Hu Y, Barmeier G and Schmidhalter U 2021. Genetic variation in grain yield and quality traits of spring malting barley. *Agronomy* **11**(6): 1177. doi: <https://doi.org/10.3390/agronomy11061177>
- Huang H, Gao X, Li Y, Tian P, Nima Y, Laba Z, Ci Z, Wei X, Qu J and Guan W 2020. Content analysis of vitamins, dietary fibers and amino acids in a wide collection of barley (*Hordeum vulgare* L.) from Tibet, China. *Bioinformation* **16**(4): 314-325. doi: <https://doi.org/10.6026/97320630016314>
- Kara B and Akman Z 2007. Correlation and path coefficient analysis in the local wheat ecotypes. *SDU J. Nat. Appl. Sci.* **1**(3): 219-224.
- Kaydan D and Yagmur M 2007. A research on yield and yield components of some two-rowed barley varieties (*Hordeum vulgare* L.) in Van ecological conditions. *J. Agric. Sci.* **13**(3): 269-278.
- Kıran AK 1999. Characterization of some barley (*Hordeum vulgare* L.) plant genetic resources material. *Anadolu J. Aegean Agric. Res. Inst.* **9**(2): 72-90.
- Monouchehr A 2006. Path analysis of barley (*Hordeum vulgare* L.) yield. *J. Agric. Sci.* **12**: 227-232.
- Shalaby E, Abd El Rahim H and Mosaad M 1988. Effect of watering regime on morphophysiological traits and harvest index and its components of wheat. *Assiut J. Agric. Sci.* **19**(5): 195-207.
- Sharafi S, Ghassemi-Golezani K, Mohammadi S, Lak S and Sorkhy B 2014. Evaluation of barley genotypes (*Hordeum vulgare* L.) by drought tolerance indices and multivariate analysis. *Int. J. Bio-Sci.* **4**(7): 117-129. doi: <https://doi.org/10.3329/jbs.v21i0.22525>
- Sonmez AC 2021. Investigation of some agricultural and quality traits of advanced barley (*Hordeum vulgare* L.) lines. *J. Tekirdag Agric. Fac.* **18**(3): 545-553. doi: <https://doi.org/10.33462/jotaf.859541>
- Tibshirani R 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B Stat. Methodol.* **58**(1): 267-288. doi: <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Yılkan Y, Ozturkcü Y, Arpalı D and Akkol S 2020. The relationships among phenological stages, grain yield and some yield components of two-rowed barley cultivars under Van ecological conditions. *YYU J. Agr. Sci.* **30**(4): 751-760. doi: <https://doi.org/10.29133/yyutbd.746577>
- Yuksel S and İncekarakaya SU 2022. Relationships between yield and yield characters of some hulless barley genotypes (*Hordeum vulgare* L. var. nudum Hook. f.) grown in dry conditions. *Turk. J. Nat. Sci.* **11**(2): 7-12. doi: <https://doi.org/10.46810/tdfd.867991>
- Zhou M 2009. Barley production and consumption. In: Zhang G and Li C (Eds), *Genetics and Improvement of Barley Malt Quality*. Springer, Berlin. pp. 1-17.

(Manuscript received on 30 March, 2026; revised on 29 August, 2026)