การประมาณผลผลิตอ้อยในภาคตะวันออกเฉียงเหนือของประเทศไทยด้วยตัวแบบ MLR

Estimation of Sugar Cane Yield in the Northeast of Thailand with MLR Model

กิดาการ สายธนู ภัฐชญา สิทธิสร และ จตุภัทร เมฆพายัพ*

Kidakan Saithanu, Pudchaya Sittisorn, Jatupat Mekparyup*

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา

Department of Mathematics, Faculty of Science, Burapha University

Received : 22 March 2017

Accepted : 8 June 2017

Published online : 22 June 2017

บทคัดย่อ

การศึกษาครั้งนี้เป็นการตรวจสอบตัวแปรอธิบายที่สำคัญซึ่งมีอิทธิพลต่อผลผลิตอ้อยในภาคตะวันออกเฉียงเหนือของประเทศไทยด้วยตัวแบบ MLR หรือตัวแบบการถดถอยเชิงเส้นพหุคูณ โดยประยุกต์เทคนิคทีละขั้นตอนและวิธีเซตย่อยที่ดีที่สุดเพื่อให้ได้สมการที่เหมาะสม และสอดคล้องกับข้อสมมุติของการถดถอย ผลการศึกษาพบว่า มีตัวแปรอิสระที่สำคัญ 5 ตัว ซึ่งถูกคัดเลือกในสมการถดถอยที่ใช้ในการประมาณค่าผลผลิตอ้อยในภาคตะวันออกเฉียงเหนือของประเทศไทย ได้แก่ พื้นที่เพาะปลูก ปริมาณอ้อยส่งไปยังโรงงาน ราคาเฉลี่ยของอ้อย อุณหภูมิต่ำสุด และจำนวนวันที่ฝนตก ซึ่งสมการถดถอยที่ได้มีค่าคลาดเคลื่อนมาตรฐานของการประมาณค่าเท่ากับ 14.7412 สมรรถนะของตัวแบบ MLR ที่ได้ตรวจสอบด้วยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ซึ่งพบว่าตัวแบบ MLR ที่ได้ประมาณผลผลิตอ้อยได้อย่างมีประสิทธิภาพด้วยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่มีค่าน้อย ซึ่งมีค่าเท่ากับ 12.7802

คำสำคัญ: ผลผลิตอ้อย ตัวแบบ MLR วิธีเซตย่อยที่ดีที่สุด การถดถอยทีละขั้นตอน

Abstract

This study investigated the significant explanatory variables influenced to the sugar cane yield in the northeast of Thailand with MLR (multiple linear regression) model. Best subsets and stepwise techniques were applied to gain the suitable MLR equation compiling with assumption of regression. The result of study revealed there were 5 important independent variables; Cultivated area, sugar cane quantity sent to factories, average price of sugar cane, minimum temperature and number of rainy days, selected in the estimated regression equation with 14.7412 for standard error of estimation. The performance of MLR model was verified with the root mean square error (RMSE). It indicated that the MLR model efficiently estimated the sugar cane yield with the small value of RMSE (12.7802).

Keywords : sugar cane yield, MLR model, best subsets, stepwise regression

*Corresponding author. E-mail : jatupat@buu.ac.th

Introduction

Sugar cane is the world's largest crop by production quantity (Food and Agriculture Organization of the United Nations, 2015). Brazil was the largest producer of sugar cane in the world. The next five major producers, in decreasing amounts of production, were India, China, Thailand, Pakistan and Mexico. Sugar cane is therefore one of the important economic crops also it turned to be the Thai commercial crop nowadays. Sugar cane predominantly grows in the tropical and subtropical regions specifically in the northeast area which is the largest producer of Thailand. Numerous variables influenced sugar cane yield; the average of rainfall, number of rainy days, maximum , minimum or average temperature, exposure duration, relative humidity, cultivated area, sugar cane variety, were reported in these works (Chimnarong, 2009), (Kapetch & Pannangpetch, 2012), (Bukate & Seresangtakul, 2013), (Office of the Cane and Sugar Board, 2011). Proper estimation of sugar cane yield is necessary for future planning and policy forming so finding the appropriate tool is essential. Both simple statistical model, like multiple linear regression or MLR (Binbol *et al*., 2006), (Chimnarong, 2009), (Xu *et al*., 2010) and advanced mathematical model, like neural network (Bukate & Seresangtakul, 2013), (Xu, *et al*., 2010) had been popularly used to forecast or estimate the sugar cane production. This study was then designed to apply MLR model following the guidance of Binbol (Binbol *et al*., 2006) also best subsets were incrementally utilized with stepwise technique basing on the crucial explanatory variables for estimation of sugar cane yield in the northeast of Thailand comprised of 15 provinces; Nakhon Ratchasima, Buri Ram, Surin, Maha Sarakham, Roi Et, Ubon Ratchathani, Mukdahan, Nakhon Phanom, Sakon Nakhon, Kalasin, Khon Kaen, Udon Thani, Chaiyaphum, Nong Khai and Loei.

Methods

The yearly data of sugar cane in the northeast of Thailand was collected since 2002 to 2013. Office of the Cane and Sugar Board, Ministry of Industry provided the sugar cane yield in Thai unit called rai or 1,600 square metres as the response variable (Y) and two explanatory variables; cultivated area (X_1: rai) and sugar cane quantity sent to factories (X_2: ton) (Office of the Cane and Sugar Board, 2011). The six remaining explanatory variables were as follows. Average price of sugar cane (X_3: baht/ton) reported from Office of Agricultural Economics, Ministry of Agriculture and Cooperatives (Office of Agricultural Economics, 2015) while maximum temperature (X_4: $^{\circ}$C), minimum temperature (X_5: $^{\circ}$C), total rainfall (X_6: mm.), number of rainy days (X_7) and maximum rainfall (X_8: mm.) were also gathered by National Statistical office (National Statistics Organization of Thailand, 2014) supported from Thai Meteorological Department. To estimate sugar cane yield with the MLR model, three steps were following performed.

1. Whole data was separated into two sets. The training data set was firstly used for training MLR model approximately 70% of whole data (145 cases). The rest of data was the validation data set later employed for testing suitability of model.

2. The estimated MLR equation was determined from the training data set. Best subset method was usually considered in the first step to roughly predetermine as few predictor variables as possible. The stepwise technique was later applied for fitting the MLR model. The estimated MLR equation was finally verified if the obtained MLR equation was suitable in accordance with four assumptions of regression analysis: (2.1) Normality of the distribution of error terms was checked with Anderson-Darling statistic (Romeu, 2003), (2.2) Independence of error terms was proved with Durbin-Watson statistic (Durbin & Watson, 1951) (2.3) Constant variance of error terms was tested with Breusch-Pagan statistic (Michael *et al*., 2005) and (2.4) Multicollinearity problem was examined whether there were correlations among explanatory variables with the Variance Inflation Factor or VIF (Michael *et al*., 2005).

3. The Root Mean Square Error computed as $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}$; Y_i be the *i*th observation of sugar cane yield, $\hat{Y}_i$ be the *i*th predicted value of sugar cane yield obtained from the MLR model and n be the total observations in the validation data set, was verified for performance of the obtained MLR model.

Results and Discussion

The MLR equation derived from the best subset method consisting five explanatory variables (X_1, X_2, X_3, X_5 and X_7) was firstly considered with rather small value of Mallows C-p (16.9). Stepwise technique was then applied for fitting the MLR model. Since each of p-value of *t* statistic for all these variables was quite small, no variable was removed from the model. When this MLR equation was checked for the normality assumption, it was violated. Transformation of sugar cane yield was required by using Box-Cox process with $\lambda = 2$. Then, the adjusted of estimated MLR equation as $\hat{Y}' = Y^2 = 1.74 - 0.00032X_1 + 0.000034X_2 + 0.02926X_3 + 3.1366X_5 + 0.37308X_7$, with 14.7412 for standard error of estimation, was retested. The appropriateness of this MLR equation was confirmatory with the p-value of *t* statistic and F statistic of ANOVA, respectively shown in Table 1 and Table 2, closed to 0. After validating of the suitability of this MLR equation, it displayed the satisfied results in agreement of four following assumptions. (1) The p-value of AD test statistic was 0.141 so the error terms normally distributed. (2) The Durbin Watson statistic value of DW=1.80305 was higher than the critical value (1.7312). Thus, the error terms were independent. (3) The Breusch-Pagan test statistic was 7.35962 smaller than the critical value (11.0706) then there was constant variance of the error terms. (4) The VIF of all these five explanatory variables were less than 5 so multicollinearity problem was not exist. The performance of the obtained MLR model was finally verified from the validation data set. It obtained small values of RMSE (12.7802).

The five significant explanatory variables affected to sugar cane yield in the northeast of Thailand were the cultivated area (X_1) and the average price of sugar cane (X_3) which gave the same result as Bukate & Seresangtakul (2013), the minimum temperature (X_5) and the number of rainy days (X_7) which obtained the same result as Binbol *et al*. (2006) and Chimnarong (2009) and the sugar cane quantity sent to factories (X_2). The estimated MLR equation obtained from this study could well estimate the sugar cane yield as seeing of small value of RMSE. Further study would consider other explanatory variables which might be influenced to the sugar cane yield such as relative humidity, sugar cane variety or even soil type, etc.

Table 1 The *t* statistic for five explanatory variables of the estimated MLR equation

Predictor	Coef	SE Coef	T	P-value
Constant	1.74000	16.19000	0.11	0.91500
X_1	-0.00032	0.00004	-8.52	0.00000
X_2	0.00003	0.00000	8.48	0.00000
X_3	0.02926	0.00802	3.65	0.00000
X_5	3.13660	0.78260	4.01	0.00000
X_7	0.37308	0.09258	4.03	0.00000

Table 2 ANOVA of the estimated MLR equation

Source	DF	SS	MS	F	P-value
Regression	5	46414.5	9282.9	42.72	0.00000
Residual Error	139	30204.9	217.3		
Total	144	76619.5			

Conclusions

The MLR equation derived from the best subsets consisting five explanatory variables (X_1, X_2, X_3, X_5 and X_7) was firstly considered with rather small value of Mallows C-p. Stepwise technique was then applied for fitting the MLR model. Since each of p-value of t statistic for all these variables was quite small, no variable was removed from the model. When this MLR equation was checked for the normality assumption, it was violated. Transformation of sugar cane yield was required by using Box-Cox process. Then, the adjusted of estimated MLR equation as $\hat{Y}' = Y^2 = 1.74 - 0.00032X_1 + 0.000034X_2 + 0.02926X_3 + 3.1366X_5 + 0.37308X_7$, with 14.7412 for standard error of estimation, was retested. The performance of the obtained MLR model was finally verified from the validation data set. It obtained small values of RMSE (12.7802).

Acknowledgements

The authors wish to thank to Office of the Cane and Sugar Board, Office of Agricultural Economics and National Statistical office, for kindly providing all data.

References

Binbol, N. L., Adebayo, A. A. & Kwon-Ndung, E. H. (2006). Influence of climatic factors on the growth and yield of sugar cane at Numan, Nigeria. *Climate Research, 32*(3), 247.

Bukate, O. & Seresangtakul, P. (2013). Sugarcane Production Forecasting Model of the Northeastern by Artificial Neural Network. *KKU Sci. J., 41*(1), 213-225. (in Thai)

Chimnarong, V. (2009). The relationship between climatic parameters and sugarcane yields: the case study of Mitr Phukieo Sugarcane Plantations. Thesis, Master of Environmental Science, Khon Kaen University. (in Thai)

Durbin, J. & Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regresiion II. *Biometrika, 38*(2), 159-177.

Food and Agriculture Organization of the United Nations. (2015). Crop production. Retrieved January 27, 2015 from web site: http://faostat3.fao.org/faostat-gateway/go/to/download/Q/QC/E

Kapetch, P. & Pannangpetch, K. (2012). The Effects of Climate Change on Sugarcane Production in Northeast of Thailand: A case Study in Kalasin Province. *Khon Kaen Agr. J., 40*, 83-91. (in Thai)

Michael, H. K., John, N., Christopher, J. N. & William Li. (2005). Applied Linear Regression Models, 5th Edition. New York: McGraw-Hill.

National Statistics Organization of Thailand. (2014). Rainfalls and Temperatures Statistics at Meteorological Department. Retrieve April 10, 2015 from web site: http://service.nso.go.th/nso/web/statseries/statseries27.html. (in Thai)

Office of Agricultural Economics. (2015). Commercial price of sugarcane. Retrieved March 30, 2015 from web site: http://www.oae.go.th/main.php?filename=index. (in Thai)

Office of the Cane and Sugar Board. (2011). Report of sugarcane cultivated area in Thailand. Retrieved 29 March, 2015 from web site: http://www.ocsb.go.th/th/cms/ detail.php?ID=923&SystemModuleKey=journal

Romeu, J. L. (2003). Anderson-Darling: a goodness of fit test for small samples assumptions. *Selected Topics in Assurance Related Technologies. 10*(5), 1-6.

Xu, Y. C., Shen, S. Q. & Chen, Z. (2010). Comparative Study of Sugarcane Average Unit Yield Prediction with Genetic BP Neural Network Algorithm. China: Department of Computer Science, Guangdong Polytechnic Institute Guangzhou, College of Engineering, South China Agricultural University.