Estimation of Flight Tickets Price using Dummy Regression Analysis

Felicia Ferren
Statistics Department
School of Computer Science,
Bina Nusantara University
Jakarta, Indonesia 11480
felicia.ferren@binus.ac.id

Ivo Herid Lesmana
Statistics Department
School of Computer Science,
Bina Nusantara University
Jakarta, Indonesia 11480
ivo.lesmana@binus.ac.id

Virgie Cecilia Johan
Statistics Department
School of Computer Science,
Bina Nusantara University
Jakarta, Indonesia 11480
Virgie.johan@binus.ac.id

Margaretha Ohyver
Statistics Department
School of Computer Science,
Bina Nusantara University
Jakarta, Indonesia 11480
mohyver@binus.ac.id

***Abstract* - On today's age and time, transportation is one of the most crucial things for a human to have. Human could go from one place to another without having the burden of tiredness and endless time of travel. One of the most versatile transportations is plane, which is a transportation by air, with it being one of the fastest and also has the most reach out of all transportation ways. Flight, however, is not cheap and customer needs to order the flight long before the flight take place. Our study aims to helps customer in predicting flight ticket price by creating a regression model capable of predicting the next flight price using the available variables in the dataset. Our study tested the data on two developed models. the final model picked is the most optimal one for our dataset, which require categorical data type – or by using Dummy Regression.**

Keywords - Dummy Regression, Multiple Linear Regression

I. INTRODUCTION

Flight has been one of ways for human to travel from one place to another since the world's first scheduled commercial passenger flight occurred in 1914. In these modern days, flight has been a mainstream task and no longer something human deemed impossible, using a vehicle called aircraft, humans are able to travel by air. Flight, however, isn't easy as it needs time and money. Flight using an aircraft has schedules and is considered tight with it having a precise time of departure and arrival. It also has a scheduled destination, so that it can save a lot of budgets with the plane having needed to travel one time per scheduled, from one city or country to another, the distance of one city or country to another is impacted upon the price of the tickets. Airline ticket prices may vary based many upon the distance, class, destination, many stops or transits, type of the airline, duration, and even how many days are left before departure when the customer bought the ticket.

As a customer, determining the minimum price to buy a ticket is the key issue (Abdella et al., 2021). However, the ticket price may be affected by several factors thus may change continuously. Moreover, airfare pricing strategy has developed, making a complex structure of sophisticated rules and mathematical models that drive the pricing strategies of airfare since the deregulation of the airline industry (Wang et al.,

2019). Formerly, several people used strategy where they bought tickets far from their departure time to acquire cheaper ticket price. However, early purchase may risk from maintaining a specific schedule to be broken, that leads to schedule change, which require more fee. In fact, that strategy does not work anymore (Abdella et al., 2021).

In this paper, we will be doing an analysis upon the data of flight booking options from 'Easemytrip' website for flights in between top 6 metro cities around India. Then, we will give a prediction for the next customer of how much would their ticket cost, by using various regression techniques and determining which one is the most optimal choice for our data.

II. METHODS

A. *Multiple Regression Analysis*

Multiple regression analysis is the study of how a dependent variable *y* is related to two or more variables, or so called the predictors. A multiple regression model is an equation which describes how the dependent variable *y* and the predictors are related. The relationship is formulated as a linear model (Chatterjee et al., 2006):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Where:

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ = regression coefficient

ε = random disturbance or error

However, this formula has assumptions that needs to be fulfilled. The assumptions underlying the structural model from the multiple linear regression are called residual assumptions, which are:

1) The errors are normally distributed,
2) The errors are heteroscedastic,
3) The errors are non-autocorrelated.

B. *Dummy Regression*

In many situations, however, there might be categorical independent variables such as gender (male, female), method of payment (cash, credit card, check), and so on, that must be handled (Anderson, 2011). Dummy Regression, which is also one of the methods to create regression model, is used to handle this type of situation.

Dummy regression is done by creating numeric dummy variable(s) to be used upon the categorical variables. However, the number of dummy variables needs to be $k - 1$, in which k is the category. For example, there is a survey about customer satisfaction with it being an ordinal of: Very Satisfied, Satisfied, Not Satisfied. We see that there are 3 categories on the survey. Then, we can create 2 dummy variables, D_1 and D_2 , where each dummy variable being coded as 0 or 1, or as 1 or 2.

The formula becomes different from multiple linear regression, as it is also necessary to count the dummy variables' values, the formula for Dummy regression is (Alkharusi, 2012):

$$Y = B_0 + B_1 X + \sum_{j}^{k-1} B_j D_j + \varepsilon_j$$

Where:

$\hat{Y}$ = dependent variables/predicted value
B_0 = intercept
X = non-categorical variable
B_j = regression coefficient
k = number of categorical variables
D = dummy variables
ε_j = error associated with each variables

Though, generally dummy variables is used to analyze and create model for categorical independent variables with numeric or continuous predicted variables, there are also some cases where not only the data has categorical independent variables, but also continuous. Meaning that it also needs to use dummy variables with corresponding Ordinary Least Squares (OLS), but keep in mind that by combining it with multiple linear, the data also needs to fulfill OLS assumptions.

C. Dataset

Dataset for this paper was obtained from Kaggle, which describes a flight travel options between top 6 metro cities in India. The dataset is collected from Ease my trip website from February 11th to March 31st, 2022, with the total of 300154 records, where each of them is distinct flight booking options from the site.

There are 11 features. The various features of the dataset are airline, flight, source city, departure time, stops, arrival time, destination city, class, duration, days left, and price. This paper will focus more on the prediction of flight ticket price by using dummy regression technique.

III. Result and Discussion

It is stated on the principle of Occam's Razor, that entities should not be multiplied without necessity (Halim et al., 2021). This means that we do not need to use a more complicated method which make similar predictions against the simpler ones.

First, we try to use basic Multiple Linear Regression into our dataset, however our dataset has an independent variable that is not continuous, it's an ordinal variable, we could try to not include the ordinal variable, therefore the Multiple Linear Model could possibly work. Our model becomes:

$$\hat{Y} = 16799{,}6 + 634{,}130 \cdot (Duration) - 140{,}73 \cdot (Days\ left)$$

Table I states the p-value for each predictor variable in t-test. All the results show that both our predictors are significant for the model.

Table I. P-Values of Each Regressor for Model using Multiple Linear Regression

Regressor Variable	***P-Value***
Duration	$< 2 * 10^{-16}$
Days_left	$< 2 * 10^{-16}$

However, this model causes the model to have terribly low in representing the entire data. It is shown by the low R-squared score: 0.04877. In other words, the model only represents 4.877% of the data. This might happens because not every predictor variable is included into the model. To handle this problem, we use those categorical variables in our data. Hence, Dummy Regression will be used together with Multiple Linear Regression, making the model for Multiple Linear capable of accepting Categorical predictor variables.

We create dummy variables for Class and Stops variables. Class variable creates one dummy variable, as it only has two categories, namely business class and economy class. Note that business class is coded 1 and economy class coded 0 for D_1. On the other side, Stops variable creates two dummy variables, as it has three different categories, namely zero stops, one, stops, and two or more stops. Note that zero stops coded 0 for both D_2 and D_3, one stops coded 0 for D_2 and 1 for D_3, and two or more stops coded 1 for both D_2 and D_3.

Table II. Dummy Coding for Stops Variable

Stops	D_2	D_3
Zero	0	0
One	0	1
Two or more	1	1

Hence, our model becomes:

$$\hat{Y} = 2487.696 + 31.9278 \cdot (Duration) - 132.4680 \cdot (Days\ left) + 45578.8077 \cdot (Class\ Dummy) + 2591.0469 \cdot (Stops\ Dummy\ 1) + 8119.6723 \cdot (Stops\ Dummy\ 2)$$

Table III states the p-value for each predictor variable in t-test. All the results show that both our predictors are significant for the model.

Table III. P-Value of Each Regressor for Model using Dummy Regression

Regressor Variable	***P-Value***
Days left	$< 2 \cdot e^{-16}$
Duration	$< 2 \cdot e^{-16}$

Class Dummy	$< 2 \cdot e^{-16}$
Stops Dummy 1	$< 2 \cdot e^{-16}$
Stops Dummy 2	$< 2 \cdot e^{-16}$

Now the Dummy Regression Model's R^2 score becomes 0,9017, which means now it represents 90,17 % of the dataset, which is a great improvement from using only adding dummy variables into our model.

IV. CONCLUSION

From our Research, we could have an understanding that Flight Ticket of a certain airport can be predicted by using Multiple Linear Regression along with Dummy Regression, where the flight ticket price as the predicted variable and days left before flight, duration of flight, class of flights, and amount of stops as the model regressors. The developed model is the combination between Multiple and Dummy Regression as by using only Multiple will have a low R-squared score, because the predicted variables is depended upon many variables including those that is a categorical, meaning that normal OLS will not include these variables into their model, thus reducing its R-squared and accuracy. It is shown in chapter 3 that multiple linear regression model has only 0.04877 while Dummy + OLS have 0.9017, therefore making it the most optimal model for predicting Flight Ticket Price.

REFERENCES

Abdella, J. A., Zaki, N. M., Shuaib, K., & Khan, F. (2021). Airline ticket price and demand prediction: A survey. *Journal of King Saud University - Computer and Information Sciences*, *33*(4), 375–391. https://doi.org/10.1016/j.jksuci.2019.02.001

Alkharusi, H. (2012). Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding. *International Journal of Education*, *4*(2), 202. https://doi.org/10.5296/ije.v4i2.1962

D. Anderson, D. Sweeney, T. Williams, N. Freeman and E. Shoesmith, Statistics for business and economics, 11th ed. 2011.

Halim, G. A., Agustin, P., Adiwijayanto, E., Ohyver, M. (2021). Estimation of Cost of Living in a Particular City using Multiple Regression Analysis and Correction of Residual Assumptions through Appropriate Methods. *2021 6th International Conference on Computer Science and Computational Intelligence (ICCSCI)*.

S. Chatterjee and A. Hadi, Regression analysis by example, 4th ed. 2006.

Wang, T., Pouyanfar, S., Tian, H., Tao, Y., Alonso, M., Luis, S., & Chen, S. C. (2019). A framework for airfare price prediction: A machine learning approach. *Proceedings - 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science, IRI 2019*, 200–207. https://doi.org/10.1109/IRI.2019.00041