Scientific Papers Series Management, Economic Engineering in Agriculture and Rural Development
Vol. 15, Issue 4, 2015
PRINT ISSN 2284-7995, E-ISSN 2285-3952

MULTIPLE CORRELATION AND REGRESSION IN PREDICTING MILK PRICE

Agatha POPESCU

University of Agricultural Sciences and Veterinary Medicine Bucharest, 59 Marasti, District 1, 11464, Bucharest, Romania, Phone: +40213182564, Fax:+40213182888, Email: agatha_popescu@yahoo.com

Corresponding author: Email: agatha_popescu@yahoo.com

Abstract

The purpose of the paper was to analyze the relationship between milk production and dairy bovine livestock and milk price, using the multiple correlation and regression models. The reference period is the 2005-2014 decade in dairy sector of Romania and the data were provided by the National Institute of Statistics. The simple correlation coefficient R_{XZ} = - 0.477 reflected a negative relationship between milk production and milk price, and the coefficient of correlation R_{YZ} = - 0.676, also reflected a negative relationship between the dairy livestock and milk price. While, the total coefficient of linear multiple correlation, $R_{Z.XY}$ = 0.771, reflected a significant positive relationship between milk price, milk production and the dairy livestock, the partial coefficient $R_{ZX.Y}$ = - 0.537 reflected a negative middle link between milk production and milk price, when the dairy livestock is constant. Also, the partial coefficient of multiple correlation, $R_{ZY.X}$ = - 0.709, reflected a strong negative influence of milk production on the pair milk price and dairy livestock. The linear multiple regression had the formula: Z= - 0.00349 X + 0.08305 Y + 165.68 and the width of the confidence interval, $\delta_{\alpha/2}$, was 29.08. In 2015, for X_e= 50,025.45 thousand hl milk production and Y_e= 1,251.23 thousand heads dairy livestock, the predicted milk price was 95.01 Lei/hl. Therefore, the multiple correlation and regression are important mathematical tools to describe the relationships between milk production, the dairy bovine livestock and milk price and predict milk price based on the other factors.

Key words: *dairy bovine livestock, milk price prediction, milk production, multiple correlation, multiple regression*

INTRODUCTION

Milk production should be analyzed in close connection with milk price and the dairy livestock, because it depends on the milked animals and their milk yield, and milk price is determined by demand/supply ratio. However, dairy farmers are complaining that farm gate milk price is frequently very small, affecting profitability [12].

For this reason, the applied statistics is helpful to better understand the relationship between milk production, dairy livestock and milk price. In this respect, multiple correlation and multiple regression are mathematical tools which allow to identify the linear simple correlation between the variables as well as the multiple correlation coefficients which may reflect the influence of one or the combined influence of two factors on the 3rd one. [1, 8, 9, 10]

The significance of the coefficient of correlation is tested using Fisher's test, and also Student's test, which involve to compare the determined values with the quantiles presented in tables, for specific probabilities and degrees of freedom and finally to decide to accept or not the H hypothesis [2, 3, 4,7].

The multiple regression allows to predict the estimated values of a dependent factor based on the values of the independent factors. [5, 6]

In this context, the paper aimed to make an analysis of the relationship between milk production, dairy bovine livestock and milk price in Romania, based on the empirical data provided by the National Institute of Statistics [11] and using the multiple correlation and multiple regression as processing methods and basic tools in predicting milk price.

MATERIALS AND METHODS

The economic indicators taken into account were: milk production obtained from dairy cows and buffalos (thousand hl), the X variable, the number of dairy bovine livestock

Scientific Papers Series Management, Economic Engineering in Agriculture and Rural Development
Vol. 15, Issue 4, 2015
PRINT ISSN 2284-7995, E-ISSN 2285-3952

(thousand heads), the Y variable, and milk price(lei/hl), the Z variable.
The data were collected from the National Institute of Statistics for the period 2005-2014 [11]
The following statistical parameters were determined for each variable: mean, standard deviation, and coefficient of variation according to the formulas:
Mean:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Standard deviation,

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Coefficient of variation,

V% = (S/ $\bar{X}$) * 100.
Also, for each pair of variables, there were calculated: the covariance, and coefficients of simple linear correlation, according to the formulas:
Covariance:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Coefficient of correlation, r_{xy}:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

The coefficients of multiple linear correlation have been determined as follows:
-The total coefficient:

$$R_{z,xy} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xz}r_{yz}r_{xy}}{1 - r_{xy}^2}}$$

-The partial coefficients:

$$R_{ZX.Y} = \frac{R_{ZX} - R_{ZY}R_{XY}}{\sqrt{(1 - R_{ZY}^2)(1 - R_{XY}^2)}}$$

$$R_{ZY.X} = \frac{R_{ZY} - R_{ZX}R_{XY}}{\sqrt{(1 - R_{ZY}^2)(1 - R_{XY}^2)}}$$

The correlations were tested using Fisher Test, according to the formula:

$F_{(XY)} = \frac{R_{Z.XY}^2}{1 - R_{Z.XY}^2} : \frac{2}{n-3}$ where F is Fisher variable with the degrees of freedom [2; n-3] DOF.

The significance of correlation was established comparing F calculated values with F from tables $F_{0.05}$, $F_{0.01}$ and $F_{0.001}$.
The testing of the hypothesis H compared to the alternative hypothesis, $\overline{H}$ was based on the comparison t Student Test, the calculation of t being made using the formula:

$$t_X = \frac{R_{ZX.Y}^2}{\sqrt{1 - R_{ZX.Y}^2}}\sqrt{n-3}.$$

The calculated t was compared to t values from tables for 7 DOF for $t_{0.025}$, $t_{0.05}$ and $t_{0.005}$.
The A variation contribution X, Y, the X.Y interaction and E to the Z variation are:

$A_{(X,Y)} = R^2_{Z.XY}$
$A_X = A_{(X.Y)} - A_Y = R^2_{Z.XY} - R^2_{ZY}$
$A_Y = A_{(X.Y)} - A_X = R^2_{Z.XY} - R^2_{ZX}$
$A_{(X.Y)} = A_{(X,Y)} - A_X - A_Y = R^2_{ZX} + R^2_{ZY} - R^2_{Z.XY}$
$A_E = 1 - A_{(X,Y)} = 1 - R^2_{Z.XY}$

The linear multiple regression was determined using the formula: $Z = B_0 + B_1X + B_2Y$, where B_1 and B_2 are the solutions of the system of linear equations:

$B_1S^2_X + B_2S_{XY} = S_{XZ}$
$B_1S_Y + B_2S^2_Y = S_{YZ}$
and $B_0 = \bar{Z} - B_1\bar{X} - B_2\bar{Y}$.

The width of the confidence interval was calculated according to the formula:

$\delta_{\alpha/2} = \sqrt{(n-1)(1 - R^2_{Z.XY})/[n(n-3)]} \cdot S_Z \cdot t_{\alpha/2}$ for (n-3) DOF.

RESULTS AND DISCUSSIONS

The basic data for the variables taken into account are presented in Table 1.
The statistical parameters: means and standard deviations for the three variables are presented in Table 2.
The covariance and coefficients of simple linear correlation are given in Table 3.
As one can see, the simple correlation coefficient between Milk production (X) and Milk price (Z), registered a negative value,

R_{XZ} = - 0.477, reflecting that between the two variables is not a positive relationship, meaning that when milk production is higher, milk price will decline, and the lower milk production, the higher milk price.
Also, between the Dairy bovine livestock (Y) and Milk price (Z), the correlation coefficient had a negative value, R_{YZ} = - 0.676, signifying that when the livestock declines, the milk price will increase, and, the higher the livestock in close connection with a higher production, the lower the milk price.

Table 1.Milk production, Dairy Bovine Livestock and Milk Price, Romania, 2005-2014

Year	Milk Production (thou hl) X	Dairy Bovine Livestock (thou heads) Y	Milk Price (Lei/hl) Z
2005	55,121	1,812	64
2006	58,307	1,810	67
2007	54,875	1,732	75
2008	53,089	1,639	88
2009	50,570	1,569	90
2010	42,824	1,299	94
2011	43,947	1,266	108
2012	42,036	1,265	111
2013	42,593	1,279	211
2014	50,535	1,307.3	209

Source: National Institute of Statistics, 2006-2015 [11]

Table 2. The average values, the standard deviations and the coefficients of variation for milk production, dairy bovine livestock and milk price

Statistical parameter	Milk production, X	Dairy bovine livestock, Y	Milk price, Z
Mean	$\bar{X}$= 49,389.7	$\bar{Y}$= 1,497.83	$\bar{Z}$= 111.7
Standard deviation	S_X = 6,072.22	S_Y = 237.50	S_Z = 54.06
Coefficient of variation	V_X% = 12.29 %	V_Y% = 15.85	V_Z% = 48.39

Source: Own calculations.

Table 3. The values of covariance and the simple correlation coefficients between milk production (X), dairy bovine livestock (Y) and milk price (Z)

Statistical parameter	Milk production(X) and Dairy bovine livestock (Y)	Milk production(X) and Milk price (Z)	Dairy bovine livestock (Y) and Milk price (Z)
Covariance	S_{XY} = 1,202,354.99	S_{XZ} = - 140,855.89	S_{YZ} = 7,815.24
Correlation coefficient	R_{XY} =0.926	R_{XZ} = - 0.477	R_{YZ} = - 0.676

Source: Own calculations.

The vector of the means, the matrix of covariance and the matrix of the linear correlations between the variables are presented below.
The vector of the means is:
($\bar{X}$= 49,389.7 ; $\bar{Y}$= 1,497.83 ; $\bar{Z}$= 111.7).
The matrix of the covariance is:

S^2_X = 36,871,855.72	S_{XY} = 1,202,354.99	S_{XZ} = - 140,855.89
S_{XY} = 1,202,354.99	S^2_Y = 56,406.25	S_{YZ} = 7,815.24
S_{XZ} = - 140,855.89	S_{YZ} = 7,815.24	S^2_Z = 2,922.48

The matrix of the linear correlation is**:**

1	R_{XY} =0.926	R_{XZ} = - 0.477
R_{XY} =0.926	1	R_{YZ} = -0.676
R_{XZ} = - 0.477	R_{YZ} = -0.676	1

The coefficients of linear multiple correlation are presented in Table 4. The total coefficient of linear multiple correlation, $R_{Z.XY}$, reflects the intensity of the statistical connection between Z, milk price, and the other two variables, XY, milk production and dairy livestock. One can notice that its value, $R_{Z.XY}$ = 0.771, is enough high and positive, therefore milk price is influenced by the both variables.
The partial coefficient of linear multiple correlation, $R_{ZX.Y}$, reflects the intensity of the statistical link between Milk production, X

and Milk price, Z, when the dairy livestock Y= constant. The value of $R_{ZX.Y}$ = -0.537 shows that, when the dairy livestock is constant, between milk production and milk price is a negative middle correlation.

The partial coefficient of linear multiple correlation, $R_{ZY.X}$, gives information about the statistical relationship between Milk price, Z., and the Dairy Livestock, Y, when Milk production X= constant. One can see that its value was $R_{ZY.X}$ = -0.709, reflecting a strong negative influence of milk production on the pair milk price and dairy livestock.

Table 4. The coefficients of linear multiple correlation

The total coefficient of linear multiple correlation $R_{Z.XY}$	The partial coefficient of linear multiple correlation $R_{ZX.Y}$	The partial coefficient of linear multiple correlation $R_{ZY.X}$
$R_{Z.XY}$ = 0.771	$R_{ZX.Y}$ = -0.537	$R_{ZY.X}$ = -0.709

Source: Own calculations.

The testing of the coefficients of linear multiple correlation.

The total coefficient of linear multiple correlation, $R_{Z.XY}$, varies from a sample to another around the total unknown coefficient of correlation, $p_{Z.XY}$.

According to its formula, the value of $F_{(XY)} = \frac{R^2_{Z.XY}}{1-R^2_{Z.XY}} : \frac{2}{n-3}$ = 5.12

From the Fisher's tables, for (2:n-3) DOF, the found values for $F_{(XY)}$ are: $F_{5\%}$ =4.74, $F_{1\%}$ =9.55 and $F_{0 \cdot 1\%}$ =21.69. The interpretation is the following one: $F_{5\%} = 4.74 < F_{(XY)} = 5.12 < F_{1\%} = 9.55$, therefore the hypothesis H is not accepted, $p_{Z.XY} \neq 0$, and Z and the pair (X,Y) are significantly linear correlated.

The partial coefficients of linear multiple correlation, $R_{ZX.Y}$ and $R_{ZY.X}$, varies from a sample to another around the unknown partial coefficient of correlation, $p_{ZX.Y}$ and $p_{ZY.X}$.

Using the formula given below, it was calculated t_X:

$t_X = \frac{R^2_{ZX.Y}}{\sqrt{1-R^2_{ZX.Y}}}\sqrt{n-3}$, and it was obtained t_X = 0.901.

From the Student's tables, for (n-3=7) DOF, the found values for t_X are: $t_{0.025\%}$ =2.365, $t_{0.0025\%}$ =3.499 and $t_{0.0005\ \%}$ =5.408. The interpretation is the following one: Because t_X = 0.901 < $t_{0.025\%}$ =2.365, the hypothesis H is accepted, $p_{ZX.Y}$ = 0, and Z and X are not linear correlated, when Y= constant.

Using the formula given below, it was calculated t_Y:

$t_Y = \frac{R^2_{ZY.X}}{\sqrt{1-R^2_{ZY.X}}}\sqrt{n-3}$, and it was obtained t_Y = 1.879.

From the Student's tables, for (n-3=7) DOF, the found values for t_Y are: $t_{0.025\%}$ =2.365, $t_{0.0025\%}$ =3.499 and $t_{0.0005\ \%}$ =5.408. The interpretation is the following one: Because t_Y = 1.879 < $t_{0.025\%}$ =2.365, the hypothesis H is accepted, $p_{Z.XY}$ = 0, and Z and Y are not linear correlated, when X= constant.(Table 5).

Table 5. Testing the coefficient linear of multiple correlation

	$F_{(XY)}$	t_X	t_Y
	5.12	0.901	1.879
Tabled values	$F_{5\%}$ =4.74, $F_{1\%}$ =9.55 $F_{0 \cdot 1\%}$ =21.69	$t_{0.025\%}$ =2.365, $t_{0.0025\%}$ =3.499 $t_{0.0005\ \%}$ =5.408	$t_{0.025\%}$ =2.365, $t_{0.0025\%}$ =3.499 $t_{0.0005\ \%}$ =5.408
Decision on the H hypothesis	H hypothesis can not be accepted, $p_{Z.XY} \neq 0$	H hypothesis must be accepted, $p_{ZX.Y}$= 0	H hypothesis must be accepted, $p_{ZY.X}$= 0
Conclusion	Z and the pair (X,Y) are significantly linear correlated.	Z and X are not linear correlated, when Y= constant.	Z and Y are not linear correlated, when X= constant

Source: Own calculations.

The contribution of X,Y, the X.Y interaction and E, error due to other factors to the Z variation is presented in Table 6.

Therefore, the total contribution of the variation of the pair of factors milk production and dairy livestock, $A_{(X,Y)}$, on milk price, Z, is 59.44 %.

The partial contribution $A_{(X)}$ of the variation of milk production, X, when the dairy livestock Y= constant on Z, was 13.80 %.

The partial contribution $A_{(Y)}$ of the variation of the dairy livestock, Y, when milk production, X=constant was 36.7 %.

The partial contribution $A_{(X.Y)}$ reflecting the interaction of milk production, X, and the dairy livestock, Y, on Z, was 8.90 %.

The contribution of the error, E, representing other factors of influence on milk price, Z, was 40.60 %.

Table 6.The contributions $A_{(X,Y)}$, A_X, A_Y, $A_{(X.Y)}$, and A_E to the variation of Z, Milk price

$A_{(X,Y)}$	A_X	A_Y	$A_{(X.Y)}$	A_E
$=R^2_{Z.XY}$	$=R^2_{Z.XY}-R^2_{ZY}$	$=R^2_{Z.XY}-R^2_{ZX}$	$=R^2_{ZX}+R^2_{ZY}-R^2_{Z.XY}$	$=1-R^2_{Z.XY}$
0.594= 59.44 %	0.138= 13.80 %	0.367= 36.7 %	0.089= 8.90 %	0.406= 40.60 %

Source: Own calculations.

The linear multiple regression was determined because the coefficient $R_{Z.XY}$ had a significant value.

The formula of the linear multiple regression is:

$Z= B_0+ B_1X+B_2Y$,

where B_0, B_1 and B_2 are the solutions of the system of linear equations:

$$B_1 \sum x_i^2 + B_2 \sum x_i y_i + B_0 \sum x_i = \sum x_i z_i$$
$$B_1 \sum x_i y_i + B_2 \sum y_i^2 + B_0 \sum y_i = \sum y_i z_i$$
$$B_1 \sum x_i + B_2 \sum y_i + n B_0 = \sum z_i$$

The formulas used to calculate B_0, B_1 and B_2 were:

$B_1 = Rzx.y\ (S_{Z.Y}/S_{X.Y})$

$B_2 = Rzy.x\ (S_{Z.X}/S_{Y.X})$ and

$B_0 = \bar{Z} - B_1\bar{X} - B_2\bar{Y}$

Therefore, solving the system of normal equations, we found the following solutions:

B_1 = - 0.00349

B_2 = 0.08305 and

B_0 = 165.68.

The B_1 value could be interpreted as follows: milk price will decline by - 0.00349 Lei per 1 hl less of milk production, when dairy livestock is constant.

The B_2 value could be interpreted: milk price will increase by 0.08305 Lei/hl per an increase of one thousand heads of dairy livestock, when milk production is constant.

Therefore, the linear multiple regression will have the formula:

Z= - 0.00349 X + 0.08305 Y + 165.68.

The prediction of Milk Price. Taking into account the average annual gain/loss of milk production ΔX= - 509.55 thousand hl, and the average annual gain/loss of dairy bovine livestock ΔY = - 56.07 thousand heads, registered in the period 2005-2014, we are expecting that in the year 2015, the milk production to be X_e= 50,025.45 thousand hl (50,535 thousand hl registered in the last year of the analysis, that is in 2014 - 509.55 thousand hl), and the dairy bovine livestock to become Y_e= 1,251.23 thousand heads (1,307.3 thousand heads in 2014 - 56.07 thousand heads).

Replacing X and Y in the formula of the linear multiple regression, we will obtain the expected milk price (Z_e), as follows:

Z_e = - 0.00349 *50,025.45 + 0.08305* 1,251.23 + 165.68 = 95.01 Lei/hl in the year 2015.

The width of the confidence interval, $\delta_{\alpha/2}$, was 29.08, according to the formula:

$$\delta_{\alpha/2}=\sqrt{(n-1)(1-R^2_{Z.XY})/[n(n-3)]}\,.\,S_Z\,.\,t_{\alpha/2}$$

for (n-3=7) DOF. For α= 5 %, in the Table it was found $t_{0.025}$ =2.36 for 7 DOF, so that $\delta_{2.5\%}$ = 29.08.

The regression plan with its width of the confidence interval will be:

Z= - 0.00349 X + 0.08305 Y + 165.68 ± 29.08.

Therefore, the maximum value of milk price will be 95.01 + 29.08 = 124.09 Lei/hl, and its minimum value will be 95.01 - 29.08 = 65.93 Lei/hl.

In Table 7 are given the values x_i, y_i, and z_i, the expected values z_e and the differences $\Delta z_i= z_i- z_e$.

Table 7. The x_i, y_i, and z_i values, the expected values z_e and the differences $\Delta z_i = z_i - z_e$

x_i	y_i	z_i	z_e	$\Delta z_i = z_i - z_e$
55,121	1,812	64	112	-48
58,307	1,810	67	105	-38
54,875	1,732	75	110	-35
53,089	1,639	88	109	-21
50,570	1,569	90	114	-24
42,824	1,299	94	116	-22
43,947	1,266	108	117	-9
42,036	1,265	111	118	-7
42,593	1,279	211	113	+98
50,535	1,307.3	209	103	+106

Source: Own calculations

CONCLUSIONS

Between Milk production and Milk price, the correlation coefficient, R_{XZ} = - 0.477, reflected a negative relationship, that is, the higher the milk production, the lower the milk price.

Between the Dairy bovine livestock and Milk price, the correlation coefficient R_{YZ} = - 0.676, also reflected a negative relationship, meaning the lower the dairy livestock declines, the higher milk price.

The total coefficient of linear multiple correlation, $R_{Z.XY}$ = 0.771, reflected that between milk price and milk production and the dairy livestock is a significant positive relationship.

The partial coefficient of linear multiple correlation $R_{ZX.Y}$ = - 0.537 showed that, when the dairy livestock is constant, between milk production and milk price is a negative middle correlation.

The partial coefficient of linear multiple correlation, $R_{ZY.X}$ = - 0.709, reflected a strong negative influence of milk production on the pair milk price and dairy livestock.

The total contribution of the variation of the pair of factors milk production and dairy livestock, $A_{(X,Y)}$, on milk price, Z, was 59.44 %.

The partial contribution $A_{(X)}$ of the variation of milk production, X, when the dairy livestock Y= constant on Z, was 13.80 %.

The partial contribution $A_{(Y)}$ of the variation of the dairy livestock, Y, when milk production, X=constant was 36.7 %. The partial contribution $A_{(X.Y)}$ reflecting the interaction of milk production, X, and the dairy livestock, Y, on Z, was 8.90 %. The contribution of the error, E, representing other factors of influence on milk price, Z, was 40.60 %.

The linear multiple regression had the formula: Z= - 0.00349 X + 0.08305 Y + 165.68 and the width of the confidence interval, $\delta_{\alpha/2}$, was 29.08, resulting Z= - 0.00349 X + 0.08305 Y + 165.68 ± 29.08.

In 2015, for X_e= 50,025.45 thousand hl milk production and Y_e= 1,251.23 thousand heads dairy livestock, the predicted milk price is 95.01 Lei/hl.

As a conclusion, the multiple correlation and regression could be successfully used to analyze the relationships between the milk production, the dairy bovine livestock and the milk price. Milk price could be easily predicted using the linear multiple regression function.

REFERENCES

[1] Abdi, H., 2007, Multiple Correlation Coefficient, In: Neil Salkind (Ed.) (2007). Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage

[2]Aldrich, J., 2005, Fisher and Regression, Statistical Science 20 (4): 401–417

[3]Anghelache, C, 2008, Treatise of Theoretical and Economic Statistics, the Economic Publishing House, Bucharest

[4]Armstrong, J. S., 2012,Illusions in Regression Analysis, International Journal of Forecasting (forthcoming), 28 (3): 689.

[5]Draper, N.R., Smith, H., 1998, Applied Regression Analysis (3rd ed.). John Wiley.

[6]Ene, D., 2012, Applied mathematics and statistics in agriculture, Bucharest, pp.184-193

[7]Frost Jim, 2014, Regression Analysis: How to Interpret S, the Standard Error of the Regression, http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-to-interpret-s-the-standard-error-of-the-regression

[8]Higgins, J., 2005, Introduction to multiple regression, Chapter 4, pp.111-115, http://www.biddle.com/documents/bcg_comp_chapter4.pdf

[9]Kennedy, J. B., Neville, A.M., 1986, Basic Statistical Methods for Engineers and Scientists, 3rd Ed., Harper and Row

[10]Montgomery, D.C., Runger, D.C., 1994, Applied statistics and probability for engineers, J.Wiley

[11]National Institute of Statistics, 2015, Romania, Statistical Yearbook, Chapter 14, Agriculture.
[12]Popescu Agatha, 2014, Research on profit variation depending on marketed milk and production cost in dairy farming, Scientific Papers Series Management, Economic Engineering in Agriculture and Rural Development, Vol.14(2):223-229