MULTICOLLINEARITY IN REGRESSION ANALYSIS OF WHEAT GLUTEN INDICATOR DURINGITS STORAGE

Vira Malkina[1], Serhii Kiurchev[1], Valentuna Verkholantseva[1], Viktor Dubik[2]
[1]Tavria State Agrotechnological University, Ukraine;
[2]State Agrarian and Engineering University in Podilya, Ukraine
vira.malkina@tsatu.edu.ua, dakgps@pdatu.edu.ua

Abstract. When constructing regression models describing the storage processes of agricultural products, researchers often ignore the effect of multicollinearity factors (the multicollinearity factor effect). In this case, the interpretations of the estimates of the regression model parameters are not adequate. It is suggested to perform multicollinearity diagnostics as one of the steps of the regression analysis. In the case of correlation of factors, it is proposed to construct a regression model by regularizing the LASSO parameters. In this paper a regression model is built, that describes the effect on the quality indicators of wheat grain (gluten) of storage parameters in the granary during cooling ("Average grain temperature in the granary, ºC", "Humidity of the grain, %", "Air temperature in the granary, ºC", "Refrigerant temperature, ºC" and "Volume air flow, $m^3 \cdot h^{-1}$"). The parameters of the constructed model are analyzed. The conditions for the long-term storage of cereals are the synthesis of active ventilation by the convective stream of the coolant, turning of loose mass. The developed diagrams of granaries give an opportunity to realize the complex technological processing of products at the expense of efficient distribution of the cooled air flow passing through the raw materials, regulating its gas composition in intergranular spaces, humidity, temperature and intensity of physiological and microbiological processes in these layers.

Keywords: multicollinearity, regression model, LASSO, gluten indicator, significance of factors.

Introduction

The agricultural industry is one of the most important in the economy of any country. The most important stage in providing food along with the crop cultivation is the storage organization. For efficient wheat storage, special granaries are used. The grain quality depends on the storage conditions of grain in the granary. The main indicators that influence the formation of the grain class are the values of gluten percentage, gluten deformation index and grain moisture. In this article, we look at changes in the gluten index. One of the grain quality indicators is the gluten index. The wheat gluten index varies during storage. According to studies [1], when storing grain, there is a change in various indicators that affect the quality composition of the grain, including the percentage of gluten. It is known that the gluten index value depends on storage conditions, namely, grain temperature and humidity. To obtain optimal values of these indicators, competent organization of conditions in the granary is necessary [2].

In a number of works [2; 3], to ensure effective storage conditions, it is proposed to use a special mode of grain cooling with the help of the grain mass agitating (pneumatic impulse bubbling). Since grain temperature affects the activity of enzymes, and these, in turn, affect the main indicators of grain quality (in our case gluten), it can be argued that changes in the gluten index depend on the grain temperature during storage [1; 4].

To analyze the grain storage process, it is proposed to apply the methods of the correlation and regression analysis based on experimental studies of the indicators of grain storage and its quality as a result of such storage.

On the basis of the regression model, it is proposed to analyze the degree of influence of each of the factors of external conditions that affect the gluten index of wheat during storage. This allows to identify the most relevant factors and determine their degree of influence.

Thus, the purpose of the present study is to determine, on the basis of a regression analysis, the dependence of the change in gluten content in wheat in storage on external storage factors, the degree of influence of each of the factors on the gluten index.

In modern studies, methods of the correlation and regression analysis are often used. Calculation of correlation coefficients allows revealing the tightness and direction of association of the studied indicators. The regression analysis, which is a continuation of the correlation analysis, allows us to determine the analytical expression of the relationship of the resulting value with factor indicators. Using the regression model lets us come to the important conclusions about the impact of each factor

DOI:10.22616/ERDev.2020.19.TF233

on the response variable and evaluate the degree of this impact, predict the results of managerial impact on the change in the values of the factors of this model, analyze and determine the values of factors to ensure the optimal value of the response variables.

An effective tool for analyzing the wheat storage process is the regression model, which is built on the basis of collected experimental observations.

As it is known, in order to obtain unbiased consistent efficient estimates of the parameters of the regression model, the conditions of the Gauss–Markov theorem must be satisfied [5].

Materials and methods

One of the conditions of the theorem is that vector factors must be linearly independent. A serious obstacle to constructing an adequate regression model may be the effect of correlation of influencing factors, i.e. multicollinearity.

The effect of multicollenearity means that at least two factors (independent variables) that participate in the model are associated with close correlation dependence between themselves. Many publications deal with the complexity and negative impact of multicollenearity on the entire research process [6-9]. As described in the literature, the main problem in the manifestation of multicollenearity is biased estimate, which means large p-values for assessing the statistical significance of the regression model parameters. As a result, it leads to inconsistent estimator [7; 8; 10].

Multicollinearity among independent variable factors makes it almost impossible to adequately interpret the identification and assessment of each factor influence on the response variables based on the regression model. This is due to the effect of overlapping information. As it is known, the coefficients of the regression model are used to interpret the indicator of the expected value changes of the response variable due to the factor value increase by one with constant values of other variables. With the manifestation of the effect of multicollinearity, such an interpretation does not give an adequate result.

Therefore, two problems should be considered – how to identify the effect of multicollinearity and how to build an effective regression model in the multicollinearity conditions.

To solve the first problem, there are many ways to detect multicollinearity. One of the widely used methods for detecting multicollinearity is the method using the variance inflation factor (*VIF*) [9].

It is believed [10] that if $VIF > 10$, then multicollinearity is present. A sign of multicollinearity is the large values of the coefficient of determination, the adequacy of the model (which is confirmed by the Fisher's criterion), and at the same time a large number of statistical significance of the model parameters (which is checked by the Student's criterion). In order to make the estimates of the regression model parameters, in the case of multicollinearity, effective and reliable, it is proposed to use regularization methods that correct deviations from the normal distribution of residues. This method is the LASSO method (Least Absolute Shrinkage and Selection Operator) [10].

To study the quality of wheat storage and the influence of environmental parameters (storage conditions) in the granary on the quality indicators of wheat (gluten), experimental studies have been conducted.

In the laboratory of the Dmitry Motornyi Tavria State Agrotechnological University (Melitopol, Ukraine), an investigation of the process of wheat cooling using various storage modes has been carried out in the experimental granary, and wheat quality indicators have been established during storage [3; 11].

In the process of the research, measurements were made of various indicators of the chemical composition of the grain, such as protein content (24-29 %), starch (48-62 %), enzyme activity. However, the main indicators that influence the formation of the grain class are the percentage of gluten, the gluten deformation index and the moisture content of the grain. In this article, we consider changes in the gluten index. Among the criteria for evaluating the quality of the grain products to be stored, the quantity and quality of gluten, which is understood to mean a highly hydrated protein substance consisting mainly of gliadin and gluten, were chosen. Manual gluten washing was used in the research [12].

To measure the index of deformation of gluten the device IDG-1 was used, the principle and method of operation, which are based on the measurement of the amount of residual deformation of the gluten sample after the action of the tare punch load for a specified control time, in particular 30 s.

The experimental setup for studying the process of grain cooling (Fig. 1) consists of a grain ministorage, cooling devices, and temperature sensors.

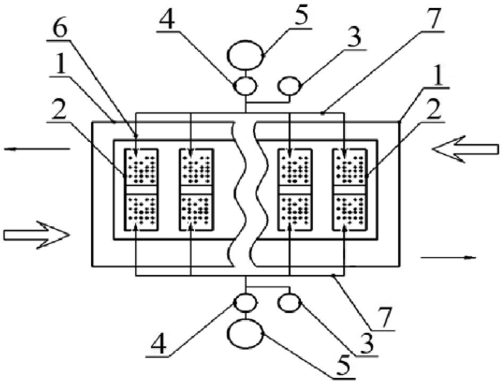

Fig. 1. **Scheme of the experimental installation for studying the process of grain cooling:**
1 – granary; 2 – pallets; 3 – fan; 4 –pneumatic impulse bubbler; 5 – coolers;
6 – collectors; 7 – pipelines

Changes in refrigerant feeding have been made by controlling the rotational speed of the refrigerant fan using the AOSN-20-220-75 autotransformer, which contains a movable flux-film contact of the graphite roller shape, which makes it possible to change the voltage value. The fan drive shaft speed has been registered using a wireless tachometer UNI-T UT372. To determine the mass, a JADEVERSNUGII-150 laboratory balance has been used with a measuring range of 0-500g.Refrigerant temperature has been evaluated with a certified Infrared Thermometer integrated into the SPLIT temperature controller. To control grain moisture, a Wile-55 moisture meter was used. Grain feed has been conducted with the point sampling method using a laboratory sampler.

The measurement results in normalized values are shown in Table 1. Figure 2 shows the graphs of the change in indicators obtained experimentally on a normalized scale, where x_1 is the average grain temperature in the granary, x_2 is the moisture in the grain, x_3 is the air temperature in the granary, x_4 is the volumetric air supply, x_5 is the refrigerant temperature.

Normalized values of factors are calculated by the formula

$$x_{ij}=\frac{z_{ij}-\bar{z}_i}{\bar{\sigma}_i} \,, \tag{1}$$

where z_{ij} – real values of factors;
$\bar{z}_i$ – mean values of factors;
$\bar{\sigma}_i$ – variance of factors.

Table 1 shows the normalized values $\bar{z}_i$ and $\bar{\sigma}_i$.

Table 1

Values $\bar{z}_i$ and $\bar{\sigma}_i$ factors

Factors	**Average grain temperature in the granary, ºC**	**Moisture content of the grain, %**	**Air temperature in the granary, ºC**	**Volumetric air supply, m^3·h^{-1}**	**Refrigerant temperature, ºC**
Notation of factors	x_1	x_2	x_3	x_4	x_5
$\bar{z}_i$	10.8	12.4	7.8	17160.7	3.0
$\bar{\sigma}_i$	1.87	0.33	0.97	1.87	2.56

Fig. 2 shows the dynamics of changes in each indicator by weeks in normalized values, where x_1 is the average grain temperature in the granary, x_2 is the moisture in the grain, x_3 is the air temperature in the granary, x_4 is the volumetric air supply, x_5 is the refrigerant temperature, y is gluten of grain, %.

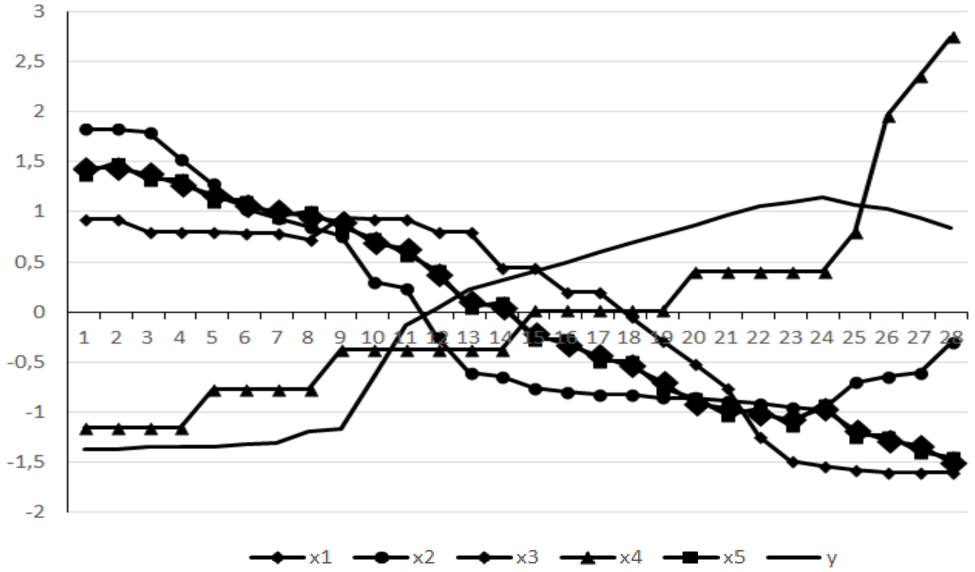

Fig. 2. **Dynamics of the indicators given in the normalized scale**

Results and discussion

On the basis of the data obtained as a result of measuring the values of the wheat storage factors and the gluten index, a correlation analysis was performed.

Before constructing the regression model, the data have been analyzed for the presence of the multicollinearity effect. All partial correlation coefficients between the factors have values close in magnitude to unity, which already indicates the high level of correlation between them.

We construct a linear regression model of the form

$$y=43.51-0.3941x_1-0.952x_2-2.34\cdot10^{-6}x_3-63.7\cdot10^{-3}x_4-29.17\cdot10^{-3}x_5\ . \tag{2}$$

All factors except b_0 are not significant, and the determination coefficient is 0,976. Moreover, the model is adequate, according to the Fisher's criterion, at a significance level of 0.05. This indicates the presence of a strong correlation of factors. *VIF* values for each factor are shown in Table 2.

Table 2

***VIF* values for each factor**

Factors	**Average grain temperature in the granary, ºC**	**Moisture content of the grain, %**	**Air temperature in the granary, ºC**	**Volumetric air supply, $m^3 \cdot h^{-1}$**	**Refrigerant temperature, ºC**
VIF	497.53	26.92	345.67	20.12	8.38

As we can see, all values of the *VIF* indicator exceed 5, which indicates the strong correlation of the factors among themselves. In other words, there is an effect of multicollinearity.

Many works deal with the problem of constructing an adequate regression model under conditions of multicollinearity [6-8]. Most often, it is proposed to remove one of the correlating factors from the analysis. However, this will lead to information loss. Based on such a model, we cannot assess the degree of influence on the response variable of those factors that we have removed from consideration. In addition, in our case, practically all factors correlate among themselves, and their removal from the model will lead to the inefficiency of the entire study.

To construct the regression model under conditions of multicollinearity, it is proposed to use the LASSO method. The LASSO method regularizes parameters and overcomes the disadvantages of the least squared error method [10].

The essence of regularization is that to minimize the function

$$L=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{i=1}^{n}|\beta_i|, \tag{3}$$

where λ – set parameter (penalty);
β_i – regression model coefficients;
y_i – experimental values of the indicator;
$\hat{y}$ – theoretical values of the indicator, calculated on the basis of the regression equation.

That is, a penalty is introduced for unreasonably large values of the β_i parameters. Moreover, the value of this penalty is proportional to the value of the λ parameter, with which we can choose a more stable solution.

After constructing the regression model by the LASSO method, the regression analysis has been performed for various values of the parameter λ. Using the results of cross-validation, we determine the appropriate values of the λ parameter. According to the analysis, the following values of λ have been selected: λ_{min} = 0.001651433 and λ_{1se} = 0.08219219.

Then the regression equation constructed by the LASSO method at the value of λ_{min} = 0.001651433 has the form:

$$y=44.9106-0.2901x_1-1.0469x_2-57.4\cdot10^{-3}x_3-98.9\cdot10^{-3}x_4-8.17\cdot10^{-6}x_5\,. \tag{4}$$

We give the estimates of the error variance for a linear regression constructed by the least squared error method and the LASSO method. In the first case (for the least-squared error method), the error variance estimate is 0.02274. For the regression constructed by the LASSO method with λ_{min} = 0.001651433, the error variance estimate is 0.02278. For the regression constructed by the LASSO method at λ_{1se} = 0.08219219, the error variance estimate is 0.02587.

The analysis of the constructed regression model allows to determine the priority of factors of external conditions during storage of wheat and the degree of their influence on changes of the gluten index. To assess the degree of influence of each factor on the gluten index during storage, we find R^2 determination coefficient, E_i elasticity coefficients, β_i coefficients and Δ_i, which are calculated based on the parameters of the constructed regression model. The values of the calculated indicators E_i, β_i and Δ_i for each coefficient of the regression model are shown in Table 3.

Table 3

Elasticity coefficients

Coefficent	Average grain temperature in the granary, ºC	Moisture content of the grain, %	Air temperature in the granary, ºC	Volumetric air supply, $m^3\cdot h^{-1}$	Refrigerant temperature, ºC
E_i	$-116.2\cdot10^{-3}$	$-481.1\cdot10^{-3}$	$-028.6\cdot10^{-3}$	$-5.2\cdot10^{-3}$	$-6.3\cdot10^{-3}$
β_i	$-559.2\cdot10^{-3}$	$-353.1\cdot10^{-3}$	$-190.4\cdot10^{-3}$	$-34.9\cdot10^{-3}$	$-151.0\cdot10^{-3}$
Δ_i	$554.6\cdot10^{-3}$	$344.7\cdot10^{-3}$	$188.6\cdot10^{-3}$	$29.3\cdot10^{-3}$	$-116.6\cdot10^{-3}$

Based on the coefficient of elasticity, we make the following conclusion. The increase in the "Average grain temperature in a granary, ºC" factor by 1 % is the reason of the decrease of the "Gluten, %" indicator by 0.116 %. The increase in the "Moisture, %" factor by 1 % leads to the decrease of the "Gluten, %" indicator by 0.48 %. The increase in the "Air temperature in the granary, ºC" factor by 1 % is the reason of the decrease of the "Gluten, %" indicator by 0.02 %. The increase in the "Refrigerant temperature, ºC" factor by 1 % causes the decrease in the value of the "Gluten, %" indicator by 0.063 %. The increase in the "Volumetric air supply, $m^3\cdot h^{-1}$"factor by 1 % is the reason of the decrease of the value of the "Gluten, %" indicator by 0.052 %.

On the grounds of the β_i coefficients, we put the factors in order according to the degree of their influence on the gluten index. The "Average grain temperature in the granary, ºC" factor has the greatest impact, then "Grain moisture, %", "Air temperature in the granary, ºC", "Refrigerant temperature, ºC" and "Volumetric air supply, $m^3\cdot h^{-1}$" follow.

Using the Δ_i, coefficients we estimate the proportion of the influence of each factor on the gluten index. The share of the "Average grain temperature in the granary, ºC" factor is 55 %, the share of the "Humidity of grain, %" factor is 34 %, the share of the "Air temperature in the granary, ºC" factor is 18 %, the share of the "Refrigerant temperature, ºC" factor is 11 % and the share of the "Volumetric air supply, $m^3 \cdot h^{-1}$" factor is 2.9 %.

Conclusions

Often, when constructing regression models that describe the processes of storage and processing of agricultural products, the correlation effect of factors is not taken into account. The adverse multicollinearity effect negatively affects the interpretation of the constructed model, namely, when analyzing the degree of the influence of each individual factor on the studied indicator. The proposed model describes the effect of grain storage indicators (grain mass temperature, grain moisture, air temperature in the granary, refrigerant temperature and air supply volume) on the quality characteristics of grain (gluten). When analyzing data, using the partial correlation coefficients and *VIF* index, the effect of multicollinearity has been revealed. In this case, using the least squared error method is not effective. The regression model has been built based on the LASSO method, which allows constructing effective estimates of regression parameters. According to the constructed model, the factors affecting the gluten index have been analyzed and it was shown that the most significant factor is "Average grain temperature in the granary, ºC", then "Humidity of grain, %", "Air temperature in the granary, ºC", "Refrigerant temperature ,ºC" and "Volumetric air supply, $m^3 \cdot h^{-1}$" follow. The indicators of the share of the each factor influence on the wheat gluten index during storage have been determined. The share of the "Average grain temperature in the granary, ºC" factor is 55 %, the share of the "Humidity of grain, %" factor is 34 %, the share of the "Air temperature in the granary, ºC" factor is 18 %, the share of the "Refrigerant temperature, ºC" factor is 11 % and the share of the "Volumetric air supply, $m^3 \cdot h^{-1}$" factor is 2.9 %.

References

[1] Баум А. Е. Применение искусственно охлажденного воздуха при хранении зерна за рубежом (Baum A.E. The use of artificially cooled air when storing grain abroad. Series "Elevator Industry") Moscow: Central Scientific Research Institute for Technology and Economics of the Ministry of the USSR, 1977. 28 p. (In Russian).

[2] Skaletska L.V., Dukhovskaya T.M., Senkov A.M. Technology of storage and processing of crop production. Workshop. Kiev. High school, 1994. 330 p.

[3] Palamarchuk I., Kiurchev S., Verkholantseva V., Palianychka N., Hryhorenko O. Optimization of the Parameters for the Process of Grain Cooling. Renewable Energy Sources: Engineering, Technology, Innovation, Springer Proceedings in Energy, ICORES 2018. Chapter No. 94. pp 981-988 .

[4] Трисвятський Л.А. Хранение зерна. (Trisvyatsky L.A. Grain storage). Moscow: Agropromizdat, 1986. 352p. (In Russian).

[5] Damodar N. Gujarati. Basic Econometrics. Fourth edition. The McGraw-Hill Companies, 2004. 1002 p.

[6] Aiken L.S., West S.G. Multiple regression: Testing and interpreting interaction. Newbury Park, CA: SAGE, 1991. 212 p.

[7] Masom G. Coping with multicollinearity. The Canadian Journal of program evaluation, vol. 2. 1987, pp. 87-93

[8] Mela C.F., Kopalle Praveen K. The impact of collinearity on analysis: the asymmetric effect of negative correlations. Applied Economics, vol. 34. 2002, pp. 667-677.

[9] Kutner M. Nachtsheim C. Neter J. Applied Linear Statistical Models. Fourth edition. McGraw-Hill/ Irwin, 2004. 1396 p.

[10] Hastie T., Tibshirani R., Wainwright M. Statistical Learning with Sparsity: The Lasso and Generalization/ Chapman and Hall/CRC, 2015, 362 p.

[11] Kiurchev S., Vercholantseva V. Linear and nonlinear relationship of wheat storage characteristics. Canadian Scientific Journal. ISSUE 1. 2015. VOL. 2. pp. 10-15.

[12] State standard 13586.1-68. Corn. Methods for determining the quantity and quality of gluten. Moscow: Gosstandart, 1968. 4 p.