CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	A study of airline route data
Authors	Yu, Xiaochen
Publication date	2020
Original Citation	Yu, X. 2020. A study of airline route data. MRes Thesis, University College Cork.
Type of publication	Masters thesis (Research)
Rights	© 2020, Xiaochen Yu. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-06 15:39:39
Item downloaded from	https://hdl.handle.net/10468/11329



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh



A Study of Airline Route Data

by

Xiaochen Yu

Student number: 114113830

Presented for the degree of

Master of Science in Statistics

Under the Supervision of:
Eric Wolsztynski, PhD and Janet O'Sullivan, MSc

2020

DECLARATION

This is to certify that the work I am submitting for a **Master of Science in Statistics,** entitled **"**A Study of Airline Route Data**",** under the direction and supervision of Dr. Eric Wolsztynski and Ms. Janet O'Sullivan, University College Cork, Ireland, is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

Name: Xiaochen Yu
Student no: 114113830
Signature:

Abstract

For decades, with the advancement of airline information system construction, the aviation industry has successfully built a number of information systems. An enormous amount of data has been accumulated through the successful operation of these systems for the aviation sector. The effective use of these invaluable data assets has increasingly become a requirement for the relevant airline departments, and the focus of aviation industry. Revenue management is crucial for measuring the operational success of the airlines. However, the traditional forecasting methods cannot support processing of the underlying data that keeps changing over time, which have impaired the accuracy of the forecast, and thus, the credibility. The new airline passenger ticket revenue pricing methods proposed in this thesis have explored the possibility of solving the existing problems through advanced modeling techniques, and thus provide a view for better airline route planning and optimisation. First of all, the airline data is classified in a targeted manner. The factors of available seat kilometers, revenue passenger kilometers, load factors, total number of passengers, average fares, etc. collected from different time periods were used to establish a multiple linear regression model in the statistical software, R. Through empirical analysis it is found that the factors affecting the passenger ticket revenue, over different time periods of the same company, are different. Therefore, a multivariate linear regression model was established which was based on the data of different airlines in one specific time period. It was empirically found through this approach that different airlines had different factors affecting ticket revenue in the same time period. A multivariate linear regression model was established for analysing the data of the head office and the various branches at the same time period and through this it was found that the factors affecting the passenger ticket revenue of the head office and the branches at the same time period were different. The research results of this dissertation can provide scientific evidence that airlines should consider analysing real-time ticket data for ticket price and flight plan collected from the IT system to maximise the revenue.

Key Words: Revenue management; multivariate linear regression; stepwise regression method; airline ticket pricing; available seat kilometers; revenue passenger kilometers.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 6

1.1 RESEARCH BACKGROUND OF STATISTICAL ANALYSIS OF FARES 6
1.2 SIGNIFICANCE OF STATISTICAL ANALYSIS OF FARES 8
1.3 MULTIVARIATE STATISTICAL ANALYSIS 9
1.4 THESIS LAYOUT 10

CHAPTER 2: MULTIPLE LINEAR REGRESSION ANALYSIS 12

2.1 MATHEMATICAL MODEL 12
2.2 BASIC ASSUMPTIONS OF THE MODEL 13
2.3 PARAMETER ESTIMATION 14
2.4 SIGNIFICANCE TEST 15
2.4.1 SIGNIFICANCE TEST OF REGRESSION EQUATION -- F TEST 15
2.4.2 SIGNIFICANCE TEST OF REGRESSION PARAMETERS -- T TEST 16
2.4.3 SIGNIFICANCE TEST OF LINEAR REGRESSION EQUATION . 17
2.5 RESIDUAL ANALYSIS 18
2.6 FORECAST 19
2.7 DATA PREPROCESSING 19

CHAPTER 3: INDEPENDENT VARIABLE SELECTION AND STEPWISE REGRESSION 22

3.1 SELECTION CRITERIA FOR INDEPENDENT VARIABLES 22
3.2 STEPWISE REGRESSION 24

CHAPTER 4: EMPIRICAL ANALYSIS 27

4.1 DATA SPECIFICATION 27
4.2 AIRLINE A MODELS 34
4.3 ANALYSIS OF AIRLINE B TICKET INCOME FACTORS 42
4.4 ANALYSIS OF THE REGRESSION RESULTS 49

CHAPTER 5: CONCLUSION 52

REFERENCE 53

ACKNOWLEDGEMENTS 55

Chapter 1: Introduction

1.1 Research background of statistical analysis of fares

People in the airline industry who frequently travel to the United States often experience the following scenario: in order to save on the cost of travel, they have to apply to the cooperative airlines for free tickets in the US domestic section before departure. There is a high flight density- generally there is one flight every half an hour for an airline in some large airports. However, at the airport, there could be long queues at the desks, and one can find that empty flights seats are no longer available that day. In order to reach their destination on time, customers finally must buy a ticket for the same route at several times the usual price. After boarding, people in the industry will naturally think that all the encounters are due to the impact on the airline's "revenue management". The high price tickets contribute a lot of revenue to the corresponding airline's revenue management. At the same time, it is precisely because of the revenue management that they can ensure that they can buy tickets on the spot and go to the destination.

We can get some relevant information from the development of the pricing mechanism for the US aviation industry. In 1978, the U.S. Aviation Administration issued the act of deregulation, which gave airlines more rights in their pricing mechanism and market access; it was the trigger of a global air transport revolution. At that time, many proponents of this act thought that after the price of airline fares was deregulated people would buy tickets more easily and quickly, that the price discrimination would disappear, and people could easily buy tickets as easily as buying coke in front of vending machines. However, the reality did not conform to the analysis and prediction. When the civil aviation administration of various countries loosened the control of ticket prices, large airlines developed CRSs (computerised reservation systems) and GDSs (global distribution systems) in a short time, using their own technology and capital advantages. These systems allow airlines to make use of advanced algorithms to price tickets scientifically after analysing several exogenous and internal factors that affect the price of tickets. The purpose is to maximise airline revenue.

China's civil aviation industry had clearly recognised this development. Although it started later than Europe and America, it has become the second largest aviation market in the world since 2005 after decades of development. However, our research on the pricing mechanism of air tickets is still very obscure compared with developed countries, which is reflected in the fact that the discounts for air tickets is not large enough compared with the international aviation giants. At the same time, even if domestic airlines sell cheaper tickets, consumers often do not have access to information in this regard. This shows that China’s airlines have much work to do in marketing management and marketing promotion.

Experts and scholars pointed out that China's airlines must learn from large airlines in the world's top 500 in order to achieve leapfrog development. After decades of development and reform, large international airlines have formed a very complete and systematic operation mode, including human resource management, operation cost control, air ticket pricing mechanism, etc. It is absolutely necesary for Chinese airlines to learn from these advanced experiences.

As the world economic growth slows down, so does the domestic economic growth. At the same time, the rapid development of other transportation industries, such as railway transportation, high-speed rail operation and highway construction, has seriously affected the air transport industry. The market competition is becoming more and more intense, but the total profit in the aviation industry has declined year on year. Therefore, airlines must strengthen internal management to avoid potential financial crisis and operational risks and other issues. Only by consolidating the internal management can we accept the external impact and realise the sustainable development of the enterprise.

Foreign countries have decades of experience in the analysis of airline passenger fares, and have a relatively complete and mature system. Ganjiang mentioned in 2003 that the United States promulgated The Airline Deregulation Act on October 24, 1978. Since then, the government has no longer regulated the routes and fares [1]. Fan Wei summed up the international common pricing methods in 2006. On the one hand, he collected the history and real-time data of flights, established mathematical models, and made comprehensive analyses with the data. On the other hand, he studied other factors beyond the control of airlines, such as opportunity market, tourist composition, light and peak season, and predicted the future changes of the market based on the actual situation[2]. In foreign countries, whether the pricing results are reasonable mainly refers to the two indicators of "Available Seat Kilometers" (ASK) and "Revenue Passenger Kilometers (RPK)". Available seat kilometers refer to the number of seats (carrying capacity) provided by airlines multiplied by flight mileage, which is an indicator to quantify capital investment. Revenue passenger kilometers refer to the number of passengers transported multiplied by flight mileage, which is an indicator to quantify the production of airlines. The above two indicators intuitively reflect the quality of airline operation and management, which has always been an important reference data for pricing decisions of various companies.

Based on the relatively successful experience abroad, revenue management was used by domestic Chinese airlines in 1998 [3]. Because of the continuous development of other transportation industries, Wang Fenglin pointed out in 2010 that the completion of the high-speed railway had a significant diversion effect on the aviation industry [4]. As early as 2003, Wu Tong, Li Dong and others pointed out that airlines should try to improve the management of business processes. It can effectively avoid the impact of the development of other transportation industries on the aviation industry. [5]In 2006, a large domestic airline such as Air China, China Southern Airlines and China Eastern Airlines, has set up a network revenue department to manage flights

through demand, supply analysis, forecast revenue, evaluation cost and profits. At present, the influencing factors of ticket revenue, such as the weak season or the sensitivity of passengers to the price and the information of competitors, still rely on the subjective analysis of the staff of the sales department. In fact, for the prediction of a certain variable, we should use the method of subjective and objective analysis.

The traditional pricing method of ticket revenue is too simple; the personal experience of pricing personnel is no longer suitable for the rapid development of the big data era, especially for different dimensions of ticket pricing system. Ticket revenue is the most important part of airline revenue, and it is an index to measure the financial situation of the company. Analysing the influencing factors of ticket revenue and adopting the appropriate pricing method are the primary research objectives of airlines to maximise the benefits and improve the overall benefit level, which is also an important research direction of this paper.

In the analysis of airlines' financial status and operating results, revenue settlement is a major link in the internal control system of airlines. This thesis starts from the settlement links of airlines to explore the internal control problems, which is conducive to improving the financial management level of enterprises, preventing operational risks, improving the competitiveness and sustainable development ability of airlines, and has certain practical significance.

1.2 Significance of statistical analysis of fares

Through sampling surveys, it was found that the most important concerns of passengers in the choice of travel tools are: time, price, and the comfort level. Therefore, in the process of the reform and development of air transport, the focus of people's attention has always been the price of air transportation. In the period of planned economy, the government made a direct price for air tickets, and all major airlines were prohibited from offering preferential policies to attract passengers. This kind of practice, which did not encourage airlines to improve their service level and profitability through competition, was gradually abandoned after reform and development, so that airlines now have greater pricing power. To some extent, this improved the overall quality of the airlines in the early stage.

In the early days, all airlines were not allowed discounted sales in the course of their airline ticket pricing. However, this regulation obviously hurt the interests of consumers and seriously violated the basic operation rules of the market economy. It was widely criticised by all circles of the society. Therefore, since 2000, China's Civil Aviation Administration (CAAC) has gradually relaxed the price control of air routes and group fares between some cities. In 2003, the National Development and Reform Commission issued the "CAAC domestic air transport price reform plan". According to the plan, the maximum increase of ticket price for domestic flights shall not exceed 25% of the benchmark price (0.7 yen per passenger per kilometer on average), and the minimum ticket price shall not be lower than 45% of the benchmark price. However,

there are no restrictions on the ticket price of 242 domestic tourist routes and 225 airline holiday management routes; in 2013 restrictions were further relaxed. For some civil aviation routes with fierce competition with land and sea transportation industries, and domestic routes operated by more than two airlines, the adjustment of passenger transportation fare can be determined by market price, i.e. airlines have more independent pricing rights. The promulgation of this regulation marks that the pricing mechanism of China's civil aviation passenger fares has gradually been converted from the guidance price of the government, to airlines adjusting their own fares according to the actual situation, which is conducive to improving the service level and profitability of our major airlines through competition.

As the CAAC has not announced the pricing mechanism of government guidance price and the independent pricing mechanism of major airlines, ordinary consumers are still curious about the price of air tickets.

Based on the long-term observation and data collection from multiple routes, this thesis analyses the relationship between ticket price and multiple independent variables of air fare, which have a significant impact on ticket prices. This study is based on the actual data of multiple domestic airlines, which enables ordinary consumers to have a more rational understanding of the pricing mechanism of air tickets for civil aviation in China. It provides some valuable information for the airlines and even the civil aviation administration staff, and, provides some references for academic research.

1.3 Multivariate statistical analysis

After collecting the data, we have to establish an appropriate mathematical model to describe the relationship among these variables. Often choosing variables of interest is based on practical experience.

The parameter estimation of classical multiple regression model mostly uses the ordinary least square method. After the unknown parameters are estimated it is necessary to test the statistical significance and economic significance of the model. Through factor analysis of the independent variables, according to the regression coefficient of the model, the structural relationship between the independent variables can be found, and some quantitative basis for evaluation can be given. At the same time, the regression model reveals the possibly causal relationship between variables, and the explanatory variables can be controlled by the value of the explanatory variables considered by the model.

So far, the application of regression analysis in the airline industry has not been widely applied in the pricing of ticket revenue, but it has been widely applied in other forms of transportation. In 2019, Jia Xinru applied the correlation analysis method to study the influence rule of passenger train's seat rate, mileage and train ticket revenue on passenger trains and proved that based on the historical data, the calculation table of unit mileage income under different types of train's seat rate was calculated. Compared with the actual data, the calculation of train income based on the calculation table is

simple and accurate. [6]. Lv Xiaoyan took the marketing demand of railway passenger transportation as the research background in 2007, and using time series analysis, combined with the characteristics of railway ticket data, the key factors that affect the sequence change are identified, and the obtained model has better practicability [7].

This thesis selects the data of a certain airline in January 2018 and January 2019 as samples, and applies multiple regression analysis method to analyse the influencing factors of airline ticket revenue on airline revenue. The impact of air passenger ticket revenue is analysed from four aspects: passenger volume, civil aviation route mileage, inbound tourists and gross national income.

1.4 Thesis Layout

This thesis studies and analyses the influencing factors of airline ticket revenue through the following five chapters:

The first chapter introduces the research background and significance of the thesis. Ticket revenue is an indicator to measure the operating capacity of airlines. This work uses multivariate linear regression to analyse the factors influencing airline ticket revenue.

The second chapter introduces the theory and method of the multivariate linear regression model, the basic assumptions of the model, the method of estimating the unknown parameters - the least square method, and the goodness of fit of the models to be concerned with after fitting the model, and makes a significance test for unknown parameters and regression models. The purpose of residual analysis is to test whether the data meet the basic assumptions. If results are acceptable, we can use the model to make some predictions. Finally, we propose preprocessing the data before regression.

The third chapter introduces in detail how to choose the independent variables and the criteria for selecting independent variables when applying multiple regression models. When the number of variables is large, the computation is large and difficult to complete, so a simple method of stepwise regression is introduced.

The fourth chapter is the data regression analysis of an airline's two subsidiaries and head office. Using the stepwise regression method to analyse the data, from the perspective of the factors such as seat kilometers, passenger kilometers, seat rate, total number of people, and average ticket price we find that the influencing factors of the ticket income of the same flight in different time periods and different flights in the same time periods are different, and the same influencing factors will have different interpretations. At the same time, it also demonstrates that the influence factors of the ticket revenue of the head office and the branch office are different in the same period, and the same influence factors will have different degrees of importance.

The fifth chapter analyses and compares the application of multiple regression models in airline ticket revenue through the combination of theory and practice. It concludes that different airlines need to establish different models and analyse their influencing factors, so as to formulate different company strategies.

Chapter 2: Multiple linear regression analysis

The development of a business' economy is always closely related to certain economic variables. Most of the traditional pricing methods of passenger ticket revenue are based on relevant experience to analyse the close degree of the factors and factors that affect the development of the aviation industry. The main object of regression analysis is the statistical relationship among objective variables. The regression model is used mainly to reveal the quantitative relationship between related variables. We should set the dependent variable y according to the purpose of the study, and then select the variables that have statistical relations with y as independent variables to determine the causal or association relationship between various factors from a qualitative perspective. The key to the correct choice of variables lies in whether the researchers can correctly grasp the economic connotation of the economic activities they are studying, which requires the researchers to have enough knowledge of the economic problems and their backgrounds studied.

Linear regression based on least squares method is the classical regression model. The multivariate linear regression model is a model that explains dependent variables by two or more than two explanatory variables. There is often more than one factor that affects a variable in practical problems, which is difficult to judge directly with the help of graphs, so multivariate data analysis is needed.

The regression analysis method was put forward by Gauss in the early 19th century (1809). There are several important stages in the establishment and analysis of regression models of practical problems. According to specific (socio-economic) problems > setting up index variables > collecting and sorting out data, > constructing theoretical model > estimating model parameters > model testing > model application (economic factor analysis, economic variable control, economic decision-making and prediction). In the above steps, if the model test fails to pass the inspection, it is necessary to revise the data or indicators from the second step (setting up index variables) again until the model is acceptable.

This chapter introduces the multiple linear regression model applied to airline ticket analysis, focusing on the basic assumptions of the multivariate linear regression models, the estimation of unknown parameters in data analysis, nature and significance test.

In the fourth chapter we will analyse the relationship between the ticket revenue of a company with the influencing factors, using the multivariate regression model.

2.1 Mathematical Model

Let y be dependent variable, X_1, X_2……X_p is P the explanatory variable for y, then

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \quad p \geq 2, \qquad (2\text{-}1)$$

It is called multiple linear regression model [15], which$\varepsilon \sim N(0,\sigma^2)$, $\beta_0,\beta_1,\ldots\ldots,\beta_p$ and σ are unknown numbers, β_0 as a regression constant. $\beta_1,\ldots\ldots,\beta_p$ are as regression parameters. A random disturbance term ε is introduced into the model. In order to reflect the unknown factors and individual factors that affect the explanatory variables, representing the data observation error.

Set vectors （$x_{i1},x_{i2},\ldots,x_{ip},y_i$）, $i=1,2,\ldots,n,$ are （ X_1， X_2,……,X_p, Y）n independent observation of the sample value, then the model (2-1) can be expressed as

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i, i=1,2,\ldots,n, \tag{2-2}$$

Which $\varepsilon_i \sim N(0,\sigma^2)$ is independent and identically distributed.

For simplicity, matrix models are often used to represent models (2-2), let

$$y=\begin{pmatrix} y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix},\beta=\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_p\end{pmatrix},\mathrm{X}=\begin{pmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ 1 & x_{31} & \cdots & x_{3p}\\ \vdots & \vdots & \cdots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{pmatrix},\varepsilon=\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}, \tag{2-3}$$

Then the multivariate linear model (2-3) can be expressed as

$$\mathrm{Y}=\mathrm{X}\beta+\varepsilon, \tag{2-4}$$

Which Y is the n-dimensional vector composed of dependent variables, X is $n\times(p+1)$ order matrix, β is $p+1$ dimensional vector, ε is n-dimensional vector, and satisfies

$$E(\varepsilon)=0, Var(\varepsilon)=\sigma^2. \tag{2-5}$$

2.2 Basic Assumptions of the Model

The linear model places very strong constraints and assumptions on the relationship in the data. We assess whether the assumptions are reasonable mainly through analysis of residuals from the model.

In order to estimate the parameters of the model, the regression equation (2-4) should satisfy the following basic assumptions [16].

1. Explanatory variables $x_1,x_2,\ldots,x_p$ are deterministic variables, not random variables, and requires $\mathrm{rank}(X)=p+1<n$. The purpose of this hypothesis is to guarantee independent variables $x_1,x_2,\ldots,x_p$ with no correlation between them, that is, there is no multicollinearity, and the sample capacity (i.e. the number of observed values) should be greater than the number of independent variables. The matrix X is a full rank matrix.

2. The random error term satisfies Gauss Markov curve.

$$E(\varepsilon_i)=0, \qquad i=1,2,\ldots,n$$

and

$$cov(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2, i=j\\ 0, i\neq j\end{cases}\ (i,j=1,2,\ldots,n), \tag{2-6}$$

$E(\varepsilon_i) = 0$ is required that there is no systematic error between the observed values, and the mean value of the error is zero. The covariance is zero, indicating that the different samples are uncorrelated (independent of the assumption of normal distribution). The sequence is irrelevant and has the same accuracy.

3. The assumption that the random error term satisfies normal distribution.

$$\varepsilon_i \sim N(0, \sigma^2), i = 1,2, \dots, n, \quad \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \text{ are i.id.(independently and identically distributed)} \qquad (2\text{-}7)$$

The matrix form is

$$\varepsilon \sim N(0, \sigma^2 I_n).$$

Because the linear combination of random variables with independent normal distribution still obeys normal distribution, the expectation and variance vectors of regression model (2-4) can be obtained as

$$\mathrm{E(Y)} = \mathrm{X}\beta, \qquad (2\text{-}8)$$

$$\mathrm{Var(Y)} = \sigma^2 I_n, \qquad (2\text{-}9)$$

thus $\mathrm{Y} \sim N(\mathrm{X}\beta, \sigma^2 I_n)$.

2.3 Parameter Estimation

With the assumed linear model, how does one estimate the parameters of the model? Different criteria or requirements will get different conclusions. There are three kinds of methods for estimation of parameter β in multivariate regression model (2-3). Generally, least square method is used in classical model, and maximum likelihood method and moment estimation method are used in non classical model.

The least squares method is to find parameters $\beta_0, \beta_1, \dots \dots, \beta_p$ estimated value $\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}$. to minimise the sum of squares of deviations [17].

$$\mathrm{Q}(\beta_0, \beta_1, \dots \dots, \beta_p) = \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \dots - \widehat{\beta_p} x_{ip})^2$$

$$= \min_{\beta_0, \beta_1, \dots\dots, \beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2 \qquad (2\text{-}10)$$

Then get $\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}$, which are the least squares estimate of the unknown parameters $\beta_0, \beta_1, \dots \dots, \beta_p$. (2-4) is expressed in matrix form as

$$\mathrm{Q}(\beta) = (y - X\beta)^T (y - X\beta),$$

Where T is a transpose in the matrix, so the empirical regression equation can be obtained.

$$\hat{Y} = \hat{\beta} X. \qquad (2\text{-}11)$$

call

$$\hat{\varepsilon} = y - \hat{Y} = y - \hat{\beta} X$$

It is a residual variable.

The ordinary least squares estimator of variance of random error items in multivariate linear regression model is

$$\hat{\sigma} = \sqrt[2]{\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p-1}}, \quad (2\text{-}12)$$

It is an unbiased estimate.

The least squares method is only one of the most widely used parameter estimation methods, and it is also the oldest and the most classical method. Under the condition of satisfying the basic assumption, the ordinary least squares estimator of parameter β is linear (the estimators are linear combinations of the observed variables) and unbiased (the mathematical expectation of observations is equal to the true value of the estimator) and the validity (the variance of the estimator is the smallest among all linear unbiased estimators).

2.4 Significance Test

When the parameters of multiple linear regression models are estimated, that is to say, after the sample regression function is calculated, the sample regression function needs further statistical testing to determine the reliability of the estimation. There are two or three kinds of tests, one is the significance test of the regression coefficient, that is, to test the aleatory variable X_i whether the coefficient is 0. The second test is a significant test of the regression equation, that is, to test whether the data set is suitable for linear regression. The third test is a goodness of fit test to measure the degree of regression fitting.

2.4.1 Significance test of regression equation----F test

We know that the t- test of regression parameters check each explanatory variable x_j whether the corresponding variable y is significant, but we also need to check whether all the explanatory variables are united, and whether the corresponding variable y is also significant. This is the f- test to be studied in this section.

The method to test the significance of the overall linear regression equation is called F test, which is to test the significance of regression equation directly from the regression effect according to the square sum decomposition formula [18].

The sum of squares decomposition is

$$\sum_{i=1}^{n}(y-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}-\bar{y})^2 + \sum_{i=1}^{n}(\bar{y}-\hat{y})^2, \quad (2\text{-}13)$$

$\sum_{i=1}^{n}(y-\bar{y})^2$ is the total sum of squares, the abbreviation is SST (Sum of Squares for Total). $\sum_{i=1}^{n}(\hat{y}-\bar{y})^2$ is the sum of squares, the abbreviation is SSR (Sum of Squares

for Regression). $\sum_{i=1}^{n}(\bar{y}-\hat{y})^2$ is the sum of squares for residuals, the abbreviation is SSE (Sum of Squares for Error).

The total sum of squares can reflect the uncertainty of the y of the dependent variable, that is, the degree of fluctuation. After establishing the linear regression of Y and X, the total square sum (SST) can be decomposed into two parts: the sum of squares of regression (SSR) and the sum of squares of residuals (SSE). Here, the regression squared sum (SSR) is determined by the regression equation, which is related to X of independent variables, and is caused by the fluctuation of X. While the residual sum of squares of residuals (SSE) is a wave that can not be explained by X, which is caused by factors beyond x control.

In a regression equation, the part that can be explained by the independent variable x is SSR, and the part that can not be explained by the independent variable is called SSE. Thus, we can get that the larger the regression squared sum (SSR), the more the X can explain, the better the effect of regression.

Constructing F test statistics

$$\mathrm{F}=\frac{SSR}{SSE/(n-2)} \tag{2-14}$$

Under the assumption that this is the normal distribution, and the original hypothesis H_0 is established, F was subordinate to freedom degree of $(1,n-2)$ F distribution. When F is greater than the critical value $F_\alpha(1,n-2)$, the original hypothesis is rejected, it shows that the regression equation is significant, and X has a significant linear relationship with Y.

2.4.2 Significance test of regression parameters---T test

As we all know, the regression equation is significant, that is, F test does not show that each independent variable X has a very significant impact on Y. Because collecting data in real life is a very labor-intensive and financial matter, when collecting data indicators, we should try to delete those variables that do not affect the dependent variable Y, so as to establish a relatively simple regression equation, so we need to test the significance of each variable. In statistical inference, t test is mainly used to determine the significance of regression coefficient [19,20].

The original hypothesis of t- test is

$$H_0:\beta_0=0, \tag{2-15}$$

That is, assuming that the parameter is not significant. So the opposite hypothesis is

$$H_1:\beta_0\neq 0. \tag{2-16}$$

The purpose of the significance test of regression coefficient is to test whether the influence of independent variable x on dependent variable y is significant. If the original hypothesis is established, there is no linear relationship between Y and X.

Generally, it can be judged by the two parameters of T and P, because when judging by T value, it will be more difficult to judge according to the number of

variables. P value is a probability value, which is usually judged by significance test. When the probability value P is smaller than or equal to the significance level α, the coefficient of the independent variable is significantly non-zero, the independent variable has a significant effect on the dependent variable, and the original hypothesis H_0 should be rejected. That is to say, the independent variable X has a linear effect on the dependent variable Y; otherwise, there is no significant difference between the coefficient and 0 (essentially equal to 0); the independent variable has no significant effect on the dependent variable, and the original hypothesis H_0 should be accepted. That is to say that the independent variable x has no linear effect on the dependent variable Y. It should be deleted from the equation and reestimated the equation. However, only the least significant (with the highest corresponding probability) can be deleted at one time. Delete one at a time until all are significant

2.4.3 Significance test of linear regression equation

As the multiple linear regression equations discuss the linear relationship between the independent variable X and dependent variable Y, we can use the correlation coefficient between X and Y to test the significance of the regression equation.

Suppose $x_{i1}, x_{i2}, \dots, x_{ip}, y_i, i = 1,2, \dots, n$ are the N observations of variables X and Y, let [20]

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt[2]{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2-17}$$

be the correlation coefficient between X and Y. Numerical r is the degree of closeness linear relation between X and Y. The range of values is $|r| \leq 1$. When $r = 1$, there is a completely positive correlation between X and Y. When $r = -1$, there is a complete negative correlation between X and Y. When $r = 0$, there is no correlation between X and Y. In the statistical analysis, the above three cases are generally hard to meet, and most of them are $0 < |r| < 1$. The closer the r to the 1, the closer the linear relationship between X and Y is, the better our regression results.

In section 2.3.2, we introduce the significance of regression squared sum (SSR) and the sum of squares of residuals (SSE). If the sum of the sum of squares of regression squares (SSR) in the total square sum (SST) is greater, the effect of linear regression is better, that is, the goodness of fit of regression models and sample observations is better; if the sum of residual squares and SSE is larger, That is to say, the goodness of fit between regression model and sample value is worse. The ratio of regression squared sum (SSR) to total square sum (SST) is defined as the coefficient of sample determination.

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \tag{2-18}$$

Coefficient of determination r^2 is a relative index used to judge the goodness of fit between a regression model and a sample. It reflects the ratio of independent variables

in the fluctuation of dependent variables. The value of r^2 is always between 0 and 1. The closer the value of r^2 gets to zero, the worse the fitting of the regression model and the sample observations indicates that the selected variable X is not sufficient for the interpretation of Y. The closer the value of r^2 gets to 1, the better the regression model fits with the sample observations. This indicates that the independent variable X can explain most of Y's information.

In general, we divided the strength of the two variables into the following categories: when r^2 greater than 0.8, it is considered highly correlated. When the r^2 value ranges from 0.5 to 0.8, it is considered to be moderately correlated. When the r^2 value ranges from 0.3 to 0.5, it is regarded as low correlation. When r^2 less than 0.3, it indicates that the correlation between two variables is extremely weak and can be regarded as irrelevant in practical application.

2.5 Residual Analysis

After considering the T test and F test, the multivariate linear equation can only show that the linear relationship between variables X and Y is significant, and the regression equation is effective, but it cannot guarantee that the data fit well. Only when the data from the residuals of the model meets the relevant assumptions, can we can conclude that the regression model is useful. Before making use of this regression model for analysis and prediction, we should plot the residuals to diagnose whether the model meets the basic assumptions, so as to further modify and adjust the model [24-25].

The residual value, $e_i = y_i - \hat{y}_i$, represents the difference between the actual observed value and the regression value obtained through the regression model, and the residual term e_i can also be regarded as an estimate value of error ε_i. The residual graph is based on the independent variable X as the horizontal axis and residual e_i as a vertical axis, the corresponding points are drawn in the Cartesian coordinate system, so as to make some analysis of the data quality. Figure 2.1 gives some common residual plots, according to the difference of residual plots, the data are also different.

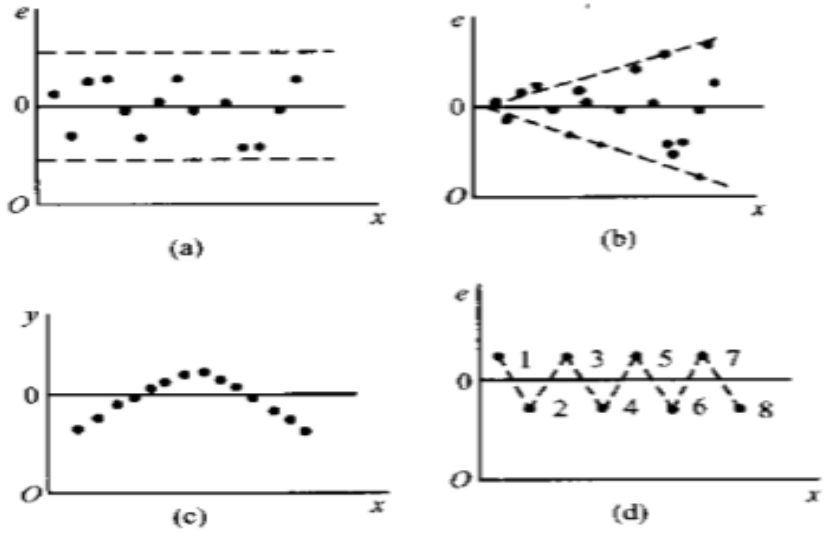


Fig. 2.1 residual graph

In general, when a regression model satisfies all the basic assumptions, all residual points should be randomly dropped near $e = 0$. In general, it satisfies the shape presented in Figure 2.1 (a). If the residual graph appears in Figure 2.1 (b), it shows that the variance of the observed value of Y is not the same, but increases with the increase of X. In this case, we can process logarithmic data and eliminate the effect of heteroscedasticity. If the residual graph appears in Figure 2.1 (c), it shows that there is a nonlinear relationship between X and Y. It may be curvilinear and need to consider other methods to fit the sample. Figure 2.1 (d) is a well-known cobweb phenomenon, indicating that X has strong autocorrelation.

2.6 Forecasting

The purpose of establishing regression equations is not only to analyse the influencing factors, but also to apply them. It is often used for prediction.

Prediction is divided into single value prediction and interval prediction. Single value prediction is to use single value as the prediction variable of dependent variable. For example, when we study the relationship between independent variable x and dependent variable y, we can get the sample value of X first, and use regression model $\hat{Y} = \hat{\beta}X$, that is the predicted single value of Y.

Because single value prediction $\hat{Y}$ is only an approximate value, only knowing this value is of a little significance. What we want to know more is the accuracy of prediction. For solving this problem, we need to do interval prediction, that is, a range of predictive values for Y [21-23]. A given range is relatively more reliable than a given number. For a given significance level α, when the actual value of Y for a given X is contained with the interval (T_1, T_2) with probability $1 - \alpha$. This interval is the prediction interval, expressed by formula

$$P(T_1 < Y < T_2) = 1 - \alpha. \tag{2-19}$$

Under normal circumstances, the shorter the confidence interval, the higher the accuracy of prediction is. In order to improve the prediction accuracy, data acquisition can be done, that is, the larger the sample size n, the better the data acquisition. But the collection data $x_1, x_2, \ldots, x_n$ should not be too concentrated, The X value of prediction can not be deviated from the sample mean $\bar{X}$ too far when forecasting, and the prediction result will not be good when it is too far.

2.7 Data Preprocessing

In multivariate linear regression analysis, because there are many independent variables involved, the difference of the units of the independent variables will bring some complications to the analysis. In addition, because of the large amount of data

involved in multivariate regression, rounding error may cause some problems, so we need to deal with the raw data and avoid this situation as much as possible.

There are two reasons for rounding errors. One is that the data level of the independent variables is very different. For example, 80000 and 0.5, which are very different. The other is that there is a certain correlation between matrix X. Therefore, we need to process some data to reduce the above two situations. There are two methods used - centralisation and standardisation.

Centralisation is to transfer the origin of coordinates to the sample center, that is, coordinate transformation.

$$x'_{ij} = x_{ij} - \overline{x_j}, i = 1,2,\dots,n; j = 1,2,\dots,p \quad (2\text{-}20)$$

$$y'_i = y_i - \bar{y}, i = 1,2,\dots,n; \quad (2\text{-}21)$$

The empirical regression equation can be derived from the original

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \cdots + \widehat{\beta_p} x_{ip} \quad (2\text{-}22)$$

Convert to

$$\widehat{y_i'} = \widehat{\beta_1} x'_{i1} + \cdots + \widehat{\beta_p} x'_{ip}. \quad (2\text{-}23)$$

Type (2-23) is the centralisation adjusted regression equation. The constant term of the centralised empirical regression equation is 0, but the estimation value of other unknown parameters has not changed since the translation transformation of the coordinate system can only change the intercept between the straight line and the axis, but it does not change the slope.

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_0} - \widehat{\beta_1}\overline{x_1} - \cdots - \widehat{\beta_p}\overline{x_p}. \quad (2\text{-}24)$$

Standardisation is a method of further processing data on the basis of centralisation. Most of the units used in the X are very different. The size of the data is too large to be placed on the same standard for comparison. In order to eliminate the effects caused by magnitude differences and dimensional differences, data need to be standardized.

The formula for standardization of sample data is

$$x_{ij}^* = \frac{x_{ij} - \overline{x_j}}{s}, s = \sqrt[2]{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}, i = 1,2,\dots,n; j = 1,2,\dots,p \quad (2\text{-}25)$$

$$y_i^* = \frac{y_j - \bar{y}}{s}, s = \sqrt[2]{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}, i = 1,2,\dots,n \quad (2\text{-}26)$$

The regression equation of the normalized sample data $(x_{i1}^*, x_{i2}^*, \dots, x_{ip}^*, y_i^*)$ is obtained by using the least square method, denoted as

$$\widehat{y_i^*} = \widehat{\beta_1}^* x_{i1}^* + \cdots + \widehat{\beta_p}^* x_{ip}^* \quad (2\text{-}27)$$

Where $\widehat{\beta_1}^*, \widehat{\beta_2}^*, \dots, \widehat{\beta_p}^*$ is the standardized regression coefficient of Y for independent variables $x_1, x_2, \dots, x_{p.}$ Because standardisation is based on centralisation, the regression constant is still zero. There is a relationship between standardised regression coefficient and ordinary least squares regression coefficient

$$\widehat{\beta_J}^* = \frac{\sqrt[2]{\sum_{i=1}^{n}(x_{ij}-\overline{x_J})^2}}{\sqrt[2]{\sum_{j=1}^{n}(y_j-\bar{y})^2}}\widehat{\beta}_J. \quad (2\text{-}28)$$

Chapter 3: Independent Variable Selection and Stepwise Regression

When establishing a regression model, the first problem is how to determine the regression independent variables. If some important variables are omitted, the fitting effect of the regression equation will not be good. If we consider too many independent variables, some variables may not be relevant to the research questions, and there could be a large degree of information overlap for some useful variables. At the same time, too many variables in the regression equation will make it difficult to collect data, operate the system and explain the model and also the stability of regression equation will also be poor, which will directly affect the application of regression equation.

How to simplify the model and choose the explanatory variables well is an important problem for regression analysis.

3.1 Selection criteria for independent variables

Consider m possible variables of interest, $x_1, x_2, \dots, x_m$. Because each variable has two parts: participation and non participation, there are $2^m - 1$ possible regression equations that are related to these m independent variables. The minus one represents the condition of subtracting only constant terms in the model, that is, the model without independent variables. In terms of the number of independent variables p that can be included in the model, there are m+1 different cases from 0 to m. There are C_m^p different combinations of independent variables P. Therefore, the number of all selected models is 2^m. This section will focus on one approach to selecting a regression set from these possible regression sets and give the criteria for measuring the models.

In the second chapter, from the point of view of the goodness of fit between sample and model, we evaluate the model by considering the sum of squares of residuals (SSE), or the r^2, but these two methods have obvious deficiencies.

We assume that the sum of squares of residuals of a regression model with p independent variables is SSE_p. The determination coefficient is

$$r_p{}^2 = 1 - \frac{SSE_p}{\mathrm{SST}}.$$

Next we introduce a new independent variable x_{p+1}. The sum of squares of residuals also becomes SSE_{p+1}, the determination coefficient becomes

$$r_{p+1}{}^2 = 1 - \frac{SSE_{p+1}}{\mathrm{SST}}.$$

According to the principle of least squares estimation, the sum of squares of residuals will be reduced when the independent variables are added, so that [20]

$$SSE_{p+1} \le SSE_p, r_{p+1}{}^2 \ge r_p{}^2. \quad (3\text{-}1)$$

As shown in Equation (3-1), when the subset of independent variables keeps increasing, the sum of squares of residual error decreases correspondingly, and the coefficient of determination increases correspondingly. If independent variables are selected according to the above criteria, all the collected variables would be used for regression. Due to the existence of multicollinearity among variables, the regression model would be unstable and the estimated error of the model would increase. Therefore, the residual sum of squares and the determination coefficient are not suitable for the criteria of selecting variables. There are two common criteria for information selection.

Criterion I: the coefficient of freedom adjustment decides the maximum coefficient. When the model independent variable increases, the decision coefficient will increase, and the cost is the reduction of residual degrees of freedom. The degree of freedom represents the reliability of estimation and prediction. The smaller the degree of freedom is, the smaller the reliability is. The greater the degree of freedom is, the greater the dependability is. In order to overcome the disadvantage of decision coefficient, we can modify the r^2 appropriately. Only when the independent variables which have influence on the dependent variables are added to the model, the modified decision coefficient will increase.

The revised coefficient of determination [20,25].

$$(r')^2 = 1 - \frac{n-1}{n-p-1}(1-r^2), \tag{3-2}$$

It can be seen that $(r')^2 \leq r^2$, the increase of independent variables will not lead to increase $(r')^2$. By the above formula, though $1-r^2$ decreases with the increase of independent variables, but because of the coefficient$1-\frac{n-1}{n-p-1}$ control, making $(r')^2$ does not necessarily increase with the increase of independent variables. When the contribution of increased independent variables to regression is small，$(r')^2$ may also decrease. So we can take the revised determination coefficient $(r')^2$ as a criterion for judging regression subset.

Criterion II: Akaike information criterion, AIC reaches the minimum [30,31]. In 1974, Japanese statistician Akaike proposed a more general model selection criterion based on the principle of maximum likelihood estimation. It was called the AIC criterion (Akaike Information Criterion). AIC criterion, which can be used not only for the selection of independent variables of regression equations, but also for the autoregressive model of time series analysis.

In general, suppose the likelihood function of the model is $L(\theta, x)$, the dimension of θ is p, x is a random sample, thus

$$AIC = -2lnL(\widehat{\theta_L}, x) + 2p, \tag{3-3}$$

Which $\widehat{\theta_L}$ is the maximum likelihood estimation of θ, p is the number of unknown parameters (independent variables). We know that when we do multiple linear regression, residual items ε_i is required to obey normal distribution, that is $\varepsilon_i \sim N(0, \sigma^2)$. Therefore,

$$lnL_{max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_L^2) - \frac{1}{2\hat{\sigma}_L^2}SSE. \quad (3\text{-}4)$$

We substitute $\hat{\sigma}_L^2 = \frac{SSE}{n}$ into (3-4) and get

$$lnL_{max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\frac{SSE}{n}) - \frac{n}{2}. \quad (3\text{-}5)$$

Substituting (3-5) in (3-3), the number of unknown parameters is $p + 2$, get the regression model

$$AIC = nln(SSE) + 2p. \quad (3\text{-}6)$$

In the modeling process of regression analysis, AIC is calculated for each regression subset, and the AIC minimum model corresponds to the relative optimal model.

Criterion III: Minimise the statistic C_p [32]. In 1964, Mallows propose a statistic C_p which can be used to select independent variables from the perspective of forecasting. Considering n sample points, we constructed the statistic C_p.

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p$$

$$= (n - m - 1)\frac{SSE_p}{SSE_m} - n + 2p, \quad (3\text{-}7)$$

Which $\hat{\sigma}^2 = \frac{1}{(n-m-1)}SSE_m$ is the unbiased estimation of model σ^2.

The criteria for selecting the variable $C_p$is: The regression equation corresponding to the subset of independent variable that minimises C_p is the relative optimal regression equation.

The above three criteria are the criteria for judging the set of independent variables from three angles. When the independent variables are selected, the corresponding criteria can be selected according to their own modeling needs.

3.2 Stepwise Regression

In regression modeling, not all independent variables X have significant effects on dependent variable y. Therefore, selecting independent variables that have significant effects on dependent variables has always been the focus of scholars' research. All variables can form $2^m - 1$ equations. When there are many independent variables m, it is difficult to get the model by experiment, and then choose the best model through the above three criteria.

Now the stepwise regression methods commonly used include progressive, backward and stepwise regression methods. A partial F test is used to increase or decrease variables.

The idea of the progressive method is to change from less to more, and each time a new independent variable is introduced until no variable is introduced.

The m linear regression equation is established for the dependent variable y, and the F test value of the m linear regression equation is calculated. The largest one is [20,26].

$$F_j^1 = \max\{F_1^1, F_2^1, \dots, F_m^1\}. \quad (3\text{-}8)$$

Given significance level α, if $F_j^1 > F_\alpha(1, n-2)$ exist, then select the maximum F value corresponding to the independent variable as the first dependent variable. Next, the other M-1 variables and corresponding y are established into m-1 bivariate linear equations and their F test values are calculated. Denote as $F_2^2, F_3^2, \dots, F_m^2$, and choose the largest as

$$F_j^2 = \max\{F_2^2, F_3^2, \dots, F_m^2\}. \quad (3\text{-}9)$$

If $F_j^2 > F_\alpha(1, n-3)$, then select the maximum F value corresponding to the independent variable as second dependent variables. This is repeated until the F test values of all the independent variables not introduced into the equation is less than $F_\alpha(1, n-p-1)$. When it is finished, the regression equation obtained is the relative optimal equation. The critical value $F_\alpha(1, n-p-1)$ obtained in each test is related to the number of independent variables p. When we use software to calculate, we can focus directly on significance P value as a test criterion.

In contrast to the forward method, the backward method first establishes an m degree equation of y about x, then, selects an independent variable with the smallest influence to eliminate. That is, the one with the smallest F-test value in the regression coefficient of the independent variable is eliminated.

Then, the independent variable with the least influence is selected, that is, the F test value in the regression coefficient of the independent variable is eliminated

Suppose F test for m regression coefficient and record F value as

$$F_j^m = \min\{F_1^m, F_2^m, \dots, F_m^m\}. \quad (3\text{-}10)$$

Given significance level α, if exist $F_j^m < F_\alpha(1, n-m-1)$, the corresponding variables are removed from the regression equation. An m-1 degree equation for y about x is reestablished for the m-1 independent variables in its domain. The significance test of the regression coefficient is calculated and F_j^{m-1} is calculated as above. If exist

$$F_j^{m-1} < F_\alpha(1, n-(m-1)-1),$$

Then, the corresponding variables are eliminated, continue and until the F test values of the remaining P independent variables in the regression equation are all larger than the critical value $F_\alpha(1, n-p-1)$. That is, there is no independent variable that can be eliminated before stopping. At this time, the relative optimal regression model is obtained.

The disadvantage of the forward and backward methods is that once introduced or eliminated, it will no longer be changed. Since an independent variable may be significant at the beginning, it will become insignificant after introducing other variables, and there is no way to eliminate this variable when using the forward method. In the beginning, all variables are introduced into the regression equation, and the computation is very large. If some independent variables are not very important, they can not be introduced at the beginning, so as to reduce some computation. Once again, it is no longer in the regression equation as the forward method. However, because of the different variables in the model, the influence of dependent variables will be different. Both considerations are very incomplete. Therefore, a stepwise regression method is introduced to overcome their shortcomings.

Stepwise regression is the most widely used method at present. Its basic idea is to introduce variables one by one. After each variable is introduced, all variables that have been selected in the regression model are tested. If the original variables are not significantly affected by the newly introduced variables, they will be kicked out. Therefore, every step needs F test to ensure that the whole regression model is significant after every change. This process is repeated until there is no significant independent variable to choose the regression equation, and there is no significant independent variable [27-29] in the regression equation.

Stepwise regression requires the introduction of independent variables and the elimination of independent variables, the significance level can not be the same, and to ensure that the significance level of the introduction of independent variables is less than the significance level of the independent variables, that is $\alpha_{in} \geq \alpha_{\text{out}}$. Otherwise, it will fall into a dead loop.

Chapter 4: Empirical analysis

4.1 Data specification

The core of civil aviation ticket pricing lies in anticipation of the market. The airline administration can understand the basis of this through studying historical data. Using historical data and statistical models they can potentially predict the passenger rate and the fare level according to the company's internal flight information and market changes. Whether the civil aviation passenger ticket pricing is appropriate (competitive but still maximally profitable) can be reflected in the daily operation report data of the airline. The commonly used measurement indexes (and associated variable names) that we consider in this chapter are defined below.

1) Actual number of passengers (Total passenger number): refers to the number of passengers who purchase the tickets and actually take the flight. This data is collected in the airport departure system, that is, the check-in procedures at the airport, and the number of successful passengers.

2) Passenger rate (passenger rate): an index reflecting the proportion of the airlines' input in production. Production means manufacturing or mining or growing something for sale. The formula is: Passenger rate = the number of passengers booking / the number of seats. The number of seats in a flight is the number of seats that a flight can provide.

3) Average fare (average fare): refers to the average price of each passenger ticket from the total revenue of the flight. The formula is: average fare = revenue / number of people actually gone.

4) Fuel income (fuel income):revenue from surcharges charged by airlines to reflect changes in fuel prices.

5) Passenger kilometer (passenger km):the number of kilometers multiplied by the number of passenger.

6) Available seat kilometers (available seat km), which is the product of seats and flight mileage {mileage is a poor choice - things are measured in kilometers}. It is an index

reflecting the carrying capacity, Available seat kilometers = (available seats * segment distance).

7) Passenger transport revenue (passenger revenue) refers to the income of passengers, goods, mail, baggage, and the duties of carrying out charter or charter flights on the international and domestic routes that the airlines fly on. They are paid according to the stipulated tariff rates or the agreement rates between the international aviation industry or the freight rates issued by the International Civil Aviation Organisation.

Ticket revenue is one of the main data sources for airline revenue audits. In order to study the factors that affect airline ticket revenue, this analysis uses ticket revenue (average fare) as a dependent variable. We had 5 datasets available to us: the variables listed above were in all 5 datasets. The data was from the main office of an airline group in January 2019 and two subsidiary airlines for January 2018 and January 2019. Multiple regression analysis was conducted to predict average fare using the independent variables available seat kilometers, passenger rate, passenger kilometer, total passenger number, passenger revenue, fuel income, and airline revenue. All variables are scaled in the regression analysis so that the coefficients are easier to interpret.

We first study the data from the main office of the airline company. The data is from January, 2019. Means and Standard Deviations are given in the following table:

Airline main office: Summary Statistics for all variables

Variable	**Mean**	**SD**	**UNIT**
Average fare	898.13	403.11	China Yuan
Available seat km	142875.27	126194.95	number multiply seat-kilometer
Passenger rate	0.64	0.39	percentage
Passenger km	126467.26	109203.47	Passenger-kilometer
Total passenger number	96.27	62.62	headcount

Passenger revenue	74602.24	54566.73	China Yuan
Fuel Income	5338.02	14644.95	China Yuan
Airline revenue	69208.69	51384.39	China Yuan

Table 4.1-1: Summary statistics for all variables in the 'training dataset' used for the regression analysis.

For the regression analysis we employed the ‘step’ procedure for backwards elimination in R software[8-14]. This procedure chooses the model that minimises the AIC (Section 3.1). Average fare is the dependent variable; independent variables are: available seat kilometers, passenger rate, passenger kilometer, total passenger number, fuel income, and airline revenue. The data for the stepwise procedure was a randomly chosen ‘training’ set comprising of 75% of the data for the airline main office. Table 4.1-2 gives details of the model.The remaining 25% of the data was then used as a ‘test’ dataset to check the predictive capability of the model. Using the stepwise regression method, i.e. judging by the AIC, we find that a model with the six variables available seat kilometers, passenger rate, passenger kilometer, total passenger number, passenger revenue, fuel income, and airline revenue gives the most reliability by that measure (only passenger revenue is removed). Other measures of fit, e.g. Mallow's Cp, the Bayesian information criterion and the Adjusted R-square can also be used to judge the reliability of models: we compared this result with those in Figure 4.1-1. It can be seen that all results (i.e the model chosen to minimise (or maximise) the criterion) are the same (model '120').

Airline Main Office - Model Measures of Reliability

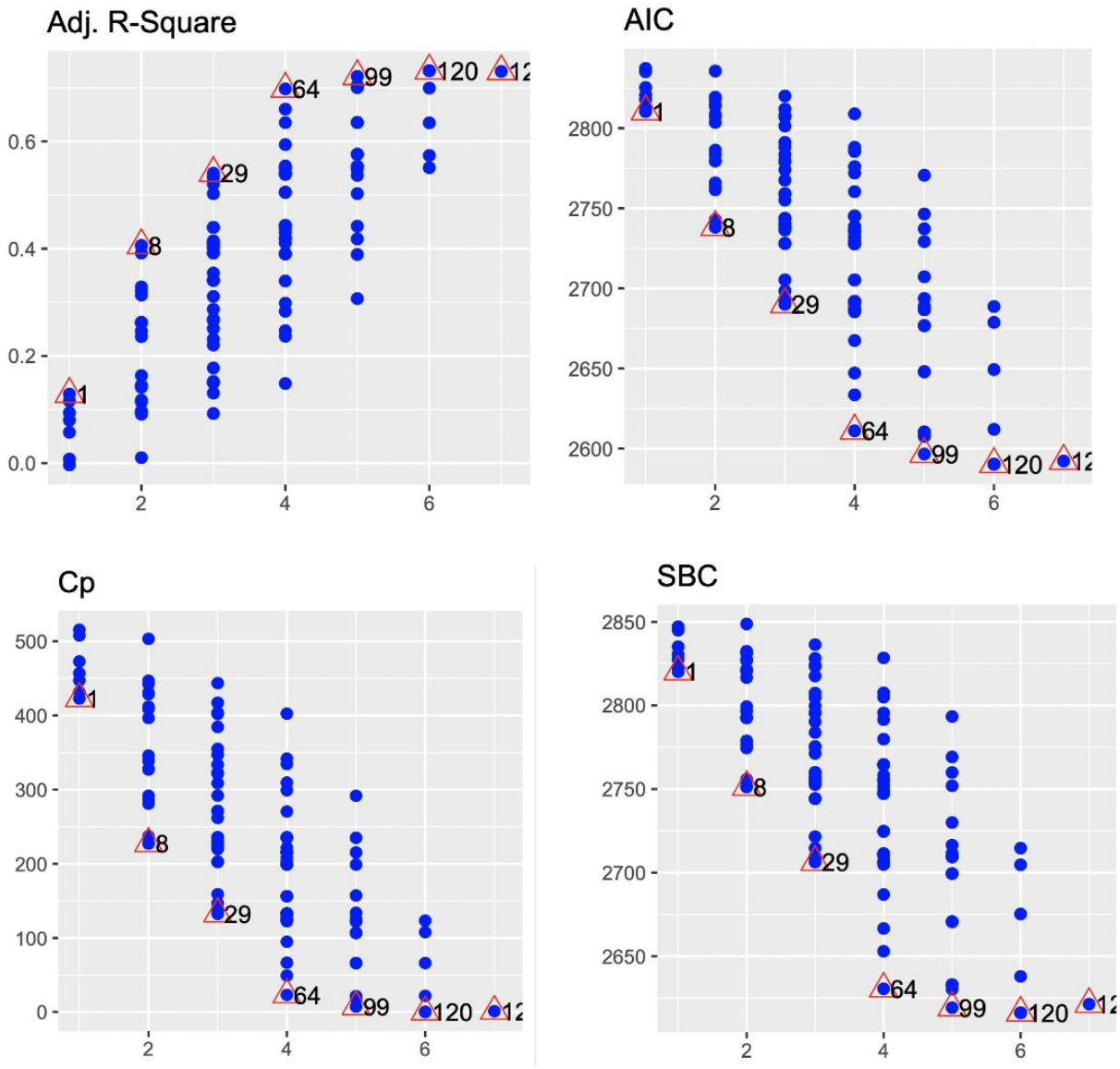


Figure 4.1-1:Measures of reliability of models

Airline Main Office Ticket Income Model

Predictor Variable [standardized]	Estimate	Std. Error	t value	Pr(>|t|)
available.seat.kilometers	1.4306	0.1284	11.1427	<.0001
passenger.rate	1.2805	0.27	4.7434	<.0001
passenger.kilometer	-1.4979	0.1438	-10.4165	<.0001
total.passenger.number	-2.7124	0.3302	-8.2155	<.0001
fuel.income	0.2898	0.0654	4.4295	<.0001
airline.revenue	1.5054	0.0878	17.1547	<.0001

Table 4.1-2: Model results from the stepwise procedure. Adj R^2 = .74. This is output from the model chosen with the stepwise procedure applied to the training set of data. Coefficient estimates are for scaled variables so that relative importance can more easily be recognised.

Airline Main Office Average Fare Model

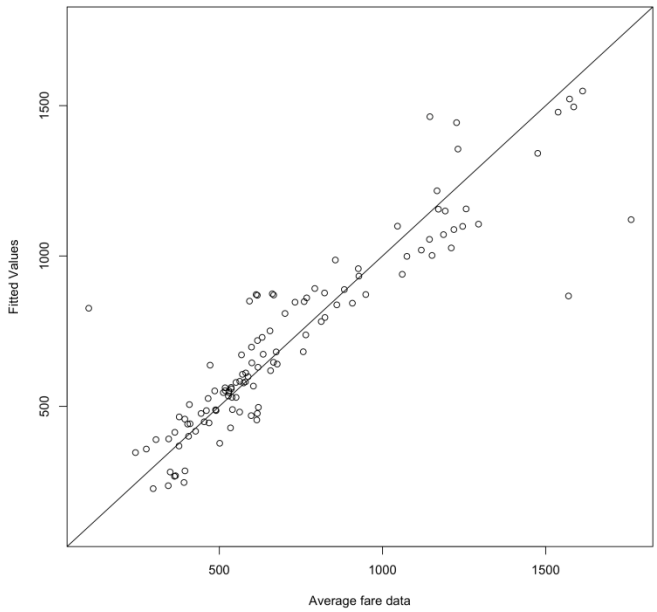


Figure 4.1-2 Average fares vs Fitted values, Main Office, January 2019 Training Dataset

Airline Main Office Average Fare Model Residuals

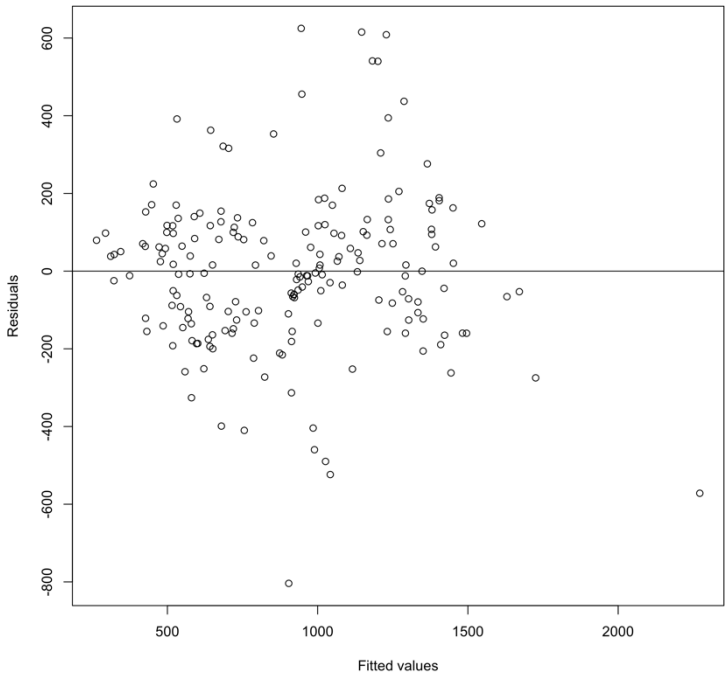


Figure 4.1-3 Fitted values vs Standardised residuals, Main Office, January 2019 Training Dataset

In Figure 4.1-2 we can see that although there is a good linear trend there are some points which are well off the curve. It may be that recalculating the model with these points removed would be helpful, but for this thesis we chose to use all data that weren't obvious errors. Figure 4.1-3 is the standardised residual scatter plot drawn by the model established for the Airline Main office in January 2019. We can see that since the residuals are scattered in the near zero that the model appears to satisfy the assumptions.

When the model from the training dataset is applied to the test dataset the fit is rather disappointing. The ratio of the root mean square error (RMSE) to the mean of the data is 24%.

The fitted values are plotted against the data in figure 4.1-4. There are no major outliers.

Airline Main Office Average Fare Predictions

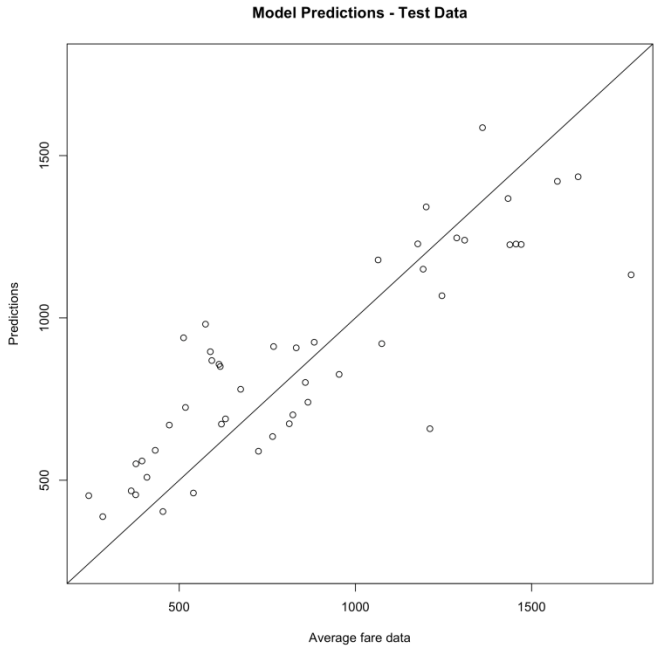


Figure 4.1-4 Average fares vs Predicted values, Main Office, January 2019 Test Dataset

In practical applications, even for the same airline, yearly data variation may affect the factors or the size of the ticket revenue. We are also interested in whether this model can be used for the two airlines subsidiaries (Airline A and Airline B). In the next section (4.2) we use R software to carry out prediction for the data of Airlines A, in January 2018 and 2019 using the model established for the main office. In section 4.3 we will apply the model to Airline B for January 2018 and 2019 data. It is hoped that a general model could be developed for the airline conglomerate.

4.2 Analysis of Airline A Ticket Income Factors

Means and Standard Deviations for this data are given in the following table:

Airline A January 2019 Summary Statistics for all variables

Variable n=115	**Mean**	**SD**	**UNIT**
average.fare	722.83	348.21	China Yuan
available.seat.kilometers	144749.57	126536.1	number multiply seat-kilometer
passenger.rate	0.65	0.41	percentage
passenger.kilometer	130249.6	105207.7	Passenger-kilometer
total.passenger.number	101.37	65.55	headcount
passenger.revenue	71703.35	48196.62	China Yuan
fuel.income	10569.66	19856.33	China Yuan
airline.revenue	61097.45	38901.89	China Yuan

Table 4.2-1: Airline A in January 2019- Summary statistics for all variables studied.

Using the model established from the main office data, we have passenger ticket income as the dependent variable, and available seat kilometers, passenger rate, passenger kilometer, total passenger number, passenger revenue, fuel income, and airline revenue as the independent variables. Coefficients for those variables are re-calculated for the predictions. We see that the predictions are generally a good fit to the data in Figure 4.2-1. The ratio of the root mean square error (RMSE) to the mean of the data is .21. Two of the variables, available seat kilometers and total passenger number, do not appear to have independent prediction capabilities for this model.

Airline A January 2019 Ticket Income Model

	Estimate	Std. Error	t value	Pr(>|t|)
available.seat.kilometers	-0.2097	0.244	-0.8596	0.3919
passenger.rate	-1.5648	0.6138	-2.5492	0.0122
passenger.kilometer	-0.83	0.2601	-3.1912	0.0019
total.passenger.number	0.889	0.7782	1.1423	0.2558
fuel.income	0.4331	0.1592	2.7199	0.0076
airline.revenue	1.1568	0.1027	11.2622	<.0001

Table 4.2-2: Results from the model established using the stepwise procedure the January 2019 data for Airline A. Adj R^2 = .82. Coefficient estimates are for scaled variables so that relative importance can more easily be recognised.

Residuals (Figure 4.2-2) do not suggest any departures from the assumptions.

Airline A January 2019 Ticket Income Model Predictions

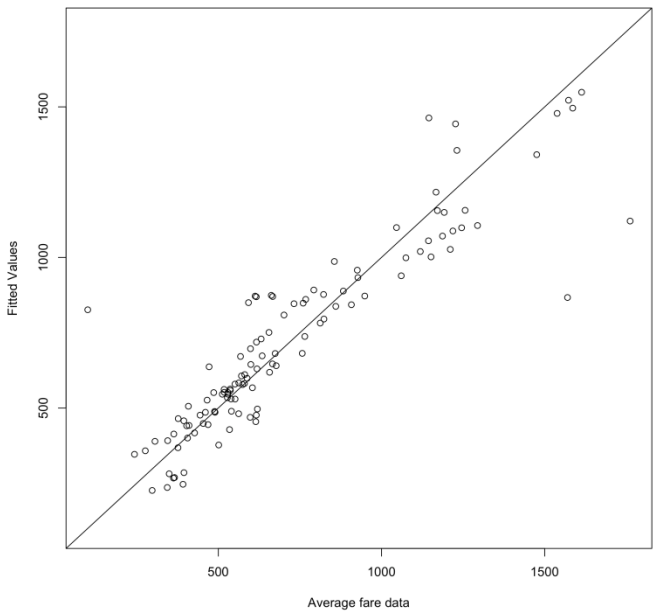


Figure 4.2-1: Average fares vs fitted values for the Airline A data for January 2019. Model coefficients have been recalculated for the variables

Airline A January 2019 Ticket Income Model Residuals

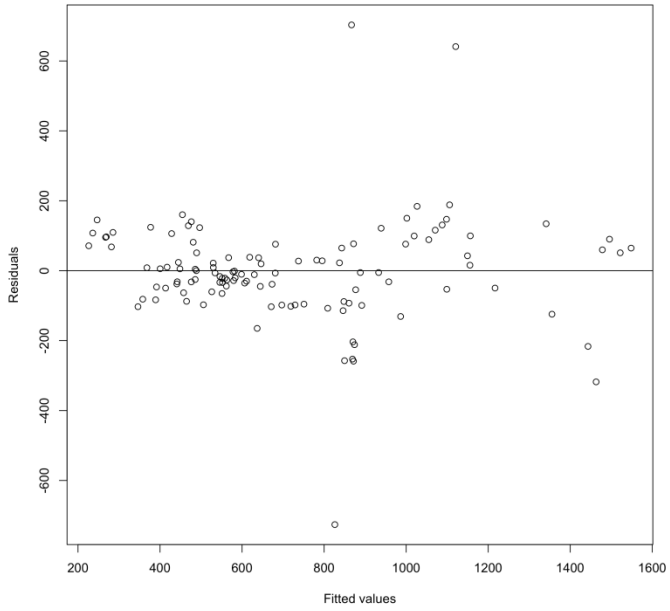


Figure 4.2-2: Fitted values vs Residuals for the Airline A data for January 2019.

In the preceding model coefficients have been re-calculated for the model using the 6 covariates. It is also possible to use the *predict* function in R to force the coefficients that were calculated for the original model for the main office. We show the plot for the fit for this model in Figure 4.2-3. The ratio of the root mean square error (RMSE) to the mean of the data is .28. The plots show that the fit is quite similar.

Airline A January 2019 Ticket Income Model Predictions from Original Model

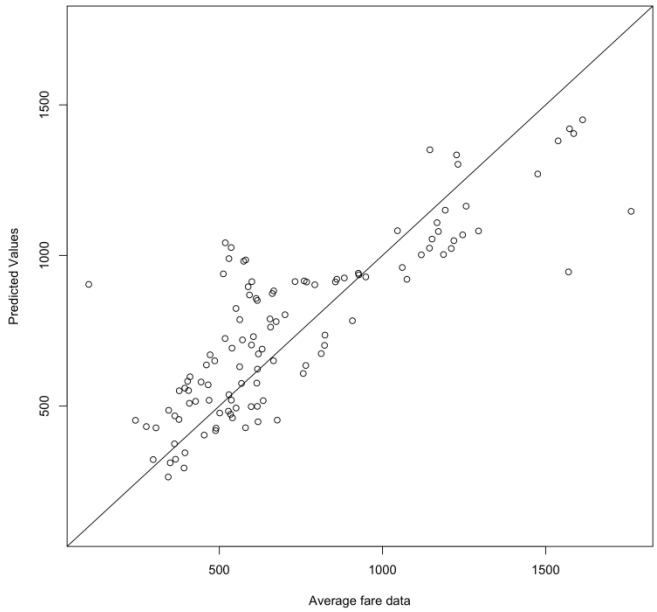


Figure 4.2-3: Average fares vs predicted values for the Airline A data for January 2019. Coefficients are as originally calculated for the Main Office.

Data is also available for January, 2018. Since it is known that yearly effects are common in the industry, we analyse this dataset separately.

Means and Standard Deviations for this data are given in the following table:

Airline A January 2018 Summary Statistics for all variables

Variable n=112	**Mean**	**SD**	**UNIT**
average.fare	588.92	122.99	China Yuan
available.seat.kilometers	144243.46	125443.35	number multiply seat-kilometer
passenger.rate	0.65	0.4	percentage
passenger.kilometer	130937.79	99513.15	Passenger-kilometer
total.passenger.number	100.9	64.53	headcount
passenger.revenue	58228.22	35285.47	China Yuan
fuel.income	9014.05	18717.75	China Yuan
airline.revenue	49146.01	20328.62	China Yuan

Table 4.2-3: Airline A in January 2018- Summary statistics for all variables studied.

Again using the model established from the main office data, we have average fare as the dependent variable, and available seat kilometers, passenger rate, passenger kilometer, total passenger number, fuel income, and airline revenue as the independent variables. Coefficients for those variables are re-calculated for the predictions. We see that the predictions are generally a good fit to the data in Figure 4.2. The ratio of the root mean square error (RMSE) to the mean of the data is .09. Residuals (Figure 4.2-4) do not suggest any departures from the assumptions despite the presence of some large points.

Airline A January 2018 Ticket Income Model

	Estimate	Std. Error	t value	Pr(>|t|)
available.seat.kilometers	-0.3641	0.3348	-1.0874	0.2793
passenger.rate	-5.9592	0.8887	-6.7058	<.0001
passenger.kilometer	-3.2109	0.3223	-9.9611	<.0001
total.passenger.number	5.3337	1.0987	4.8545	<.0001
fuel.income	1.3019	0.184	7.0742	<.0001
airline.revenue	1.7629	0.1489	11.8413	<.0001

Table 4.2-4: Results from the model established using the stepwise procedure for the January 2018 data for Airline A. Adj R^2 = .82. Coefficient estimates are for scaled variables so that relative importance can more easily be recognised.

Airline A January 2018 Ticket Income Model Predictions

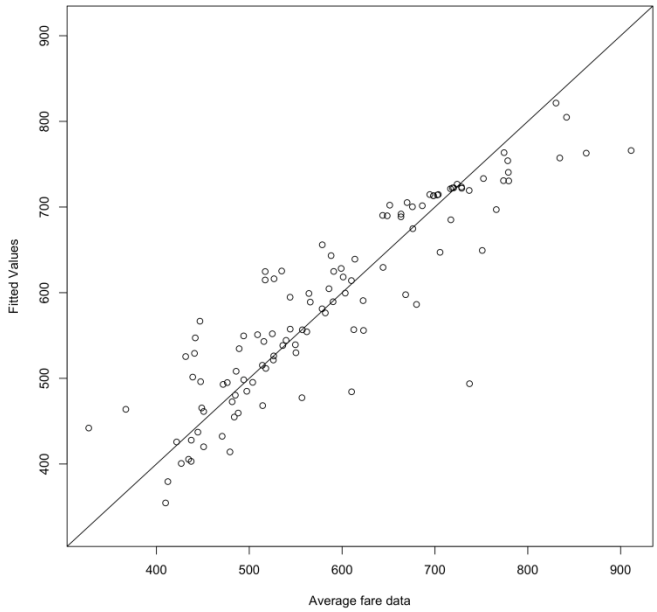


Figure 4.2-5: Average fares vs fitted values for the Airline A data for January 2018. Model coefficients have been recalculated for the variables

Airline A January 2018 Ticket Income Model Residuals

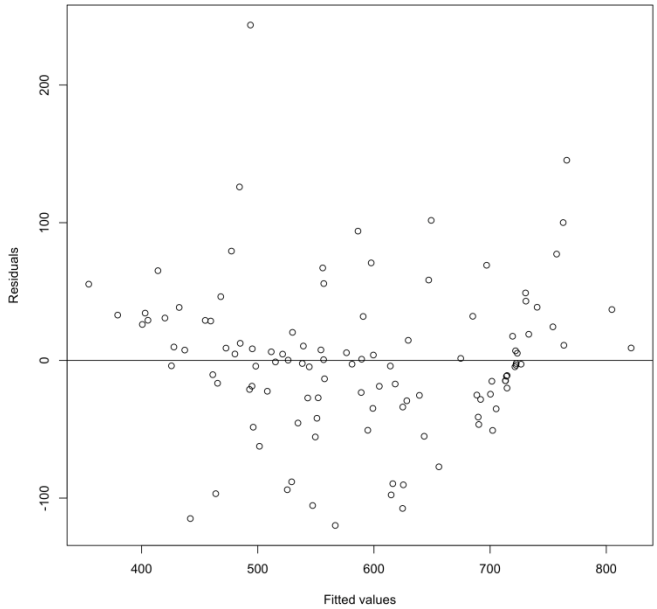


Figure 4.2-6: Fitted values vs Residuals for the Airline A data for January 2018.

Airline A January 2018 Ticket Income Model Predictions from Original Model

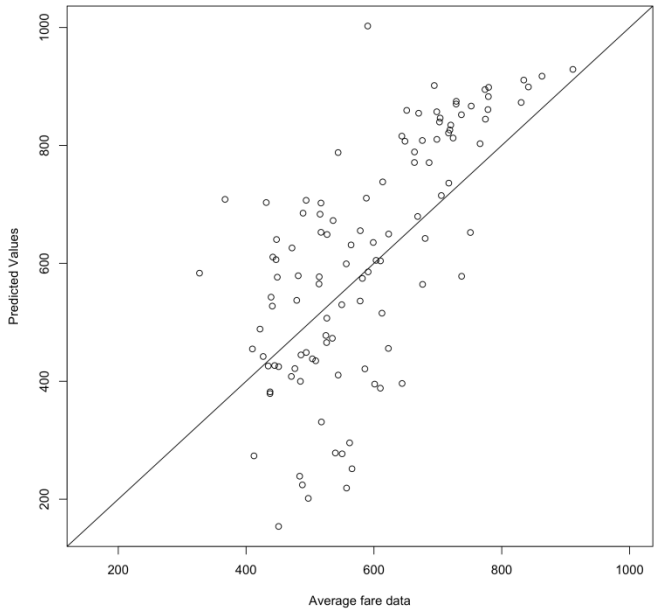


Figure 4.2-7: Average fares vs predicted values for the Airline A data for January 2018. Coefficients are as originally calculated for the Main Office.

In Table 4.2-4 coefficients were re-calculated for the model using the 6 covariates. We are also interested in how well the model with the original coefficients (from the main office) performs. The residuals and plots seem to be a reasonable fit (Figues 4.2-5 and 4.2-6). All variables have a significant contribution to the model except for available seat kilometers. Using the *predict* function in R, we apply the main office model to predict results for this dataset. We use those predictions to recalculate the performance of this model. The ratio of the root mean square error (RMSE) to the mean of the data is .25. Figure 4.2-7 shows the plot of the average fare vs the predictions. For this dataset we see quite a difference in this fit compared to the re-calculated model (Figure 4.2-5). Also notable is the change in RMSE of the fit from .09 to .25.

4.3 Analysis of Airline B Ticket Income Factors

Means and Standard Deviations for this data are given in the following table:

Airline B January 2019 Summary Statistics for all variables

Variable n=115	Mean	SD	UNIT
average.fare	1052.23	393.05	China Yuan
available.seat.kilometers	142522.1	126804.69	number multiply seat-kilometer
passenger.rate	0.62	0.38	percentage
passenger.kilometer	121975.76	113174.3	Passenger-kilometer
total.passenger.number	90.81	59.98	headcount
passenger.revenue	76674.36	60098.4	China Yuan
fuel.income	482.21	502.42	China Yuan
airline.revenue	76117.36	60049.67	China Yuan

Table 4.3-1: Airline B in January 2019- Summary statistics for all variables studied.

Using the model established from the main office data, we have passenger ticket income as the dependent variable, and available seat kilometers, passenger rate, passenger kilometer, total passenger number, passenger revenue, fuel income, and airline revenue as the independent variables. Coefficients for those variables are re-calculated for the predictions. We see that the predictions are generally a good fit to the data in Figure 4.3-1. The ratio of the root mean square error (RMSE) to the mean of the data is .25.

Airline B January 2019 Ticket Income Model

	Estimate	Std. Error	t value	Pr(>|t|)
available.seat.kilometers	-0.2097	0.244	-0.8596	0.3919
passenger.rate	-1.5648	0.6138	-2.5492	0.0122
passenger.kilometer	-0.83	0.2601	-3.1912	0.0019
total.passenger.number	0.889	0.7782	1.1423	0.2558
fuel.income	0.4331	0.1592	2.7199	0.0076
airline.revenue	1.1568	0.1027	11.2622	<0.0001

Table 4.3-2: Results from the model established using the stepwise procedure the January 2019 data for Airline B. Adj R^2 = .82. Coefficient estimates are for scaled variables so that relative importance can more easily be recognised.

The plot of the data vs fitted value has a few stray points, and the linear trend is less pronounced than for other datasets (Figure 4.3-1). Residuals (Figure 4.3-2) do not suggest any departures from the assumptions.

Airline B January 2019 Ticket Income Model Predictions

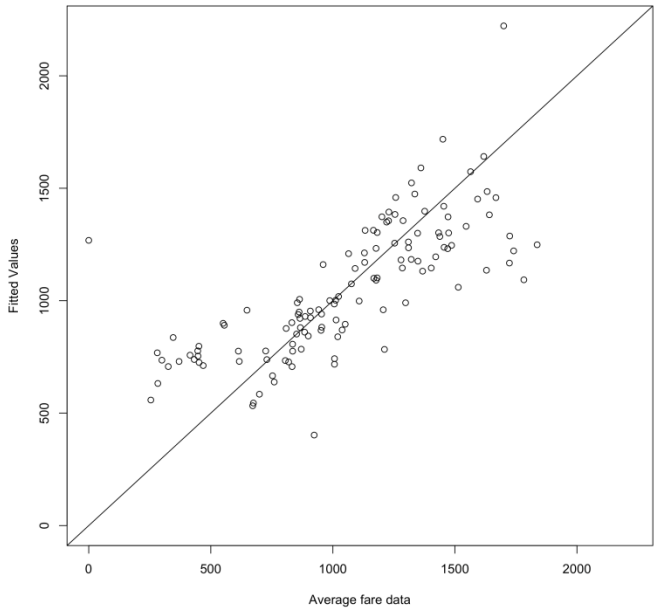


Figure 4.3-1: Average fares vs fitted values for the Airline B data for January 2019. Model coefficients have been recalculated for the variables

We see once again for this 2019 model, as in airline A, that two of the variables, available seat kilometers and total passenger number, do not appear to have independent prediction capabilities for this model.

Airline B January 2019 Ticket Income Model Residuals

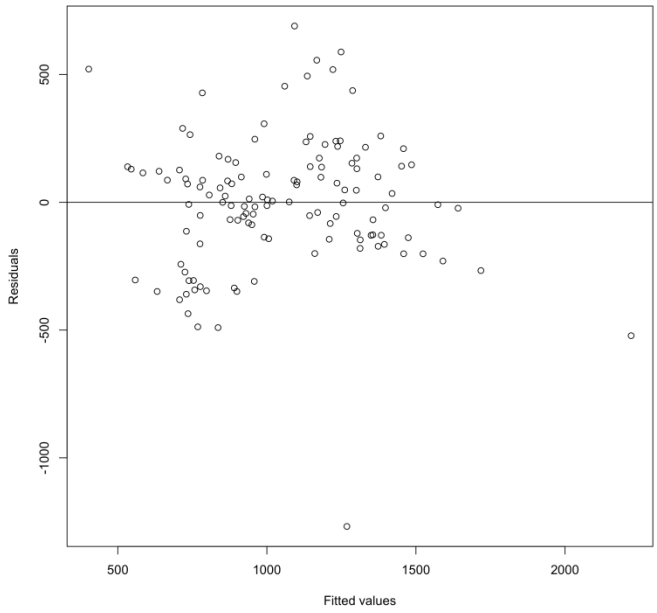


Figure 4.3-2: Fitted values vs Residuals for the Airline B data for January 2019.

In the preceding model coefficients have been re-calculated for the model using the 6 covariates. When we apply the *predict* function in R to use the coefficients that were calculated for the original model for the main office. We show the plot for the fit for this model in Figure 4.3-3. The ratio of the root mean square error (RMSE) to the mean of the data is .28. The plots show that the fit is quite similar.

Airline B January 2019 Ticket Income Model Predictions from Original Model

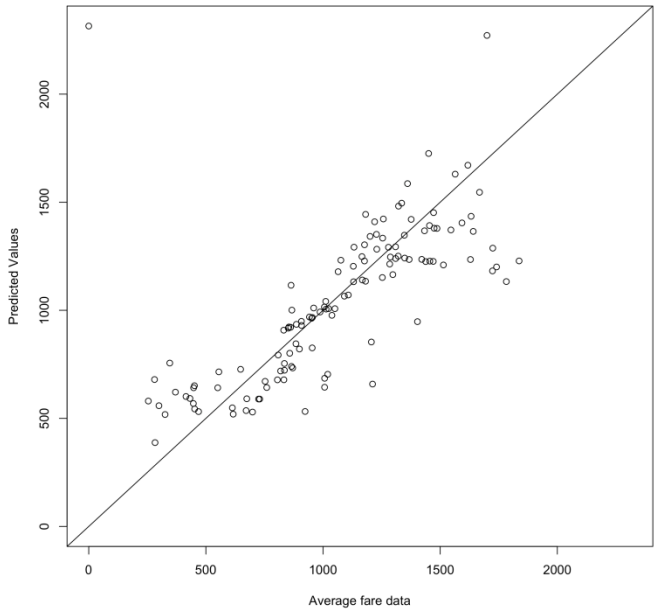


Figure 4.3-3: Data vs predicted values for the Airline B data for January 2019. Coefficients are as originally calculated for the Main Office.

Data is also available for January, 2018. Means and Standard Deviations for this data are given in the following table:

Airline B January 2018 Summary Statistics for all variables

Variable n=112	**Mean**	**SD**	**UNIT**
average.fare	859.09	254.85	China Yuan
available.seat.kilometers	152637.77	139508.79	number multiply seat-kilometer
passenger.rate	0.66	0.38	percentage
passenger.kilometer	136350.39	124127.17	Passenger-kilometer
total.passenger.number	100.45	65.21	headcount
passenger.revenue	67752.03	43233.11	China Yuan
fuel.income	144.62	293.23	China Yuan
airline.revenue	67522.27	42957.29	China Yuan

Table 4.3-3: Airline B in January 2018- Summary statistics for all variables studied.

Using the model established from the main office data, we have average fare as the dependent variable, and available seat kilometers, passenger rate, passenger kilometer, total passenger number, fuel income, and airline revenue as the independent variables. Coefficients for those variables are re-calculated for the predictions. We see that the predictions are generally a good fit to the data in Figure 4.3. The ratio of the root mean square error (RMSE) to the mean of the data is .10.

Airline B January 2018 Ticket Income Model

	Estimate	StdError	t value	Pr(>|t|)
available.seat.kilometers	2.4012	0.1323	18.1492	<.0001
passenger.rate	0.4444	0.1947	2.2828	0.0243
passenger.kilometer	-2.7349	0.1758	-15.5606	<.0001
total.passenger.number	-2.4438	0.2286	-10.6904	<.0001
fuel.income	-0.0119	0.0448	-0.2653	0.7912
airline.revenue	2.0198	0.1389	14.5409	<.0001

Table 4.3-4: Results from the model established using the stepwise procedure for the January 2018 data for Airline B. Adj R^2 = .89. Coefficient estimates are for scaled variables so that relative importance can more easily be recognised.

Airline B January 2018 Ticket Income Model Predictions

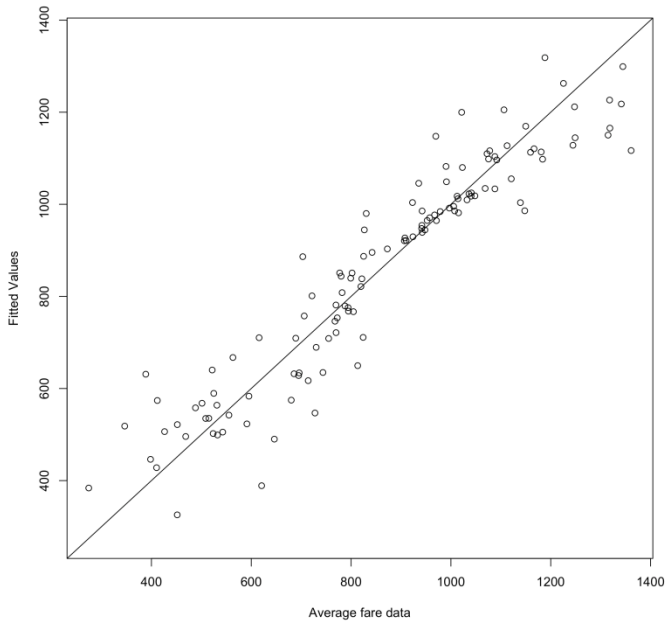


Figure 4.3-5: Data vs fitted values for the Airline B data for January 2018. Model coefficients have been recalculated for the variables

Airline B January 2018 Ticket Income Model Residuals

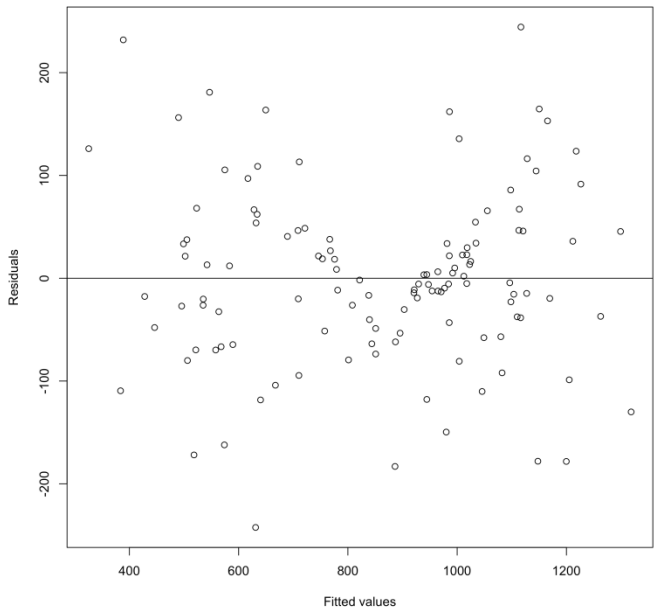


Figure 4.3-6: Data vs fitted values for the Airline B data for January 2018.

Airline B January 2018 Ticket Income Model Predictions from Original Model

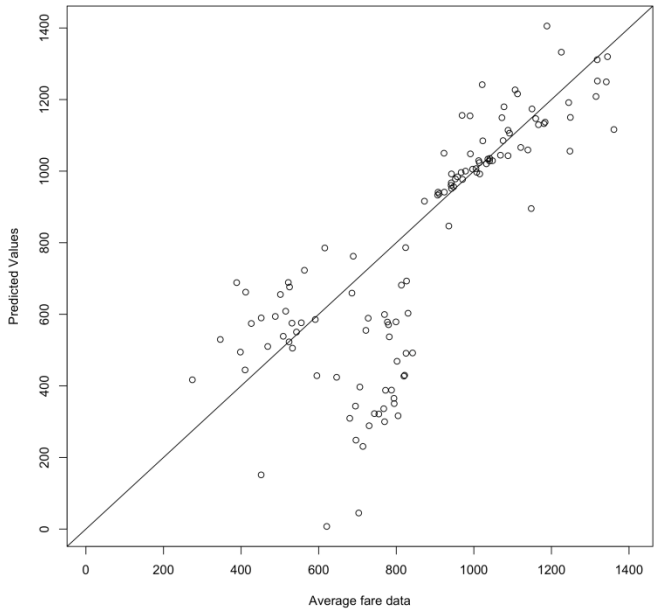


Figure 4.3-7: Data vs predicted values for the Airline B data for January 2018. Coefficients are as originally calculated for the Main Office.

In Table 4.3-4 coefficients were re-calculated for the model using the 6 covariates. All vartiables contribute significantly to the model except for fuel income. Residuals (Figure 4.3-4) do not suggest any departures from the assumptions despite the presence of some large points. Then, using the *predict* function in R, to use the coefficients that were calculated for the original model for the main office, we recalculate the performance of the model. The RMSE to the mean of the data is .25. Figure 4.3-7 shows the fit. We see quite a difference in this fit compared to Figure 4.3-5. The RMSE for the model where coefficients are recalculated is .1.

4.4 Analysis of the Regression Results

In section 4.2 and 4.3 of this chapter, we consider a multivariate linear modeling analysis of the average fare data of ‘A’ and ’B’ airlines in January 2019 and January 2018, respectively. We used the covariates that were found to be the six most important in a model for the Main Office of the airline (available seat kilometers,

passenger rate, passenger kilometer, total passenger number, fuel income, and airline revenue as the independent variables). Two regressions were considered: one with coefficients recalulated for the six covariates, and one which used the predict function in R to apply the original Min Office model to the data. A simple tablutation of the RMSE fit is shown in Table 4.4-1.

RMSE Results for all Datasets

	Optimized (Recalculated) Coefficients	Coefficients from Main Office
Airline A Jan 2019	0.21	0.28
Airline A January 2018	0.09	0.25
Airline B January 2019	0.25	0.28
Airline B January 2018	0.1	0.25

Table 4.4-1: RMSEs calculated for each model

Table of Coefficients for the 4 Models

available.seat.kilometers	-0.2097	-0.3641	0.6081	2.4012
passenger.rate	-1.5648	-5.9592	1.5562	0.4444
passenger.kilometer	-0.83	-3.2109	-1.1895	-2.7349
total.passenger.number	0.889	5.3337	-2.1696	-2.4438
fuel.income	0.4331	1.3019	0.1383	-0.0119
airline.revenue	1.1568	1.7629	1.698	2.0198

Coefficients for Airline A in January 2019 and 2018, and Airline B in January 2019 and 2018, respectively.

We can see that the model did not fit the 2019 data as well as the 2018 data in general. The RMSE values of .09 and .1, when using recalculated coefficients, are

quite good. This is especially surprising given that the Main Office data- which formed the basis of the model- was all in 2019.

Coefficients for the 4 models vary a bit more than one would hope if this model was generalisable. A summary of the changes is listed for each covariate:

- Available seat kilometers: This has a negative contribution for Airline A, both years and Airline B is positive in both years.
- Passenger rate: Similarly to available seat kilometers, This has a negative contribution for Airline A, both years and Airline B flips to positive both years.
- Passenger kilometer: This variable is always negative, but the impact is larger in 2018
- Total passenger number: This covariate has a positive influence in Airline A and a negative one in Airline B. In addition, the effect is much larger in 2018 for Airline A. But the effects are similar for both years in Airline B.
- Fuel Income: The effect is positive except for Airline B, 2018 data. The effect isn't large, although for Airline A in 2018 it increases slightly.
- Airline Revenue: This appears to be a consistent effect with a moderate impact across airlines and years.

Chapter 5: Conclusion

With the continuous development of the airline business and the continuous expansion of companies, the analysis of ticket revenue is becoming more and more important. There are many kinds of information and large amounts of data that affect airline ticket revenue. Finding the main factors that affect that revenue from the massive information available plays a decisive role in the passenger transport division of the company. It is key for transport management, and for the overall budget, to be able to predict this revenue. In this thesis we mainly explored the factors that affect airline ticket revenue, and established regression models to forecast the ticket revenue.

The empirical results show that influencing factors on ticket revenue can be useful across airlines and time periods (although the influence weight of the contribution can be different). Potentially, given more data, and possibly other variables, a useful model could be found for application to a wider set of data. The net step- given a more complete dataset- would be to create a model that incorporates airline, year and probably month or season.

The research topic of revenue management is constantly changing and improving with the improvement of theoretical models, the development of industries and enterprises, and the development of information technology. From the perspective of the air transport industry the improvement of prediction and optimisation models will bring huge economic benefits.

Reference

[1] Gan Jiang. (2003). Revenue management and marketing innovation [j]. China civil aviation.

[2] Fan Wei, &Wu Tongshui. (2006). Summary of airline revenue management research [j]. Journal of CAAC.

[3] National Development and Reform Commission, civil aviation domestic air transport price reform program hearing.

[4] Wang Fenglin.(2010).Analysis of the influence of China's railway on civil aviation operation [d]. Beijing Jiaotong University.

[5] Wu Tong, &Li Dong. (2003).The problems that airlines need to solve before implementing revenue management [j]. China civil aviation.

[6] Jia Xinru, &Yan Libin, &Zhang Yong and so on.(2019). Based on correlation analysis of passenger train ticket revenue calculation method [j]. railway computer applications.

[7] Lv Xiaoyan. (2007). Data modeling and analysis method based on decision tree and its application in railway [c]. China Academy of Railway Sciences.

[8] Gao Huixuan. (2005). Applied multivariate statistical analysis [m]. Beijing: Peking University press.

[9] Wang Xuemin. (2004) Applied multivariate statistical analysis [m]. Shanghai: Shanghai University of Finance and Economics press (Second Edition).

[10] Fan Jincheng,& Mei Chang Lin. (2002). Data analysis [m]. Beijing: Science Press.

[11] Wang Songjia, &Chen Min, &Chen Liping. (1999). Linear statistical model [m]. Beijing: Higher Education Press.

[12] Johnoson D. (2005). Applied Multivariate Methods for Data Analysts (photocopy). Beijing, Higher Education Press.

[13] Wu Guofu,& an Wanfu,& Liu Jinghai.(1992). Practical data analysis method [m]. Beijing: China Statistics Publishing House.

[14] Xue Yi, &Chen Liping. (2006). Statistical modeling and R software [m]. Beijing: Tsinghua University press.

[15] Muirhead R J.(1982). Aspcets of Mulitivariate Statistical Theory[M]. John Wiley.

[16] Kendall M. (1975).Mulitivariate Analysis[M]. Charles Criffin & Company Limited.

[17] He Xiaoqun. (1994).The Applications of Stepwise Regression Analysis in the Research of Inflation in China[J]. The Fifth Japan-China Symposion on Statistics, University Education Prese, Japan.

[18] Hoerl A E, &Kennard R W. (1970). Ridge Regression: Biased Estimation for Non-orthogonal Problems[J]. Technometries.

[19] He Xiaoqun. (1997). Discriminant Modelof Comparatively Well-off Level of Chinese City[J]. Hong Kong: Contemporary Multivariate Analysis and Its Applications.

[20] He Xiaoqun, & Liu Wenqing. (2011). Applied regression analysis [m]. applied regression analysis (Third Edition).
[21] Wang Pengyun, & Zhong Minxuan. (2020). Research on Multiple Linear Regression of Regional Income Distribution based on Cross-Section Data[E]. 2nd International Conference on Business Economics Management Science.
[22] Yin Yanqing. (2020). Model-free tests for series correlation in multivariate linear regression[J]. Journal of Statistical Planning and Inference.
[23] Eyiah-bediako F, & Bosson-amedenu S, & Otoo J. (2020). Modeling Macroeconomic Variables Using Principal Component Analysis and Multiple Linear Regression: The Case of Ghana's Economy[J]. Journal of Business and Economic Development.
[24] Zhang Yanyan, & Ma Haile, & Wang Bei. (2016). Relationships between the structure of wheat gluten and ACE inhibitory activity of hydrolysate: stepwise multiple linear regression analysis[J]. Journal of the Science of Food and Agriculture.
[25] Rose S, & McGuire T G. (2019). Limitations of P-Values and R-squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment[J]. The American statistician.
[26] Jing Xinxin, & Xiao Yao. (2019). Study on Factors Affecting Alloy Yield Based on Backward Stepwise Regression Method[J]. International Journal of Education and Economics.
[27] Yarhands D A, & Samuel A A, & Charles A. (2017). Teacher-Student Variables as Predictor of Students' Interest in Mathematics: The Use of Stepwise Multiple Linear Regression Analysis[J]. Asian Research Journal of Mathematics.
[28] Abdelrafe E, & Burairah H. (2015). Modelling and Evaluating Software Project Risks with Quantitative Analysis Techniques in Planning Software Development[J]. Journal of Computing and Information Technology.
[29] Lu Guohua. (2015). A Predictive Model of 100 Meter Women Athletes Based on Multiple Stepwise Regression Analysis[C]. 2015 5th International Conference on Education and Sports Education.
[30] Portet S. (2020). A primer on model selection using the Akaike Information Criterion[J]. Infectious Disease Modelling.
[31] Ohta M, & Mitani Y. (1995). An application of AIC information criterion to hierarchical probabilistic evaluation of acoustic environment with standard distribution form as the first expansion term[J]. The Journal of The Acoustical Societ of Japan.
[32] Vinnie K, & Nils L H. (2019). Copula information criterion for model selection with two-stage maximum likelihood estimation[J]. Econometrics and Statistics.

Acknowledgements

Graduation thesis is a very important learning task for my college life, and it is a comprehensive assessment of me. Also in order to make my comprehensive skills have a great improvement. In the process of graduation thesis, I not only reviewed the knowledge I learned before, but also learned a lot of new knowledge in order to make my knowledge structure more systematic and perfect. At the same, It also improves my ability to analyze and solve problems independently. It is far from enough for me to finish my graduation thesis by relying solely on my own efforts.

First of all, I would like to express my great respect and heartfelt thanks to my instructor. Many of the works in this paper are done by the teacher's patience guidance. From the establishment of the topic to the end of the paper, it is permeated with the efforts of the teacher. Whenever the subject encounters difficulties, the teacher gives me encouragement and guidance suggestions, so that I can overcome various difficulties and successfully complete the topic. The teacher's profound knowledge, rigorous scholarship attitude, practical and realistic style, kindly attitude and noble dedication spirit deeply moved me, and made me understand that I must be honest, serious, steadfast and conscientious. The teacher's words and deeds will benefit me all my life. I want to finish my graduation thesis successfully. Thank him for his support, encouragement, care and teaching in my scientific research and learning life. I would like to take this opportunity to express my heartfelt thanks to the teacher.

I also want to thank my classmates and roommates. Thank you for helping me with my life and study in University. We have experienced many happy and unforgettable times together. There are our laughter and laughter on the campus. In the University, we grow and progress together. Here, I wish every one of my classmates can realize their dreams and achieve success in their future life.

Secondly, I want to thank my Alma mater, especially when I am about to set foot in my post, giving me such an opportunity to exercise and study, which has deepened my understanding of previous knowledge, widened my knowledge and improved my comprehensive ability to apply my knowledge. Here, I wish my Alma mater a better development.

I am also very grateful to my parents. During my growing up, you gave me enough material and spiritual support to enable me to finish my college education successfully and find an ideal job. I will try my best to lighten the burden on my family and promise you a better life and better repay you in my future.