Proceedings of Business and Economic Studies, 2021, Volume 4, Issue 5
http://ojs.bbwpublisher.com/index.php/PBES
ISSN Online: 2209-265X
ISSN Print: 2209-2641

Developing Statistical Modellings to Investigate the Internal Drivers for the Trend of Output Values in the Manufacturing Industry: Evidence from Chinese Enterprises

Yuzhou Zhang[1,2]*, Guang Gao[2], Lidan Shou[1], Dun Wu[2], Guangping Fang[2], Hua Sun[2]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China

[2]PopSmart Technology (Zhejiang) Co., Ltd., Ningbo, China

Corresponding author: Yuzhou Zhang, zyzjacky@zju.edu.cn

Abstract: The manufacturing industry is an important pillar of the national economy. It is of vital importance to develop statistical modellings in order to quantify the relationship between potential internal drivers and the trend of output values in the manufacturing industry. However, only a few statistical modellings have been established to investigate such associations. This study developed the correlation coefficient model and generalized linear model (GLM) to measure the single and interactive effects of the internal drivers on the changes of the output values. For the GLM, different predictive variables were developed to fit into the dataset, and the performance of the models were compared using fitness parameters. Furthermore, an industry survey dataset for 1,180 manufacturing enterprises in 2020 was used to validate the models. The use of the GLM combining land area, number of employees, scientific research input, and labor productivity may have a great potential to bolster capacity in monitoring and predicting the trend of output values in the manufacture industry.

Keywords: Statistical modellings; Internal drivers; Output values; Manufacturing industry; Chinese enterprises

Publication date: October 2021; ***Online publication:*** October 29, 2021

1. Introduction

Economic forecasts are based on statistical data and economic information, starting from the status quo and laws of economic phenomena, as well as using scientific methods to predict the future development prospects of the economy. Economic forecasting is a scientific decision-making tool, and it is one of the important bases for the government to formulate economic plans as well as predict and guide the implementation of plans. There are previous studies which have investigated the internal drivers of output value trends in the manufacturing industry.

Economic forecasting generally uses two types of subdivision methods: statistical analysis and mathematical models, including adaptive filtering [1], time series forecasting [2], trend curve forecasting models [3], regression forecasting methods [4], grey forecasting models [5], Markov forecasting methods [6], etc. The generalized linear model (GLM) is an extension of the general linear model in such a way that the dependent variable is linearly related to the factors and covariates through the specified link function [7]. In addition, the model allows the dependent variable to have a non-normal distribution [7]. This study aims to develop a generalized linear model (GLM) to quantify the interactive relationship between the trends in the output values and the internal drivers of those trends in the manufacturing industry.

2. Methods

2.1. Evaluating the single effect with correlation analysis

In this study, correlation coefficient is used to evaluate the relationship between the output value trends in the manufacturing industry and the internal drivers by single effect. Assuming a sample with a sample size of *n*, *n* original data are converted into grade data, and the correlation coefficient ρ is as follows:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

The correlation symbol indicates the direction of association between *X* (independent variable) and *Y* (dependent variable). If *Y* tends to increase when *X* increases, the correlation coefficient is positive. However, if *Y* tends to decrease when *X* increases, the correlation coefficient is negative. A correlation coefficient of zero indicates that *Y* does not increase nor decrease when *X* increases. As *X* and *Y* become closer to each other's perfect monotonic function, the correlation increases in magnitude. When *X* and *Y* are monotonously correlated, the correlation coefficient becomes 1. A monotonically increasing relationship means that for any two pairs of data values, X_i Y_i and X_j Y_j, $X_i - X_j$ and $Y_i - Y_j$ always have the same sign. A monotonically decreasing relationship means that these differences always have opposite signs [8,9].

2.2. Evaluating the interactive effects with GLM

GLM has been developed using different combinations of internal drivers to better fit and predict the trends of output values considering the interactive effects [10,11]. Multicollinearity among internal drivers was checked and avoided using Spearman correlation analysis and variance inflation factor (VIF). Only one of the highly-correlated drivers (correlation coefficient > 0.6 or VIF > 5) was included in the GLM [12]. A negative binomial distribution has been assumed to allow overdispersion [13].

In the GLM, assuming the dependent variable of each result *Y* is generated from a specific distribution in the exponential family, and the mean of the distribution μ depends on the independent variable *X*, by:

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$

E(Y|X) is the expected value of *Y* conditioned on *X*; *Xβ* is the linear predictor, the linear combination of unknown parameters *β*; g is the link function. In this framework, the variance is usually a function of the mean *V*:

$$\mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \mathrm{V}(\boldsymbol{\mu}) = \mathrm{V}(g^{-1}(\mathbf{X}\boldsymbol{\beta})).$$

The unknown parameter *β* is usually estimated using maximum likelihood, maximum quasi-likelihood, or Bayesian techniques. GLM is composed of three elements:

(1) exponential family of probability distribution;

(2) linear predictor $\eta = X\beta$;

(3) the link function g makes $E(Y \mid X) = \mu = g^{-1}(\eta)$ [14, 15].

The GLM that included multiple internal drivers is listed below as an example:

$$log[E(Y)] = \beta_0 + \beta_1(V_1) + \beta_2(V_2) + \beta_3(V_3) + \beta_4(V_4) + \cdots + e$$

E(Y) represents the expected annual output value of each manufacturing company, β_0 represents the intercept, β_1 (V_1), β_2 (V_2), β_3 (V_3), and β_4 (V_4) denote the corresponding regression coefficients of the internal drivers for the trends in output values in the manufacturing industry, and *e* represents the error.

Four metrics have been reported to compare the fitness performance between the models: Bayesian information criterion (BIC), the stationary R square (R^2), the root mean square error (RMSE), and the maximum absolute percentage error (MAPE).

3. Results

A survey dataset of 1,180 manufacturing enterprises in Yinzhou, Ningbo, China in 2020 was used to validate the models, including land area, equivalent comprehensive energy consumption, number of employees, scientific research input, labor productivity, as well as the proportion of research and development input. The results from the descriptive analysis of the dataset are listed in **Table 1**.

The results from the correlation analysis indicate that the trend of the output values was positively correlated with the land area ($r = 0.57$, $P < 0.05$), followed by the number of employees ($r = 0.54$, $P < 0.05$), scientific research input ($r = 0.54$, $P < 0.05$), equivalent comprehensive energy consumption ($r = 0.41$, $P < 0.05$), and labor productivity ($r = 0.38$, $P < 0.05$). However, there was no significant relationship with the proportion of research and development input (**Figure 1**).

Table 1. Summary of the survey dataset of manufacturing enterprises in Yinzhou, Ningbo, China in 2020

	Mean	Median	Min	Max	Std. deviation
Output value (0.01 million Yuan)	7105.5	3853.3	998.8	102215.1	11362.5
Land area (square kilometer)	12.0	7.5	0.2	136.3	15.7
Equivalent comprehensive energy consumption (ton of coal)	677.1	210.8	11.1	34490.3	2611.3
Number of employees (person)	110.6	71.0	5.0	597.0	102.4
Scientific research input (0.01 million Yuan)	205.7	40.9	7.0	5284.2	470.3
Labor productivity (0.01 million Yuan/person)	13.0	11.0	0.7	82.2	9.0
Research and development input (R.D.) (%)	3.0	1.1	0.1	34.4	3.8

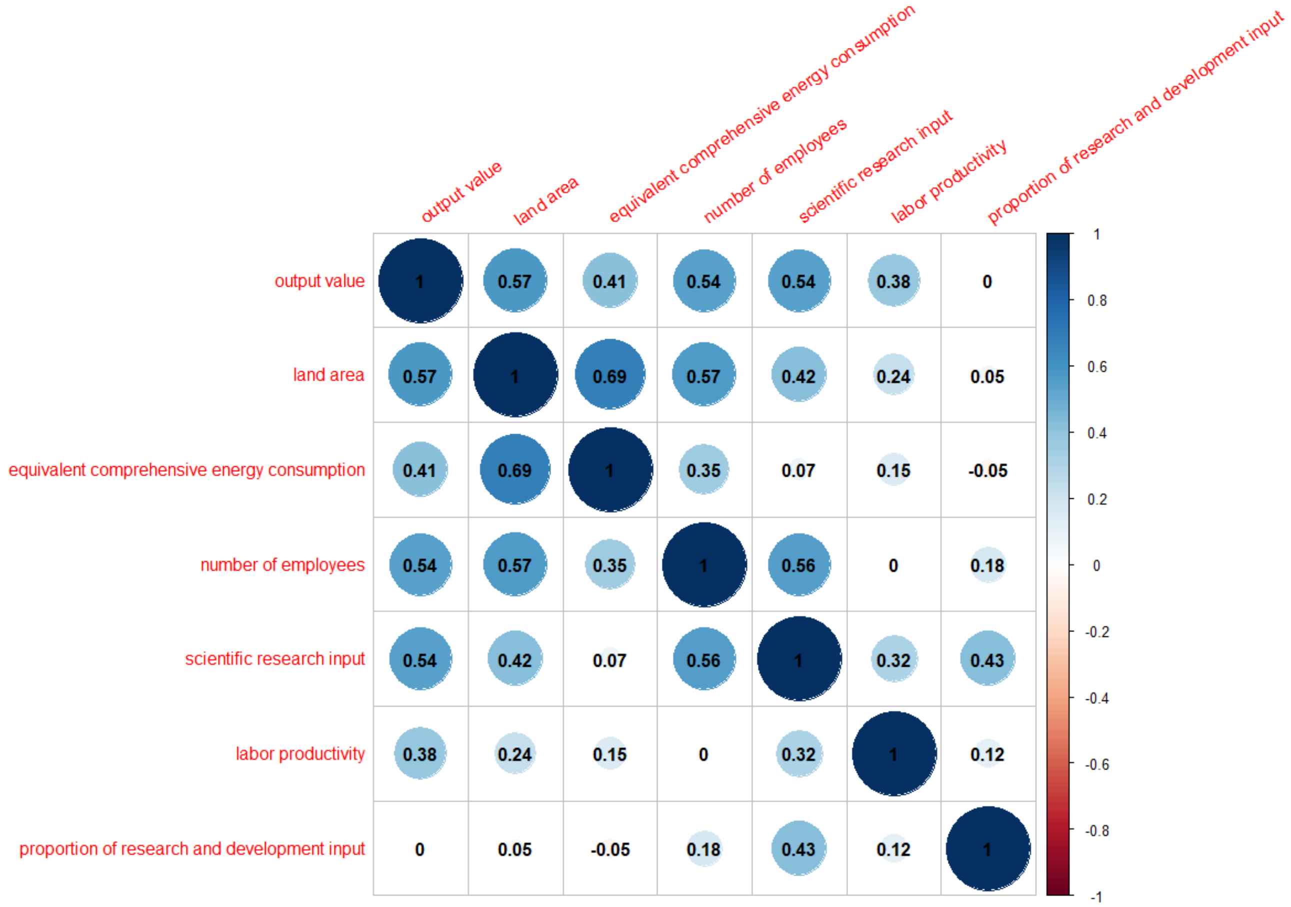

Figure 1. Correlation coefficients between the internal drivers and the trend of output values in the manufacturing industry

Different combinations of the internal drivers mentioned above were the developed to fit in the GLM and predict the trend of the output values. Multicollinearity among internal drivers was checked and avoided (**Figure 1**). Land area, number of employees, scientific research input, and labor productivity were included for further statistical modelling development. Finally, the GLM combining land area, number of employees, scientific research input, and labor productivity had better fit to the real dataset compared to the other models, with a larger R^2 value (85.6) and smaller values of BIC (36.2), RMSE (158.3), and MAPE (190.5), compared to the variables excluded models (**Table 2**).

Table 2. Goodness-of-fit of GLM using different internal drivers

	R^2	BIC	RMSE	MAPE
Model 1 (land area + number of employees)	76.8	59.9	202.2	253.6
Model 2 (land area + scientific research input)	75.4	58.6	234.1	262.7
Model 3 (land area + labor productivity)	71.5	61.9	241.2	275.3
Model 4 (number of employees + scientific research input)	68.4	66.4	288.0	337.2
Model 5 (number of employees + labor productivity)	72.1	62.4	259.5	289.2
Model 6 (scientific research input + labor productivity)	76.8	59.9	209.2	251.9
Model 7 (land area + number of employees + scientific research input)	80.3	47.0	185.6	217.6
Model 8 (land area + number of employees + labor productivity)	82.8	42.0	177.9	202.4
Model 9 (number of employees + scientific research input + labor productivity)	83.2	39.9	162.7	193.5
Model 10 (land area + number of employees + scientific research input + labor productivity)	85.6	36.2	158.3	190.5

The GLM regression analysis indicates that the land area, equivalent comprehensive energy consumption, number of employees, scientific research investment, and labor productivity had positive contributions to the output value of manufacturing enterprises, which were statistically significant ($P < 0.05$) (**Table 3**). However, R.D. had no obvious regression effect on the output value of manufacturing enterprises ($P > 0.05$).

The regression results show that the elasticity coefficient of land area is 76.8, which means that with other conditions unchanged, for every 1 unit of land area increase, the output value of manufacturing enterprises would increase by 768,000 Yuan. Similarly, for every additional 1 unit of equivalent comprehensive energy consumption, the output value of manufacturing enterprises would increase by 8,000 Yuan; for every increase in the number of employees, the output value of manufacturing enterprises would increase by 296,000 Yuan; for every increase in scientific research investment by 10,000 Yuan, the output value of manufacturing enterprises would increase by 8,000 Yuan; for every 10,000 Yuan increase in labor productivity, the output value of manufacturing enterprises would increase by 3.329 million Yuan.

Table 3. Parameters of interactive effects among internal drivers using GLM

	Estimate	Std. Error	T-value	P-value
Land area (square kilometer)	76.8	59.9	1.3	0.02
Equivalent comprehensive energy consumption (ton of coal)	0.8	0.3	1.3	0.01
Number of employees (person)	29.6	7.7	2.4	<0.01
Scientific research input (0.01 million Yuan)	0.8	0.2	3.8	<0.01
Labor productivity (0.01 million Yuan/person)	332.9	73.7	4.7	<0.01
Research and development input (R.D.) (%)	0.6	0.2	3.6	0.08

The predictive performance of the selected GLM is shown in **Figure 2**. The results indicate that the sample data fits well with the GLM comprehensive effect regression model function, and the selected development elements in the regression model function can explain 85.6% of the selected manufacturing enterprises' output value changes.

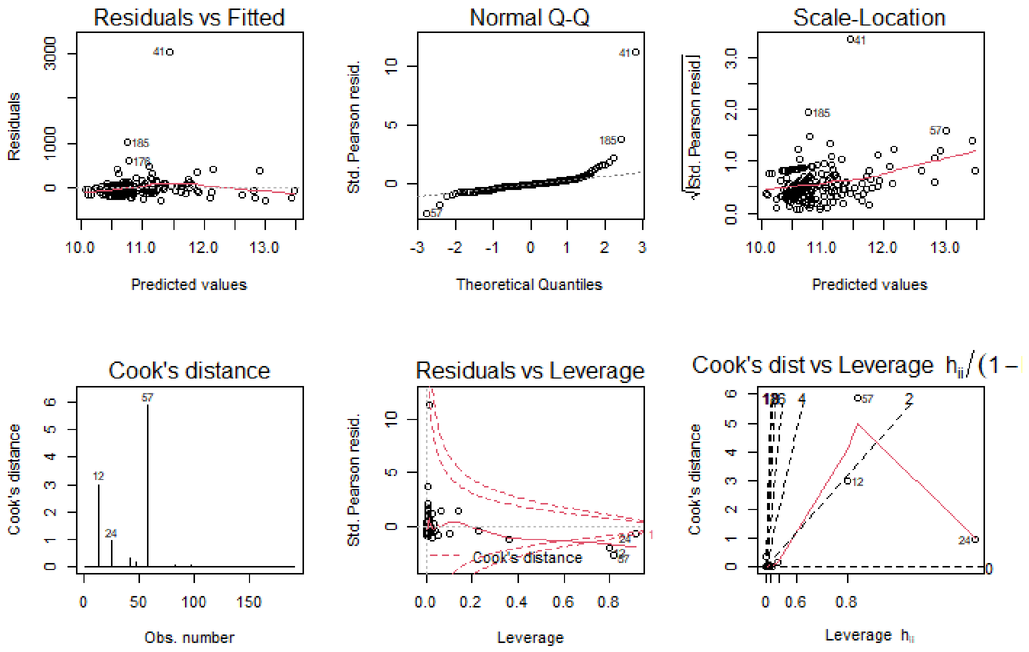

Figure 2. Predictive performance of the selected GLM

4. Discussion

This study investigated the internal drivers for the trends of the output values in the manufacturing industry. Correlation analysis and GLM-based statistical models were established to quantify the associations between the trend of the output values with the internal drivers. Using real dataset, the models were validated, and it was found that the selected model was able to fit well with the real data.

The use of GLM combining land area, number of employees, scientific research input, and labor productivity may have a great potential to bolster capacity in monitoring and predicting the trend of output values in the manufacturing industry.

Disclosure statement

The authors declare that there is no conflict of interest.

Acknowledgments

Yuzhou Zhang designed this study, collected and analyzed the data, as well as drafted the manuscript. Guang Gao, Lidan Shou, Dun Wu, Guangping Fang, and Hua Sun interpreted the results and revised the manuscript.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Garcia-Vega S, Zeng XJ, Keane J, 2020, Stock Returns Prediction Using Kernel Adaptive Filtering Within a Stock Market Interdependence Approach. Expert Systems with Applications, 160: 113668.

[2] Hanias MP, Curtis PG, Thalassinos E, 2012, Time Series Prediction with Neural Networks for the Athens Stock Exchange Indicator. European Research Studies Journal, European Research Studies Journal, 0(2): 23-32.

[3] Profillidis V, 2000, Econometric and Fuzzy Models for the Forecast of Demand in the Airport of Rhodes. Journal of Air Transport Management, 6(2): 95-100.

[4] Ismail Z, Yahya A, Shabri A, 2009, Forecasting Gold Prices Using Multiple Linear Regression Method. American Journal of Applied Sciences, 6(8): 1509.

[5] Liu L, Wang Q, Wang J, et al., 2016, A Rolling Grey Model Optimized by Particle Swarm Optimization in Economic Prediction. Computational Intelligence, 32(3): 391-419.

[6] Hassan MR, Nath B, 2005, Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA’05), 2005: Stock Market Forecasting Using Hidden Markov Model: A New Approach. IEEE, : 192-196.

[7] Madsen H, Thyregod P, 2010, Introduction to General and Generalized Linear Models, CRC Press, 32-34.

[8] Lehman A, O’Rourke N, Hatcher L, et al., 2005, Jmp for Basic Univariate and Multivariate Statistics, 67-69.

[9] Vehkalahti R, 2008, The Concise Encyclopedia of Statistics by Yadolah Dodge. International Statistical Review, 76(3): 460-461.

[10] Dobson A, 2004, The Oxford Dictionary of Statistical Terms. Yadolah Dodge (ed.), Oxford University Press, Oxford, 2003. Statistics in Medicine, 23(11): 1824-1825.

[11] Cox DR, 1984, Interaction. Revue Internationale de Statistique (International Statistical Review), 52(1): 1-24.

[12] Wu J, Tschakert P, Klutse E, et al., 2015, Buruli Ulcer Disease and Its Association with Land Cover in Southwestern Ghana. Plos Neglected Tropical Diseases, 9(6): e0003840.

[13] Wang P, Goggins WB, Chan EY, 2018, Associations of Salmonella Hospitalizations with Ambient Temperature, Humidity and Rainfall in Hong Kong. Environment International, 120: 223-230.

[14] Hastie TJ, Tibshirani RJ, 1995, Generalized Additive Models for Medical Research. Statistical Methods in Medical Research, 4(3): 187-196.

[15] Liu S, 2012, Introduction to General and Generalized Linear Models by Henrik Madsen, Poul Thyregod. International Statistical Review, 80(1): 183-184.