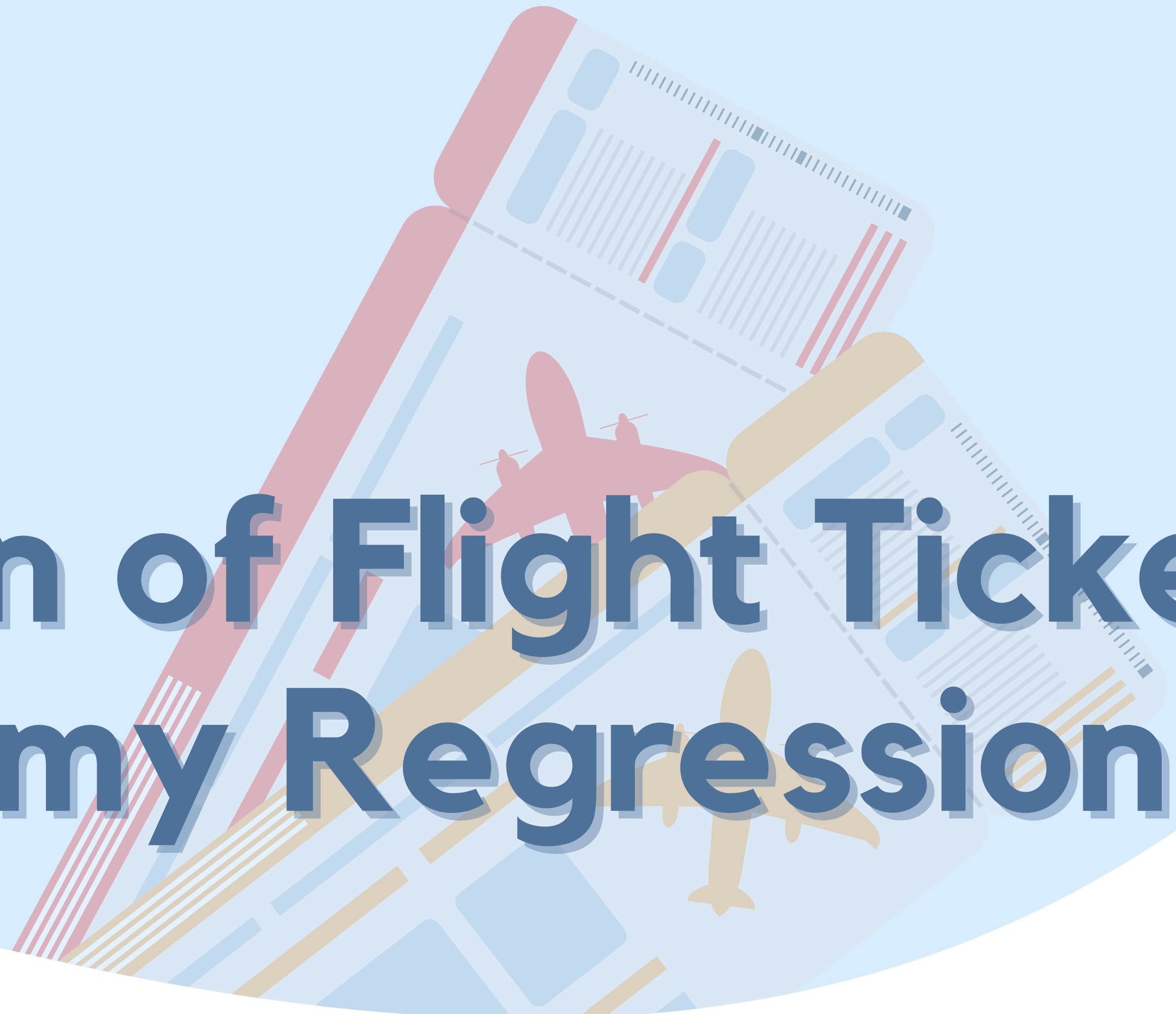

Estimation of Flight Tickets Price using Dummy Regression Analysis

Felicia Ferren - 2440013071

Ivo Herid Lesmana - 2440019844

Virgie Cecilia - 2440043433

STAT6048049 – Regression Analysis (LAB)

Introduction

Determining the minimum ticket price is essential for customers. However, flight ticket price may change continuously because of several factors. Formerly, some people used strategy where they bought tickets far away from their departure time to acquire cheaper ticket price. However, this trick doesn't work anymore.

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 the prediction for how much the ticket price would cost.

Felicia Ferren - 2440013071

Methods

The method we will be using for this paper are:

A. Multiple Linear Regression

B. Dummy Regression

Felicia Ferren - 2440013071

A. Multiple Linear Regression

The relationship of dependent variable *y* against the predictors is formulated as a linear model:

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

Where:

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

ε = random disturbance or error

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

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

Methods

B. Dummy Regression

In many situations, however, there might be categorical independent variables that must be handled. This can be handled using dummy regression. It is done by creating numeric dummy variable(s) to be used upon categorical variables. The number of dummy variable is *k - 1*, in which *k* is the number of categories.

Formula for dummy regression is:

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

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

C. Dataset

Obtained from Kaggle, and is collected from Ease my trip website from Feb 11th until March 31st 2022, with the total of 300154 records.

There are 11 features, which 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 ticked using dummy regression technique.

Felicia Ferren - 2440013071

Result and Discussion

We used R-studio for the computation. First, load and see the description and data summary:

```
library(readr)
flight <- read_csv("C:/Users/felicia ferren/OneDrive - Bina Nusantara University/!SMT 4
/[STAT6048049] Regression Analysis/datasets/flight-price-prediction/Clean_Dataset.csv")
View(flight)
summary(flight)
```

```
Data
flight              300153 obs. of 12 variables
    $ X1              : num  0 1 2 3 4 5 6 7 8 9 ...
    $ airline         : chr  "SpiceJet" "SpiceJet" "AirAsia" "Vistara" ...
    $ flight          : chr  "SG-8709" "SG-8157" "I5-764" "UK-995" ...
    $ source_city     : chr  "Delhi" "Delhi" "Delhi" "Delhi" ...
    $ departure_time  : chr  "Evening" "Early_Morning" "Early_Morning" "Morning" ...
    $ stops           : chr  "zero" "zero" "zero" "zero" ...
    $ arrival_time    : chr  "Night" "Morning" "Early_Morning" "Afternoon" ...
    $ destination_city: chr  "Mumbai" "Mumbai" "Mumbai" "Mumbai" ...
    $ class           : chr  "Economy" "Economy" "Economy" "Economy" ...
    $ duration        : num  2.17 2.33 2.17 2.25 2.33 2.33 2.08 2.17 2.17 2.25 ...
    $ days_left       : num  1 1 1 1 1 1 1 1 1 1 ...
    $ price           : num  5953 5953 5956 5955 5955 ...
```

```
> summary(flight)
       X1            airline             flight          source_city
 Min.   :     0   Length:300153      Length:300153      Length:300153
 1st Qu.: 75038   Class :character   Class :character   Class :character
 Median :150076   Mode  :character   Mode  :character   Mode  :character
 Mean   :150076
 3rd Qu.:225114
 Max.   :300152
 departure_time        stops           arrival_time
 Length:300153      Length:300153      Length:300153
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character


 destination_city      class              duration       days_left
 Length:300153      Length:300153      Min.   : 0.83   Min.   : 1
 Class :character   Class :character   1st Qu.: 6.83   1st Qu.:15
 Mode  :character   Mode  :character   Median :11.25   Median :26
                                       Mean   :12.22   Mean   :26
                                       3rd Qu.:16.17   3rd Qu.:38
                                       Max.   :49.83   Max.   :49
     price
 Min.   :  1105
 1st Qu.:  4783
 Median :  7425
 Mean   : 20890
 3rd Qu.: 42521
 Max.   :123071
```

There are 300153 observations with 12 variables.

When we checked the summary, 8 of the variables are in characters, and rest of them in numbers.
For this case, we set the price as our target variable *y* and use all of the numerical variables (duration and days_left) as the predictors.

Felicia Ferren - 2440013071

Result and Discussion

Now, develop a multiple linear regression estimation model:

```
# copy data
flight_cpy = flight

# using multiple linear regression
mlr_model = lm(price ~ duration+days_left, data = flight_cpy)
summary(mlr_model)
```

```
> summary(mlr_model)

Call:
lm(formula = price ~ duration + days_left, data = flight_cpy)

Residuals:
   Min     1Q Median     3Q    Max 
-34074 -14819 -11071  19684  98133 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16799.600    113.070  148.58   <2e-16 ***
duration      634.130      5.623  112.78   <2e-16 ***
days_left    -140.730      2.982  -47.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22140 on 300150 degrees of freedom
Multiple R-squared:  0.04877,	Adjusted R-squared:  0.04876 
F-statistic:  7694 on 2 and 300150 DF,  p-value: < 2.2e-16
```

We will be developing the model using lm().

X1 = duration

X2 = days_left

y = price

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

model format

Then, we see from the summary, that:

$$\hat{\beta}_0(intercept) = 16799.6$$

$$\hat{\beta}_1 = 634.13$$

$$\hat{\beta}_2 = -140.73$$

and our estimated regression equation will become:

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

interpreting that,

- 16799.6 is the ticket price when the ticket is bought 0 days left before the flight and 0 in duration.
- the price will increase by 634.130 every duration is longer by 1 hour.
- the price will decrease by 140.73 every addition of days left (the customer bought the ticket before flight).

Felicia Ferren - 2440013071

Result and Discussion

Then, we will be checking the overall significance and significance for each predictor using *F*-test and *t*-test:

```
> summary(mlr_model)

Call:
lm(formula = price ~ duration + days_left, data = flight_cpy)

Residuals:
   Min     1Q Median     3Q    Max 
-34074 -14819 -11071  19684  98133 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16799.600    113.070  148.58   <2e-16 ***
duration      634.130      5.623  112.78   <2e-16 ***
days_left    -140.730      2.982  -47.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22140 on 300150 degrees of freedom
Multiple R-squared:  0.04877,	Adjusted R-squared:  0.04876 
F-statistic:  7694 on 2 and 300150 DF,  p-value: < 2.2e-16
```

Using 5% level of significance,
The p-value for *F*-statistics (< 2.2e-16) shows that our model is overall significant.

The p-value for *t*-statistics on both predictors (< 2e-16) shows that both predictors are statistically significant to our model.

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

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

Then, check the R-squared score:

```
Residual standard error: 22140 on 300150 degrees of freedom
Multiple R-squared:  0.04877,	Adjusted R-squared:  0.04876 
F-statistic:  7694 on 2 and 300150 DF,  p-value: < 2.2e-16
```

This result indicates that our model only represents 4.877% of the data, which is a very bad result. In this case, we have to add more variables to the data - and because there are only categorical ones, we are going to use dummy variables.

Felicia Ferren - 2440013071

Result and Discussion

To do so, the categorical variables (except *flight*) have to be set as factors.

```
# make categorical variables into factor form
flight$airline = as.factor(flight$airline)
flight$source_city = as.factor(flight$source_city)
flight$departure_time = as.factor(flight$departure_time)
flight$stops = as.factor(flight$stops)
flight$arrival_time = as.factor(flight$arrival_time)
flight$destination_city = as.factor(flight$destination_city)
flight$class = as.factor(flight$class)
```

```
flight              300153 obs. of 12 variables
  $ X1              : num  0 1 2 3 4 5 6 7 8 9 ...
  $ airline         : Factor w/ 6 levels "Air_India","AirAsia",..: 5 5 2
  $ flight          : chr  "SG-8709" "SG-8157" "I5-764" "UK-995" ...
  $ source_city     : Factor w/ 6 levels "Bangalore","Chennai",..: 3 3 3
  $ departure_time  : Factor w/ 6 levels "Afternoon","Early_Morning",..:
  $ stops           : Factor w/ 3 levels "one","two_or_more",..: 3 3 3 3
  $ arrival_time    : Factor w/ 6 levels "Afternoon","Early_Morning",..:
  $ destination_city: Factor w/ 6 levels "Bangalore","Chennai",..: 6 6 6
  $ class           : Factor w/ 2 levels "Business","Economy": 2 2 2 2 2
  $ duration        : num  2.17 2.33 2.17 2.25 2.33 2.33 2.08 2.17 2.17 2
  $ days_left       : num  1 1 1 1 1 1 1 1 1 1 ...
  $ price           : num  5953 5953 5956 5955 5955 ...
```

This time, we will try to use two more variables with the least number of level, *stops* and *class* variable.

To make a dummy variable, we will do dummy coding.

There are only two labels on *class* variable. Hence, we will be only using one dummy variable, *D1*. *'business' class* coded 1 and *'economy' class* coded 0 in variable Dummy D1.

There are three labels on *stops* variable. Hence, we will use two dummy variables, *D2* and *D3*. For *stops* variable, *'zero' stops* coded 0 on both *D2* and *D3*, *'one' stops* coded 0 for *D2* and 1 for *D3*, and *'two_or_more' stops* coded 1 for both *D2* and *D3*.

Table II. Dummy Coding for Stops Variable

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

Result and Discussion

Then, start creating the dummy variables.

```
# copy data
flight_cpy = flight

# stops:        d1  d2
# zero          0   0
# one           0   1
# two_or_more   1   1
d1_stops = ifelse(flight_cpy$stops == 'two_or_more', 1, 0)
d2_stops = ifelse(flight_cpy$stops == 'zero', 0, 1)
flight_cpy = cbind(flight_cpy, d1_stops)
flight_cpy = cbind(flight_cpy, d2_stops)

# class:
# economy = 0; business = 1
# flight_cpy$class = ifelse(flight_cpy$class == 'Business', 1, 0)
d_class = ifelse(flight_cpy$class == 'Business', 1, 0)
flight_cpy = cbind(flight_cpy, d_class)
```

```
flight_cpy               300153 obs. of 15 variables
  $ X1              : num  0 1 2 3 4 5 6 7 8 9 ...
  $ airline         : Factor w/ 6 levels "Air_India","AirAsia",..: 5
  $ flight          : chr  "SG-8709" "SG-8157" "I5-764" "UK-995" ...
  $ source_city     : Factor w/ 6 levels "Bangalore","Chennai",..: 3
  $ departure_time  : Factor w/ 6 levels "Afternoon","Early_Morning"
  $ stops           : Factor w/ 3 levels "one","two_or_more",..: 3 3
  $ arrival_time    : Factor w/ 6 levels "Afternoon","Early_Morning"
  $ destination_city: Factor w/ 6 levels "Bangalore","Chennai",..: 6
  $ class           : Factor w/ 2 levels "Business","Economy": 2 2 2
  $ duration        : num  2.17 2.33 2.17 2.25 2.33 2.33 2.08 2.17 2
  $ days_left       : num  1 1 1 1 1 1 1 1 1 1 ...
  $ price           : num  5953 5953 5956 5955 5955 ...
  $ d1_stops        : num  0 0 0 0 0 0 0 0 0 0 ...
  $ d2_stops        : num  0 0 0 0 0 0 0 0 0 0 ...
  $ d_class         : num  0 0 0 0 0 0 0 0 0 0 ...
```

Now, we have 3 dummy variables in the dataset. Next, we will develop the regression model again using lm().

X1 = duration

X2 = days_left

D1 = class dummy

D2 = stops dummy 1

D3 = stops dummy 2

y = price

```
# using dummy regression
dummy_model = lm(price ~ duration+days_left+d_class+d1_stops+d2_stops, data = flight_cpy)
summary(dummy_model)
```

Result and Discussion

```
> summary(dummy_model)

Call:
lm(formula = price ~ duration + days_left + d_class + d1_stops +
    d2_stops, data = flight_cpy)

Residuals:
   Min     1Q Median     3Q    Max
-41029  -2909   -584   3073  66851

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2487.6965    45.7606   54.36   <2e-16 ***
duration       31.9278     2.1299   14.99   <2e-16 ***
days_left    -132.4680     0.9589 -138.15   <2e-16 ***
d_class     45578.8077    28.5297 1597.59   <2e-16 ***
d1_stops     2591.0469    63.9388   40.52   <2e-16 ***
d2_stops     8119.6723    46.6733  173.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7115 on 300147 degrees of freedom
Multiple R-squared:  0.9017,    Adjusted R-squared:  0.9017
F-statistic: 5.509e+05 on 5 and 300147 DF,  p-value: < 2.2e-16
```

From the model summary, we can see that:

$$\hat{\beta}_0(intercept) = 2487.6965$$

$$\hat{\beta}_1 = 31.9278$$

$$\hat{\beta}_2 = -132.4680$$

$$\hat{\beta}_3 = 45578.8077$$

$$\hat{\beta}_4 = 2591.0469$$

$$\hat{\beta}_5 = 8119.6723$$

and our estimation regression equation becomes:

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

Felicia Ferren - 2440013071

Result and Discussion

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

interpreting that,

- 2487.70 is the ticket price when the ticket is **economy class**, has **zero stop**, is bought 0 days left before the flight, and 0 in duration.
- 10607.67 is the ticket price when the ticket is an **economy class**, has **one stop**, bought 0 days left before the flight, and 0 in duration.
- 13198.42 is the ticket price when the ticket is an **economy class**, has **two stops**, bought 0 days left before the flight, and 0 in duration.
- 48066.51 is the ticket price when the ticket is **business class**, has **zero stop**, bought 0 days left before the flight, and 0 in duration.
- 56186.18 is the ticket price when the ticket is **business class**, has **one stop**, bought 0 days left before the flight, and 0 in duration.
- 58777.23 is the ticket price when the ticket is **business class**, has **two stops**, bought 0 days left before the flight, and 0 in duration.
- the price will increase by 31.9278 every duration is longer by 1 hour.
- the price will decrease by 132.4680 every addition of days left (the customer bought the ticket before flight).

Result and Discussion

Then, we will be checking the overall significance and significance for each predictor using *F*-test and *t*-test:

```
> summary(dummy_model)

Call:
lm(formula = price ~ duration + days_left + d_class + d1_stops +
    d2_stops, data = flight_cpy)

Residuals:
   Min     1Q Median     3Q    Max
-41029  -2909   -584   3073  66851

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2487.6965    45.7606   54.36   <2e-16 ***
duration       31.9278     2.1299   14.99   <2e-16 ***
days_left    -132.4680     0.9589 -138.15   <2e-16 ***
d_class     45578.8077    28.5297 1597.59   <2e-16 ***
d1_stops     2591.0469    63.9388   40.52   <2e-16 ***
d2_stops     8119.6723    46.6733  173.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7115 on 300147 degrees of freedom
Multiple R-squared:  0.9017,	Adjusted R-squared:  0.9017
F-statistic: 5.509e+05 on 5 and 300147 DF,  p-value: < 2.2e-16
```

Using 5% level of significance,
The p-value for *F*-statistics (< 2.2e-16) shows that our model is overall significant.

The p-value for *t*-statistics on both predictors (< 2e-16) shows that all predictors are statistically significant to our 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	$< 2 \cdot e^{-16}$

Then, check the R-squared score:

```
Residual standard error: 7115 on 300147 degrees of freedom
Multiple R-squared:  0.9017,	Adjusted R-squared:  0.9017
F-statistic: 5.509e+05 on 5 and 300147 DF,  p-value: < 2.2e-16
```

This result indicates that our model represents 90.17% of the data, which is a very good result. We achieve better model by adding the categorical variables. We use this model to predict ticket price.

Felicia Ferren - 2440013071

Conclusion

From our research, flight price can be predicted by using **multiple linear regression along with dummy variable**, 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 regressors**. Adding categorical variables greatly impacts the fitness of the model. It is shown that multiple linear regression model has only 0.04877 R-squared score while OLS+Dummy has 0.9017 R-squared score, making it **the most optimal model for predicting Flight Ticket Price**.

Felicia Ferren - 2440013071

Thank you!

Here is the link to access the explanation video:
https://bit.ly/FinalExamRegressionAnalysisLAB

Alternate link:
https://youtu.be/QWz8dYg75EM

Thank you.

Felicia Ferren - 2440013071