PROGRAMMING FOR DATA SCIENCE

ST2195 Coursework Submission

KOH CHING HUI, NICHOLAS
UOL ID: 220459392
Word Count: 2010

Part 1: Introduction

In this section, we employed the Metropolis-Hastings algorithm, specifically using a Random Walk Metropolis (RWM) variant, to simulate random numbers for a distribution defined by the probability density function $f(x) = \frac{1}{2}\, e^{-|x|}$. The algorithm was executed with N=10000 iterations and a step size (s) of 1.

To gain a visual understanding of the generated samples, a histogram and kernel density plot was generated on both R and python, in figures 1.1 and 2.1 respectively. Our algorithm's ability to capture the underlying probability density function $f(x)$ is evident from the overlay in these plots. Both histogram and kernel density plots closely mirror $f(x)$, showing that our RWM algorithm successfully approximates the target distribution.

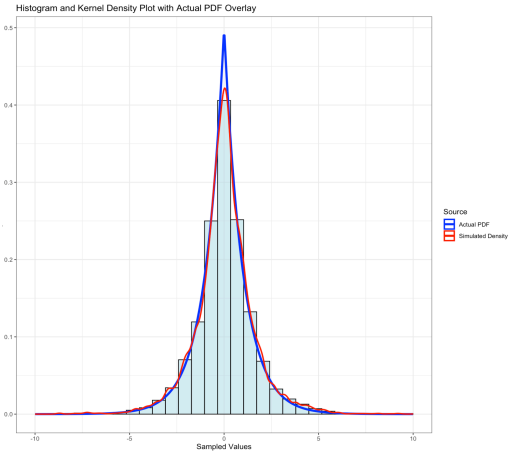

Figure 1.1 – R Overlay Plot

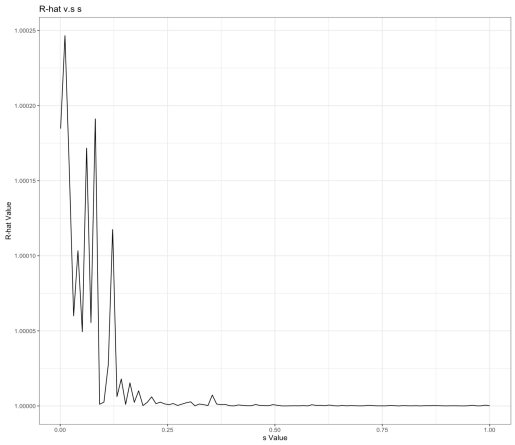

Figure 1.2 – R Grid Plot

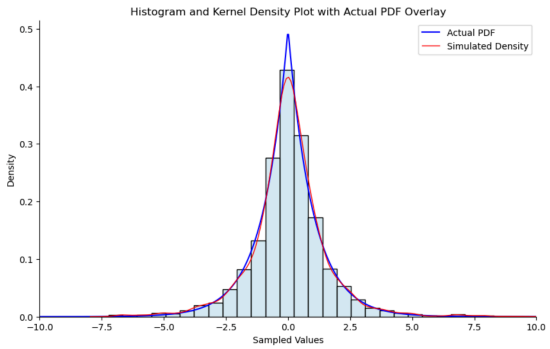

Figure 2.1 – Python Overlay Plot

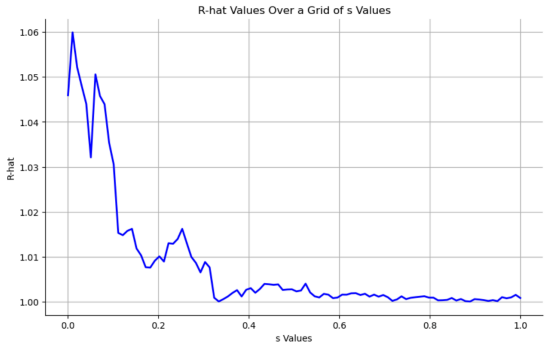

Figure 2.2 – Python Grid Plot

In figures 1.2 and 2.2, we observe that initial R-hat value starts around 1.05, indicating potential lack of convergence in chains. However, as s varies from 0.001 to 1, the R-hat value exhibit a notable decreasing trend toward 1. This decline implies improved convergence and stability in the algorithm as the tuning parameter used (s) is adjusted.

In conclusion, the observed pattern in R-hat values suggests that adjusting the tuning parameter (s) allows the algorithm to find a configuration that enhances convergence, leading to more stable and reliable estimates of the target distribution.

Results of the Monte Carlo estimates of the mean and standard deviation are as such:

```
Monte Carlo Estimate of Mean: 0.051363455826388224
Monte Carlo Estimate of Standard Deviation: 1.6304042092411584
```

Part 2: Introduction

In addressing the substantial volume of data encompassing nearly 60 million rows spanning ten consecutive years (1997 to 2006), a robust and efficient strategy is employed. To handle this large dataset, a connection to an SQLite database is established. The necessary data, extracted from CSV files, are organized into tables named as airports, carriers, planes, and flights within the SQLite database.

The utilization of a database structure ensures reliability and efficiency in managing the extensive dataset. Adopting this approach enables seamless handling of the vast amount of flight-related information over the specified timeframe.

a) What are the best times and days of the week to minimise delays each year?

In this analysis, the query is segmented into four distinct columns, each contributing to a comprehensive understanding of the data. Figures 3.1 and 3.2 shows the first and the last 18 rows of the query respectively. Given the similarity between the results obtained from R and Python, the analysis presented below focuses solely on the R implementation. The characteristics of attributes used are as follows:

1. **year**: Ranging from 1997 to 2006, capturing a decade long timeframe.
2. **DayOfWeek**: Represented by values from 1 to 7, delineating each day of the week.
3. **time_interval**: Divided into intervals of 4 hours, starting from 0000.
4. **avg_delay**: Signifying the corresponding average flight delay for a specific time interval on a particular day of the week within a given year.

In the interest of keeping our analysis straightforward and within the specified constraints, we have chosen to focus on calculating average flight delays based on departure-related timings for the entirety of this part. This approach allows us to streamline our analysis and focus on key metrics that directly impact departure punctuality and operational efficiency.

	Year	DayOfWeek	time_interval	avg_delay
1	1997	1	0000-0359	47.448799
2	1997	1	0400-0759	7.880953
3	1997	1	0800-1159	14.836612
4	1997	1	1200-1559	17.973565
5	1997	1	1600-1959	21.881888
6	1997	1	2000-2359	31.435884
7	1997	2	0000-0359	61.130523
8	1997	2	0400-0759	8.411548
9	1997	2	0800-1159	14.798000
10	1997	2	1200-1559	17.361046
11	1997	2	1600-1959	21.599748
12	1997	2	2000-2359	30.983090
13	1997	3	0000-0359	82.239881
14	1997	3	0400-0759	8.528561
15	1997	3	0800-1159	15.027874
16	1997	3	1200-1559	18.146519
17	1997	3	1600-1959	23.015936
18	1997	3	2000-2359	36.678805

Figure 3.1

	Year	DayOfWeek	time_interval	avg_delay
403	2006	5	0000-0359	106.37428
404	2006	5	0400-0759	12.60138
405	2006	5	0800-1159	21.71140
406	2006	5	1200-1559	27.65000
407	2006	5	1600-1959	35.69832
408	2006	5	2000-2359	55.11933
409	2006	6	0000-0359	66.72917
410	2006	6	0400-0759	12.83532
411	2006	6	0800-1159	20.74437
412	2006	6	1200-1559	25.99026
413	2006	6	1600-1959	32.05754
414	2006	6	2000-2359	43.56618
415	2006	7	0000-0359	95.58571
416	2006	7	0400-0759	12.90612
417	2006	7	0800-1159	19.26794
418	2006	7	1200-1559	24.42251
419	2006	7	1600-1959	31.27966
420	2006	7	2000-2359	48.97850

Figure 3.2

In the initial phase, we focus on rectifying anomalies within the dataset that could impact the accuracy of our results. Notably, we address the issue of timings exceeding '2400' in the departure and arrival delay columns. This correction is essential for ensuring the temporal integrity of the dataset in adherence to the 24-hour time format.

Additionally, values replaced by 'none' or 'null' are systematically excluded from our analysis. This step is crucial for maintaining data consistency and reliability, as these replaced values could otherwise introduce inaccuracies into our findings.

Visualisation of Distribution by Day of Week

The cleaned dataset now forms the basis for our analysis, providing a reliable foundation for uncovering the optimal times and days of the week to minimize flight delays for each year.

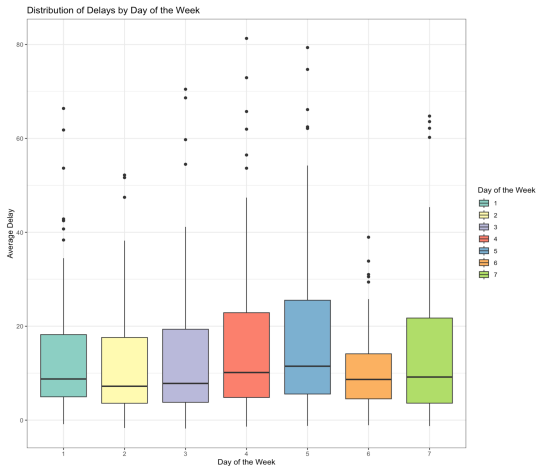

Figure 4.1 – R Boxplot of Avg Delays.

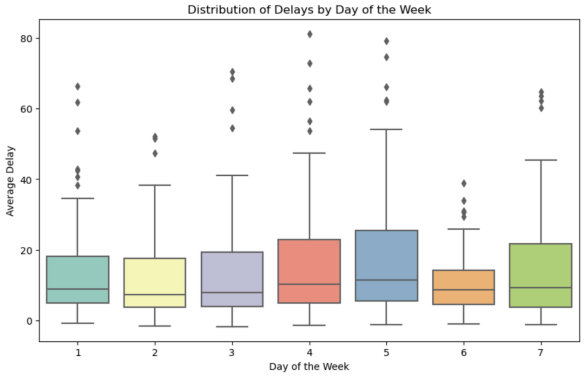

Figure 4.2 – Python Boxplot of Avg Delays

In examining the median delay from the boxplots, we observe that Day 5 exhibit the highest median delay among all days, suggesting that on average, flights on Day 5 experiences longer delays compared to other days.

Analysing the presence of outliers reveal additional insights. Day 1 displays the highest number of outliers, potentially indicating a wider range of delay occurrences compared to other days. Day 4 follows with a notable number of outliers, suggesting sporadic but significant delays on this day.

These findings have practical implications for both airlines and travellers. The high median delay on Day 5 suggests potential congestion or operational challenges specific to that day of the week, which could inform scheduling decisions and resource allocation for airlines. Similarly, the presence of outliers on Day 1 and Day 4 underscores the need for proactive measures to mitigate delays, such as adjusting flight schedules or increasing operational efficiency.

Visualisation of Delays Across Time, Day, and Year

In our comprehensive analysis of flight details spanning a decade, we sought to visualize the patterns of delays across each day of the week over the years. To achieve this, we generated a series of heatmaps, each representing a single year. Within each heatmap, individual tiles correspond to specific time intervals, days of the week, and the respective year.

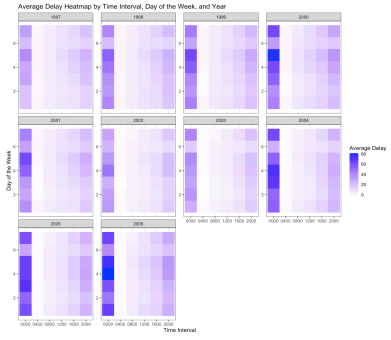

Figure 5.1 – R Heatmap

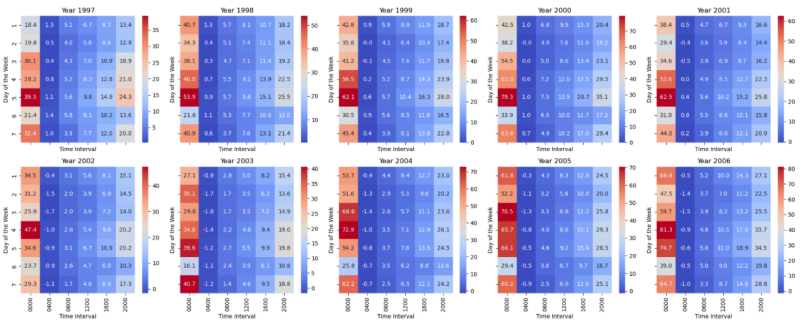

Figure 5.2 – Python Heatmap

The heatmaps above illustrates average flight delays across different time intervals throughout the day, with each plot representing a specific year. One notable trend observed in the heatmap is that the average delays are highest during the time interval of 0000-0400 and gradually decrease as the day progresses. Delays then start to pick up again towards the end of the day, showing a consistent pattern across all years.

There is a subsequent decrease in average delays as the day progresses, suggesting a temporary alleviation during the mid-morning to early afternoon hours. However, this respite is transient, as the delays tend to escalate again from the 1600hrs to 2000hrs interval. The late afternoon and early evening periods thus witness a resurgence in average delays, indicating a distinct temporal pattern in flight delays over the years.

b) Evaluate whether older planes suffer more delays on a year-to-year basis.

To address this question, we explore the relationship between the age of commercial planes and their average delay. The analysis aims to determine whether older planes experience more delays on a year-to-year basis.

Data used are from the flights and planes datasets, focusing on years 1997 to 2006. The variable 'plane age' was determined by computing the difference between the year of the flight and the year of manufacture for the specific plane involved in that flight, adhering to the context of evaluating the relationship on a year-to-year basis. Since the column 'CarrierDelay' has missing values for flights from 1997 to 2002, we opted to use the 'DepDelay' column measure for the analysis.

The scatterplot below illustrates the relationship between plane age and average delay. Upon calculation, the correlation coefficient between plane age and average delay was determined to be -0.39, indicating a moderate negative linear relationship between the two

variables. Furthermore, a trend line was fitted to the plot, revealing a discernible downward trend.

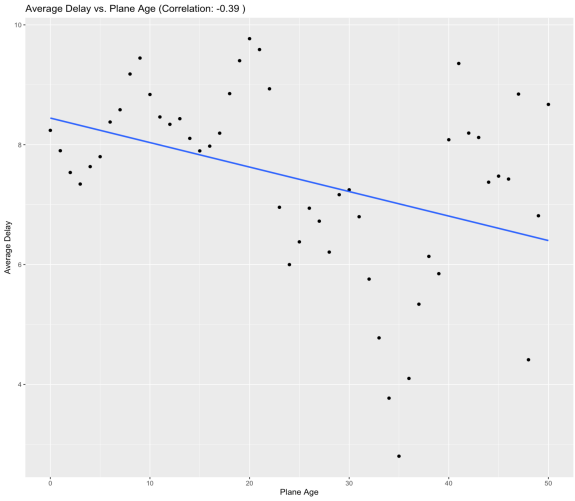

Figure 6.1 – R Scatterplot with Trend Line

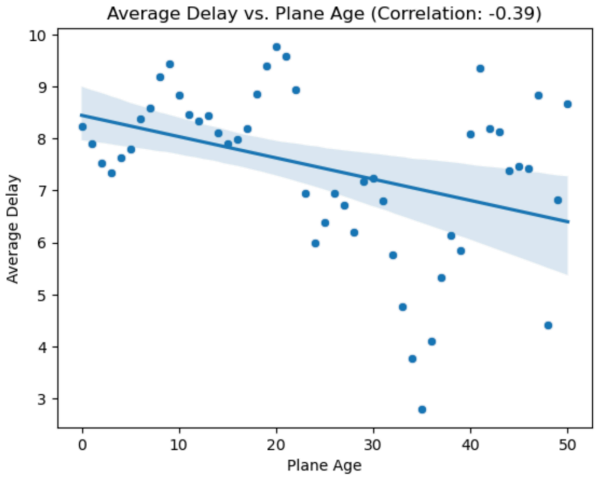

Figure 6.2 – Python Scatterplot with Trend Line

These findings suggest that as plane age increases, there is a tendency for average delay to decrease. It is worth noting that while the correlation coefficient of -0.39 denotes a moderate strength of the relationship, there remains a noticeable inclination for older planes to experience lower average delays.

In addition to the observed relationship between plane age and average delay, it is noteworthy to consider the underlying factors that may contribute to this phenomenon. Younger planes, manufactured in more recent years, may be more prone to encountering operational issues compared to their older and more seasoned counterparts. Conversely, older planes may have undergone extensive maintenance and operational refinements over the years, resulting in a more reliable and predictable performance, despite their advanced age.

Therefore, while the age of the aircraft plays a significant role in average delay trends, the operational context and maintenance history of the planes also warrant consideration in understanding the dynamics of delays in commercial aviation.

c) Logistic Regression Analysis of Diverted Flights (1997 – 2006)

Here, we delve into the analysis of flights over the span of 1997 to 2006 to understand the factors contributing to fight diversion, utilizing logistic regression modelling techniques. The fitted models aim to predict the probability of flight diversion based on selected features engineered to better capture nuances of any diversions. The selected features are as such:

- **CRSDepTime**: Planned departure time of the flight.
- **CRSArrTime**: Planned arrival time of the flight
- **Distance**: Euclidean distance between departure and arrival airports, calculated using their associated longitudinal and latitude coordinates.

Logistic regression models were fitted for each year independently. The formula used for model fitting was:

$$Diverted \sim distance + CRSDepTime + CRSArrTime$$

The models were constructed using a binary response variable, where '0' denoted no diversion and '1' represented a diverted flight. The outcomes of the fitted models for each year in R are illustrated in Figures 7.1 and 7.2, showcasing their respective coefficients. Because the output in Python mirrors these findings, it is not presented here.

To construct comprehensive models from the extensive dataset, we began by sampling a subset of data for each year. This approach ensured that we had a representative yet manageable dataset for logistic regression analysis.

```
Year: 1997
Coefficients:
CRSDepTime: 0.013828811078165722
CRSArrTime: -0.016288687455647738
Distance: -0.13647774067080315
================================================
Year: 1998
Coefficients:
CRSDepTime: -0.6469508214143364
CRSArrTime: -0.26968478433396426
Distance: 0.528071623309912
================================================
Year: 1999
Coefficients:
CRSDepTime: -0.4126815899667362
CRSArrTime: 0.44683472212415376
Distance: 0.3008755650876281
================================================
Year: 2000
Coefficients:
CRSDepTime: -0.22973625902879075
CRSArrTime: 0.5283759682526928
Distance: 0.23424329502617813
================================================
Year: 2001
Coefficients:
CRSDepTime: 0.02104406108823808
CRSArrTime: -0.38756200541777175
Distance: 0.16339689517488054
```

Figure 7.1 – R

```
Year: 2002
Coefficients:
CRSDepTime: -0.1675615507136448
CRSArrTime: 0.5617755370344758
Distance: -0.04130336029616981
================================================
Year: 2003
Coefficients:
CRSDepTime: -0.3746521823270583
CRSArrTime: -0.09127206802793122
Distance: 0.25920047894948745
================================================
Year: 2004
Coefficients:
CRSDepTime: -0.7984788477760382
CRSArrTime: 0.9900253315891862
Distance: 0.3113430659437663
================================================
Year: 2005
Coefficients:
CRSDepTime: 0.41116902231192115
CRSArrTime: -0.30489462992240823
Distance: 0.13806468455023874
================================================
Year: 2006
Coefficients:
CRSDepTime: 0.06224603674128054
CRSArrTime: 0.21662240129485502
Distance: 0.35751462424301433
```

Figure 7.2 – R

Insights and Interpretation of variables

1. CRSDepTime and CRSArrTime

The coefficient of CRSDepTime and CRSArrTime represents how the probability of flights diversion changes relative to the scheduled departure and arrival times. A negative coefficient suggests that earlier departure/arrival times are associated with a lower probability of diversion, while a positive coefficient indicates the opposite.

Across the years analysed, flights scheduled for earlier departure times tend to have a lower likelihood of diversion, possibly due to more favourable weather or less air traffic during these hours. Similarly, flights scheduled to arrive earlier are less likely to experience diversions, suggesting that timely arrivals are conducive to smoother flight operations and fewer disruptions.

2. Flight Distance

The coefficient of Distance feature reflects how changes in the distance between departure and arrival airports influence the probability of flight diversions. A positive coefficient indicates that longer flight distances are associated with a higher likelihood of diversions.

According to our analysis, flights covering longer distances have a higher probability of diversion. This may be due to increased operational complexities such as fuel management or routing challenges, potentially leading to this increased probability. In order to be certain, further research must be conducted.

Visualization of Coefficients Across Years

The coefficient plot in Figures 8.1 and 8.2 visualizes the trends of three key features – CRSArrTime, CRSDepTime, and Distance – across the years from 1997 to 2006. These features are essential indicators in logistic regression models for predicting flight diversions.

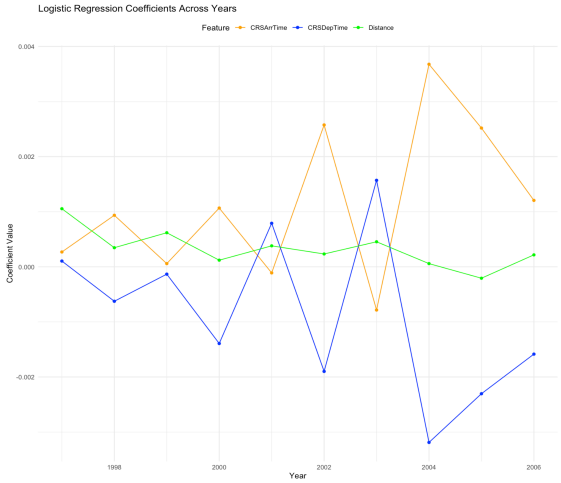

Figure 8.1 – R Coefficient Plot

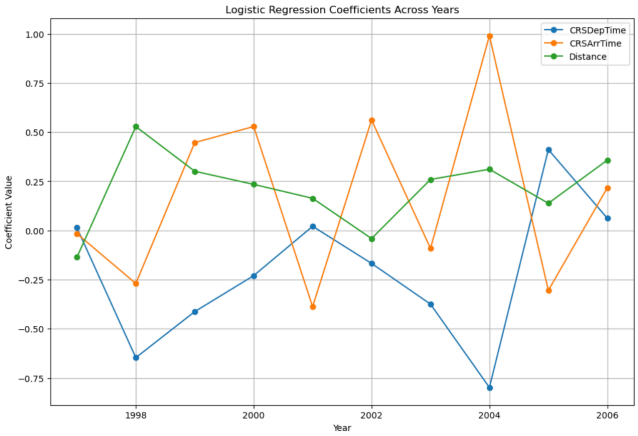

Figure 8.2 – Python Coefficient Plot

Seen from above, the inverse relationship between CRSArrTime and CRSDepTime suggests that there might be a trade-off between scheduled arrival and departure times of flights. This inverse relationship reflects the intricate dynamics between flight scheduling and diversion likelihood.

In contrast, the coefficient plot shows that the line representing Distance remains relatively stable, fluctuating around 0 across the years. This stability suggests that the distance between departure and arrival airports has a consistent, albeit moderate, influence on the likelihood of flight diversion throughout the years under consideration.

Conclusion

In summary, the coefficient plot offers valuable insights into the relationships between key features and flight diversion likelihood over the years. Understanding these trends can aid aviation stakeholders in enhancing operational efficiency, optimizing flight scheduling practices, and mitigating the risks associated with flight diversions.