Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 211-215, 2, 2024
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

IJANSER

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

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 211-215, 2, 2024
Copyright © 2024 IJANSER
Research Article

Predicting The Demand For Shared Bicycles In Seoul By Multiple Linear Regression

Yann Ling Goh^{*1}, Raymond Ling Leh Bin ^{2}

[1]Universiti Tunku Abdul Rahman, Lee Kong Chian Faculty of Engineering and Science, Malaysia, gohyl@utar.edu.my
[2]Universiti Tunku Abdul Rahman, Faculty of Accountancy and Management, Malaysia, linglb@utar.edu.my

Email of the corresponding author: (gohyl@utar.edu.my)

(Received: 04 March 2024, Accepted: 08 March 2024)

(4th International Conference on Innovative Academic Studies ICIAS 2024, March 12-13, 2024)

ATIF/REFERENCE: Goh, Y. L. & Bin, R. L. L. (2024). Predicting The Demand For Shared Bicycles In Seoul By Multiple Linear Regression. *International Journal of Advanced Natural Sciences and Engineering Researches, 8(2), 211-215.*

Abstract: The study used a multiple linear regression model to model the demand for shared bicycles and related factors in Seoul for the year 2020. Data analysis was performed to find out the influencing factors that affect the demand for shared bicycles in Seoul. Correlation analysis was carried out to check the relationship between all variables and identify the multicollinearity problem in the data. After fitting multiple linear regression, it was found that the demand for shared bicycles in Seoul was significantly affected by hour of the day, temperature, humidity, visibility, solar radiation and rainfall. Among these variables, it was found out that solar radiation is the most important factor.

Keywords- Linear Regression, Shared Bicycles Correlation, Analysis Multicollinearity.

I. INTRODUCTION

The bicycle sharing system allows users to make a one way bicycle trip within a short distance. The original bicycle sharing system began in Europe in the 1960s, but the concept was not popularized globally until the mid-2000s. Nowadays, the system is very intuitive among cities, and it is easy for users to understand even first time user. Usually these systems are operated through automatic kiosks to save manpower and reduce user waiting time. When a person rides a shared bicycle to a small town, the bicycle system may be cheaper than renting a car or using public transportation. Furthermore, bicycle sharing system helps in reducing air pollution because the use of bicycles will reduce the use of motor vehicles and hence reduce emissions of pollutants to the air [1-3].

LinkBike was launched in 2016 and was the first bicycle sharing system in Malaysia. Currently, there are 250 LinkBike bicycles and 25 stations in George Town, Penang, Malaysia. LinkBike uses two electronic payment methods - smart cards and mobile phone applications. The mobile phone app is designed as a QR code scanner with a GPS based device that can display the nearby location of LinkBike. The number of bicycles available at any specific location in real time is updated by this app. The LinkBike's smart card can be topped up by using credit card or pay cash at the LinkBike office. Each LinkBike bicycle is equipped with a light emitting diode. LinkBike bicycle will light up the front and rear of the bicycle once movement is detected in order to improve the visibility and safety of night riders [4, 5].

Multiple linear regression is a statistical method that uses a few independent variables to predict the outcome of a dependent variable. As multiple linear regression contains more than one explanatory variable, it is the extension of ordinary least-squares regression. The linear relationship between the explanatory variables and response variable is modeled by multiple linear regression [6-9]. Rule-based regression prediction model was applied to predict the demand for bicycle sharing. Cubist was categorized as rule-based learning, an advanced empirical modelling approach to improve the performance of existing learning algorithms. The study showed that cubist algorithm could be used as an effective tool for bike sharing demand forecasting. An analysis of the importance of variables was carried out to reveal the hidden relationships between the variables [10]. Regression model with spatially varying coefficients had been constructed in examining land use, social demographics and transportation influence for the need of designing bicycle sharing services in different locations [11].

Spatial analysis by analytical hierarchy process and spatial multi criteria analysis was used to find the best place for siting shared bicycle station. According to the analysis, the heart of the city of Yogyakarta and its surroundings was the most suitable place for a bicycle sharing station [12]. Dynamic bicycle sharing design model was created with demand forecasting and optimization scheduling. This study proposed a method to solve the model based on Nicked Pareto Genetic Algorithm and was confirmed by a clinical study. After the scheduled optimization was improved, the waiting, transfer, and departure behavior of users when they cannot borrow a bicycle was greatly reduced. The results showed that a dynamic and logical schedule could effectively improve resource allocation and improve system service level [13]. Furthermore, K-means clustering has been proven as a clustering algorithm for rebalancing bicycle sharing patterns. The research proposed a new comprehensive methodology for dynamic bicycle redistribution, starting with forecasting the number and position of bicycle in the area of operation of the system and ending with a system to support relocation decisions [14, 15].

II. MATERIAL AND METHOD

In this study, data of Seoul Bike Sharing Demand for the year 2020 was used. It was obtained from the UCI datasets. The dependent variable, Y was the total count of bicycles rented. The independent variables were x_1, hour of the day, x_2, temperature (Celsius), x_3, humidity (%), x_4, windspeed (m/s), x_5, visibility (10m), x_6, dew point temperature (Celsius), x_7, solar radiation (MJ/m2), x_8, rainfall (mm) and x_9, snowfall (cm).

The estimated multiple linear regression model is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

where $\hat{Y} =$ dependent variable, $x_j =$ independent variables, $\hat{\beta}_0 = Y$-intercept (constant term), $\hat{\beta}_j =$ slope coefficients for each independent variable, $j = 1,2,\cdots,k$.

Once the model is built, a testing on the significance of the partial regression coefficients will be performed. The purpose of the test is to check whether there exist predictor variables which do not contribute significantly to the model. If yes, then the particular predictor variable can be omitted from the model in order to form the reduced regression model. The hypothesis testing is set as follow with null and alternative hypothesis:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

The test statistics is:

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}; \text{Reject } H_0 \text{ when } |t_0| > t_{\frac{\alpha}{2},n-k-1}$$

where

1. $se(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$
2. n is the number of observations
3. k is the number of independent variables

If H_0 is not rejected, this indicates that the regressor x_j does not contribute significantly to the model, in other words, that particular regressor can be deleted from the model.

When there are several independent variables in the model, the risk of having the problem of multicollinearity between independent variables is always exists. In order to look for those highly correlated independent variables, a correlation matrix is being observed. The entries of the matrix should contain values between -1 and 1, the closer the entries' values to 1 or -1, the higher the correlation between the variables. Besides, variance inflation factor (VIF) is also used to determine multicollinearity between variables. In R, VIF can be easily computed by using Dr. Fox's CAR package, which gives the advanced utilities for regression modelling. In general, the VIF is computed based on a tolerance, c. The tolerance c is defined as $\mathrm{c} = 1 - \mathrm{R}_j^2$ where the R-squared value is obtained by regression the j^{th} predictor on the remaining predictors. Then the VIF is computed as $\mathrm{VIF} = \frac{1}{\mathrm{c}}$. VIF equal to 1 implies that single independent variable is not correlated with other variables. The higher the value of VIF, the larger the correlation of the variable with other variables. VIF value of 10 or more implies very large correlation.

III. RESULTS AND DİSCUSSİON

This study used 9 independent variables with 8760 observations. The dependent variable was the total count of bicycles rented. The independent variables were assumed to have effect on dependent variable. Table 1 showed the correlation matrix between independent variables. Based on Table 1, there was very strong positive correlation between x_2 and x_6 since the value was very close to +1. This fact indicated potentially to have multicollinearity problems in the data.

Table 1. Correlation matrix between independent variables

Variable		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	Hour of the day	1								
x_2	Temperature (°C)	0.124	1							
x_3	Humidity (%)	-0.241	0.159	1						
x_4	Wind speed (m/s)	0.285	-0.036	-0.337	1					
x_5	Visibility (10m)	0.099	0.035	-0.543	0.172	1				
x_6	Dew point temperature (°C)	0.003	**0.913**	0.537	-0.176	-0.177	1			
x_7	Solar radiation (MJ/m2)	0.145	0.354	-0.462	0.332	0.150	0.094	1		
x_8	Rainfall (mm)	0.009	0.050	0.236	-0.020	-0.168	0.126	-0.074	1	
x_9	Snowfall (cm)	-0.022	-0.218	0.108	-0.004	-0.122	-0.151	-0.072	0.009	1

Table 2 showed estimated regression coefficients for the Seoul Bike Sharing Demand dataset using full model of multiple linear regression. The VIF values for each independent variable were shown in the last column of Table 2. The presence of multicollinearity in the model was confirmed because of high values of VIF (more than 10) in variables x_2, x_3 and x_6. The variables x_2, x_3 and x_6 perhaps causing redundant consequences to the response variable. As variable x_6 (Dew point temperature) showed the highest VIF value (115.69) among these 3 variables, variable x_6 will be excluded from the model. Table 3 displayed the result of regression analysis when variable x_6 was excluded. The performance was improved in term of low values of VIF (less than 3), reflected multicollinearity problem was solved.

Table 2. Full model of the Seoul Bike Sharing Demand dataset

Variable		Parameter estimate	DF	Standard error	t value	p value	VIF
Intercept		548.85	1	100.05	5.49	4.23E-08	0
x_1	Hour of the day	27.32	1	0.79	34.64	1.5E-246	1.18
x_2	Temperature (°C)	26.58	1	3.92	6.78	1.28E-11	87.11
x_3	Humidity (%)	-8.81	1	1.11	-7.93	2.56E-15	20.36
x_4	Wind speed (m/s)	6.92	1	5.47	1.27	0.205662	1.28
x_5	Visibility (10m)	0.02	1	0.01	2.06	0.039275	1.57
x_6	Dew point temperature (°C)	5.41	1	4.13	1.31	0.190206	115.69
x_7	Solar radiation (MJ/m2)	-79.34	1	8.21	-9.66	5.54E-22	2.02
x_8	Rainfall (mm)	-58.81	1	4.63	-12.70	1.27E-36	1.08
x_9	Snowfall (cm)	21.08	1	12.02	1.75	0.079629	1.10

Table 3. Reduced model of the Seoul Bike Sharing Demand dataset after removing x_6

Variable		Parameter estimate	DF	Standard error	t value	p value	VIF
Intercept		426.94	1	36.75	11.62	5.66E-31	0
x_1	Hour of the day	27.27	1	0.79	34.62	3.4E-246	1.18
x_2	Temperature (°C)	31.67	1	0.54	59.16	0	1.62
x_3	Humidity (%)	-7.45	1	0.39	-18.99	7.9E-79	2.53
x_4	Wind speed (m/s)	6.64	1	5.46	1.21	0.224641	1.27
x_5	Visibility (10m)	0.02	1	0.01	2.18	0.029615	1.56
x_7	Solar radiation (MJ/m2)	-81.69	1	8.01	-10.20	2.85E-24	1.92
x_8	Rainfall (mm)	-59.49	1	4.60	-12.93	6.8E-38	1.07
x_9	Snowfall (cm)	20.02	1	12.00	1.67	0.095256	1.09

Backward elimination is the feature selection technique used for determining independent variables that significantly contributing the total count of bicycles rented. According to this technique, independent variable with highest p value will be removed from the model. From the Table 3, among all the 8 independent variables, variable x_4 (wind speed) had the highest p value of 0.224641, so this variable was withdrawn from the model. The backward elimination process was continued until all the variables had p value less than the specified alpha. The model was finalised when no more variables could be excluded from the model. Table 4 displayed the final reduced model of the Seoul Bike Sharing Demand dataset.

Table 4. Final reduced model of the Seoul Bike Sharing Demand dataset after removing all insignificant independent variables

Variable		Parameter estimate	DF	Standard error	t value	p value	VIF
Intercept		433.38	1	35.97	12.05	3.56E-33	0
x_1	Hour of the day	27.57	1	0.76	36.12	1.8E-266	1.11
x_2	Temperature (°C)	31.35	1	0.51	61.22	0	1.49
x_3	Humidity (%)	-7.38	1	0.39	-19.05	2.4E-79	2.47
x_5	Visibility (10m)	0.02	1	0.01	2.24	0.02538	1.55
x_7	Solar radiation (MJ/m2)	-77.85	1	7.71	-10.10	7.43E-24	1.78
x_8	Rainfall (mm)	-59.43	1	4.60	-12.93	6.78E-38	1.07

In the final reduced model, it was observed that variables which were significantly contributing the total count of bicycles rented, namely x_1 (hour of the day), x_2 (temperature), x_3 (humidity), x_5 (visibility), x_7 (solar radiation) and x_8 (rainfall). The estimated regression equation of the total count of bicycles rented was written as:

$$\hat{y} = 433.38 + 27.57x_1 + 31.35x_2 - 7.38x_3 + 0.02x_5 - 77.85x_7 - 59.43x_8$$

x_1, x_2, x_5 had positive effects while x_3, x_7, x_8 had negative effects on the total count of bicycles rented. The negative effects of the x_8 means that higher value of rainfall corresponds to lower value of the total count of bicycles rented. This information indicated that people in Seol tend not to rent bicycles when rainy. Furthermore, among all the independent variables which contributed significantly to the total count of bicycles rented, variable x_7 which was solar radiation had the highest effect on the total count of bicycles rented. This explained that solar radiation had great influence on the total count of bicycles rented.

IV. CONCLUSION

This study used multiple linear regression to model data of Bike Sharing Demand in Seoul. The analysis presented in this study revealed that hour of the day, temperature, humidity, visibility, solar radiation and rainfall are the major factors affect the total count of bicycles rented. The bicycle sharing system has the potential to overcome traffic jams, reduce carbon emissions, solve inadequate parking spaces, etc. The advantages of bicycle sharing system should be promoted to public to encourage more citizens to participate. The increase of utilisation of bicycle sharing will provide a healthy lifestyle for the users and offer an environmentally friendly mode of transport.

REFERENCES

1. Kamel, M. B., & Sayed, T. (2021). The impact of bike network indicators on bike kilometers traveled and bike safety: A network theory approach. Environment and Planning B: Urban Analytics and City Science, 48(7), 2055-2072.
2. Kim, K. (2021). Impact of Covid-19 on usage patterns of a bike-sharing system: case study of Seoul. Journal of Transportation Engineering, Part A: Systems, 147(10), 05021006.
3. Scott, D. M., Lu, W., & Brown, M. J. (2021). Route choice of bike share users: Leveraging GPS data to derive choice sets. Journal of Transport Geography, 90, 102903.
4. Ding, H., Sze, N. N., Li, H., & Guo, Y. (2021). Effect of London cycle hire scheme on bicycle safety. Travel Behaviour and Society, 22, 227-235.
5. Gu, T., Kim, I., & Currie, G. (2021). The two-wheeled renaissance in China—an empirical review of bicycle, E- bike, and motorbike development. International Journal of Sustainable Transportation, 15(4), 239-258.
6. Bekesiene, S., Meidute-Kavaliauskiene, I., & Vasiliauskiene, V. (2021). Accurate prediction of concentration changes in ozone as an air pollutant by multiple linear regression and artificial neural networks. Mathematics, 9(4), 356.
7. Goh, Y. L., Goh, Y. H., Raymond, L. L. B., & 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.
8. Goh, Y. L., Goh, Y. H., Yip, C. C., & Ng, K. H. (2022). Housing price prediction by divided regression analysis. Chiang Mai Journal of Science, 49(2), 1669-1682.
9. Liu, M., Hu, S., Ge, Y., Heuvelink, G. B., Ren, Z., & Huang, X. (2021). Using multiple linear regression and random forests to identify spatial poverty determinants in rural China. Spatial Statistics, 42, 100461.
10. VE, S., & Cho, Y. (2020). A rule-based model for Seoul bike sharing demand prediction using weather data. European Journal of Remote Sensing, 53(sup1), 166-183.
11. Wang, X., Cheng, Z., Trépanier, M., & Sun, L. (2021). Modeling bike-sharing demand using a regression model with spatially varying coefficients. Journal of Transport Geography, 93, 103059.
12. Kurniadhini, F., & Roychansyah, M. S. (2020). The suitability level of bike-sharing station in Yogyakarta using SMCA technique. In IOP Conference Series: Earth and Environmental Science, 451(1), 012033. IOP Publishing.
13. He, L., Guo, T., & Tang, K. (2020). Dynamic scheduling model of bike-sharing considering invalid demand. Journal of Advanced Transportation, 2020.
14. Caggiani, L., Camporeale, R., Ottomanelli, M., & Szeto, W. Y. (2018). A modeling framework for the dynamic management of free-floating bike-sharing systems. Transportation Research Part C: Emerging Technologies, 87, 159-182.
15. Goh, Y. L., Goh, Y. H., Yip, C. C., Ting, C. H., Bin, R. L. L., & Chen, K. P. (2020). Students' academic performance analysis by K-means clustering for investigating students' health conditions within clusters. Annals of Tropical Medicine and Public Health, 23(13).