Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 163-168, 4, 2023

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 163-168, 4, 2023

Research Article

Productivity Prediction of Garment Employees using Multiple Linear Regression

Yann Ling Goh [1*], Chern Long Ng [2] and Raymond Ling Leh Bin [3]

[1,2] *Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Cheras, 43000 Kajang, Selangor, Malaysia*
[3] *Faculty of Accountancy and Management, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Cheras, 43000 Kajang, Selangor, Malaysia*

[*]*(gohyl@utar.edu.my) Email of the corresponding author*

(Received: 27 April 2023, Accepted: 9 May 2023)

(DOI: 10.59287/ijanser.2023.7.4.644)

(1st International Conference on Recent Academic Studies ICRAS 2023, May 2-4, 2023)

ATIF/REFERENCE: Goh, Y. L., Ng, C. L. & Bin, R. L. L. (2023). Productivity Prediction of Garment Employees using Multiple Linear Regression. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(4), 163- 168.

Abstract **–**This paper presents a multiple linear curve fitting method for finding how the change of manipulated variables affect the final result using productivity prediction of garment employees data set. To perform fitting, a function is defined, which depends on the parameters that measures the closeness between the data and model. The simplest form of modelling is linear regression, which is the prediction of one variable from another. When the relationship between a few variables are assumed to be linear, then multiple linear regression will be used to model the relationship between a continuous response variable and continuous or categorical explanatory variables.

Keywords – Multiple Linear Regression, Multicollinearity, Variable, Statistical, Transformation

I. INTRODUCTION

Statistics is characterized as a numerical science which consists of the collection, analyzing, understanding and presenting the data. In this paper, the productivity prediction of garment employees data set is used for regression purposes by predicting the productivity range between 0 and 1. We investigate the relationship between two or more factors and a target variable by using the curve fitting method with multiple linear regression model. The model adequacy checking would be applied on multiple linear regression model. This is to ensure that the factors have linear relationship among each other.

The general formula for multiple linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

where
n = number of regressors
y = target variable (or dependent variable)
x_n = the n^{th} factor (or independent variable)
β_0 = the point where the regression line crosses the vertical axis
β_n = slope or coefficient for the n^{th} factor
ε = the model's random error term with NID(0,σ^2)

II. LITERATURE REVIEW

According to Loftus (2021), simple linear regression gives us a full understanding of the relationship between predictor and response. There is only one manipulated variable used to find or predict the target variable in simple linear regression [1].

The main idea for regression is to plot the line of best fit. The general formula for simple linear regression is $y = \beta_0 + \beta_1 x + \varepsilon$, where the gradient or the regression coefficient of the manipulated variable is β_1 and the coordinate with the line crosses the vertical axis is β_0. The y is the target to be predicted, x is the factors or independent variables, and ε is the random error term [2-5].

Multiple linear regression analysis is a technique that uses statistical analysis of the relation between multiple independent variables and a single dependent variable, in which these variables are linearly related to each other.

The concept for multiple linear regression is similar to the simple linear regression, but it contains two or more than two predictor variables. The general formula for multiple linear regression is $y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon$ where the β_i is the regression coefficient for each manipulated variable and x_i is the value for i^{th} manipulated variable.

When fitting a multiple linear regression model, we must safeguard against the problem of multicollinearity. It is important that the manipulated variables which are included in the regression analysis are not very profoundly correlated with one another [6-9].

Multicollinearity is the inordinate relationship among the manipulated variables. This supposition can be tried by calculating the VIF value, which estimates whether manipulated variables are exceptionally related with one another proposing. When distinguishing multicollinearity, the strongly correlated variables are set as a single variable, or the variables that should be included in the model are carefully considered [10-12].

R is a programming language and software environment. It is mainly used for data analysis, statistical computing, graphics and has become the reference working software tool in many fields of research and development.

R is an amazing modern calculation climate for visualization, information control and statistical computation. In R, information is treated as a matrix, vector or network which is similar to mathematics. It likewise gives different measurable techniques and functions including general statistical testing, various kinds of modelling like the linear and nonlinear model, clustering, classification, regression and so on. Moreover, R gives different graphical functions to create the higher quality and better-designed plots. Both graphical and statistical methods should be possible by means of a few lines of coding [13-14].

III. MATERIALS AND METHOD

The formula for estimated multiple linear regression is shown below. The predicted target, $\hat{y}$, also known as predicted productivity will be calculated,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_n x_n$$

A technique for recognizing multicollinearity is to calculate the variance inflation factor (VIF) value. This is an action where the standard error of the multicollinearity coefficient estimation is inflated because of multicollinearity. The VIF formula is $VIF = \frac{1}{1-R_i{}^2}$ where $R_i{}^2$ is the determination coefficients of variables *i*. A VIF of 10 and above indicates that there is multicollinearity problem.

The fitting of the multiple linear regression model is based on the following significant assumptions:

- The relationship between the target variable, *y* and manipulated variables, x_n is approximately linear.
- The errors are uncorrelated.
- The errors are normally distributed.

The validity of these assumptions is required for the outcomes to be significant. On the off chance that these assumptions are violated, the outcome can be mistaken. Assuming these departures are small, the final result may not be changed essentially. If the deviations are huge, the model obtained may become unstable in an alternate example could prompt a completely unique model with inverse

conclusions. Hence, the basic assumptions must be confirmed prior to attempt to regression modelling.

Box-Cox strategy is a technique where the target y is written in type of y^{λ}. λ and the boundaries can be assessed at the same time by utilizing the maximum likelihood technique. There is a conspicuous issue that when λ ways to zero, y^{λ} ways to 1, which is useless. The target can be transformed as

$$y^{\lambda} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ ln\, y & \lambda = 0. \end{cases}$$

Be that as it may, there is yet an issue, when λ changes, the value of $(y^{\lambda}-1)/\lambda$ change emphatically. The formula is improved as

$$y^{\lambda} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y} ln\, y & \lambda = 0. \end{cases}$$

where $\dot{y} = ln^{-1}(\frac{1}{n}\sum_{i=1}^{n} ln\, y_i)$ is the mean of the manipulated variables.

IV. Results and Discussion

The productivity prediction of garment employees data set with 8 attributes are chosen in this study. Table 1 shows the selected attributes with response *y*.

Table 1: Selected attributes with response *y*

Variable	Represented as
quarter	x_1
day	x_2
team	x_3
Target productivity	x_4
smv	x_5
incentive	x_6
no. of style change	x_7
no. of workers	x_8
actual productivity	y

The linear correlation relationship between factors is calculated using R language and shown in Table 2. According to the Table 2, there is a strong positive relationship between x_5 and x_8 (0.91).

Table 2: Linear correlation coefficients between factors

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y
x_1	1.00								
x_2	0.06	1.00							
x_3	0.02	0.00	1.00						
x_4	-0.15	-0.02	0.05	1.00					
x_5	-0.01	-0.02	-0.11	-0.10	1.00				
x_6	-0.02	-0.03	0.00	0.18	0.63	1.00			
x_7	0.21	-0.04	-0.01	-0.23	0.31	0.04	1.00		
x_8	-0.01	-0.03	-0.07	-0.11	**0.91**	0.72	0.33	1.00	
y	-0.05	-0.02	-0.11	0.37	-0.08	0.28	-0.16	-0.02	1.00

Table 3 shows the estimated parameter and VIF value for each factor. Since all the VIF values are less than 10, we assume that multicollinearity problems do not exist.

Table 3: Overall regression model for productivity prediction of garment employees data set

Variable	df	Estimated	Std. Error	t-value	p-value	VIF
(Intercept)	1	0.27575	0.06260	4.405	1.17e-05	
x_1	1	0.00046	0.00352	0.131	0.896	1.0847
x_2	1	-0.00032	0.00196	-0.165	0.869	1.0096
x_3	1	-0.00757	0.00122	-6.199	8.10e-10	1.0268
x_4	1	0.70320	0.07755	9.068	< 2e-16	1.2216
x_5	1	-0.00566	0.00093	-6.055	1.95e-09	6.1811
x_6	1	0.00258	0.00024	10.879	< 2e-16	2.6690
x_7	1	-0.00188	0.01129	-0.166	0.868	1.3189

x_8	1	0.00018	0.00055	0.334	0.739	8.6159

Based on the backward deletion method at selected alpha value of 0.05, the variable x_2 records the highest p-value (0.896) among all the factors, so x_2 will be deleted from the regression model. This step is repeated until all the VIF values are less than 10 and p-values are less than 0.05. The results after removing the insignificant coefficients are shown in Table 4, 5 and 6. The final regression model is shown in Table 7.

Table 4: Reduced regression model after removing factor "quarter"

Variable	df	Estimated	Std. Error	t-value	p-value	VIF
(Intercept)	1	0.27786	0.06048	4.595	4.86e-06	
x_2	1	-0.00031	0.00195	-0.156	0.876	1.0049
x_3	1	-0.00756	0.00122	-6.201	8.02e-10	1.0265
x_4	1	0.70184	0.07681	9.138	< 2e-16	1.1996
x_5	1	-0.00566	0.00093	-6.056	1.93e-09	6.1795
x_6	1	0.00259	0.00024	10.947	< 2e-16	2.6443
x_7	1	-0.00155	0.01100	-0.141	0.888	1.2522
x_8	1	0.00018	0.00054	0.323	0.746	8.5477

Table 5: Reduced regression model after removing factor "no. of style change"

Variable	df	Estimated	Std. Error	t-value	p-value	VIF
(Intercept)	1	0.27713	0.06023	4.601	4.70e-06	
x_2	1	-0.00029	0.00195	-0.151	0.880	1.0030
x_3	1	-0.00757	0.00122	-6.216	7.32e-10	1.0246
x_4	1	0.70301	0.07632	9.212	< 2e-16	1.1854
x_5	1	-0.00566	0.00093	-6.060	1.89e-09	6.1793
x_6	1	0.00259	0.00023	11.350	< 2e-16	2.4781
x_8	1	0.00016	0.00053	0.300	0.746	8.1567

Table 6: Reduced regression model after removing factor "day"

Variable	df	Estimated	Std. Error	t-value	p-value	VIF
(Intercept)	1	0.27572	0.05947	4.636	3.98e-06	
x_3	1	-0.00757	0.00122	-6.218	7.20e-10	1.0246
x_4	1	0.70333	0.07625	9.223	< 2e-16	1.1845
x_5	1	-0.00566	0.00093	-6.071	1.76e-09	6.1724
x_6	1	0.00259	0.00023	11.355	< 2e-16	2.4781
x_8	1	0.00016	0.00053	0.307	0.759	8.1342

Table 7: Final regression model for productivity prediction of garment employees data set

Variable	Estimated	Std. Error	SS	t-value	p-value
(Intercept)	0.2810	0.0569170	19.497	4.937	9.22e-07
x_3	-0.0076	0.0012155	0.309	-6.214	7.42e-10
x_4	0.6976	0.0739081	3.815	9.439	< 2e-16
x_5	-0.0054	0.0005079	0.108	-10.676	< 2e-16
x_6	0.0026	0.0001974	3.246	13.322	< 2e-16

The predicted regression model after removing factors x_1 (quarter), x_2 (day), x_7 (no. of style change) and x_8 (no. of workers) is:

$$\hat{y} = 0.281 - 0.0076x_3 + 0.6976x_4 - 0.0054x_5 + 0.0026x_6$$

After the regression model is fitted, some methods in model adequacy checking are used for analysis.

Figure 1 and 2 show that the data does not come from a normal distribution. This is due to the histogram has a slightly left skewed distribution while the points do not lie along a straight line in normal probability plot. Hence, data transformation will be performed to get a better regression model.

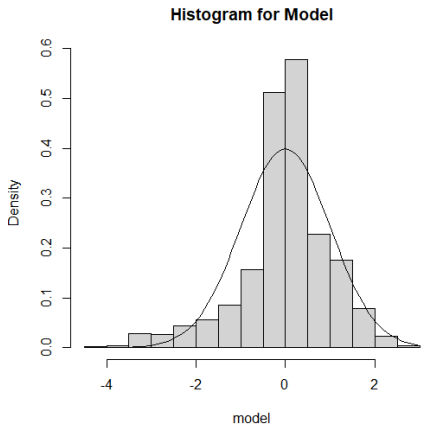

Figure 1: Histogram for the fitted model

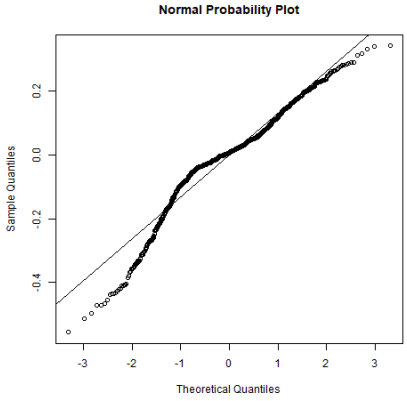

Figure 2: Normal probability plot for the fitted model

The regression model after data transformation is:

$$\hat{y}^2 = 0.0026 - 0.0132x_3 + 0.9563x_4 - 0.0042x_5 + 0.0021x_6$$

Figure 3 shows the normal probability plot after data transformation. The points lie approximately along a straight line reflects a normal distribution.

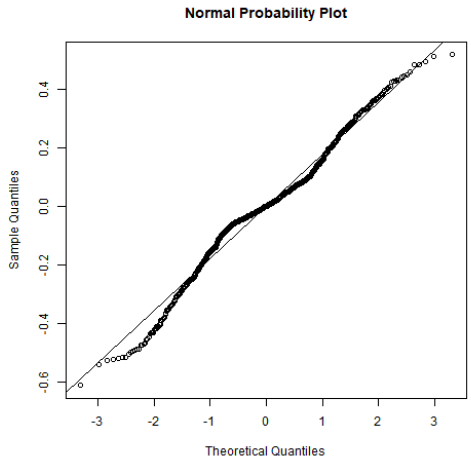

Figure 3: Normal probability plot after data transformation

V. CONCLUSION

In general, the determined multiple linear regression model is a helpful model to predict the productivity of garment employees by choosing team, target productivity, standard minute values and incentive as manipulated variables. Model adequacy checking and data transformation can assist in analyzing the data as well as validate the significant regression assumptions.

REFERENCES

[1] Bangdiwala, S. I. (2018). Regression: simple linear. *International Journal of Injury Control and Safety Promotion*, *25*(1), 113-115.

[2] Escanciano, J. C. (2018). A simple and robust estimator for linear regression models with strictly exogenous instruments. *The Econometrics Journal*, *21*(1), 36-54.

[3] Goh, Y. L., & Pooi, A. H. (2012). Confidence intervals for multivariate value at risk. *Scienceasia*, *39S*, 70-74.

[4] Laverty, W. H., & Kelly, I. W. (2021). Exploring the effects of assumption violations on simple linear regression and correlation using excel. *American Journal of Theoretical and Applied Statistics*, *10*(4), 194-201.

[5] Loftus, S. C. (2021). *Basic statistics with R: reaching decisions with data*. Academic Press.

[6] Alita, D., Putra, A. D., & Darwis, D. (2021). Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(3), 1-5.

[7] Araiza-Aguilar, J. A., Rojas-Valencia, M. N., & Aguilar-Vera, R. A. (2020). Forecast generation model of municipal solid waste using multiple linear regression. *Global Journal of Environmental Science and Management*, *6*(1), 1-14.

[8] Goh, Y. L., Goh, Y. H., Bin, R. L. L., & Chee, W. H. (2019). Predicting the performance of the players in NBA Players by divided regression analysis. *Malaysian Journal of Fundamental and Applied Sciences*, *15*(3), 441-446.

[9] Olsen, A. A., McLaughlin, J. E., & Harpe, S. E. (2020). Using multiple linear regression in pharmacy education scholarship. *Currents in Pharmacy Teaching and Learning*, *12*(10), 1258-1268.
[10] Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370-374.
[11] Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39-42.
[12] Tsagris, M., & Pandis, N. (2021). Multicollinearity. *American Journal of Orthodontics and Dentofacial Orthopedics*, *159*(5), 695-696.
[13] Chan, B. K., & Chan, B. K. (2018). Data analysis using R programming. *Biostatistics for Human Genetic Epidemiology*, 47-122.
[14] Özkaya, U., Yiğit, E., Seyfi, L., Öztürk, Ş., & Singh, D. (2021). Comparative regression analysis for estimating resonant frequency of c-like patch antennas. Mathematical Problems in Engineering, 2021, 1-8.