ISE5331 Second Assignment
Aviation Industry Monthly Data Regression Report

November 17, 2025

Ziyang ZHANG
z1yang.zhang@connect.polyu.hk

---

1 Introduction

The aviation industry is a highly data-driven sector, where operational decisions such as flight scheduling, fleet deployment and resource allocation all rely on accurate demand forecasting. Passenger throughput, measured as the total number of passengers handled in a given period, is one of the most important indicators of market demand and operational performance. Understanding how passenger throughput is influenced by operational variables such as flight movements, aircraft utilization and cargo throughput is therefore essential for both policymakers and airline operators.

In this report, we conduct a multiple linear regression analysis using monthly data of China's civil aviation industry from January 2021 to December 2023. The dependent variable is *passenger throughput* (in units of 10,000 persons), and the explanatory variables include: (1) number of flight movements (in units of 10,000 flights), (2) average daily aircraft utilization (hours per day), and (3) cargo and mail throughput (in units of 10,000 tons). The aim is to quantify the relationships between passenger throughput and these key operational variables, and to build a regression model that can be used for prediction and interpretation.

2 Data Description

This study uses officially published monthly statistics from the Civil Aviation Administration of China (CAAC) for the period from January 2021 to December 2023. Each record corresponds to one month and includes the following variables:

- **Passenger** : total number of passengers handled by civil aviation.
- **Flights**: total number of flight takeoffs and landings.
- **Util** : average daily utilization rate of aircraft in hours per day.
- **Cargo** : total cargo and mail throughput.

Table 1 presents the complete dataset used for regression analysis. Units are clearly indicated in the column headers and are consistent with CAAC official publications.

Month	Passenger (10k)	Flights (10k)	Util (h/day)	Cargo (10k t)
2021-01	6213.1	75.0	6.3	158.3
2021-02	4943.3	52.2	5.3	111.1
2021-03	9850.2	96.4	8.2	154.8
2021-04	10524.8	99.2	8.6	157.0
2021-05	10508.4	99.2	8.2	159.7
2021-06	8516.1	80.6	7.0	155.3
2021-07	10138.6	95.2	7.7	147.1
2021-08	4631.5	66.9	4.6	133.6
2021-09	7424.4	83.6	6.7	149.1
2021-10	7978.0	86.9	6.7	150.2
2021-11	4420.8	64.4	4.7	149.1
2021-12	5590.3	72.8	5.6	156.9
2022-01	6105.3	71.7	6.2	154.6
2022-02	6471.4	69.4	6.6	103.3
2022-03	3171.4	53.3	3.5	129.0
2022-04	1625.0	37.7	2.2	102.2
2022-05	2501.2	49.8	3.0	116.2
2022-06	4555.8	64.0	4.7	128.1
2022-07	7046.2	84.1	6.3	128.4
2022-08	6658.3	79.5	5.8	122.1
2022-09	4133.9	60.1	3.9	123.3
2022-10	3274.6	47.2	3.2	116.1
2022-11	2587.4	43.4	2.8	113.7
2022-12	3869.5	51.5	3.9	114.9
2023-01	8183.9	72.9	6.7	113.6
2023-02	8894.6	85.6	7.6	105.4
2023-03	9373.8	98.1	7.5	130.8
2023-04	10279.1	97.9	8.2	126.7
2023-05	10549.7	103.5	8.2	133.9
2023-06	10803.3	100.4	8.3	146.2
2023-07	12670.4	110.1	9.1	138.0
2023-08	12944.4	111.0	9.2	144.3
2023-09	10833.8	101.3	8.4	157.9
2023-10	11353.5	101.1	8.2	152.0
2023-11	9901.0	93.9	7.9	164.3
2023-12	10163.1	94.1	8.0	164.9

Table 1: Monthly aviation data from January 2021 to December 2023.

The data clearly reflect the impact of the COVID-19 pandemic and subsequent recovery. For example, passenger throughput and aircraft utilization dropped significantly in some months of 2022, while both indicators rebounded strongly in 2023 with the recovery of domestic and international travel demand.

3 Code

The full runnable code used in this assignment is also available in an online GitHub repository[1].

In this section, we present the complete Python code used for the regression analysis. The code is written in a clear and modular way, with comments explaining each major step. The implementation relies on the standard data analysis stack in Python: `numpy`, `pandas`, and `statsmodels`.

```
# Import required libraries
# numpy and pandas for data handling; statsmodels for regression;
# matplotlib for visualization.
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

# ------------------------------
# 1. Construct the monthly dataset (2021-01 to 2023-12)
# ------------------------------
data = {
    'Month': pd.date_range("2021-01", periods=36, freq='M'),
    'Passenger': [6213.1, 4943.3, 9850.2, 10524.8, 10508.4, 8516.1,
                  10138.6, 4631.5, 7424.4, 7978.0, 4420.8, 5590.3,
                  6105.3, 6471.4, 3171.4, 1625.0, 2501.2, 4555.8,
                  7046.2, 6658.3, 4133.9, 3274.6, 2587.4, 3869.5,
                  8183.9, 8894.6, 9373.8, 10279.1, 10549.7, 10803.3,
                  12670.4, 12944.4, 10833.8, 11353.5, 9901.0, 10163.1],
    'Flights': [75.0, 52.2, 96.4, 99.2, 99.2, 80.6,
                95.2, 66.9, 83.6, 86.9, 64.4, 72.8,
                71.7, 69.4, 53.3, 37.7, 49.8, 64.0,
                84.1, 79.5, 60.1, 47.2, 43.4, 51.5,
                72.9, 85.6, 98.1, 97.9, 103.5, 100.4,
                110.1, 111.0, 101.3, 101.1, 93.9, 94.1],
    'Util': [6.3, 5.3, 8.2, 8.6, 8.2, 7.0,
             7.7, 4.6, 6.7, 6.7, 4.7, 5.6,
             6.2, 6.6, 3.5, 2.2, 3.0, 4.7,
             6.3, 5.8, 3.9, 3.2, 2.8, 3.9,
             6.7, 7.6, 7.5, 8.2, 8.2, 8.3,
             9.1, 9.2, 8.4, 8.2, 7.9, 8.0],
    'Cargo': [158.3, 111.1, 154.8, 157.0, 159.7, 155.3,
              147.1, 133.6, 149.1, 150.2, 149.1, 156.9,
              154.6, 103.3, 129.0, 102.2, 116.2, 128.1,
              128.4, 122.1, 123.3, 116.1, 113.7, 114.9,
```

[1] GitHub: `https://github.com/SOMNAMBULI1ST/ISE5331_Assignment2Code/tree/main`

```
            113.6, 105.4, 130.8, 126.7, 133.9, 146.2,
            138.0, 144.3, 157.9, 152.0, 164.3, 164.9]
}

# Put the data into a pandas DataFrame
df = pd.DataFrame(data)

# ----------------------------
# 2. Build the multiple linear regression model
#    Dependent variable: Passenger
#    Independent variables: Flights, Util, Cargo
# ----------------------------
X = df[['Flights', 'Util', 'Cargo']]  # design matrix without intercept
y = df['Passenger']                   # response vector

# Add constant term (intercept) to X
X = sm.add_constant(X)

# Fit the OLS regression model
model = sm.OLS(y, X).fit()

# Print full regression summary:
# includes coefficients, standard errors, t-stats, p-values, R-squared, etc.
print(model.summary())

# Also extract predicted values for later plotting
y_pred = model.predict(X)

# ----------------------------
# 3. Visualization 1:
#    Actual vs. Predicted Passenger Throughput
# ----------------------------
plt.figure()
plt.scatter(df['Passenger'], y_pred, label='Predicted vs Actual')

# Plot a 45-degree reference line
min_val = min(df['Passenger'].min(), y_pred.min())
max_val = max(df['Passenger'].max(), y_pred.max())
plt.plot([min_val, max_val], [min_val, max_val],
         linestyle='--', label='Perfect fit line')

plt.xlabel("Actual Passenger Throughput (10k persons)")
plt.ylabel("Predicted Passenger Throughput (10k persons)")
plt.title("Actual vs Predicted Passenger Throughput")
plt.legend()
```

```
plt.tight_layout()

# Save figure to file (used in LaTeX as fig_actual_vs_pred.png)
plt.savefig("fig_actual_vs_pred.png", dpi=300)
plt.close()

# ----------------------------
# 4. Visualization 2:
#    Flights vs. Passenger Throughput (with simple linear fit)
# ----------------------------
plt.figure()
plt.scatter(df['Flights'], df['Passenger'], label='Observed points')

# Fit a simple linear trend line for visualization:
coef = np.polyfit(df['Flights'], df['Passenger'], 1)
x_line = np.linspace(df['Flights'].min(), df['Flights'].max(), 100)
y_line = np.polyval(coef, x_line)
plt.plot(x_line, y_line,
         linestyle='--', label='Linear fit')

plt.xlabel("Flights (10k flights)")
plt.ylabel("Passenger Throughput (10k persons)")
plt.title("Flights vs Passenger Throughput")
plt.legend()
plt.tight_layout()

# Save figure to file (used in LaTeX as fig_flights_vs_passenger.png)
plt.savefig("fig_flights_vs_passenger.png", dpi=300)
```

4 Results and Visualization

In this section, we present the estimation results of the multiple linear regression model and visualize the model performance as well as the relationship between key variables.

Based on the estimated model, the regression equation for passenger throughput (`Passenger`) is:

$$\text{Passenger} = -2530.0236 + 68.4317 \times \text{Flights} + 922.8727 \times \text{Util} - 9.5419 \times \text{Cargo}. \quad (1)$$

The coefficient of determination is $R^2 = 0.976$, indicating that the model explains about 97.6% of the variance in passenger throughput. This suggests an excellent goodness-of-fit.

Coefficient Estimates

Table 2 summarizes the estimated coefficients, standard errors, t-statistics and p-values for each parameter in the model.

Parameter	Coefficient	Std. Error	t-Statistic	$P > \|t\|$
Constant	-2530.0236	635.100	-3.984	0.000
Flights	68.4317	17.151	3.990	0.000
Util	922.8727	172.032	5.365	0.000
Cargo	-9.5419	5.686	-1.678	0.103

Table 2: Estimated regression coefficients for the multiple linear regression model (dependent variable: Passenger).

From Table 2, we can see that both `Flights` and `Util` have positive and highly significant coefficients (p-values < 0.01), implying that an increase in the number of flights or in aircraft daily utilization is associated with a significant increase in passenger throughput. By contrast, the coefficient on `Cargo` is negative but statistically insignificant at the 5% level ($p \approx 0.103$), suggesting that cargo throughput does not have a clear linear effect on passenger throughput in this sample.

Actual vs Predicted Passenger Throughput

To evaluate the overall predictive performance of the model, Figure 1 compares the actual passenger throughput with the model-predicted values.

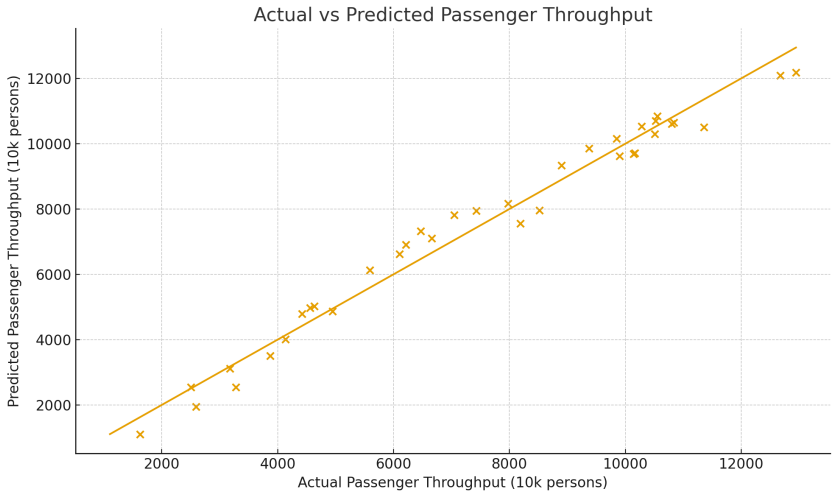

Figure 1: Actual vs. predicted passenger throughput (10,000 persons). The dashed line indicates the perfect-fit reference.

In Figure 1, each point represents one month. The dashed diagonal line corresponds to the ideal situation where prediction equals observation. Most points lie close to this line, confirming that the model's predictions are very close to the actual passenger throughput values.

Relationship between Flights and Passenger Throughput

Beyond overall fit, it is also informative to visualize the relationship between one of the key explanatory variables and the response. Figure 2 plots the number of flights against passenger throughput, together with a fitted linear trend line.

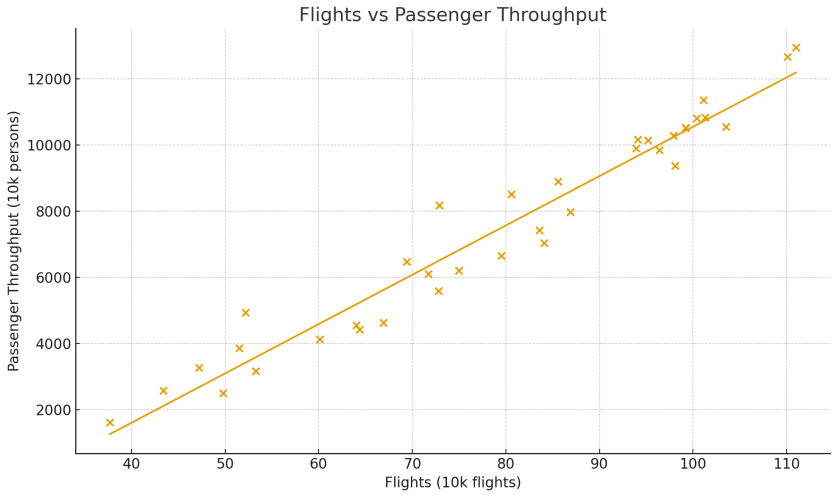

Figure 2: Scatter plot of flights vs. passenger throughput with fitted linear trend line.

As shown in Figure 2, there is a clear positive relationship between flight movements and passenger throughput: months with more flights tend to handle more passengers. This pattern is consistent with the positive and significant coefficient on `Flights` in Table 2.

In conclusion, the multiple linear regression model constructed in this report successfully captures the main linear relationships between passenger throughput and key operational variables in the Chinese civil aviation industry from 2021 to 2023. The analysis confirms that increasing the number of flights and improving aircraft utilization are effective ways to boost passenger throughput, while cargo throughput does not exert a strong linear effect on passenger traffic over the studied period.

References

[1] Civil Aviation Administration of China (CAAC). *Monthly Transport Statistics – Main Production Indicators*. Available at the statistics directory: https://www.caac.gov.cn/XXGK/XXGK/TJSJ/

[2] Civil Aviation Administration of China (CAAC). *Main Production Indicators of China Civil Aviation, May 2022* https://www.caac.gov.cn/XXGK/XXGK/TJSJ/202206/P020220620544668274582.pdf

[3] Civil Aviation Administration of China (CAAC). *Main Production Indicators of China Civil Aviation, December 2022* https://www.caac.gov.cn/XXGK/XXGK/TJSJ/202301/P020230120399887457361.pdf

[4] Ministry of Transport of the People's Republic of China. *Main Production Indicators of China Civil Aviation, April 2022* https://www.mot.gov.cn/tongjishuju/minhang/202206/P020220607379341075082.pdf

[5] Ministry of Transport of the People's Republic of China. *2021 Civil Aviation Industry Development Statistical Bulletin* https://www.mot.gov.cn/tongjishuju/minhang/202206/P020220607377281705999.pdf

[6] Ministry of Transport of the People's Republic of China. *Civil Aviation Statistical Data Portal*. https://www.mot.gov.cn/tongjishuju/minhang/

Details of using the generative AI

- **ChatGPT5.1 Thinking :** Assist in data analysis workflow design, regression interpretation, and LaTeX report drafting (bilingual).