Analysis on the Application of Linear Regression in Various Fields

Xiyu Xie[1,a,*]

[1]*Pacifica Christian High School, 1730 Wilshire Blvd, CA, USA*
a. xxie21@pacificachristian.org
**corresponding author*

Keywords: linear regression, multiple regression, fuzzy neural network.

Abstract: In order to find out the application of linear regression, several researches on different fields using linear regression were analyzed. The results suggested that linear regression is successful for the wide uses in fields related to predicting multiple component content in food. Additionally, linear regression does not require complex calculation and much time. However, polynomial regression has better prediction accuracy than linear regression and fuzzy neural network has a better prediction accuracy than statistical regression.

1. Introduction

Linear regression [16,17] models the relationship between one dependent variable and one or more independent variables by fitting a line that best represents the data. In order to predict the future outcomes, researchers try to find the best-fitting line for the observed data using linear regression model. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = b*x+c, in which y represents dependent variable, x represents independent variable, b represents regression coefficient, and c is a constant. The most common method for fitting a regression line is the method of least-squares, which minimizes the sum of the squares of the vertical deviations from each data point to the line. In actual application, linear regression can be applied in various fields to solve different questions.

2. Literature Review

In 1997, Shen, Long, and Wang [1] discussed about the methods of typhoon forecasting in the Northern part of Fujian province published in ACTA Oenological Sinica. They fitted the variety of air pressure a day before typhoon, the central pressure a day before typhoon, and the maximum wind speed at typhoon center a day before typhoon with respect to predicted value using linear regression method. In conclusion, based on data from 1949 to 1991, they built two forecast models for two stations respectively that had an accuracy rate of 97.4% and 88%.

Moreover, in 2005, Huang, Chen, and Liu [2] talked about the problem of temperature change between Chaidamu basin and northeast of Qinghai province from 1999 to 1957. Using linear regression method, they fitted data of the average annual temperature in Chaidamu basin, the average temperature in four seasons and the change of the average annual temperature in the two areas with respect to the year and drew the conclusion that the temperature increasing trend was different in ten years, it increased 0.367 ℃ in Chaidamu basin and 0.293 ℃ in northeast of Qinghai province, this trend was more significant since 1990.

What's more, in 2005, Wu, Yin, Zheng, and Yang [3] discussed about climate changes in the Tibetan Plateau during the last thirty years, and the data of average annual temperature, precipitation, maximum potential evapotranspiration and dryness are fitted using linear regression. The result shows that the main trends of climate change are temperature rise and precipitation increase; potential evapotranspiration decreased and most of the areas were ascending to more humid status.

Also, in 2008, Zhang, Zhao, and Niu [4] analyzed the climatic change of the Loess Plateau of Shanxi using the annual precipitation and annual average temperature data of four representative weather stations in Shanxi with respect to the year, and they applied linear regression. The conclusion was that the climate in Shanxi was becoming warmer and drier for the last 50 years.

In another study in 2017, Liu, Liu, Xu, and Ge [5] discussed about the change of geographical environment in Southern Margin of Tibetan Plateau since 1980s. They fitted the data of monthly and annual precipitation, monthly mean, maximum and minimum air temperature and annual mean, maximum and minimum air temperature at various elevations with respect to time. The results indicate that the vegetation condition in the east is better than that in the central and western regions, and the temperature has increased significantly.

In a study that researched about the factors of savings of Chinese residents, Wang and Liu [6] in 2009 analyzed data of one-year deposit rate in China, commodity housing price index, stock index yield, urban residents' wage income and the exchange rate of RMB against the US dollar with respect to the balance of household savings using linear regression and found that deposit rate, income and the exchange rate had a positive correlation with the balance of household savings.

In 2008, Wang and Jiang [7] studied about factors that influence the development of household biogas in rural areas of Shaanxi province between 1980 and 2005. They used linear regression to analyze the reserve of biomass resource, the household net income, the change of annual average temperature, literacy level of rural residents and the government investment with respect to the annual output of biogas. The results implied that the reserve of biomass resource and government investment both have a strong positive correlation with annual output of biogas, while the change of annual temperature, the household net income, and literacy level of rural residents do not have a notable effect on the annual output of biogas.

In 2006, Shu, Zhu, Chen and Wu [8] analyzed the evaluation and prediction of water quality status in Yangtze river. By applying linear regression model, they fitted data between 1995 and 2004 of the proportion of class I, II, III, IV, V water and inferior class V water with respect to sewage discharge. In conclusion, the proportion of class I water decreased significantly, from 25.8% to 1.2%, while the proportion of class V water and inferior class V water increased overtime. They predicted that in the future 10 years the water quality of the Yangtze river will be seriously deteriorated.

In 2013, Xie, A, and Wei [9] predicted the operating speed of vehicle, they analyzed the curve radius and slope gathered from 232 samples on the second-level highway with respect to driving speed. The conclusion was that the curve radius had a 0.0221 correlation with the driving speed and the slope had a 0.4224 correlation with the driving speed using the method of linear regression.

In 1994, Chen, Zhang, Wang, Dong, He and Chen [10] investigated about the factors influencing seed yield of CMS92-63. They used linear regression to find the correlation of actual length of stem

inflorescence, secondary branching number, total flower number per plant, single fruit number, straw weight, and single seed number with respect to the seed yield per plant. The results showed that these factors all had positive correlation with the seed yield of CMS92-63.

In 2012, Li, Tu, Hu and Zhang [11] studied the influence of precipitation change on water and sediment evolution in Poyang lake basin. Using the linear regression method, they fitted the annual precipitation between 1961 and 2006 with respect to the annual runoff and sediment transport of Poyang lake. They concluded that annual precipitation and annual runoff had a positive correlation with a coefficient of 0.922, and the annual precipitation and sediment transport had a positive correlation with a coefficient of 0.066.

In 2005, Wang and Zhou [12] discussed about the factors that widened Chinese residents' income gap. They applied linear regression to analyze data of the proportion of employment in the tertiary industry in the labor force, the consumer price index and the unemployment rate with respect to the national residents' income gap. In conclusion, the consumer price index and the income gap had a negative correlation of -0.071, while the proportion of employment in the tertiary industry in the labor force and the unemployment rate both had positive correlation with respect to the national income gap.

In 2010, Wu, Wang, Zhao, Liu, Zhu, Zhang, Li, Jin, Yu, and Li [13] discussed about the predictive value of hsCRP level in the progression to hypertension in prehypertensive population. They gathered 2441 samples to find the correlation between the levkiel of lg hsCPR with respect to the contraction pressure and relaxation pressure. They concluded that there was a positive correlation of 0.39 between lg hsCPR and the contraction pressure.

In 2012, Ma, Wang, Chen, Du, Li, Huang, Shi, Yin, Zhang, A, Dong, and Wu [14] researched about the X-ray measurement and WOMAC scores of knee osteoarthritis. They collected 250 samples and applied linear regression to analyze data of femoral angle, tibial angle, femorotibial angle, joint gap angle with respect to WOMAC scores. They found that the correlation between tibial angle, joint gap angle on antero-posterior X-ray and WOMAC scores is significant, which is Y= 125.616+3.079X (X represents joint gap angle) for the right knee and Y=132.587-1.163X (X represents tibial angle) for the left knee.

Lastly, in 1995, Gu and Wang [15] researched about the measure of the oil content in the rapeseeds. By applying linear regression model, they analyze the light absorption of the main components of rapeseeds with respect to the content of oil. The result showed that the correlation is significant, and the model is successful since the measurement error is less than 1.3%.

3. Conclusions

Linear regression is widely used in every field. The advantages of it are that linear regression does not require complex calculation, therefore even when there is large amount of data, the model still runs fast. Also, it can give interpretation of each variable based on the coefficients. The research on the measure of the oil content in the rapeseeds shows that the measurement error is less than 1.3%. The researchers conclude that linear regression not only saves time, but is also successful and powerful for the wide uses in fields related to predicting multiple component content in food. The research The Appraisal and Forecast of Yangtze's Water Quality applied to two models: linear regression model and the ARIMA model. The study suggests that linear regression is not suitable when analyzing the sum of proportions of two classes of water overtime; instead, the researchers choose ARIMA model. In another study, researchers compared polynomial regression with linear regression and concluded that polynomial regression is superior to that of the linear regression. Also, they compared the fuzzy neural network with regression models and found that the fuzzy neural

network can get a better prediction accuracy than that by statistical regression, so it is of more potential in the application than the regression models.

References

[1] *Shen Rusong, long Baosen, & Wang Zhong. (1997). A regression model for forecasting typhoon gales in northern fujian province. Acta oceanologica sinica, 19(2), 32-37.*

[2] *Huang Yong, Chen Zongyan, & Liu Chune. (2005). Comparison and analysis of temperature change between Chaidamu basin and northeast of Qinghai province in 43 years. Journal of Qinghai University (natural science edition)(03), 56-60.*

[3] *Wu Shao-hong, Yin Yun-he, Zheng Du, & Yang Qin-ye. (2005). Climate changes in the Tibetan plateau during the last three decades. Acta geographica sinica (01), 2-10.*

[4] *Zhang Chunlin, Zhao Jingbo, & Niu Junjie. (2008). Study on warming and drying climate of Shanxi loess plateau in recent 50 years. Journal of arid land resources and environment (02), 72-76.*

[5] *Liu Ronggao, Liu Yang, Xu Xinliang, & Ge Quansheng. (2017). Study on the geographical environment and changes of the southern edge of qinghai-tibet plateau in recent 30 years. Proceedings of the Chinese Academy of Sciences (09), 91-101.*

[6] *WANG Yao-qing, & LIU Wei-qi. (2009). Inluential factors analysis of Chinese residents deposits. Application of statistics and management 28(6), 951-957.*

[7] *Wang Lijia, Jiang Zhide. (0). Analysis on the influenced factors of rural biogas development in Shanxi. Rural economics (11), 12-15.*

[8] *Shu Xiaohui, Zhu Honghong, Chen Xichun, & Wu Hongliang. (2006). The appraisal and forecast of Yangtze's water quality. Journal of Huaihua University (natural science), 25(2), 35-44.*

[9] *Xie Shaobo, Abidan, & Wei Lang. (2013). Comparative analysis of prediction model for vehicle operating speed. Journal of chang 'an university (natural science edition)(05), 85-89.*

[10] *Chen Zhujun, Zhang Mingfang, Wang Bingliang, Dong Weimin, He Jun, & Chen Wenjun. (1994). Correlation analysis of seed yield components of cytoplasmic male sterile lines in mustard. Acta agriculturae zhejiangensis (4), 241-246.*

[11] *Li Ying [1], Tu Anguo [1], Hu Genhua [1], & Zhang Huaming [1]. (2012). Study on the influence of precipitation change on water and sediment evolution in poyang lake basin. China soil and water conservation (4), 25-28.*

[12] *Wang Peigang, & Zhou Changcheng. (2005). Empirical analysis and dynamic study on the widening income gap in China: an interpretation based on multiple linear regression model. Management world (11).*

[13] *Wu Shouling, Wang Na, Zhao Haiyan, Liu Yi, Zhu Feng, & Zhang Ziqiang, et al. (2010). Risk prediction of high sensitivity C-reactive protein on the incidence of hypertension in prehypertensive population. Chin J hypertens (04), 97-101.*

[14] *Ma Yufeng, Wang Qingfu, Chen Zhaojun, Du Chunlin, Li Junhai, & Huang Hu, et al. (2012). Multiple linear regression analysis of X-ray measurement and WOMAC scores of knee osteoarthritis. China J Orthop Trauma (05), 23-26.*

[15] *Gu Weizhu, & Wang Yanxiang. (1995). Analysis of rapeseed oil content by multiple linear regression analysis. Journal of the Chinese Cereals and Oils Association (02), 57-64.*

[16] *"Linear Regression." Linear Regression, www.stat.yale.edu/Courses/1997-98/101/linreg.htm.*

[17] *"What Is Linear Regression?" Statistics Solutions, www.statisticssolutions.com/what-is-linear-regression/*