REGRESSION MODELLING IN PREDICTING MILK PRODUCTION DEPENDING ON DAIRY BOVINE LIVESTOCK

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 goal of the paper was to chose the best regression model for milk production, the dependent variable, Y, and dairy bovine livestock, the independent variable, X, testing two polynomial regressions: linear regression and quadratic regression. The data were collected from the National Institute of Statistics for the period 2007-2014. The discrimination between models was based on the standard error of each type of regression, choosing the one which assured the highest accuracy of the prediction. As a conclusion, between milk production and dairy bovine livestock, it is a strong and positive correlation, r= 0.884, and about 78.30 % of milk production is determined by the milked livestock. As long as the Regression Standard Error was smaller in case of the Linear Regression, σ_{est} = 2,286.028830 than in case of the Quadratic Regression, σ_{est} = 2,336.915726, the linear regression having the formula y =23.974 x + 13,527 proved to better fit the data.

Key words: *dairy bovine livestock, milk production, regression modelling, standard error*

INTRODUCTION

Milk production depends on the milked livestock and average milk production per animal. The main part of the milk production, more exactly 87.40 % of milk produced in Romania comes from dairy cows and buffalos. [9]

As long as in Romania milk yield is still very small, just about 3,364 liters per year and cow, it is obvious that the dairy bovine livestock has an important contribution to milk production.

The evolution and the relationship between milk production and dairy livestock could be studied using the statistical descriptive analysis and also the regression functions, because they are two variables closely connected, milk production being the dependent variable and dairy livestock the independent one.

Regression modeling and the value of the regression standard error is a useful tool to establish more accurately which is the most suitable regression type for a pair of variables. [4]

Polynomial regressions are most commonly used in such a case, and the calculation of standard error for each type of regression could discriminate them, and finally the researcher to chose that regression type whose standard error is the smallest one, because it assures the highest precision. [6]

In this context, the purpose of this paper was to test the most suitable polynomial regression**,** in two variants: linear regression and quadratic regression**,** for the pair of variables: milk production, the dependent variable Y, depending on dairy bovine livestock, the independent variable, X, using as discrimination tool the standard error of each regression type.

MATERIALS AND METHODS

The main indicators considered in this study have been: milk production coming from dairy cows and buffalos, the predictor Y, and the number of dairy bovine livestock, X.

The data were collected for the period 2007-2014 from Romania's Statistical Yearbook, provided by the National Institute of Statistics. [8]

For each variable, there were calculated the following statistical parameters: mean, median, variance, standard deviation, coefficient of variation, minimum and maximum value, according to the well known mathematical formulas. [1]

For the pair of variables taken into consideration, there were designed the scatter plots and determined two types of polynomial regressions: the linear regression and the quadratic regression, with their specific parameters, a and b in case of linear regression and a, b and c in case of the parabolic fit, the coefficient determination and standard error, using the facilities offered by Excel Data Analysis.

The regression functions had the well known formulas: y= a + bx in case of the linear regression and $y = ax^2 + bx + c$ in case the quadratic regression. [3, 7]

For each type of regression it was calculated the correlation coefficient, r, which measures how well a regression type or another fits the data. The interpretation was given depending on the r value as follows: if r value is close to 1, it indicates a good fit of the regression model, if r value is small, it is consider that the regression model is not appropriate.

The Regression Standard Error was used to measure the accuracy of the predicted Y, reflecting how much is the deviation of the Y observed values from the theoretical values Y_i' are situated on the regression line.

The calculation of the standard error supposed: (i) to give values to X in the linear regression formula in order to obtain the predicted Y', (ii) to make the difference between the observed values of Y and the predicted values Y', that is to determine the errors of prediction, Y- Y', (iii) to calculate the squared errors of prediction, $(Y- Y')^2$, because the regression line is the one which minimizes the sum of squared deviation of prediction, also called the sum of squared errors.

The standard error was calculated using the formula:

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

where σ_{est} is the standard error of the prediction, Y is the observed values of the variable Y, Y' is the predicted value of Y, and N is the number of pairs of variables. The numerator is the sum of squared errors, reflecting the differences between the actual scores and the predicted scores. [2, 5]

RESULTS AND DISCUSSIONS

The statistical parameters of the number of dairy bovine livestock and milk production.

The number of dairy livestock including cows and buffalos registered a decline from 1,732 thousand heads in the year 2007 to 1,307.3 thousand heads in the year 2012, de reduction accounting for 24.53 %.

The average number of dairy bovine livestock was 1,419.538 thousand heads with a standard deviation of 193.6376 and a coefficient of variation of 13.64 %. The maximum livestock was noticed in the year 2007 and the minimum level was registered in the year 2012.(Table 1).

Table 1. Descriptive statistics of Number of dairy livestock and Milk production, Romania, 2007-2014

Year	Milk production (thousand hl) **Y**	No. of dairy cows and buffalos (Thousand heads) **X**
2007	54,875	1,732
2008	53,089	1,639
2009	50,570	1,569
2010	42,824	1,299
2011	43,947	1,266
2012	42,036	1,265
2013	42,593	1,279
2014	50,535	1,307.3
Mean	47,558.63	1,419.538
Median	47,241	1,303.15
Standard Deviation	5,246.208	193.6376
Variance	27,522,700.84	37,495.51
Coefficient of variation (%)	11.03	13.64
Maximum	54,875	1,732
Minimum	42,036	1,265

Source: National Institute of Statistics, 2015. [9] Own calculation.

The milk production obtained from dairy cows and buffalos also registered a decline, from 54,875 thousand hl in the year 2007 to 50,535 thousand hl in the year 2014, the reduction accounting for 7.91 %. In the analyzed period, the maximum production was noticed in the year 2007 and the minimum one in the year 2012. The average milk production was 47,558.63 thousand hl, with a standard deviation of 5,246.208 and a variation coefficient of 11.03 % (Table 1).

Testing the Linear Regression.

Firstly, it was established a scatter plot of the empirical data, using the linear regression function.(Fig.1.)

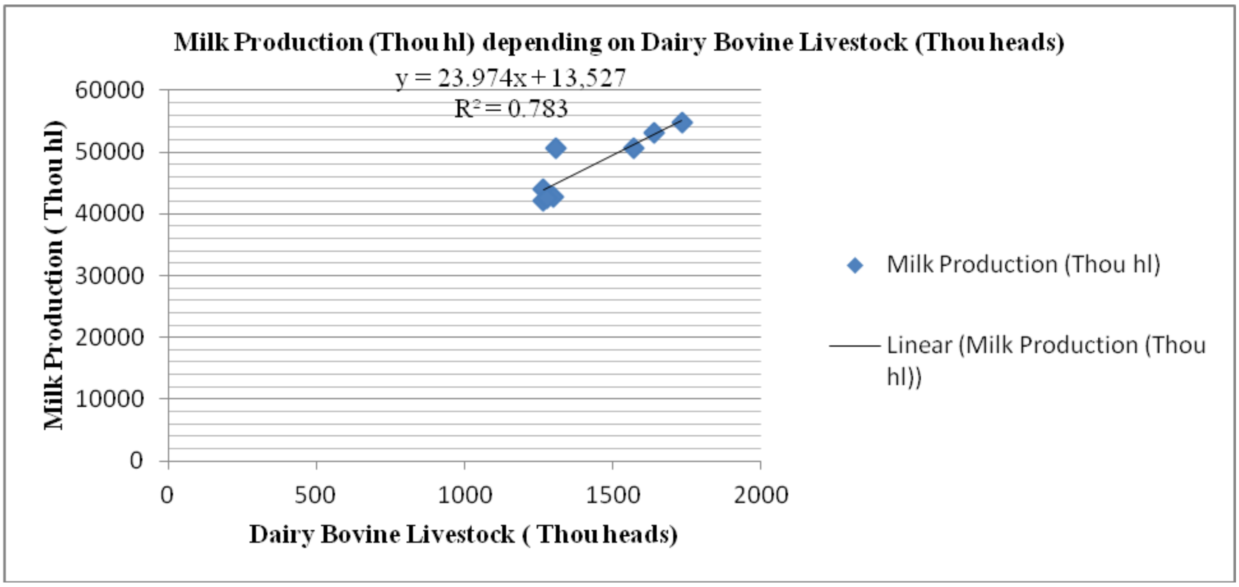

Fig.1.Milk Production evolution depending on Dairy Bovine Livestock, Linear Regression Line. Own design.

The linear regression resulting from the use of Excel Data Analysis facilities was:

$$y = 23.974\ x + 13{,}527$$

where Y is the predicted milk production, the dependent variable, depending on milk production, X, the vector of the independent variable, a = 13,527 is the intercept, a constant, and b=23.974 is the regression slope.

The Multiple R, that is the coefficient of simple correlation between X, the dairy bovine livestock and Y, the milk production is R= 0.884871581, reflecting a strong positive relationship between the two variables.

The R^2 or R Square value was R^2= 0.782997716. The determination coefficient reflects that 78.30 % of the variation of milk production is determined by the dairy bovine livestock.

The standard error of the linear regression was σ_{est} = 2,286.028830, being calculated as the square root of the ratio between the sum of the squared errors of the prediction and the number of variables as presented in Table 2.

The diagram from Fig.2. shows that between the independent variable, Dairy Bovine Livestock and Residuals, that is the errors of prediction, there is no correlation. Therefore, the regression model is good.

Testing the Polynomial Regression or Parabolic Fit

Firstly, it was established a scatter plot of the empirical data, using the parabolic fit regression function.(Fig.3.)

The polynomial regression of the 2nd degree or the Parabolic Fit, resulting from the use of Excel Data Analysis facilities was:

$$y = -0.0237x^2 + 93.997x - 37{,}392$$

where Y is the predicted milk production, the dependent variable, depending on milk production, X, the vector of the independent variable, a = - 13,527 is the intercept, a constant, b=93.997, c = - 0.0237.

The Multiple R, that is the coefficient of simple correlation between X, the dairy bovine livestock and Y, the milk production is R= 0.8880, reflecting a strong positive relationship between the two variables.

Table 2. The calculation of the Linear Regression Standard Error

Observed Y	Predicted Y'	The errors of prediction, Y -Y'	The squared errors of prediction $(Y - Y')^2$
54,875	55,049.52562	-174.5256	30,459.185055
53,089	52,819.96605	269.03395	72,379.266252
50,570	51,141.80293	-571.8029	326,958.556448
42,824	44,668.88804	-1,844.888	3,403,611.732544
43,947	43,877.754	69.2459	4,794.994666
42,036	43,853.78024	-1,817.78	3,304,324.1284
42,593	44,189.41287	-1,596.413	2,548,534.466569
50,535	44,867.87024	5,667.1298	32,116,360.170048
			$\sum (Y - Y')^2 =$ 41,807,422.499982
The linear Regression Standard error			σ_{est} = 2,286.028830

Source: Own calculation.

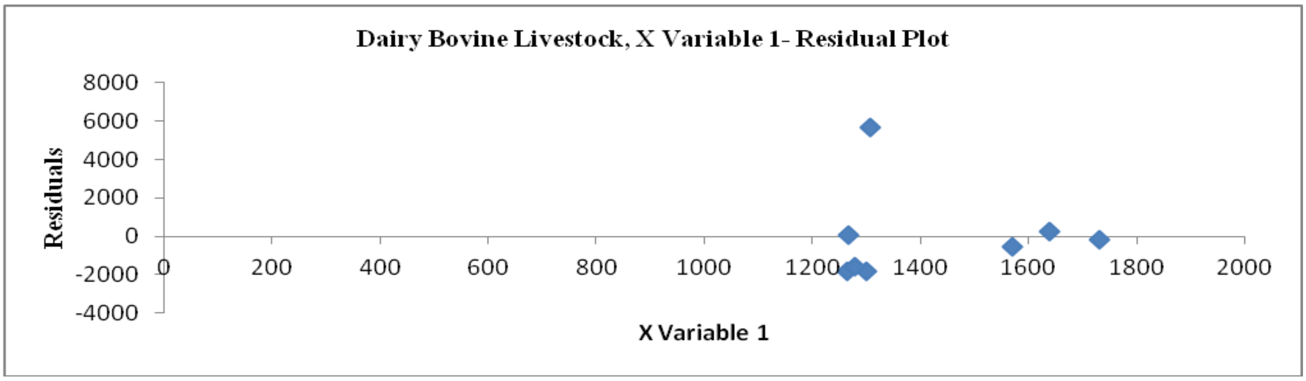

Fig.2.The diagram of the independent variable X, Dairy Bovine Livestock, and Residuals. Own design.

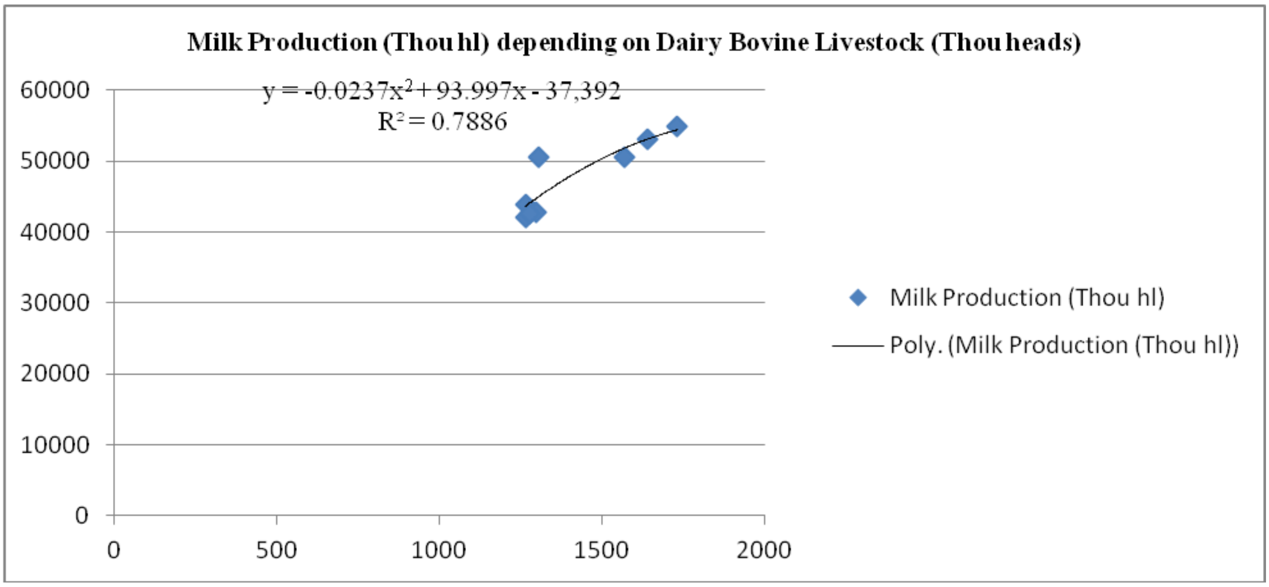

Fig.3.Milk Production evolution depending on Dairy Bovine Livestock, Parabolic Regression Line. Own design.

The R^2 or R Square value was R^2= 0.7886. The determination coefficient reflects that 78.86 % of the variation of milk production is determined by the dairy bovine livestock.

The standard error of the linear regression was σ_{est} = 2,336.915726, being calculated as the square root of the ratio between the sum of the squared errors of the prediction and the number of variables as presented in Table 3.

Table 3. The calculation of the Quadratic Regression Standard Error

Observed Y	Predicted Y'	The errors of prediction, Y -Y'	The squared errors of prediction (Y -Y')^2
54,875	54,314.9752	560.0248	313,627.776615
53,089	53,003.2753	85.7247	7,348.724190
50,570	51,745.5573	-1,175.5573	1,381,934.965583
42,824	45,378.6993	-2,554.6993	6,526,488.513420
43,947	43,622.8848	324.1152	105,050.662871
42,036	43,588.8725	-1,552.8725	2,411,413.001256
42,593	44,060.7313	-1,467.7313	2,154,235.168999
50,535	44,986.189127	5,548.810873	30,789,302.104323
			∑ (Y -Y')^2= 43,689,400.917258
The Quadratic Regression Standard error			σ_{est} = 2,336.915726

Source: Own calculation.

CONCLUSIONS

The study pointed out the importance to comparatively analyze the standard error of various regression models, and finally to decide to chose the regression type whose standard error is the smallest one, as the standard error is a measure of the accuracy of the prediction.

Taking into account that the Regression Standard Error was smaller in case of the Linear Regression, σ_{est} = 2,286.028830 than in case of the Quadratic Regression, σ_{est} = 2,336.915726, the conclusion is that the linear regression fits better to the evolution of milk production depending on the number of dairy bovine livestock.

Therefore, for the period 2007-2014, the linear regression function reflecting the dynamics of milk production depending on the dairy bovine livestock was: y =23.974 x + 13,527.

REFERENCES

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

[2]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

[3]Emerson, E, 2015, Quadratic Regression Calculator, http://www.had2know.com/academics/quadratic-regression-calculator.html

[4]Everitt, B.S., 2003, The Cambridge Dictionary of Statistics

[5]Jianqing, F, 1996, Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC

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

[7]Lane, D.M., Introduction to Linear Regression, http://onlinesstatbook.com/2/regression/intro.html

[8]National Institute of Statistics, 2015, Romania, Statistical Yearbook, Chapter 14, Agriculture.

[9] National Institute of Statistics, 2014, Animal Production 2013, Press Release No.152, June 30, 2014