Midterm project

Longhao

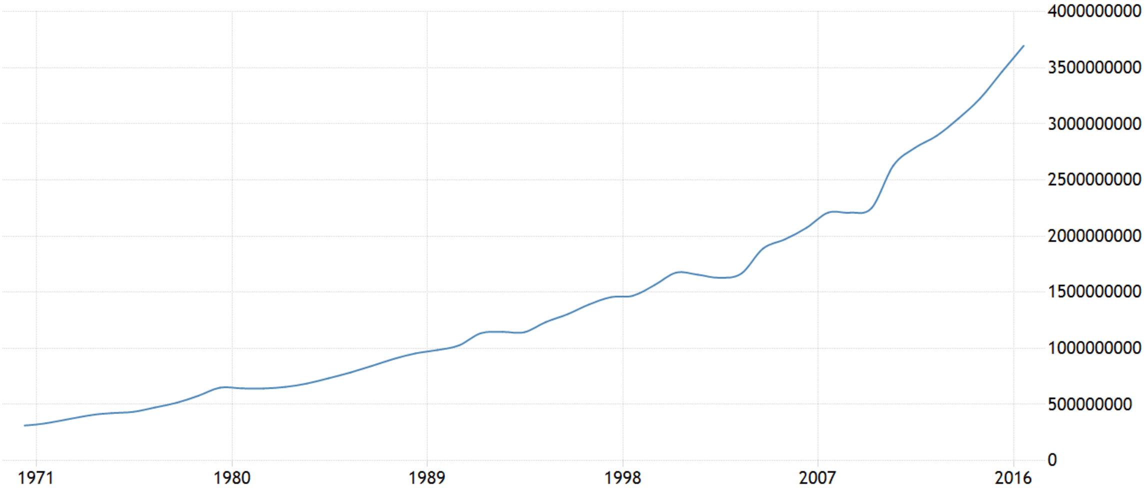
11/14/2018

Abstract

Forcasts are important for making business decisions such as knowing how much to produce. For many aircraft manufacture companies, even a 1% increase in forcasting accuracy means an increase of revenue for millions of dollars. In this project, two models are developed as a stepstone to predict the number of delievered aircrafts in the future. The first model uses time series to predict the yearly new aircrafts export values. The second one applies macro economic variables to predict the yearly carried passenger for each country by using multi level mixed effect model.

Introduction

By any measure, the commercial aviation sector is soaring. More people are taking to the air than ever before, as aviation industry has now recorded eight straight years of steady and above-trend growth. In the plot below, we can see the number of passengers carried in a certain year grew faster and faster, especially after 2010.



WORLDBANK | TRADINGECONOMICS.COM

This is the link to the image https://tradingeconomics.com/world/air-transport-passengers-carried-wb-data.html

In the past few months, there was a earthshaking trade war between China and United States. Among with many other counterattack policy published, one of them was a 25 percent tariff on U.S. civil aircraft with an empty weight of 15,000-45,000 kilograms, which is targeting below the Boeing model line. This is more like a warning shot to the U.S. Administration to proceed no further but it stired some concerns in aviation industry.

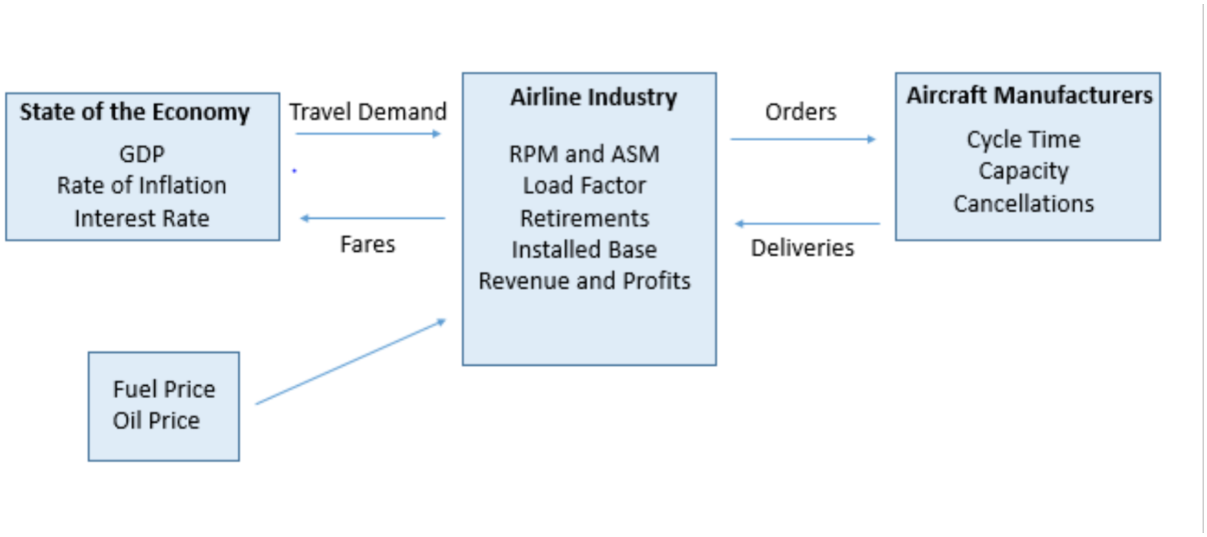


Figure 1: Number of passengers carried by aircrafts

Methods

As a former commercial pilot. I am interested to find out if there is any potensial impact of this tariff by using time seires prediction. It is important to notice that aircraft deliever has a very long cycle. For example, an order for four aircrafts can take several years before the last aircraft is delievered. Therefore, it is reletively not easy to draw a conclusion on the impact of tariff given the time elapsed between the Chinese government published its policy and now(only several months).

After the time series prediction, I shall shrift the focus to multilevel prediction model on number of passengers carried each year for an individual country. To keep this mid-term project relatively sussinct, I choose not to predict the number of aircrafts delievered each year but number of passengers each year. This is because predicting the number of aircrafts delievered entails factors like the size of aircraft, retirements, cancellation of order etc. What I want to focus on is the travelling demand. Like the graph below indicats, there is a dynamic relationship between state of the economy and airline industry.

Materials(Input variables)

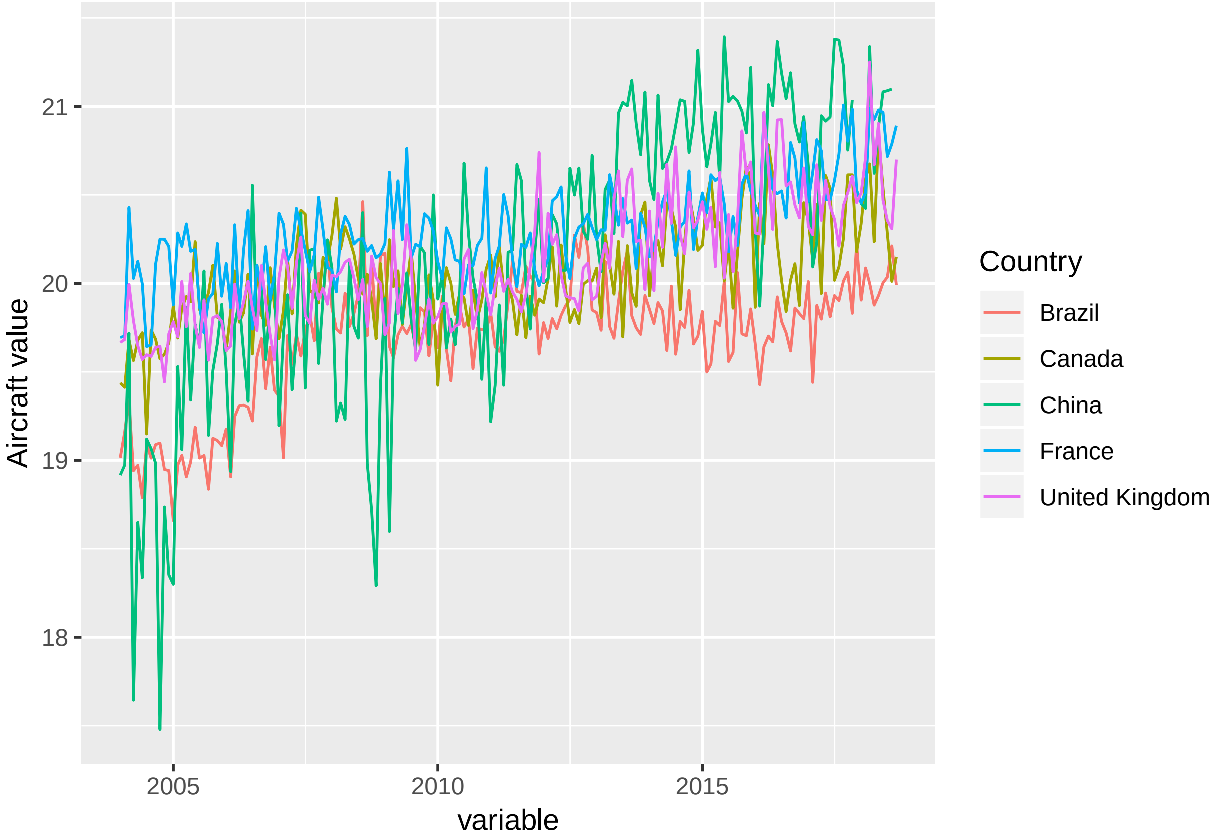
Since we are only interested in the traveling demand of a certain country. We will look at the following variables: GDP-worldwide per capita, GDP growth rate, interest rate, jet fuel price, crude oil price, and rate of inflation. Besides these variables, I would also add a group indicators for different income group countries from high income to low imcome according to world bank website. Another group indicator is region which describes the geographical position of the country. One of interesting predictor variables is the coastline length in the model. Generally, a country with longer coastline length has bigger land areas. Therefore, people are more likely to take commercial aircraft since other transportation could be slow or inconvenient. For example, cars or trains can not trespass over the sea.

The data used to construct the model analysis are from four sources, U.S. Census Bureau, World Bank open Data , U.S. Energy Information, and central intelligence agency.

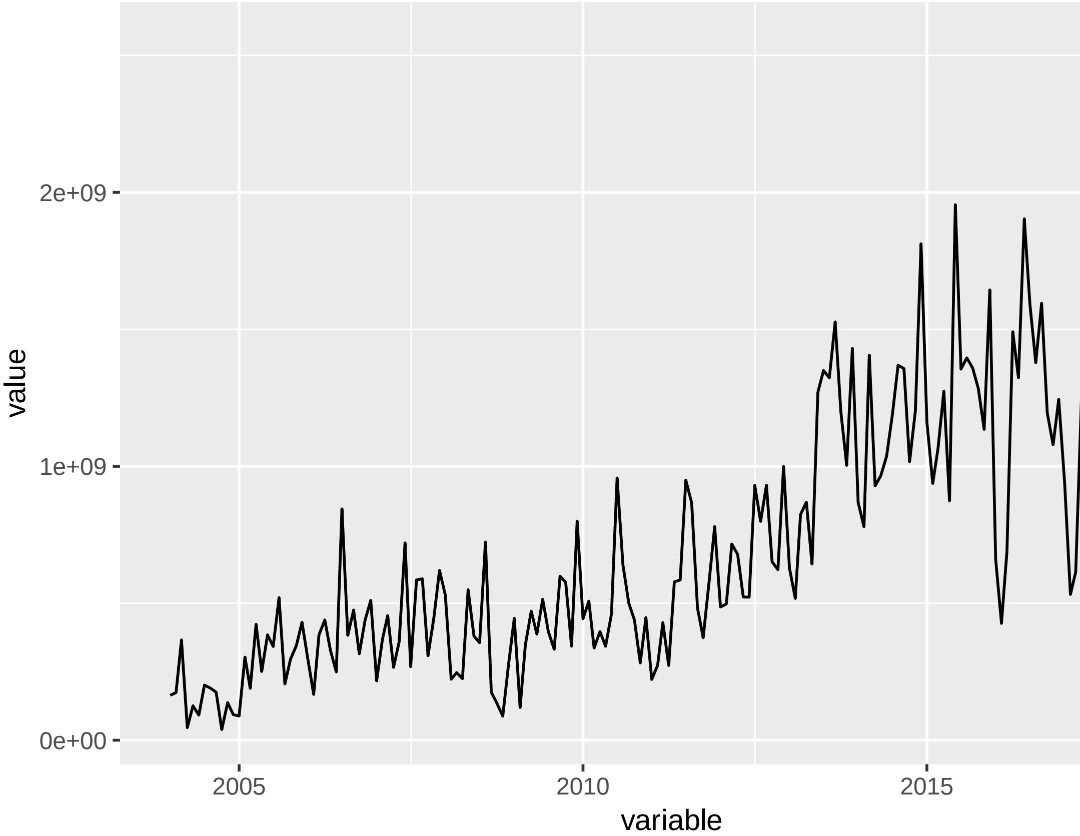
```
## Warning: Missing column names filled in: 'X5' [5]
```

```
## Warning: 2 parsing failures.
## row # A tibble: 2 x 5 col     row col          expected   actual   file
```

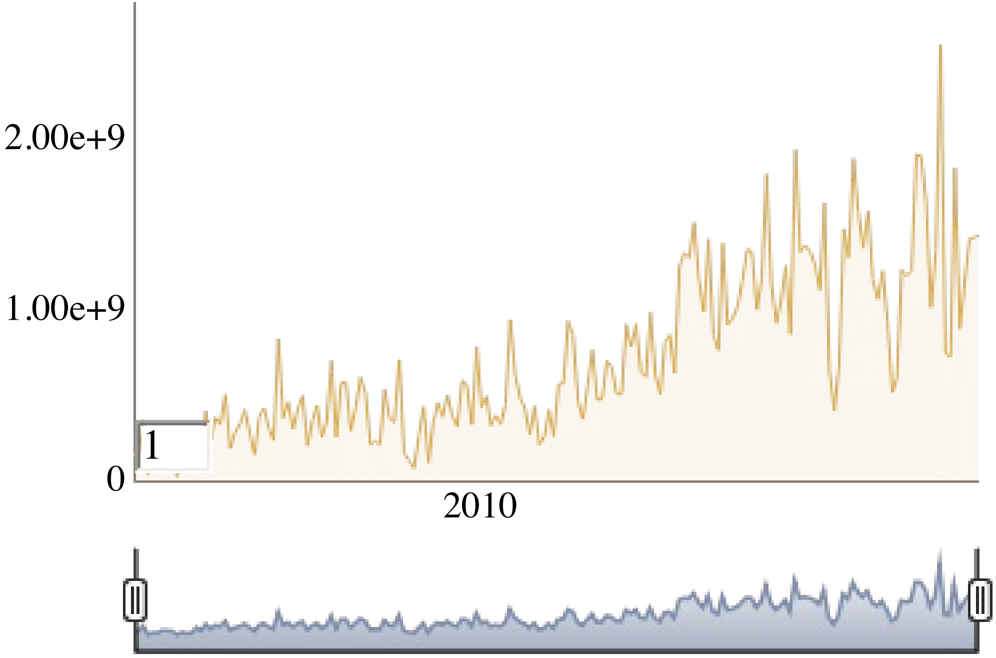
Next let's take a look at the export values to top 5 countries on January 2014. They are France, United Kingdom, Canada, Brazil, China. We can see an overall increase of export aircraft values for these countries as well as some seasonality in aviation industry.



In the next step, we take a closer look at the export aircraft value to China. I have made an interactive ggplot so

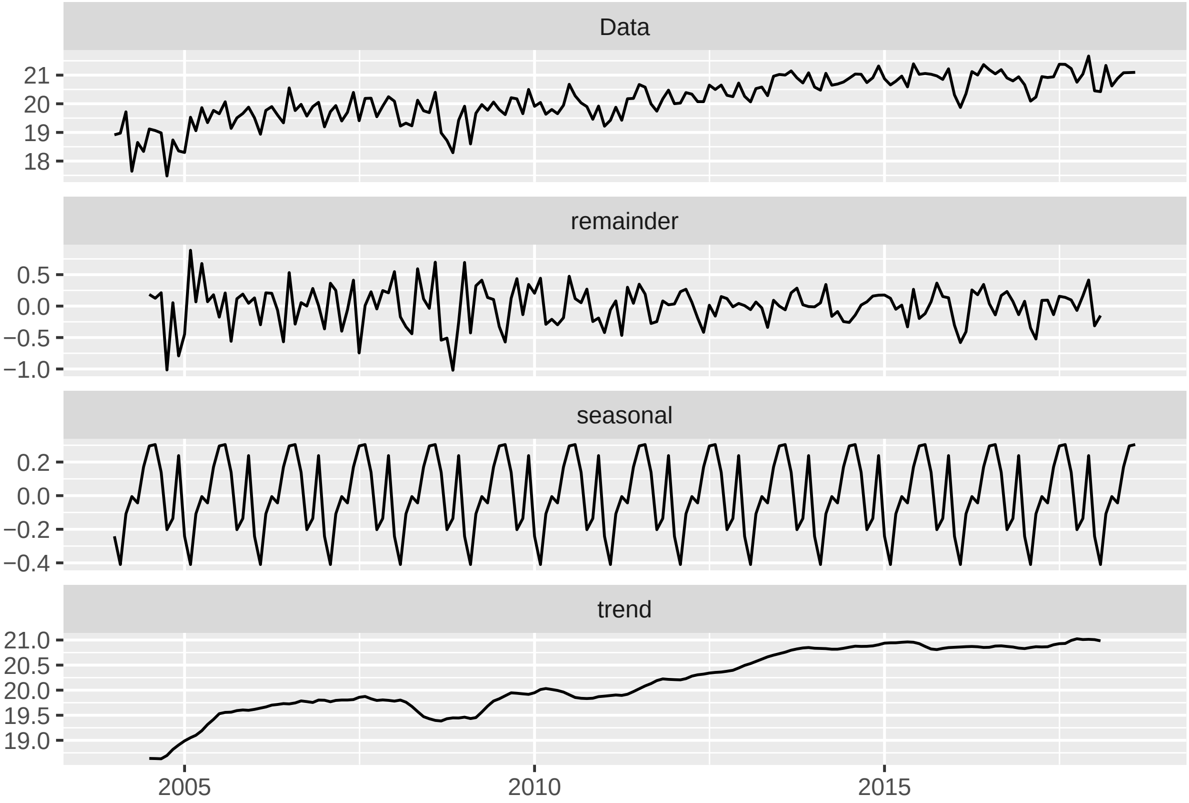


that customers can play with it.

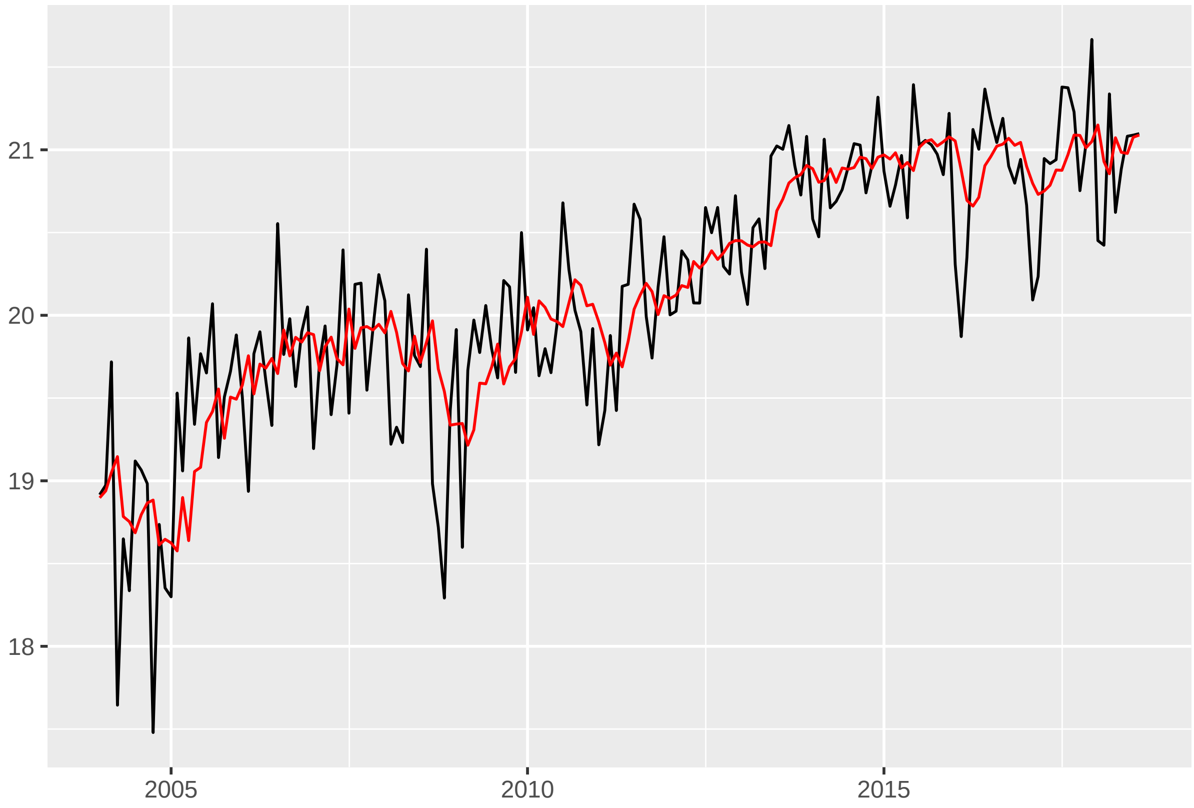


It looks pretty chaos, so I'd like to use time series to decompose. From these fours parts, we can see that the remainder fluctuates around 0 value and the trend increases from 2005 to 2018. There is a slightly drop of trend around 2008. This coincodes with the finacial crisis of 2007-2008.

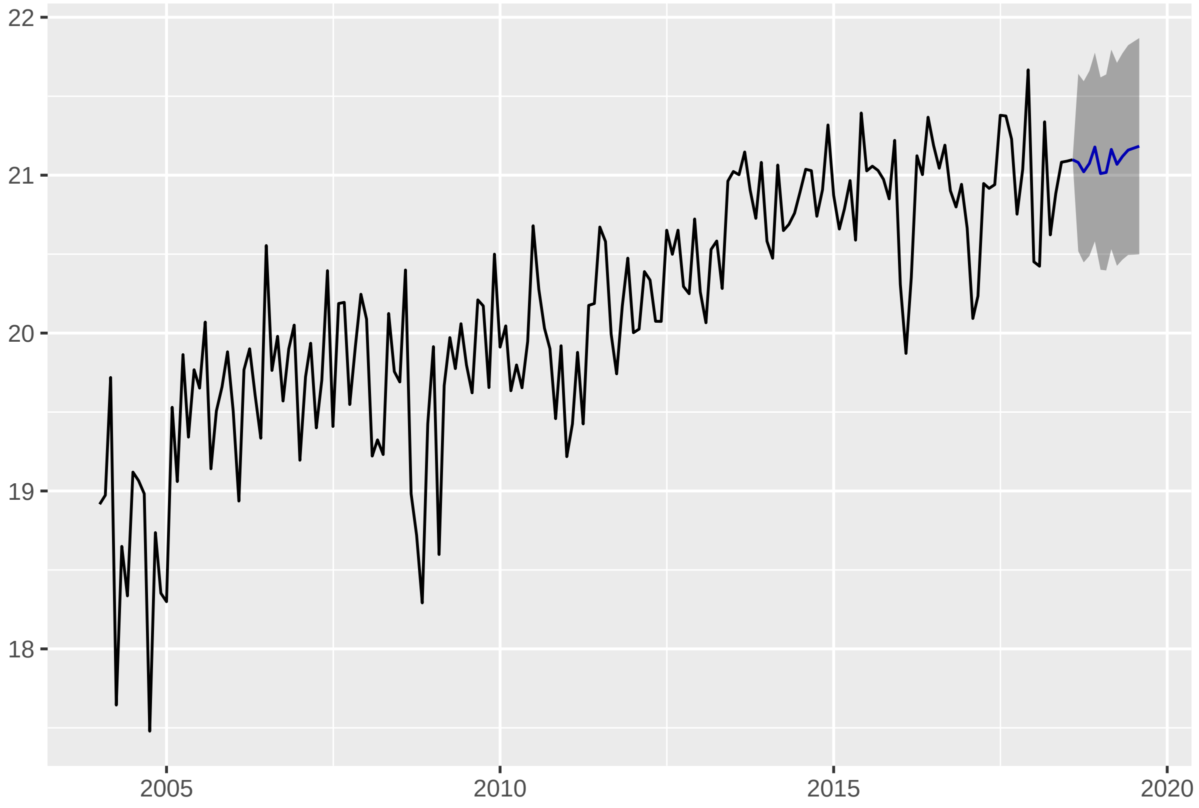
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
## Warning: Removed 24 rows containing missing values (geom_path).
```



We are interested in forcast the same period from 2004 January to 2018 September by time-series using auto.arima function. As we can see the red line is the prediction whereas the black line are actual values of aircraft exporting to China. In general, our prediction matches with the trend of export aircraft values.

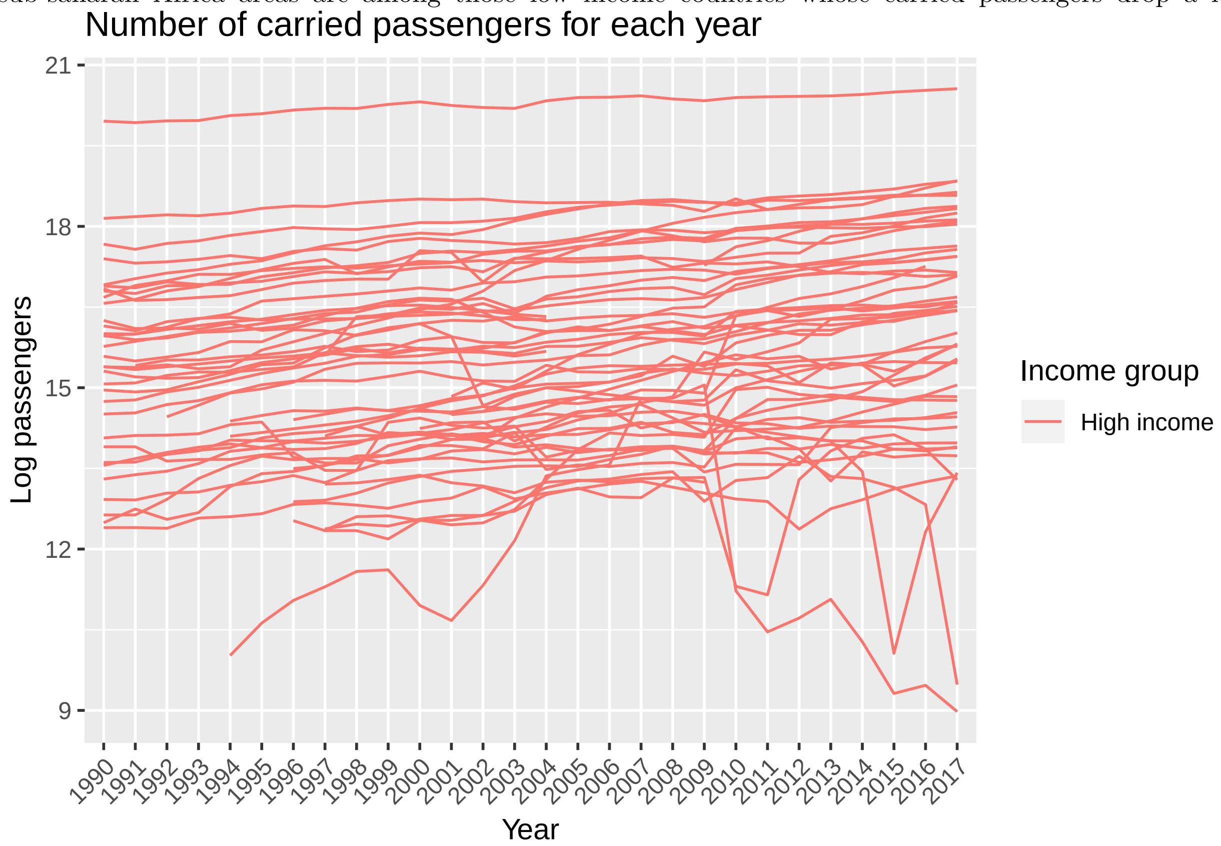


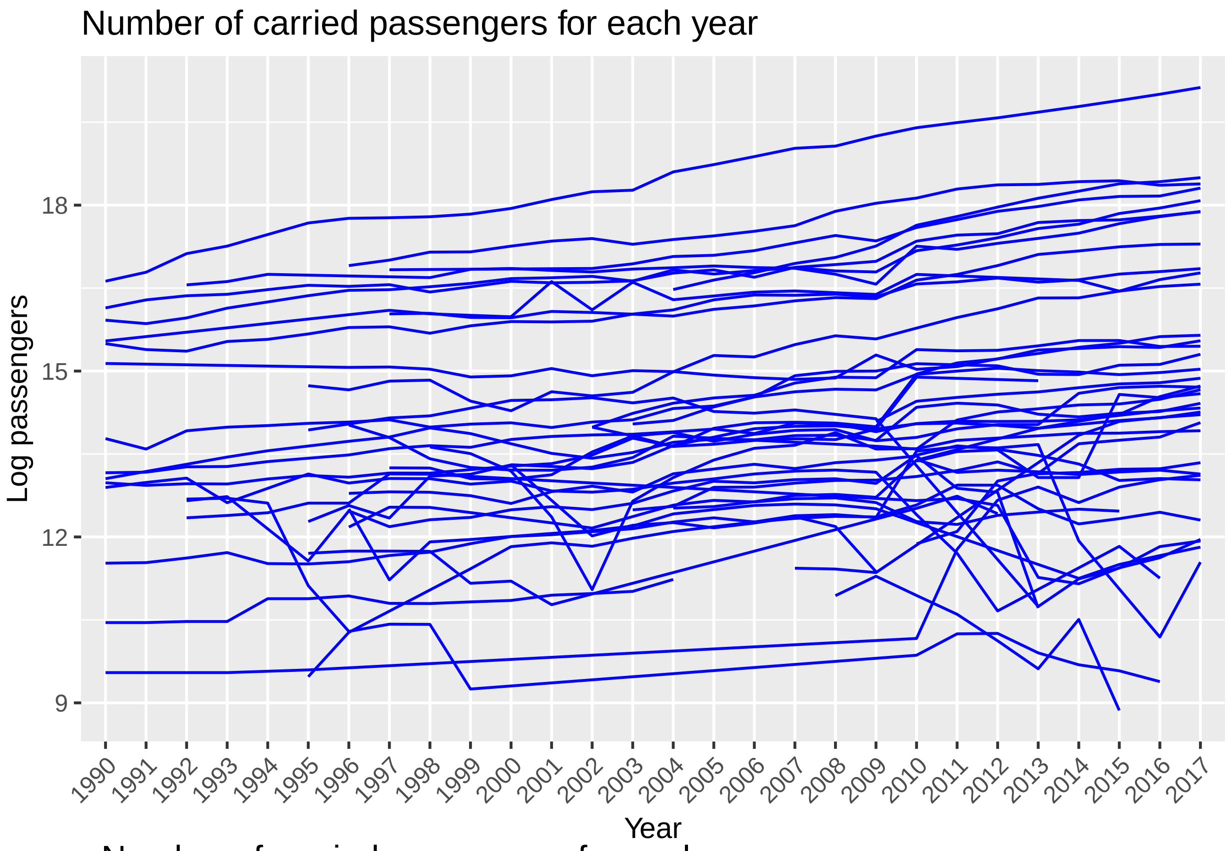
In the plot below, we can see the prediction of value of aircrafts exporting to China in the next 12 months. The grey area is the 95% confidence interval area.

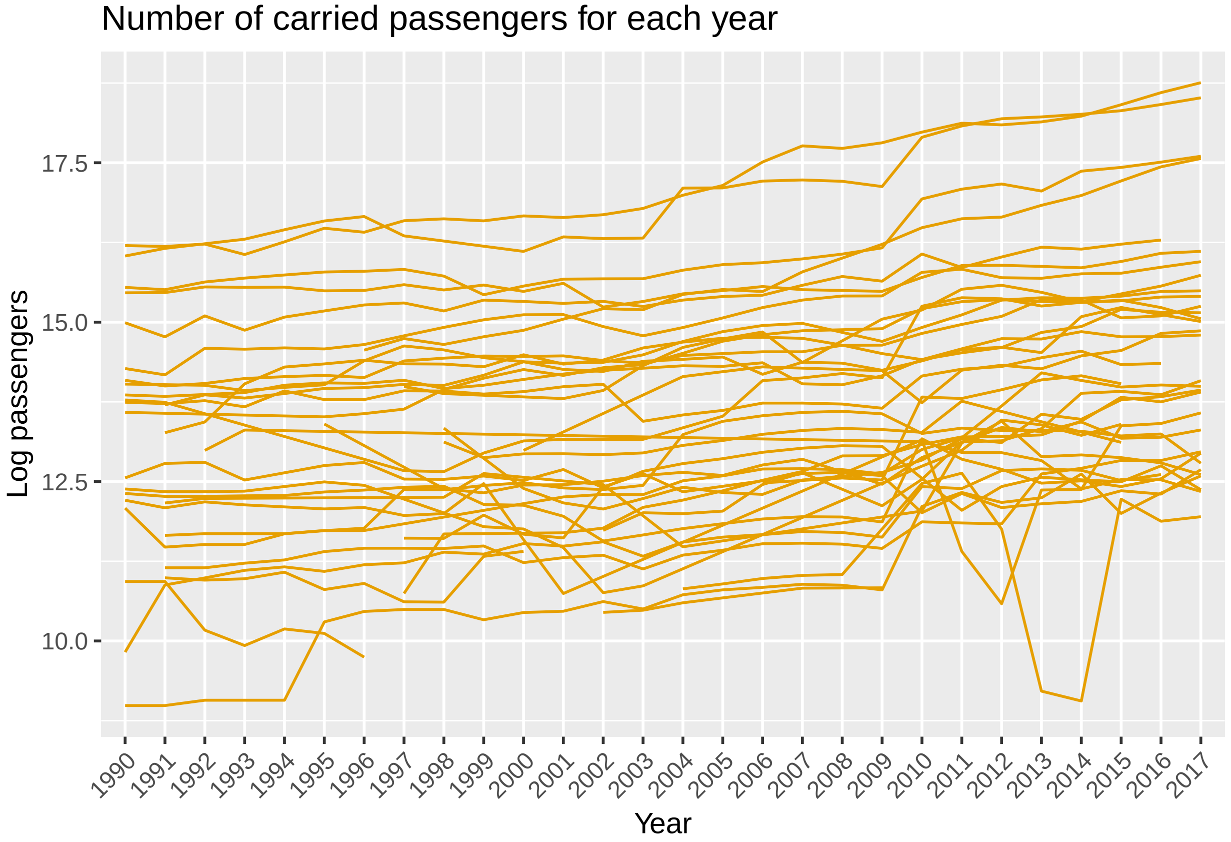


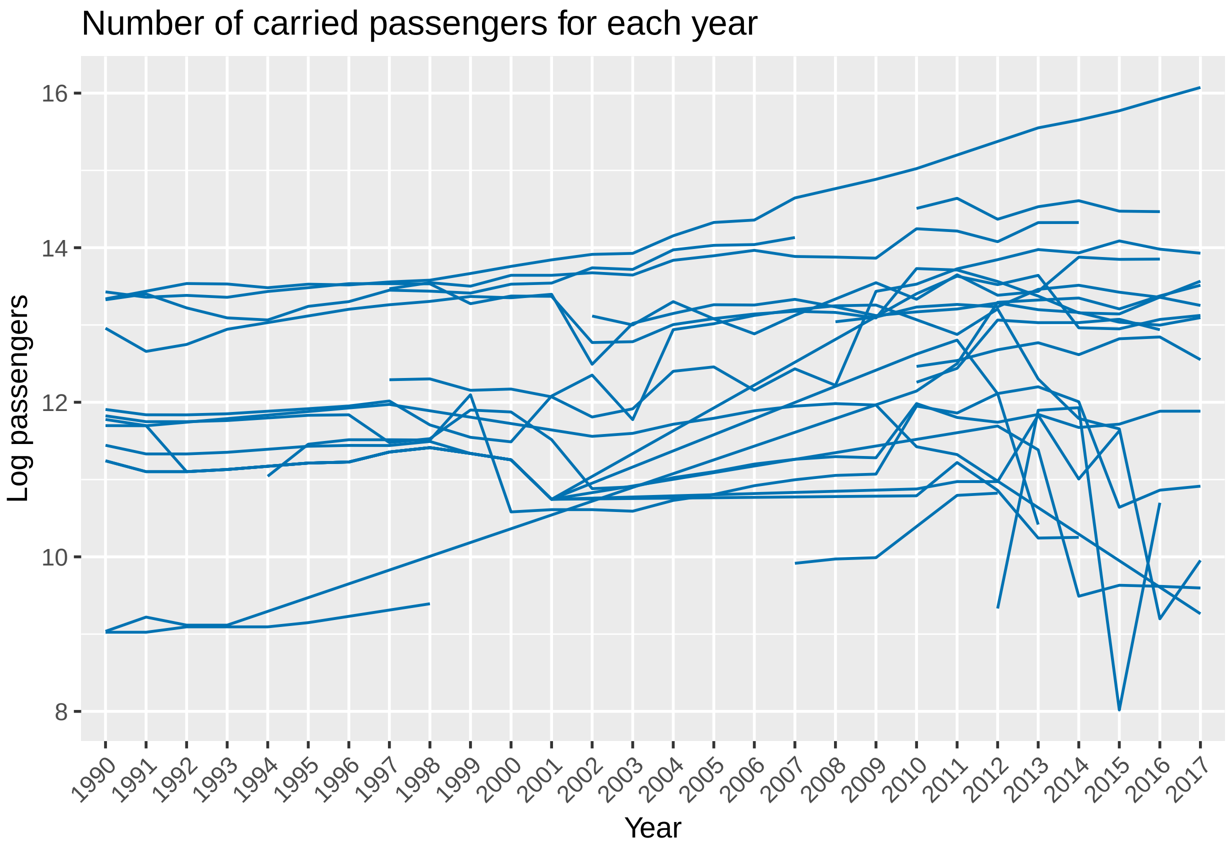
```
## Warning: Missing column names filled in: 'X63' [63]

## Warning: Missing column names filled in: 'X63' [63]

## Warning: Missing column names filled in: 'X63' [63]

## Warning: Missing column names filled in: 'X63' [63]
## Warning: Column `variable` joining factors with different levels, coercing
## to character vector
## Warning: Column `variable` joining character vector and factor, coercing
## into character vector

## Warning: Column `variable` joining character vector and factor, coercing
## into character vector

## Warning: Column `variable` joining character vector and factor, coercing
## into character vector
```

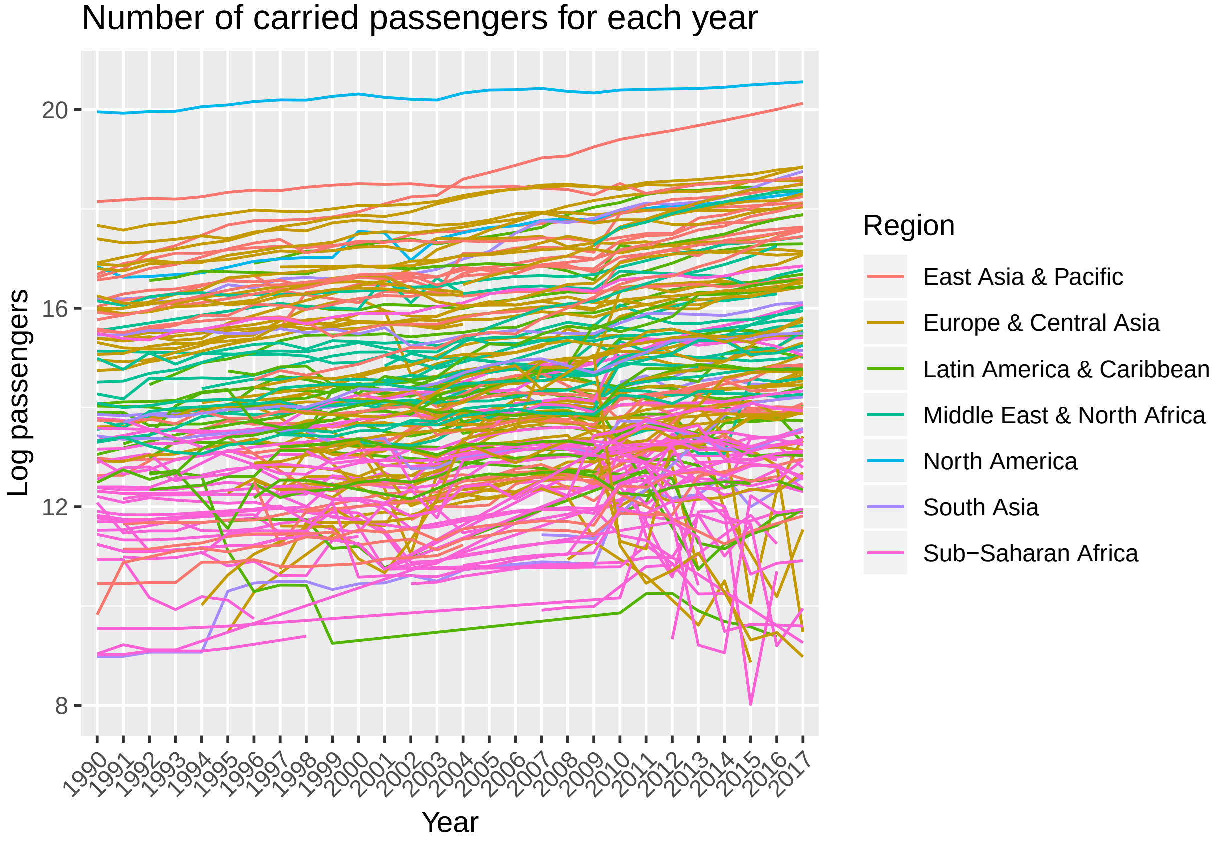
After checking the export values of aircraft, we will look at the relationship between time and passengers from the plot. Most countries experience an increase of carried passengers over the past 30 years. Some low income and lower middle income countries experienced a drop of carried passengers after 2012. With further investigation we can see from the second plot that countries from sub-saharan Africa areas are among those low income countries whose carried passengers drop a lot.



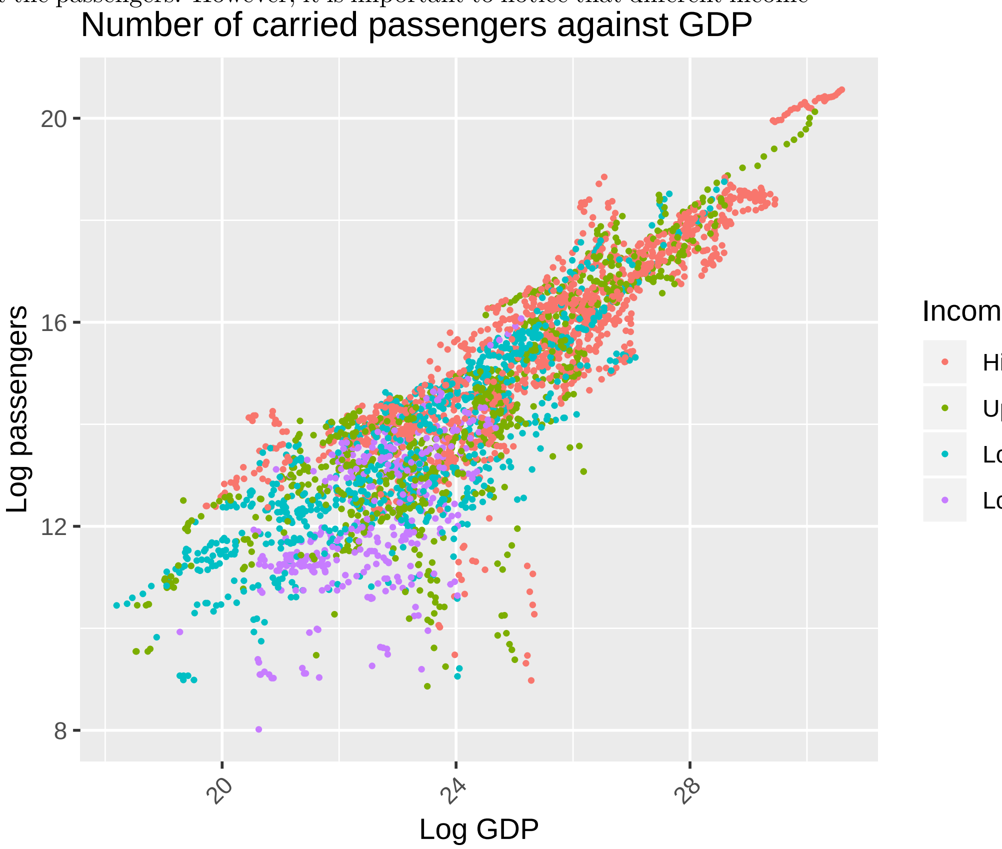

Number of carried passengers for each year
Log passengers
18
15
12
9
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year


Number of carried passengers for each year
Log passengers
17.5
15.0
12.5
10.0
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year

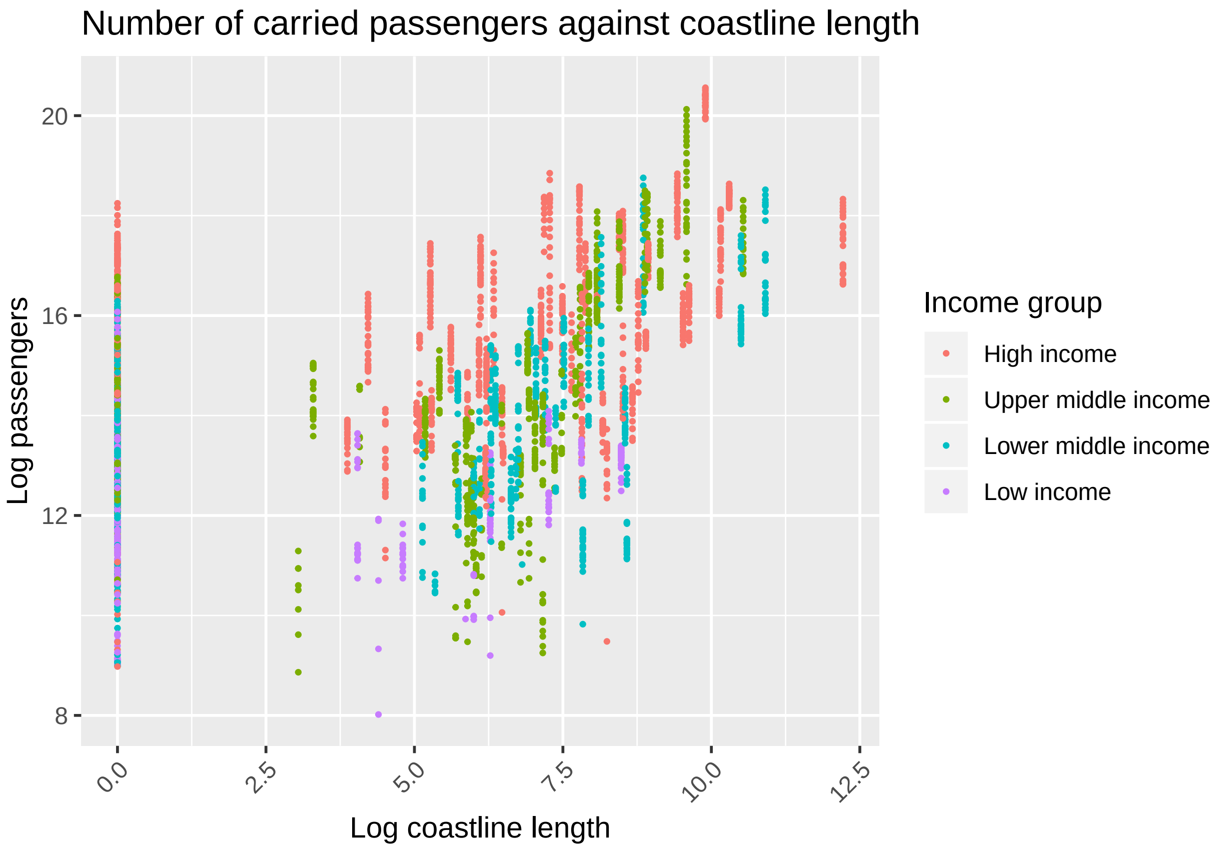

Number of carried passengers for each year


Number of carried passengers for each year

From this plot we can see a strong positive relationship between GDP and passengers. Therefore, it is tentative to use linear regression model to predict the passengers. However, it is important to notice that different income groups of countries behave differently.



This plot describes a relationship between coastline length and number of passengers in a certain year. We can also see a positive relationship between coastline length and passengers.



Since Airbus and Boeing company keep their predicting models as a commercial secret, our only available recources are from previous research done on forecasting commercial airline demand. For example, Jacobson(1970) used a linear regression model to predict trips starting from an airport through two independent variables: average income and airfare. He prediction has an R squared value of 0.82. Another example was from Haney who used some socioeconomic variables to represent the city surrending the airport. He used population, total personal income, fares, distance, time, highway miles, passenger originations.

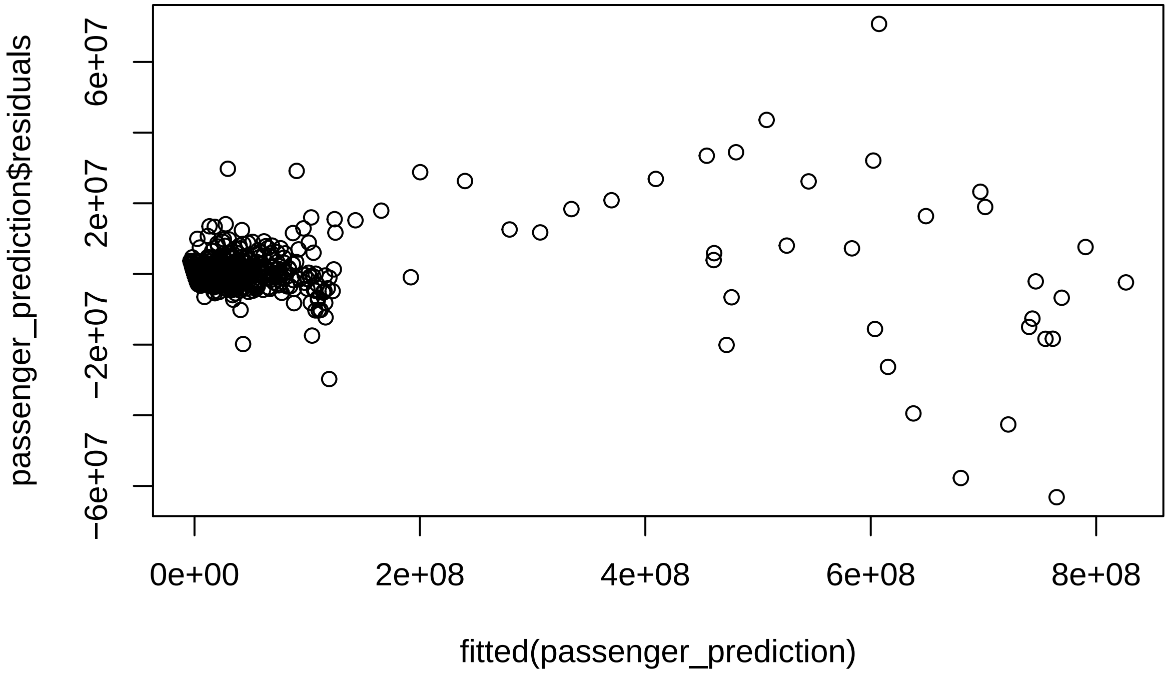
The forecasting model that is of first interest is linear regression model. We will first look at the linear regression without log transformation.

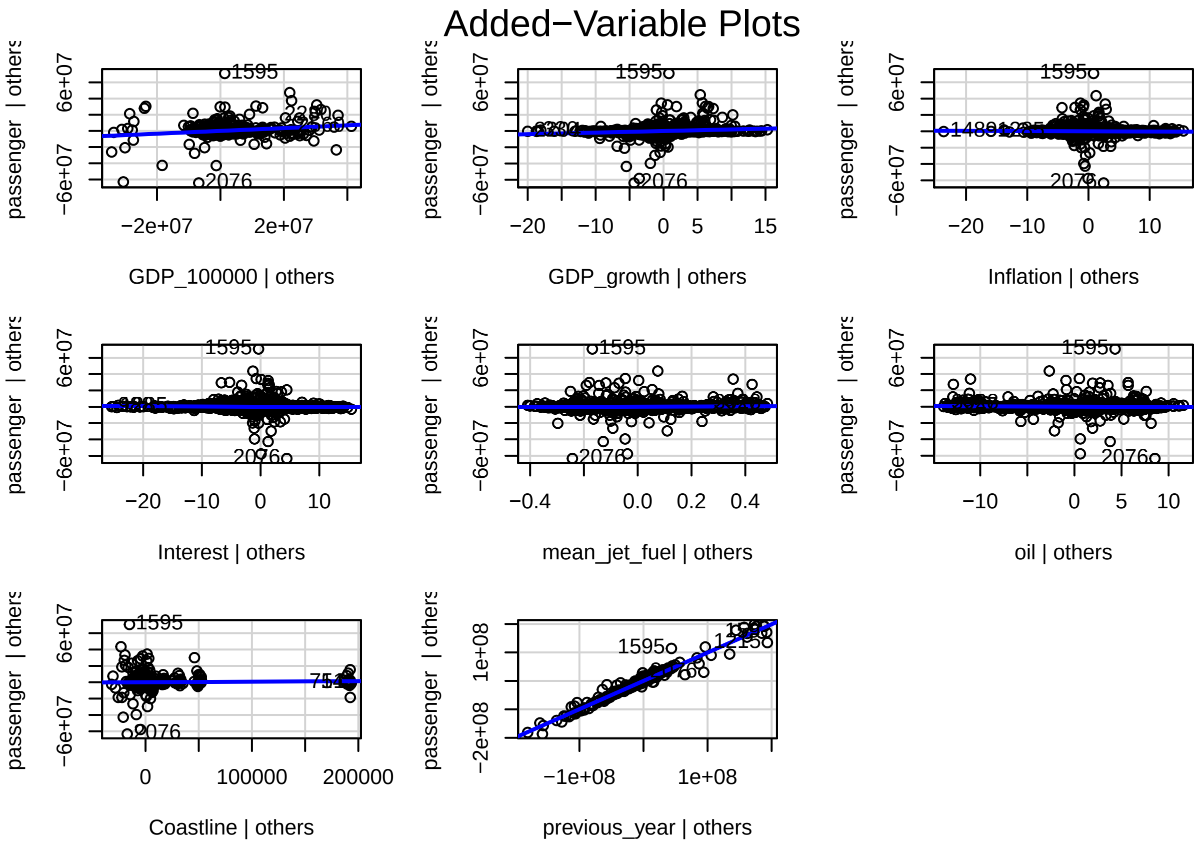
We can see a strong relationship between GDP and passengers. Every 100,000 increase of total GDP is associated with an increase of 4.6 passengers. Likewise, every 1 percent increase of GDP growth rate is associated with an increase of around 380000 passengers in one years. Both inflation and interest rate are associated with an increase of passengers. These results fall into my expectation since I would expect that inflation encourages people to spend more money and interest rate indicates that the economy is in healthy condition. However, the two standard error for both variables cross 0. Jet fuel price has a positive relationship with number of passengers. This does not agree with our intuition since we would expect them to have negative relationship. Crude oil price has a negative relationship with the number of passengers with a coefficient of 1.371e+05. The last variable is log of coastline. Unfortunately, the coeffecient again does not meet my expectation because it indicates a negative relationship with number of passengers. This model has an adjusted R-sauqred value of 0.9175.

```
##
## Call:
## lm(formula = passenger ~ GDP_100000 + GDP_growth + Inflation +
##     Interest + mean_jet_fuel + oil + Coastline + previous_year,
##     data = a)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -63178133   -737902   -232577    393263  70758101
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.743e+05  3.625e+05   0.757   0.4493
## GDP_100000     1.727e-01  2.000e-02   8.634  < 2e-16 ***
## GDP_growth     1.957e+05  2.602e+04   7.520 8.01e-14 ***
## Inflation     -2.565e+04  2.255e+04  -1.138   0.2555
## Interest      -3.276e+04  1.702e+04  -1.924   0.0545 .
## mean_jet_fuel  9.142e+05  5.643e+05   1.620   0.1054
## oil           -3.379e+04  1.976e+04  -1.710   0.0875 .
## Coastline      6.659e+00  4.058e+00   1.641   0.1009
## previous_year  9.944e-01  4.271e-03 232.829  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4370000 on 2123 degrees of freedom
##   (1186 observations deleted due to missingness)
## Multiple R-squared:  0.9969, Adjusted R-squared:  0.9969
## F-statistic: 8.533e+04 on 8 and 2123 DF,  p-value: < 2.2e-16
```

This is to look at the marginal effect of the predictors. We can see that GDP and passenger follows a good linear relationship with each other. There are some input variables that do not follow a linear relationship such as interest rate.





By taking the log of some imput variables and output variable, we get a different coefficient values after remodelling. GDP growth rate along with inflation and interest rate all indicate a negative relationship with number of passengers, which do not make much sense. Jet fuel has a negative relationship with number of passengers. This model is probably not as good as the one before because by taking the log of GDP and passenger, we highlight the importance of other variables.

```
##
## Call:
## lm(formula = log_passenger ~ log_GDP + GDP_growth + Inflation +
##     Interest + mean_jet_fuel + oil + log_coastline, data = a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7644 -0.5198  0.0977  0.6716  2.5105
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.751e+00  2.663e-01 -14.089  < 2e-16 ***
## log_GDP        7.485e-01  1.041e-02  71.936  < 2e-16 ***
## GDP_growth    -7.704e-04  5.850e-03  -0.132    0.895
## Inflation     -3.526e-02  5.079e-03  -6.942 5.11e-12 ***
## Interest      -1.990e-02  3.917e-03  -5.079 4.12e-07 ***
## mean_jet_fuel -1.250e-01  1.271e-01  -0.983    0.325
## oil            8.376e-05  4.434e-03   0.019    0.985
## log_coastline  9.464e-02  6.392e-03  14.806  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9873 on 2144 degrees of freedom
##   (1166 observations deleted due to missingness)
## Multiple R-squared:  0.796,  Adjusted R-squared:  0.7953
## F-statistic:  1195 on 7 and 2144 DF,  p-value: < 2.2e-16
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: passenger ~ GDP_100000 + GDP_growth + Inflation + Interest +
##     mean_jet_fuel + oil + IncomeGroup + (1 | Region)
##    Data: a
##
## REML criterion at convergence: 78289.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.9745 -0.1225  0.0247  0.2275  7.9944
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  Region   (Intercept) 8.332e+14 28864590
##  Residual             4.076e+14 20187959
## Number of obs: 2152, groups:  Region, 7
##
## Fixed effects:
##                                  Estimate Std. Error t value
## (Intercept)                     7.330e+06  1.106e+07   0.663
## GDP_100000                      4.199e+00  3.419e-02 122.828
## GDP_growth                      6.585e+05  1.220e+05   5.398
## Inflation                       1.863e+05  1.135e+05   1.641
## Interest                        3.930e+04  8.368e+04   0.470
## mean_jet_fuel                   4.395e+06  2.621e+06   1.677
## oil                            -2.011e+05  9.113e+04  -2.207
## IncomeGroupUpper middle income  5.241e+06  1.200e+06   4.366
## IncomeGroupLower middle income  4.699e+06  1.439e+06   3.266
## IncomeGroupLow income           4.234e+06  1.991e+06   2.127
##
## Correlation of Fixed Effects:
##             (Intr) GDP_10 GDP_gr Infltn Intrst mn_jt_ oil    InGUmi InGLmi
## GDP_100000  -0.045
## GDP_growth  -0.025  0.025
## Inflation   -0.022  0.037 -0.064
## Interest    -0.062  0.042 -0.008  0.209
## mean_jet_fl  0.087 -0.055  0.085  0.192 -0.005
## oil         -0.101  0.040 -0.096 -0.178  0.052 -0.982
## IncmGrpUpmi -0.031  0.055 -0.045 -0.233 -0.138 -0.105  0.076
## IncmGrpLwmi -0.030  0.113 -0.072 -0.356 -0.215 -0.114  0.088  0.515
## IncmGrpLwin -0.026  0.073 -0.092 -0.157 -0.109 -0.098  0.073  0.405  0.456
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

Result:

Discussion

This mid term project implemented two methods in predicting the aircraft export values to China and expected passengers for each country in one year. The multiple regression model used the past behavior of macro economic indicators for prediction. It was important to examine the input variables in multiple regression model because of the correlations between input variables. Highly correlated input variables can make prediction less accurate.

Overall, these two models provide sufficient accuracy in prediction. They are easy to use, and user friendly.

The limitation of regression model is that it does not cover the prediction of aircraft orders and deliveries, which is the primary concern for aviation industry. Another limitation is some of the input variables such as airfare are not incorporated in the prediction model because such data are not easily accessible through internet. Fitting more variables into the prediction model will greatly improve the accuracy. In the future, the next move would be to build on this model to predict the number of ordered aircrafts.

Data scources: https://data.worldbank.org/ https://www.census.gov/ https://www.eia.gov/ https://www.cia.gov/index.html

bibliography Monahan, Kayla M., "Aircraft Demand Forecasting" (2016). Masters Theses. 329. https://scholarworks.umass.edu/masters_theses_2/329 Haney, D. (1975). Review of aviation forecasting methodology. Rep. dot-40176, 6. Boeing (2018) Current Market Outlook 2018-2037