Causal Analysis in the Airline Industry

Likhith Karanuthala, Manaswini Rekula, Pankhuri Tyagi, Sheetal Pasam

MSBA 6440

Group 9 - Likhith Karanuthala, Manaswini Rekula, Pankhuri Tyagi, Sheetal Pasam

Executive Summary

The airline industry generates billions of dollars in revenue each year and employs hundreds of thousands of people. However, it operates in a very fragile ecosystem that is easily affected by various variables Thus, we want to take a closer look at the factors which may affect this widespread industry.

One of the key metrics to measure airline performance is the number of passengers flying on average per day in a given route.

This metric can help us in optimizing resources such as:

- Number of flights scheduled on that route
- Staffs deployed
- New promotions to be applied etc.

Given our panel data, we want to understand what factors affect our metric of interest.

Our data spans over 4 years (1997 - 2000) and has 4596 observations spread over 7 features. The unit of observation is the combination of origin and destination airport cities, target variable is Number of passengers flying on average per day and the Independent variables are Fare (average one way fare), and Market Share (market fraction of the biggest carrier).

The primary objective of this analysis is to test the causal relationship of the air fare of that route with the average number of passengers per day and the biggest carrier's market share for that route. The secondary objective of this analysis is to recommend a business strategy based on the results of the primary objective.

Since the data is observational without any scope of randomly assigning routes into treatment and control groups, methods such as Instrument Variable, Fixed Effect Regression and Random Effect Regression were used to establish any causality between the dependent variable and the independent variables.

Our conclusion is that with each dollar increase in the airline fare, the passenger count for the trip decreases by 0.45% and With each percent increase in market share by the biggest carrier, the passenger count decreases by 7.2%

Threats to causal inference

- **Omitted Variable:** Unknown variables such as the population of the destination city can affect the passengers traveling on a route

- **Measurement Error:** We assume that there is no measurement error and outlier situations such as system glitches are disregarded for our analysis
- **Simultaneity:** The fare and number of passengers can affect each other simultaneously. If the fare increases then the number of passengers will decrease which will prompt the fare to decrease
- **Selection Bias:** We assume that data collection has been done randomly and is representative of the real world behavior

Assumptions

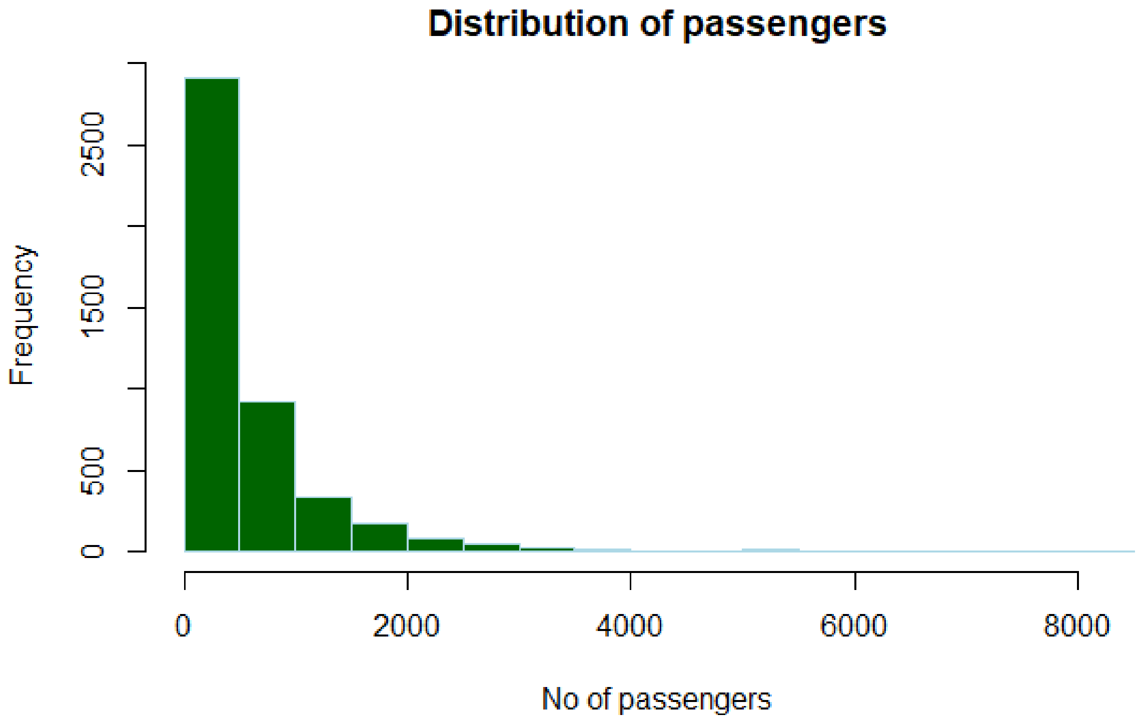
- We assume that our dataset only contains time invariant confounding variables
- We use an instrument variable approach, which has the following assumptions:
  - **Relevance criterion:** The independent variable and Instrument should be correlated, which is called as relevance criterion
  - **Exclusion criterion:** The error term and Instrument should not be correlated. Note that while the relevance criterion can be tested, exclusion criterion cannot be tested.

Data Summary

The dataset used is a panel dataset spanning four years (1997-2000), collected from data sets used in 'Econometric Analysis of Cross Section and Panel Data' By Jeffrey Wooldridge.

It contains 4,596 observations over 7 features such as 'Year', 'Origin Airport', 'Distance' and so on (Appendix 1.1).

The unit of observation is the combination of origin and destination airport cities. The target variable is Number of passengers flying on average per day. The Independent variables are Fare (average one way fare), and Market Share (market fraction of the biggest carrier).



As the target variable is right skewed, we will be taking the log of no. of passengers in the analysis (Appendix 1.2). The column ‘lpassen’ contains the log of the passenger variable.

Analysis

OLS Regression

First, we performed a simple OLS regression to understand how number of passengers is related to fare and biggest carrier’s market share in that route (Appendix 2.1).

However, as the independent variable “fare” and the dependent variable “number of passengers” can affect each other, potential simultaneity is a major endogeneity issue prior to establishing causality in our project. To address this, we are taking an instrumental variable as it can overcome correlation from other source of endogeneity in X and resolve simultaneity and possible omitted variable bias and measurement error.

Instrumental Variables Approach

A good instrument is an exogenous variable that strongly influences the independent variable (variable of interest) and does not impact the final outcome of interest except through the interested independent variable. Therefore, the instrument should not be correlated with the dependent variable Y.

Variable “Distance” which indicates the distance between the origin and destination of the flight from the dataset is observed to be a good instrument based on its high correlation with price (0.63) and no or low correlation with the number of passengers (-0.09) (Appendix 1.3).

2SLS: Two-stage least squares (Appendix 2.2.1)

1. We regress fare onto the distance variable, recover the predicted values, and use them for the second stage of regression

$$\textbf{price}=\pi_0 + \pi_1 * \textbf{ distance } +\varepsilon$$

2. Regress number of passengers onto predicted values of price

$$\textbf{number of passengers}=\alpha+\beta_1*\textbf{biggest carrier market share }+\beta_2*\textbf{price}^{\wedge} + \varepsilon_1$$

The high F score of 2923 in the first regression indicates that distance is indeed a strong instrument.

Note: This gives a consistent estimate of the beta value but the standard errors are wrong as the software doesn’t recognize the usage of predicted price variable

Prior to interpreting the effect of price on the number of passengers, the instrument should be evaluated and this has been done by IV reg diagnostics. The test statistics of the IV reg are described below (Appendix 2.2.2)

Weak IV Test:

This is typically based on first-stage F-stat. P value of <2e-16 ***, implies we reject the null of weak instruments with statistical significance

Wu-Hausman Test:

The null hypothesis is that IV yields equivalent estimates to OLS. The P value of 0.127 implies that we don’t have enough evidence to reject the null hypothesis with statistical significance. Thus we should not use instrument distance because we are losing power by doing so, and getting wider standard errors than necessary

Fixed Effect and Random Effect Regression:

Both methods are good panel data estimation techniques. There might be variables other than the ones present in this dataset that might affect causality. Such confounding variables that may affect causality are dealt with in both the methods but in different ways.

Consider the regression equation -

Passengersit = b0 + b1*Fareit + b2*Mkt. Shareit + b3*Confoundit + eit

where ‘i’ represents each route in the dataset; ‘t’ represents each time period

Fixed Effect Regression: The new intercept becomes (b0 + b3*Confoundi). In fixed effect, the confound term may or may not be correlated with the independent variables (fare and biggest carrier’s market share in this case).

Passengersit = (b0 + b3*Confoundi) + b1*Fareit + b2*Mkt. Shareit + eit

Random Effect Regression: The assumption in this method is that the confounds are time invariant for each unit of observation. The confound term in the above regression equation is treated as a random variable and the confound term is now a part of the error term. The confound term in the above equation is assumed to be uncorrelated with the independent variables (fare and biggest carrier’s market share in this case).

Passengersit = b0 + b1*Fareit + b2*Mkt. Shareit + (eit + b3*Confoundi)

There are three different techniques in fixed effect through which we can eliminate the time invariant confounds − demeaning, dummy variables and first difference. We used demeaning for this project. This technique eliminates the confound variables by subtracting the means from every term of the regression equation, therefore there will be no more intercept term in the regression results.

The resulting equation is as follows:

Fixed Effects (Appendix 2.3.1):

No. of passengers = -0.0045*Fare - 0.072*Biggest Carrier’s Mkt Share + e

Random Effects (Appendix 2.3.2):

No. of passengers = 6.84 - 0.0041*Fare - 0.125*Biggest Carrier’s Mkt Share + e

To decide which method to use to find the casual relationship, Hausman’s test was conducted. The test gave a p-value of 1.281e-12 (Appendix 2.3.3), which means the null hypothesis is rejected, which means that Fixed Effect Regression works best in this case.

Conclusion

From the fixed effect regression results, the observations were:

- With each dollar increase in the airline fare, the passenger count for the trip to be decreased by 0.45%
- With each percent increase in market share by the biggest carrier, the passenger count to decrease by 7.2%

Recommendation

Niche or upcoming airlines should have competitive pricing to increase their passenger volume, to counter the airlines with a bigger market share.

This will have more impact if implemented on routes in which the biggest market share value is high.

Limitations

- Time variant confounding variables could be present which are not addressed
- Data is old and might not be completely applicable to today’s airline industry. It would be helpful to perform this analysis with newer data set
- Since the dataset only spans over 4 years, we might miss out on industry changes that happened over longer periods of time (such as oil wars, market collapses etc.) Thus, in future we can perform this analysis using a longer ranged data set

```
# APPENDIX
# 1. Loading the Data
library(haven)
```

```
## Warning: package ’haven’ was built under R version 4.1.3
```

```
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
library(plm)
```

```
## Warning: package ’plm’ was built under R version 4.1.3

##
## Attaching package: ’plm’

## The following objects are masked from ’package:dplyr’:
##
##     between, lag, lead
```

```
library(dplyr)
library(MASS)
```

```
##
## Attaching package: ’MASS’

## The following object is masked from ’package:dplyr’:
##
##     select
```

```
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.1.3

## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 4.1.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.1.3

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Warning: package 'sandwich' was built under R version 4.1.3

## Loading required package: survival
```

```
library(ggplot2)

set.seed(2020)

# 1. Loading the data

df=read_dta('C:/Users/pankh/Downloads/Causal Inference/airfare (1).dta')
attach(df)

# 1.1 Data description
sapply(df, n_distinct)
```

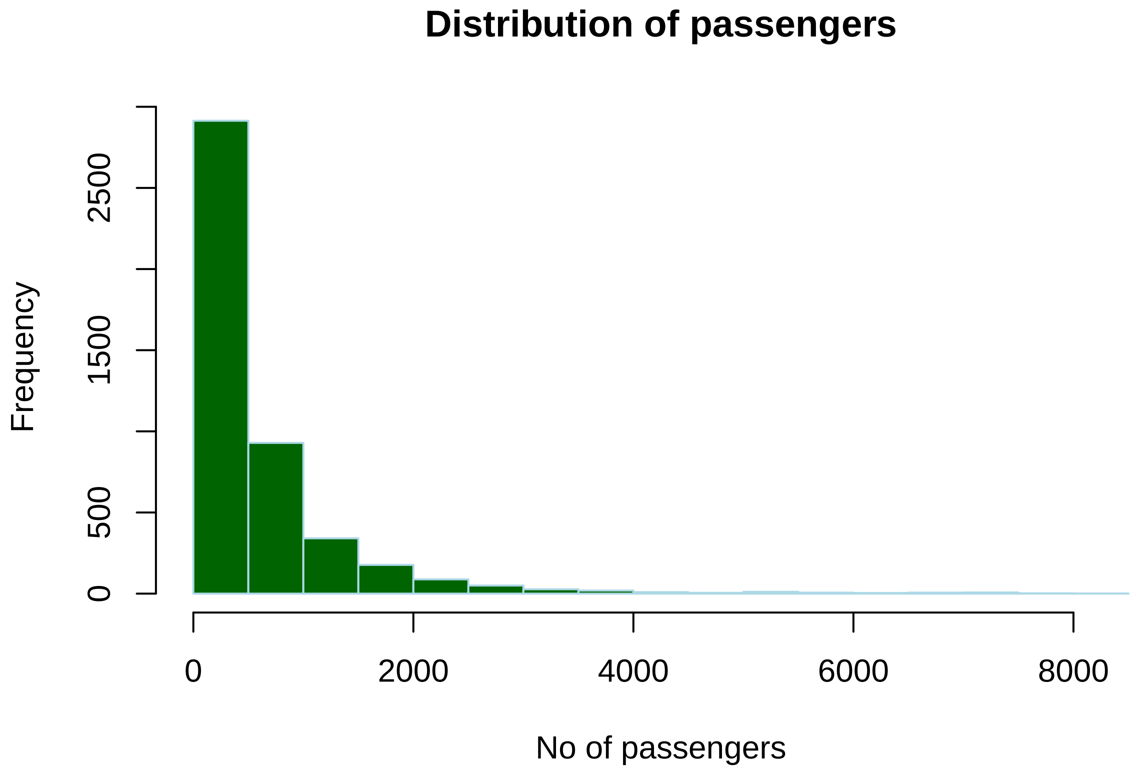
```
##    year  origin  destin      id    dist  passen    fare bmktshr   ldist     y98
##       4      94      97    1149     846    1491     363    3397     846       2
##     y99     y00   lfare ldistsq  concen lpassen
##       2       2     363     846    3397    1491
```

```
nrow(df)
```

```
## [1] 4596
```

```
# 1.2 Passenger variable distribution

hist(passen,xlab = "No of passengers",col = "darkgreen", border = "lightblue",main="Distribution of pass
```



```
# 1.3 Checking the correlation between fare & distance and log of passengers and distance.

cor(df$fare,df$dist) # 0.62559
```

```
## [1] 0.623569
```

```
cor(df$lpassen,df$dist) #-0.09122831
```

```
## [1] -0.09122831
```

```
# 2. Analysis
# 2.1 Running a basic OLS regression.

model1 = lm(lpassen ~ bmktshr+ fare, data = df)
summary(model1)
```

```
##
## Call:
## lm(formula = lpassen ~ bmktshr + fare, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1793 -0.6261 -0.1050  0.5558  3.1447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.6367281  0.0570620 116.307  < 2e-16 ***
## bmktshr     -0.3606081  0.0663977  -5.431 5.89e-08 ***
## fare        -0.0022354  0.0001742 -12.834  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8679 on 4593 degrees of freedom
## Multiple R-squared:  0.03649,    Adjusted R-squared:  0.03607
## F-statistic: 86.97 on 2 and 4593 DF,  p-value: < 2.2e-16
```

```
# But there this does not take the inherent endogeneity into account. Thus, we use the instrumental var

# 2.2 IV Approach

# 2.2.1 Let's first conduct 2SLS manually.
x_simul <- lm(fare ~ dist)$fitted.values
summary(lm(fare ~ dist))
```

```
##
## Call:
## lm(formula = fare ~ dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -127.27  -47.08  -15.47   41.52  233.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.033e+02  1.643e+00   62.87   <2e-16 ***
## dist        7.632e-02  1.412e-03   54.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.55 on 4594 degrees of freedom
## Multiple R-squared:  0.3888, Adjusted R-squared:  0.3887
## F-statistic:  2923 on 1 and 4594 DF,  p-value: < 2.2e-16
```

```
ols_corrected <- lm(lpassen ~ x_simul)
summary(ols_corrected)
```

```
##
## Call:
## lm(formula = lpassen ~ x_simul)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3876 -0.6444 -0.1217  0.5594  3.1187
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.3258265  0.0513996 123.071  < 2e-16 ***
## x_simul     -0.0017271  0.0002781  -6.209 5.79e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8804 on 4594 degrees of freedom
## Multiple R-squared:  0.008323,   Adjusted R-squared:  0.008107
## F-statistic: 38.55 on 1 and 4594 DF,  p-value: 5.793e-10
```

```
# 2.2.2 Using the ivreg package. Note that the beta values are correctly estimated but the standard err

iv_test <- ivreg(formula=lpassen ~ fare | dist)
summary(iv_test,diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = lpassen ~ fare | dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2389 -0.6292 -0.1183  0.5556  3.1659
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.3258265  0.0508470 124.409  < 2e-16 ***
## fare        -0.0017271  0.0002752  -6.277 3.78e-10 ***
##
## Diagnostic tests:
##                     df1  df2 statistic p-value
## Weak instruments    1 4594  2922.832  <2e-16 ***
## Wu-Hausman          1 4593     2.326   0.127
## Sargan              0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8709 on 4594 degrees of freedom
## Multiple R-Squared: 0.02953, Adjusted R-squared: 0.02932
## Wald test:  39.4 on 1 and 4594 DF,  p-value: 3.776e-10
```

```
# Strong instrument but we fail to reject the Hausman test. Therefore, we do not not use IV as the resu

# 2.3 Fixed Effects and Random Effects

# 2.3.1 Fixed effect
within_reg = plm(lpassen ~ fare + bmktshr, data = df, index="id", effect='individual', model="within")
summary(within_reg)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lpassen ~ fare + bmktshr, data = df, effect = "individual",
##     model = "within", index = "id")
##
## Balanced Panel: n = 1149, T = 4, N = 4596
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -2.1197997 -0.0613967 -0.0024255  0.0578353  1.9839303
##
## Coefficients:
##            Estimate  Std. Error t-value Pr(>|t|)
## fare    -0.00454284  0.00014086 -32.252   <2e-16 ***
## bmktshr -0.07116323  0.04618081  -1.541   0.1234
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Total Sum of Squares:    128.1
## Residual Sum of Squares: 98.24
## R-Squared:      0.23309
## Adj. R-Squared: -0.022913
## F-statistic: 523.536 on 2 and 3445 DF, p-value: < 2.22e-16
```

```
# 2.3.2 Random effect

random_reg <- plm(lpassen ~ fare + bmktshr, data = df, index="id", effect="individual", model="random")
summary(random_reg)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora’s transformation)
##
## Call:
## plm(formula = lpassen ~ fare + bmktshr, data = df, effect = "individual",
##     model = "random", index = "id")
##
## Balanced Panel: n = 1149, T = 4, N = 4596
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.02852 0.16887 0.038
## individual    0.72419 0.85099 0.962
## theta: 0.9013
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -2.4054084 -0.0827201 -0.0042812  0.0841679  1.6904628
##
## Coefficients:
##                Estimate  Std. Error  z-value Pr(>|z|)
## (Intercept)  6.83935402  0.04363492 156.7404  < 2e-16 ***
## fare        -0.00417470  0.00013114 -31.8339  < 2e-16 ***
## bmktshr     -0.12440872  0.04385923  -2.8365  0.00456 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    161.85
## Residual Sum of Squares: 132.39
## R-Squared:      0.18204
## Adj. R-Squared: 0.18169
## Chisq: 1022.2 on 2 DF, p-value: < 2.22e-16
```

```
# 2.3.3 Wu Hausman test

phtest(within_reg, random_reg)
```

```
##
##  Hausman Test
##
## data:  lpassen ~ fare + bmktshr
## chisq = 54.766, df = 2, p-value = 1.281e-12
## alternative hypothesis: one model is inconsistent
```