

From Basic to Advanced: Linear Regression Approaches for Analyzing Transportation Data

Ruilin Cao^{1*}, Qiao Hu², Zhongyi Yu³

¹*Shenzhen College of International Education, Shenzhen, China*

²*Wuhan Britan-China School, Wuhan, China*

³*Beijing 21st Century School, Beijing, China*

**Corresponding Author. Email: s22227.cao@stu.scie.com.cn*

Abstract: This paper employs both simple linear regression and multiple linear regression models to analyze transportation operation-related data. It primarily addresses three key issues: civil aviation business evaluation and passenger flow prediction, fuel consumption and payload prediction in aviation, and traffic accident impact forecasting. For civil aviation business evaluation and passenger flow prediction, the study applies multiple linear regression to assess the influence of factors such as national income on air passenger volume. While the constructed model demonstrates good fitting performance, limitations include a small sample size and multicollinearity among independent variables. Regarding traffic accident impact prediction, the research focuses on highway accidents by integrating VISSIM simulation, nonlinear regression, and neural networks to forecast spatiotemporal accident effects. Results indicate that the neural network achieves higher accuracy, whereas nonlinear regression offers the advantage of formula interpretability. Concerning fuel consumption and payload prediction, a simple linear regression model is established to quantify the relationship between aircraft payload and fuel usage. This provides dispatchers with actionable insights for fuel management and can be extended to different aircraft types and flight segments. Collectively, these studies offer significant methodological frameworks and decision-support tools for management across various transportation sectors.

Keywords: Regression Analysis, Least Squares, Regression Equation, Significance Test

1. Introduction

In real life, it is common to encounter the study of the relationship between a dependent variable and its corresponding multiple independent variables. For instance, the demand for a commodity is not only influenced by its price but also significantly impacted by factors such as people's average wages, standard of living, and the availability of substitute products [1]. To accurately explore and quantify such relationships, researchers commonly employ multiple regression analysis, which is a robust statistical method capable of analyzing how several independent variables collectively affect a single dependent variable. This technique has been instrumental in various fields, from economics to social sciences, in effectively forecasting trends and formulating strategic decisions based on multifaceted data.

With the rapid development of the global economy, human needs have transitioned substantially from satisfying basic necessities to pursuing various recreational products and services. Within this

context, tourism stands out prominently, reflecting a significant shift toward leisure-oriented consumption [2]. Consequently, the booming demand for tourism services has directly spurred the rapid growth of the civil aviation industry. As a critical infrastructure supporting tourism, civil aviation has experienced an unprecedented surge in passenger volumes, making the accurate analysis and prediction of passenger traffic vital for airline companies [3]. Airlines increasingly focus on understanding the underlying trends and dynamics influencing passenger demand to optimize operational efficiency and strategic planning. Existing research highlights that civil aviation passenger traffic is impacted not only by straightforward economic indicators such as national income levels and consumer spending but also by political stability, climate conditions, and seasonal variations. These factors, individually and collectively, create a complex environment that significantly influences passenger traffic volumes and their fluctuating patterns. Addressing these variables comprehensively through multiple regression analysis is therefore critical to accurately forecasting future trends and enhancing industry responsiveness [4].

This paper seeks to deepen the understanding of civil aviation passenger traffic dynamics by focusing on key influencing factors, including national income, consumption expenditures, railway passenger volumes, and the mileage of civil aviation routes. Adopting a comprehensive multiple regression analysis framework, this research systematically investigates how these various factors individually and collectively affect civil aviation passenger trends. The next sections are organised as follows: the first section reviews the existing theoretical foundations on linear regression, provides background information for this study and outlines the research methodology, which includes multivariate linear regression as well as uni-variate linear regression. The variable selection criteria and the analytical techniques employed are described in detail. The second section is a practical application, where the theoretical equations are obtained by performing regression analysis on the raw data.

2. Simple linear regression model

2.1. Basic theory

By studying the simple linear regression model, a predictable quantitative solution is provided for dispatchers. Since dispatchers need to adjust the fuel usage of aircraft plans based on the changes in payload in a short period of time, it is necessary to provide them with a quantifiable reference solution.

Univariate linear regression can be used to analyze the linear relationship between two variables. The basic application method is the following. It involves collecting paired data and checking whether the data meets the linear trend, such as by drawing a scatter plot for preliminary observation. Then, it uses the least squares method to estimate the parameter value by minimizing the sum of squared errors, which is usually simulated by Excel or Python. Next, it employs R to explain the degree of fit of the model to the data variation and checks whether the residuals are randomly distributed to verify the linear hypothesis. Finally, it substitutes new values to calculate the predicted value. Usually, the univariate linear regression model is used in business analysis, economics, etc. to study the linear relationship between two variables.

Supposed that one can determine the independent variable x and dependent variable Y , then one can make predictions by the linear regression below:

$$Y = a + bx \quad (1)$$

Using the least squares method to solve the model parameters a and b , it is found that

$$a = \frac{(\sum Y)(\sum x^2) - (\sum x)(\sum xY)}{n(\sum x^2) - (\sum x)^2}, b = \frac{n(\sum xY) - (\sum x)(\sum Y)}{n(\sum x^2) - (\sum x)^2} \quad (2)$$

By minimizing the sum of the squares of the differences R between the actual value and the calculated value Y , one can get the calculation formula for a and b , and also

$$R = \frac{n \sum xY - (\sum x)(\sum Y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum Y^2 - (\sum Y)^2]}} \quad (3)$$

At the significance level α , the critical value table of the correlation coefficient is shown. If the $R > R_{\alpha}(n - 2)$ is satisfied, the correlation coefficient is significant and the regression equation is meaningful [5].

2.2. Flight load and fuel quantity model establishment

The authors shall select the Boeing B737-800W model and the Guangzhou to Zhengzhou flight segment [6]. Then, it shall change the payload, use the LIDO system to calculate the payload and fuel data, and draw a scatter plot. The data are shown in Table 1.

Table 1: The payload (x) and fuel (Y) of Boeing B737-800W [6]

times	Payload (x)	Fuel (Y)	times	Payload (x)	Fuel (Y)	times	Payload (x)	Fuel (Y)
1	5000	9985	4	8300	10266	7	11320	10534
2	6200	10089	5	9500	10371	8	13250	10709
3	6900	10147	6	10000	10416	9	14750	10847

In this data shown in Table 1, it can be inferred that there is an obvious linear relationship between payload x and fuel Y . People can substitute the data into the formula to calculate $b = 0.088$, $a = 9538.155$, and get the linear regression equation $Y = 0.088x + 9538.155$ between the payload x and the fuel volume Y . Besides, one can substitute the data into the formula to calculate $R = 0.9999$, at the significant level $R_{\alpha=0.05}(n - 2) = R_{\alpha=0.05}(0.088x + 9538.155) = 0.6021$, which is satisfied equation $R > R_{\alpha}(n - 2)$. Therefore, this shows that the regression equation is valid.

3. Multiple linear regression model

3.1. Basic theory

Let the number of independent variables affecting the dependent variable y be m , denoted as $x_1, x_2, \dots, x_m$, the multivariate linear model means that the effect of these independent variables on y is linear, i.e., the relation equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \varepsilon \quad (4)$$

where: $\beta_0, \beta_1, \beta_2, \dots, \beta_m$ is $m + 1$ unknown parameters, β_0 is constant term, $\beta_1, \beta_2, \dots, \beta_m$ are called the regression coefficient and are m general variables that can be given precise values and can be controlled in general, called explanatory variables [7]. Call y a linear regression function on the independent variables $x_1, x_2, \dots, x_m$. ε is the random error and is usually considered as $\varepsilon \sim N(0, \sigma^2)$.

In the practical problem, n sets of observations on $(x_1, x_2, \dots, x_m, y_i)$ ($i = 1, 2, \dots, n$) are obtained, and the

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1m} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2m} + \varepsilon_2 \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{nm} + \varepsilon_n \end{cases} \quad (5)$$

where $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \sim N(0, \sigma^2)$ and $i = 1, 2, \dots, n$. This model is called a multiple linear regression model. Let

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (6)$$

Then the matrix form of the above mathematical model is

$$Y = X\beta + \varepsilon \quad (7)$$

where ε is an n –dimensional random vector whose components are independent and identically distributed [8].

To proceed, there are some basic assumptions of the multiple linear regression model. It is generally accepted that regression models should satisfy the following basic assumptions: 1) Explanatory variables $x_1, x_2, \dots, x_m$ is a random variable. observed value $x_{i1}, x_{i2}, \dots, x_{im}$ is a constant. 2) The assumptions for variance chi-square and uncorrelation are $E(\varepsilon_i) = 0, i = 1, 2, \dots, n$ and $\text{CoV}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} (i, j = 1, 2, \dots, n)$. This is known as the Gauss-Markov condition, abbreviated as the G-M condition. In this condition, some important qualities of the regression coefficient equation can be obtained, for example, the least squares estimate of the regression coefficient is obtained.

For example, the least squares estimate of the regression coefficients is obtained as a linear unbiased estimate of the regression coefficients with minimum variance, minimum variance linear unbiased estimation, etc. The normal distribution is assumed to be: $\varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \dots, n$, in which $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are mutually independent. Further results on the estimation of the regression coefficients and the estimation of σ^2 can be obtained, e.g., they are unbiased estimates of the minimum variance of the regression coefficients and σ^2 , respectively, and regression significance tests and interval estimations can be performed.

Table 2: Statistical data for 1995-2010 [9]

years	y	x ₁	x ₂	x ₃	x ₄	x ₅
1995	277	3162	2266	97789	18	217
1996	358	4020	2634	103667	19	504
1997	412	4426	3037	110645	23	684
1998	481	4729	3359	114360	26	932
1999	534	5110	3665	119906	28	951
2000	469	5683	4030	127253	27	1137
2001	665	6782	4686	132424	31	1542
2002	893	8424	5855	134532	33	2140
2003	1196	9431	6662	130295	39	2738
2004	1572	11176	7663	134915	47	3228

Table 2: (continued).

2005	1730	14086	9646	147174	45	3803
2006	1540	15811	10806	136568	57	2940
2007	1992	17261	11596	114854	61	3295
2008	2614	19868	13163	114097	67	4003
2009	3463	24268	15582	119632	100	3974
2010	4060	29858	19139	126550	115	4983

3.2. Civil aviation passenger capacity modelling and solving

The forecast value y denotes the number of civil aviation passengers (10,000 persons), denotes the gross national income (100 million yuan), x_2 denotes the amount of consumption (100 million yuan), x_3 denotes the railway capacity (10,000 persons), x_4 denotes the distance of civil aviation routes (10,000 kilometres), and x_5 denotes the number of foreign travellers (10,000 persons). Statistical data for 1995-2010 were obtained from the Statistical Abstract 2010 and are shown in Table 2 [9].

A multiple linear regression model of y with the respective variable x_i , $1 \leq i \leq 5$, is established as follows

$$y = \beta_0 + \beta_i \sum_{i=1}^5 x_i + \varepsilon (i = 1, 2, 3, 4, 5) \quad (8)$$

where $\varepsilon \sim N(0, \sigma^2)$ and $E(\varepsilon_i) = 0 (i = 1, 2, 3, 4, 5)$. Using the original data, the correlation matrix is calculated in R language, and the output results can be obtained, from which the correlation analysis is made. It is found that $R = 0.999$, $R^2 = 0.998$, adjusted $R = 0.9971$, and standard error equals to 62.32. From the correlation matrix, it can be seen that the correlation coefficients between y and x_1, x_2, x_3, x_4, x_5 are relatively high, above 0.9, indicating that the selected independent variables are highly linearly correlated with y , and it is feasible to use y and the independent variables x_1, x_2, x_3, x_4, x_5 to do multivariate linear regression. The correlation coefficient between y and x_3 , $r_{y3} = 0.227$, is small, and x_3 is the amount of railway passengers, which indicates that the amount of railway passengers has no significant effect on the amount of civil aviation passenger traffic. There is no significant effect of railway passenger traffic on civil aviation passenger traffic.

The regression equation is

$$\hat{y} = 714.0404 + 0.3963x_1 - 0.6179x_2 - 0.0082x_3 + 19.7288x_4 + 0.4295x_5 \quad (9)$$

The complex correlation coefficient $R = 0.999$, coefficient of determination $R^2 = 0.9981$, F-test: $F = 1025$, P-value = 3.28×10^{-13} , significant at 1% or stricter level, indicating that the overall regression equation is highly significant.

3.3. GDP of Gansu Province

Gansu Province is a very important location. It is a central transport station in YiDaiYiLu, connecting Asia and Europe. Gansu has invested intensive aviation building development by expand airport and flight route to boost the economy and attract tourist. Governors need to understand air travel is demanding due to the rising competition between railroads and roads. This research utilizes MLR (multiple linear regression) to estimate the variables that could affect civil aviation passenger statistics in Gansu Province as China's airline turns busy due to the new infrastructure and strategic programs-China's Belt and Road. Investigating in these elements' relationship that contributes to passengers

can enable policy makers to efficiently allocate resources and improve Gansu's aviation market to be more competitive.

The study uses data from year 2006 to 2016 [10]. Gansu's civil aviation passenger volume is the dependent variable(y). Six potential influencing factors are selected as independent variables(x): Provincial GDP, Population size, Resident consumption level, Civil aviation route mileage, Tourist arrivals, and Railway passenger volume.

During the process of applying Multiple Linear Regression, Multi-collinearity was detected, with variance inflation factors all exceeding 10. Non-significant factors (like GDP and Population) were automatically eliminated using backward elimination. So, the refined equation contains two significant predictors (explanatory variable x_4 and x_6):

$$y = 50.321 + 567.757x_4(\text{route mileage}) + 0.04x_6(\text{tourism numbers}) \quad (10)$$

For every 1,000 km added to the route mileage, an approximate value of 567,757 passengers was added, while an increase of 10,000 tourists added close to 400 passengers. As for the result, it is found that R^2 is 0.85, means that 85% of variance in y can be explained by the predictors x . ANOVA: $P < 0.001$ confirms that the model is statistically significant.

4. Conclusion

To summarize, this work considers the applications of univariate and multiple linear regression models. Firstly, the univariate linear regression model provides a reliable method to explain the relationship between aviation fuel and payload. The linear relationship between fuel and payload also confirms the mutual influence between the two. As shown in the model, for every 1,000kg increase in payload, the fuel volume also increases by 88kg. The staff can more accurately predict the specific increase and decrease in fuel demand when the payload changes within a small range, thereby reducing the workload and saving resources. This method can also be applied to other different aircraft models and flight segments, increasing or reducing fuel in advance according to the changing trend of payload, reducing temporary adjustments before takeoff, and reducing operational risks. Secondly, the statistical test of the multiple linear regression model is highly significant and the R^2 is extremely high, indicating a good fit to the data within the sample period. However, due to the small sample size (only 16 periods) and the high correlation of the independent variables, such a high correlation indicates the risk of multicollinearity, which can be considered by eliminating or combining the highly correlated variables, or by using methods such as principal component regression or ridge regression. Finally, the investigation determined that aviation route mileage together with tourism numbers serve as fundamental elements which affect civil aviation passenger volume in Gansu Province. The critical drivers to increase air travel demand include expanding flight routes together with tourism promotion efforts. The Silk Road Economic Belt positioning creates possibilities to construct fresh airports and enhance present facilities for airports like Lanzhou Zhongchuan International Airport. Enhancing airport transportation and facilities will both service increasing customer requirements and turn Gansu into a major air transport center in Northwest China.

In addition, there is also a nonlinear regression analysis which uses predictors through curved or non-straight-line patterns. Nonlinear regression deviates from linear regression by using mathematical equations which include polynomials and exponentials and logarithmic functions among others. The technique finds widespread applications in biological studies where it evaluates growth rates and economic investigations track the decline of returns. The implementation method relies on an assumption that the data's underlying pattern conforms precisely to the selected mathematical equation which restricts potential model adjustments. Future efforts can be made to apply this nonlinear regression analysis to specific examples.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] Maganioti, A.E., Chrissanthi, H.D., Charalabos, P.C., Andreas, R.D., George, P.N. and Christos, C.N. (2010) Cointegration of Event-Related Potential (ERP) Signals in Experiments with Different Electromagnetic Field (EMF) Conditions. *Health*, 2, 400-406.
- [2] Botorabi, F., Haapasalo, J., Smith, E., Haapasalo, H. and Parkkila, S. (2011) Carbonic Anhydrase VII—A Potential Prognostic Marker in Gliomas. *Health*, 3, 6-12.
- [3] Ross, S. M. (2019). *A first course in probability* (10th ed.). Pearson Education.
- [4] Yang You & Li Xiao hong. (2006). *Instance Analysis of Multi-Collinearity by Stepwise Regression*. *Journal of Chongqing Three Gorges University*, (03), 39-41.
- [5] Zhang, J., Yang, J., Liu, F., et al. (2024). Analysis of multiple linear regression prediction algorithm in teaching optimization. *Electronic Technology*, 53(12), 40-41.
- [6] Deng, C. Z. (2017). Study on the relationship between flight payload and fuel based on simple linear regression model. *Journal of Civil Aviation*, 1(01), 13-14+68.
- [7] XIE Chunyu & WAN Wenjun. (2022). Analysis on Influnec Factors of Stock Price in Secondary Market—Multiple Linear Regression Based on R. *China Academic Journal*, (15), 99-102.
- [8] Soffritti, G., & Galimberti, G. (2010). Multivariate linear regression with non-normal errors: A solution based on mixture models. *Statistics and Computing*, 21(4), 523–536.
- [9] Ma, Y., & Tang, Y. T. (2020). Analysis of the influencing factors of civil aviation passenger traffic in Gansu Province based on multiple linear regression model. *Gansu Science and Technology*, 36(09), 46-49+13.
- [10] Zhang, Y., Miao, G., & Li, Y. K. (2012). Multiple linear regression analysis of civil aviation passenger traffic. *Journal of Sichuan Ordnance*, 33(08), 81-84+91.