STAA57 W21 Draft Report

Group 4 (Adham F, Jason Y, Mohamed T, Wesley M)

Link to RStudio Cloud shared project: https://rstudio.cloud/spaces/115177/project/2202760

Introduction

The topic of this paper is to investigate the aircraft operating cost of DFC, particularly the Cost per Session. There are two main factors that affect the aircraft operating cost which are jet fuel price and the duration of the session.It is intuitive to think that the duration of a flight will be affected by the weather. Factors such as cloud coverage, wind and precipitation can affect an aircraft/pilot's performance, thus the duration of flight/lesson. With an increase in flight duration comes an increase in maintenance and fuel costs. Therefore, we should expect that there will be weather factors in our data that can be significant predictors for the aircraft operating cost to a certain extent. It is also known that jet fuel price does change monthly/seasonally depending on seasonal travel, such as during Christmas and Summer. Therefore, we should expect that there are seasonal changes to the aircraft operating costs. Since seasonal changes in jet fuel costs are fairly obvious and weather data is closely correlated to seasonal weather changes, we expect the weather factors that we are using to be able to predict the aircraft operating cost. In addition, since COVID-19 has dramatically impacted the Airline industry, which is closely related to jet fuel, we will also be investigating COVID-19 being a factor in predicting the aircraft operating costs.

Questions:

1. What are the trends in jet fuel costs over different time units (months/season)?
    - For this question, web scrapping was used to see how fuel price changes over time and attempted to see how these changes affect the Cost per Session for DFC
2. What factors affect the duration of the sessions?
    - A linear regression model was used to assess the significant covariates associated with the duration of the sessions. Since duration can act as a predictor of cost, the statiscally significant covariates from this model will be used to in the next linear regression model.
3. What factors affect cost in a significant matter?
    - Another linear regression model is used in a similar fashion as the method to answer the question above. It will be using the covariates that were deemed statistically significant in the previous model.

Data

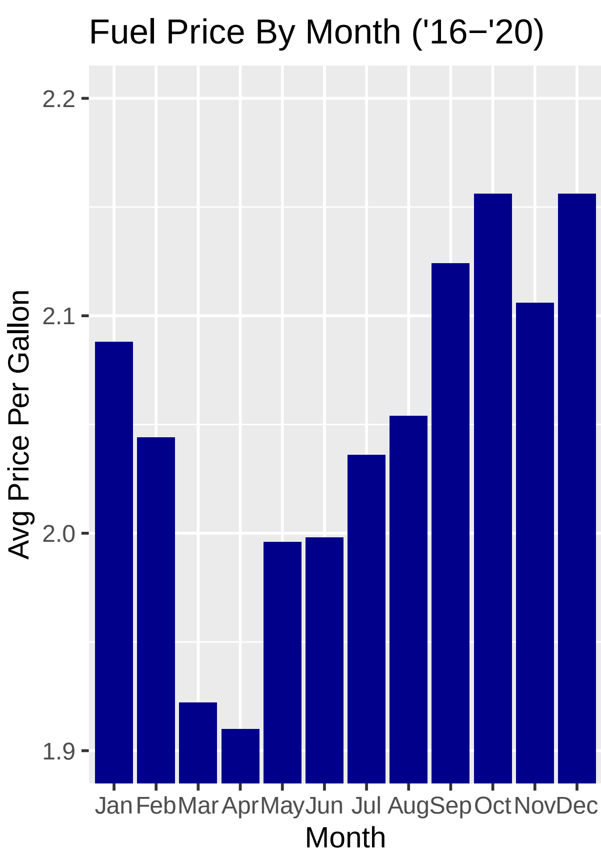
The external data that is used in this project were gathered from the following sources:

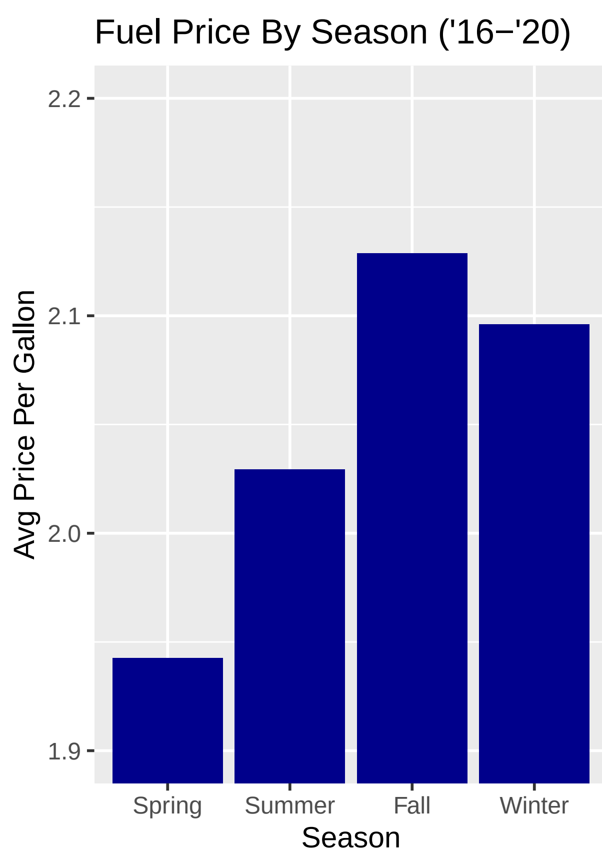
- Fuel (https://www.indexmundi.com/commodities/?commodity=jet-fuel&months=240&currency=cad)
    - Historical fuel data by month, including the cost/gallon and %change
    - Assuming that all planes used in the flight school use the same fuel

  - From preliminary research, it was found that C-172, C-150 and C-152 aircraft models only consume jet fuel
- Maintenance costs and repairs (https://cessna150152club.org/Costs):
  - The annual operating cost of a Cessna plane model for maintenance and repairs
  - Based on this research, an assumed 15 dollars static cost per hour was added to each training session cost
  - In addition, the fuel consumption per hour is estimated to be 6 gallons
- Weather (https://oshawa.weatherstats.ca/download.html):
  - Historical weather data by day, including temp, precipitation, wind, etc.
  - Gathered data from Oshawa location

Analysis

Question 1: What are the trends in jet fuel costs over different time units (months/season)?

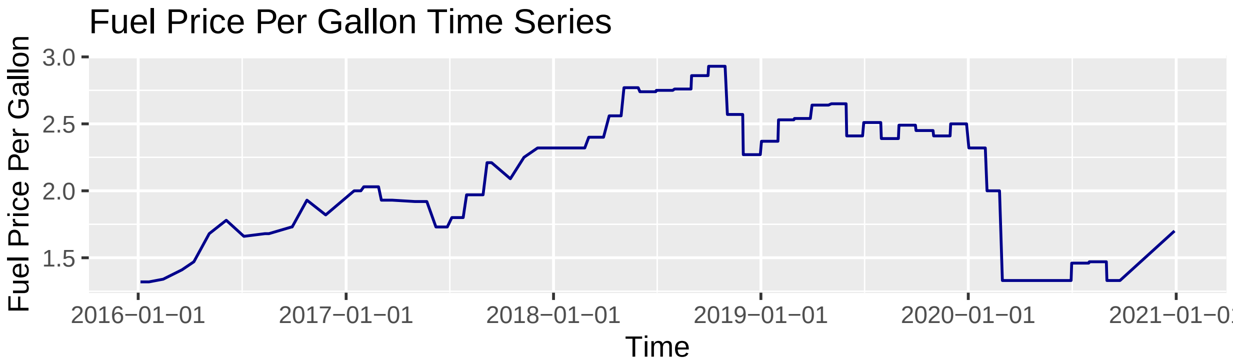




From the monthly data on the left, we can see that in March and April the fuel price is significantly cheaper than any other month at a price around \$1.91/gallon. Following that, it steadily increases and peaks around \$2.15/gallon around October and December.

For the seasonal data on the right, we can see that spring is significantly cheaper than any other season with a price at around \$1.94/gallon. On the other hand, fuel price is most expensive in the fall season with a price of about \$2.13/gallon.

Therefore it should be anticipated for the Cost per Session to change respectively during these noted months and seasons.



From the above graph, it can be seen there was a drastic drop in the jet fuel price per gallon in 2020, leading to lower aircraft operating costs. Therefore, we expect that the Cost per Session to be lower during 2020, i.e. during the pandemic.

Question 2: What factors affect the duration of sessions?

How was the data prepared?

For sessions that perform a standard set of exercises, it's intuitive to anticipate that their total duration should be around the average total duration for that same set of exercises unless there are external factors that affect it. Thus a multiple linear regression model can be utilized to show any significance for external factors against duration.

To prepare the data for our model, the data was grouped by standard sets of exercises. Noting that there are a lot of unique exercise combinations, the most repeated set of exercises were chosen as it had enough data for analysis.

Since a single session might have multiple training types resulting in multiple rows, the total duration for a single session was computed by adding the duration of the rows with the same Session_ID using grouping.

As a minor note, ground sessions were removed as the model's purpose is to measure duration per session with aircrafts against external factors.

Lastly, to add on the external factors the weather data was joined with the processed data set.

We identified an initial set of weather factors that were deemed to have a strong effect on the duration of a session. It should be mentioned that the min_windchill column from the weather data was removed since it contained many null values, practically removing around 70% of our data when including it in the linear regression model. The relevant factors were:

- `avg_relative_humidity`
- `avg_dew_point`
- `avg_pressure_sea`
- `avg_pressure_station`
- `avg_visibility`
- `avg_health_index`
- `precipitation`

- `avg_cloud_cover_4`
- `avg_temperature`

Selecting the most popular exercise routine

Table 1: Most Frequest Group of Exercises

Exercises	count
16,17,18,30	664
16,17,18	136
16,17,18,23,30	89

Linear regression model for Duration per Session:

Refer to *Figure 1*

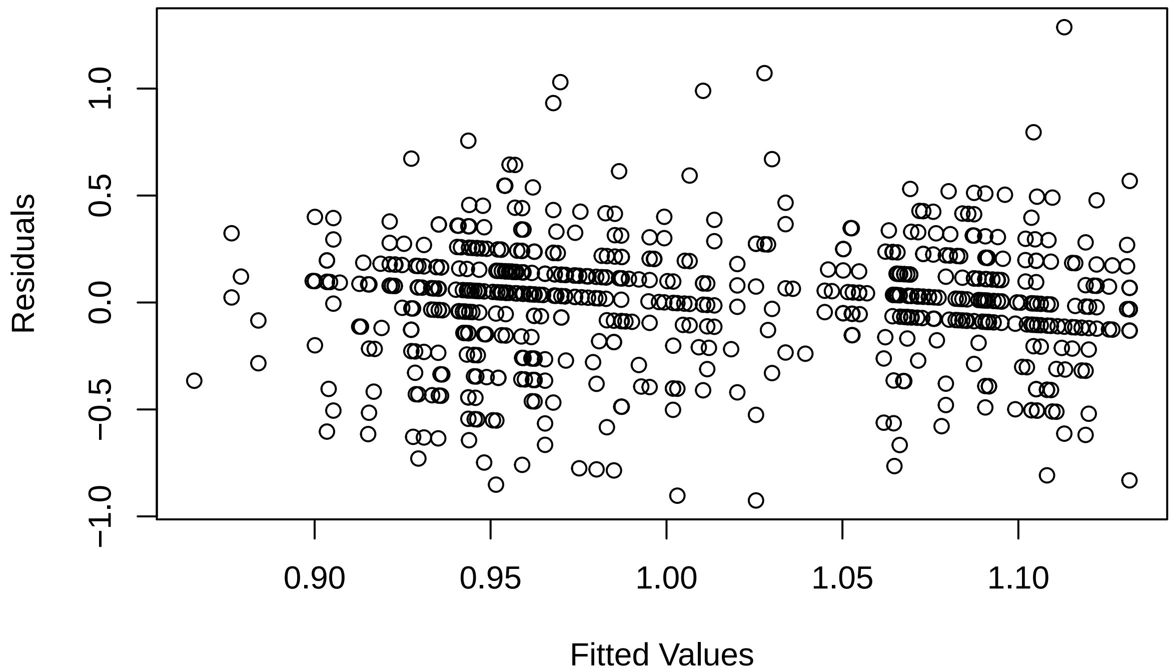
Analysis for Duration per Session linear regression model

It can be seen from the summary function that the p-values for all the covariates are high except for Summer, despite having a high R-squared of 0.89. Having a high R-squared is good for a model since most of the variation in the response variable can be explained through the model. However, coupled with high p-values for almost all covariates renders the model useless as it implies that almost none of the covariates have a satistical significance in association with the total duration. Intuitively since there are multiple covariates from weather data, there might be multiple collinearities causing the high p-values. Hence, a correlation matrix for the weather data was calculated.

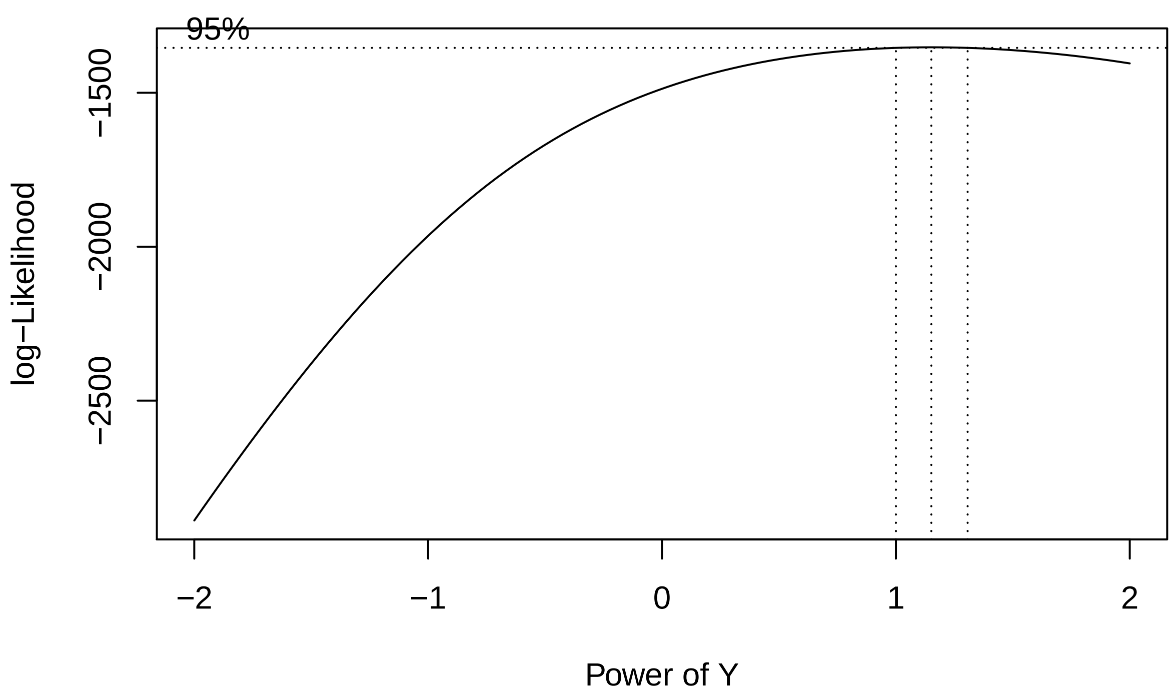
Correlation matrix for the weather covariates Refer to *Figure 2*

From the correlation matrix, it can be seen that avg_pressure_sea and avg_pressure_station have a correlation of almost 1. avg_pressure_station was chosen to be removed as avg_pressure_sea measures are used as a standard during aircraft takeoffs. Also, avg_dew_point and avg_temperature have a correlation of 0.9. From background research, it was found that dew_point is used as an indication of humidity and visibility, thus it was chosen to be removed from the model since humidity and visibility are both given. A new linear regression model was fit disregarding these two covariates.

Refer to *Figure 3*



Unfortunately, despite attempting to remove variables with high correlation, the model still performed poorly with high p-values. A further look into the residual plot gave a clear indication that the linear regression model assumption of having a constant variation among residuals was being violated, giving a further indication that a linear model is not a good fit for the data. To remedy this, a boxcox transformation was attempted to see if a power of the response variable could improve the linear model fit.



Unfortunately, it seemed that the power of 1 is almost the best transformation, therefore putting the response variable to a different power would not improve the model. Another attempt at taking the log function of the response variable was undertaken instead.

Refer to *Figure 4*

However, the R-squared drastically decreased to 0.061 and the p-values were still high therefore this model is still insufficient and can be safely discarded.

Summary of Duration Per Session Model

Overall, it seemed that using the linear regression model to fit a model over the data is not the appropriate model for this data due to the high p-values and thus due to the low statistical significance of the majority of the covariates in predicting the response variable. Our initial plan was to fit all duration vs all the possible covariates and choosing the most statistically significant covariates to use in the following Cost per Session model. However, since Summer is the only one, we will be fitting the Cost per Session model using all the covariates.

Question 3: What factors affect cost in a significant matter?

How was the data prepared?

The aircraft operating cost per session was calculated using the equation below:

```
CostPerSession = (Duration * (FuelPricePerGallon*GallonsPerHr + OperatingCostPerHr))
```

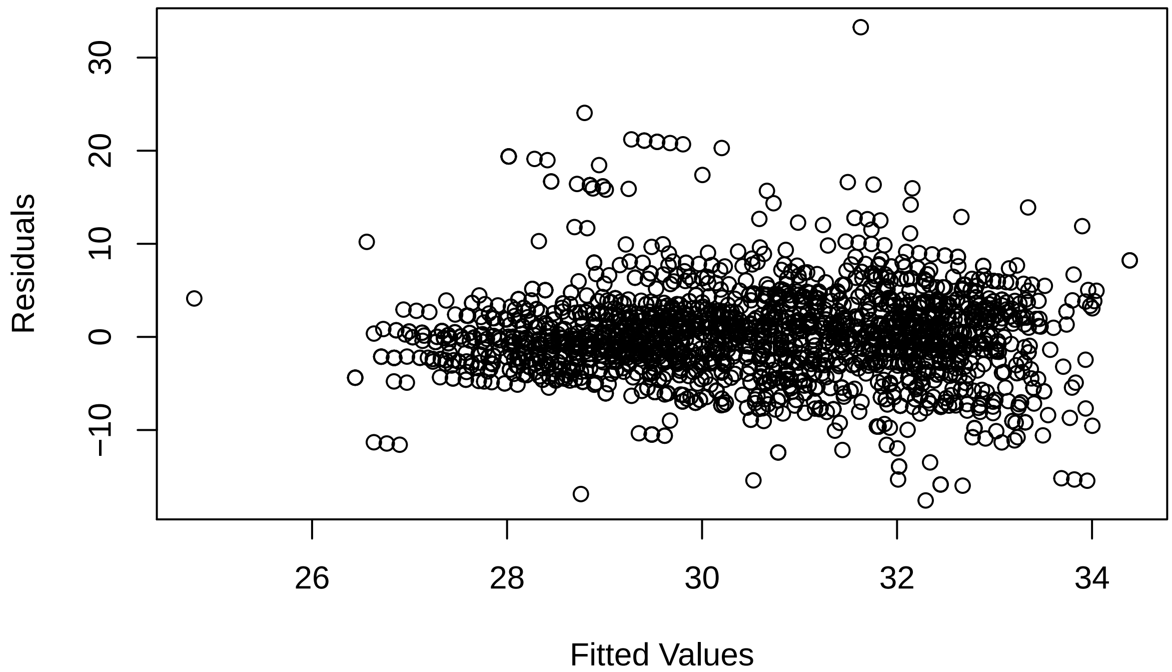
The aircraft operating cost equation uses duration, fuel price per gallon, and an average number of gallons consumed per hour (6) as well as an estimated maintenance and repairs cost per hour (15). After the Cost per Session column has been calculated, the data frame is then grouped by date and an Average Cost per Session is calculated for each date to smoothly join the data frame with the daily weather data. A linear regression model can then be built using this data frame.

Since the cost of each session is dependent on the duration, it was observed that session duration can vary due to the exercises being performed as well as the number of exercises. Some sessions might have more exercises and thus have a higher cost. Due to the lack of information on the expected individual exercise duration, it was assumed that each exercise will have the same duration. Since the cost of a session is related to the exercises being performed, we added a column representing the number of exercises in a session and added that to our model.

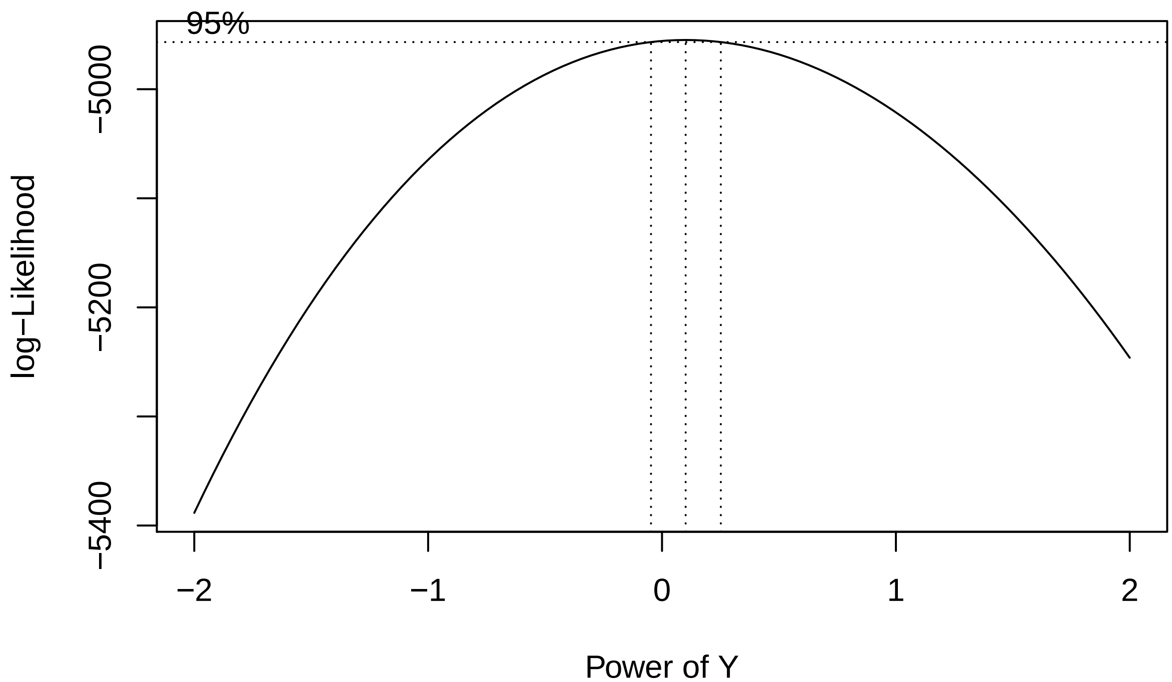
Refer to *Figure 5*

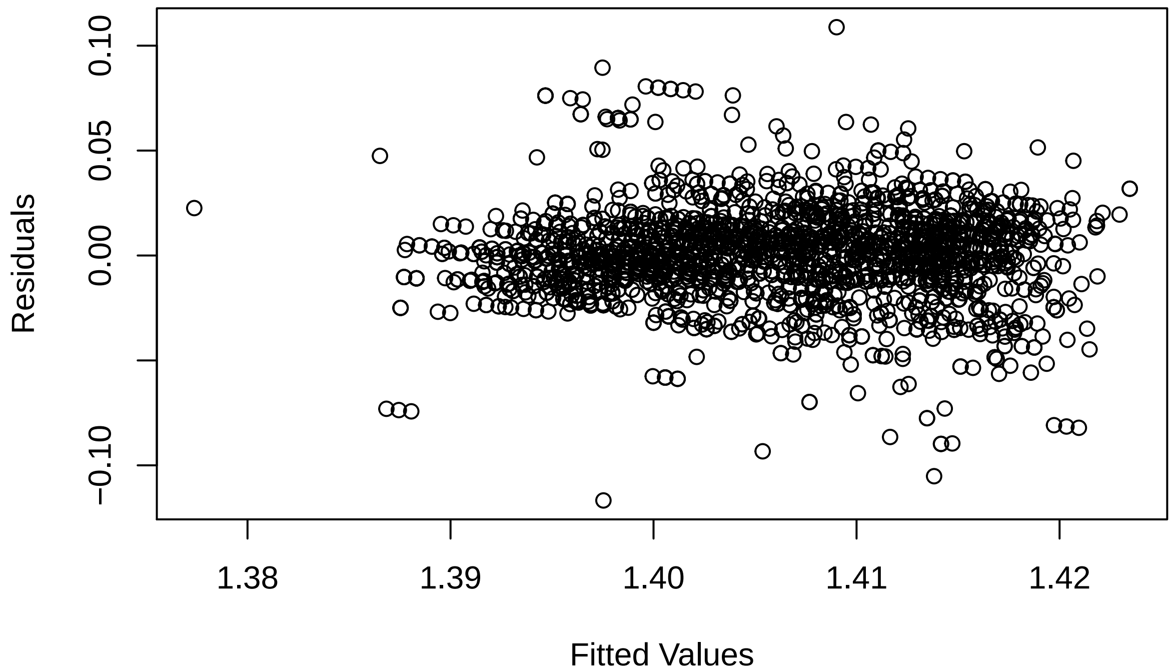
Analysis of the Cost per Session Regression Model

Looking at the summary for the cost model, there are covariates with p-values below 0.05: `SeasonSummer`, `SeasonWinter`, `avg_health_index`, `n_exercises`, `avg_pressure_sea`, `avg_pressure_station`, `precipitation`, and `avg_cloud_cover_4`. This indicates that there is a statistical significance for these covariates when predicting the Cost per Session for a day, as they all were able to reject the null hypothesis for the t-test with an alpha of 0.05. Therefore, we can conclude that these covariates are associated with the Cost per Session with 95% confidence. In addition, the Cost per Session model has a very high R-squared value of 0.98 so this shows the model fits the data well and can explain 98% of the variation in the Cost per Session for each day. Before interpreting the fitted linear regression equation, a residual analysis is undertaken.



Analyzing the residual plot, the residuals are not randomly scattered, rather they are centered around 0 forming an oval shape. This conveys that the Cost per Session model can be improved by transformations of the response and covariates as the errors do not follow a constant variance. So again, a boxcox transformation was utilized to attempt to improve the model.





Refer to *Figure 6*

From the boxcox transformation plot, it can be deduced that the best transformation is to raise the power of the response variable to 0.1. The covariates that reject the null hypothesis of the t-test with an alpha of 0.05 are: `avg_pressure_sea`, `avg_pressure_station`, `avg_health_index`, `precipitation`, `COVIDFALSE`, `COVIDTRUE`, `SeasonSummer` and `n_exercises`. Unlike the previous model without transformation of the response variable, `SeasonWinter` and `avg_cloud_cover_4` are no longer statistically significant but COVID-19 is now statistically significant. Thus, these covariates are associated with the Cost per Session with 95% confidence. However, even after running the boxcox transformation we still see a non-random pattern in the residuals plot and so the quality of our model is about the same as before in terms of residual plot analysis. For the sake of interpretation simplicity, the first model is chosen as the linear regression fit.

Interpreting the coefficients statistically significant covariates

The analysis below for the COVID-19 covariates assumes that every other variable is held constant:

- `SeasonSummer` - during summer the cost increases by 1 dollar per session
- `SeasonWinter` - during winter the cost increases by 2.26 dollars per session
- `avg_health_index` - for each unit increase in the `avg_health_index` the cost increases by 0.74 dollars per session
- `n_exercises` - for each additional exercise the cost decreases by 0.13 dollars per session
- `avg_pressure_sea` - for each unit increase in the `avg_pressure_sea` the cost decreases by 31 dollars per session
- `avg_pressure_station` - for each unit increase in the `avg_pressure_station` the cost increases by 32 dollars per session

- `precipitation` - for each unit increase in the precipitation the cost decreases by 0.07 dollars per session
- `avg_cloud_cover_4` - for each unit increase in the precipitation the cost decreases by 0.34 dollars per session

An interesting observation is that as the number of exercises increasing is associated with the cost decreasing - this does not mean causation, rather just association, meaning that increasing the number of exercises doesn't result in a decrease in cost. A possible explanation is that there can be a lot of easy exercises for beginner sessions that don't take long to complete therefore lowering the duration of the session thus lowering the cost per session.

Unlike what we anticipated from question 1 that Fall would have higher cost, it turned out to be lower than the usual cost as the cofficient for the fall was used as the base state for the equation and the other seasons such as Summer and Winter had a statistically significant coefficients. It was during the Winter that the price per session increases by 2.26 dollars as anticipated to have a spike around December in the jet fuel analysis. Unintuitively, this model shows that COVID-19 had no effect on the price per session due to its high p-value. However, it is important to keep in mind that association does not mean causation, therefore this does not rule out that COVID-19 cannot affect the price per session.

Summary

Overall, the Duration per Session model was not as useful as anticipated due to linear regression not being a good fit for the data, and that can be the result of choosing the incorrect model type. Nonetheless, the second model did indeed provide some useful insight on how the Cost per Session changes depending on the statistically significant covariates, despite going against some anticipated patterns found using the jet fuel analysis. It is important to note that the residual plot for the Cost per Session did not appear to follow a random pattern thus there can be a better model for the data.

Appendix

Figure 1: Summary of duration LR model (All factors)

```
##
## Call:
## lm(formula = Total_Duration ~ avg_relative_humidity + avg_dew_point +
##     avg_pressure_sea + avg_pressure_station + avg_visibility +
##     avg_health_index + avg_cloud_cover_4 + avg_temperature +
##     COVID + precipitation + Season - 1, data = duration.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92023 -0.12848  0.03313  0.16232  1.28188
##
## Coefficients:
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## avg_relative_humidity -1.870e-03  3.210e-03  -0.583  0.56030
## avg_dew_point          9.073e-03  1.224e-02   0.741  0.45885
## avg_pressure_sea       2.607e+00  1.803e+00   1.446  0.14879
## avg_pressure_station  -2.618e+00  1.836e+00  -1.426  0.15438
## avg_visibility         2.508e-06  5.537e-06   0.453  0.65078
## avg_health_index      -4.093e-03  2.950e-02  -0.139  0.88970
## avg_cloud_cover_4      1.379e-02  2.075e-02   0.664  0.50665
## avg_temperature        4.587e-04  1.168e-02   0.039  0.96868
## COVIDFALSE            -2.375e+00  2.260e+00  -1.051  0.29366
## COVIDTRUE             -2.340e+00  2.260e+00  -1.035  0.30086
## precipitation          3.609e-04  3.883e-03   0.093  0.92599
## SeasonSpring          -4.245e-03  5.268e-02  -0.081  0.93580
## SeasonSummer           1.296e-01  4.339e-02   2.988  0.00292 **
## SeasonWinter           1.941e-02  4.942e-02   0.393  0.69458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2967 on 641 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.9226, Adjusted R-squared:  0.9209
## F-statistic: 545.6 on 14 and 641 DF,  p-value: < 2.2e-16
```

Figure 2: Weather factor correlation matrix

```
##                       avg_relative_humidity avg_dew_point avg_pressure_sea
## avg_relative_humidity             1.0000000     0.5205644      -0.31752124
## avg_dew_point                     0.5205644     1.0000000      -0.38949256
## avg_pressure_sea                 -0.3175212    -0.3894926       1.00000000
## avg_pressure_station             -0.3166981    -0.3712001       0.99971608
## avg_visibility                   -0.5786371    -0.1708754       0.32336137
## avg_health_index                 -0.1063215    -0.1195438       0.07779705
## precipitation                     0.3246735     0.2533825      -0.26405725
## avg_cloud_cover_4                 0.4869239     0.2761234      -0.34845257
## avg_temperature                   0.1628771     0.9076234      -0.31691614
##                       avg_pressure_station avg_visibility avg_health_index
## avg_relative_humidity          -0.31669809    -0.57863712      -0.10632147
## avg_dew_point                  -0.37120007    -0.17087539      -0.11954378
## avg_pressure_sea                0.99971608     0.32336137       0.07779705
## avg_pressure_station            1.00000000     0.32768446       0.07597322
## avg_visibility                  0.32768446     1.00000000      -0.06999009
## avg_health_index                0.07597322    -0.06999009       1.00000000
## precipitation                  -0.26253135    -0.32768768      -0.01866383
## avg_cloud_cover_4              -0.34900166    -0.35403107      -0.10594446
## avg_temperature                -0.29615833     0.05126322      -0.09290758
##                       precipitation avg_cloud_cover_4 avg_temperature
## avg_relative_humidity    0.32467352         0.4869239      0.16287706
## avg_dew_point            0.25338250         0.2761234      0.90762337
## avg_pressure_sea        -0.26405725        -0.3484526     -0.31691614
```

```
## avg_pressure_station    -0.26253135       -0.3490017     -0.29615833
## avg_visibility          -0.32768768       -0.3540311      0.05126322
## avg_health_index        -0.01866383       -0.1059445     -0.09290758
## precipitation            1.00000000        0.3321391      0.15428709
## avg_cloud_cover_4        0.33213906        1.0000000      0.11097086
## avg_temperature          0.15428709        0.1109709      1.00000000
```

Figure 3: Summary of duration LR model (Filtered factors)

```
##
## Call:
## lm(formula = Total_Duration ~ avg_relative_humidity + avg_pressure_sea +
##     avg_visibility + avg_health_index + avg_cloud_cover_4 + avg_temperature +
##     COVID + precipitation + Season - 1, data = duration.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92543 -0.12816  0.03484  0.16148  1.28695
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## avg_relative_humidity  2.419e-04  1.830e-03   0.132  0.89485
## avg_pressure_sea       3.573e-02  2.161e-02   1.653  0.09879 .
## avg_visibility         2.085e-06  5.516e-06   0.378  0.70560
## avg_health_index       3.530e-03  2.898e-02   0.122  0.90306
## avg_cloud_cover_4      1.506e-02  2.073e-02   0.727  0.46773
## avg_temperature        9.880e-04  2.299e-03   0.430  0.66757
## COVIDFALSE            -2.788e+00  2.235e+00  -1.247  0.21272
## COVIDTRUE             -2.754e+00  2.236e+00  -1.232  0.21849
## precipitation          3.074e-04  3.807e-03   0.081  0.93566
## SeasonSpring          -1.963e-03  5.237e-02  -0.037  0.97011
## SeasonSummer           1.341e-01  4.329e-02   3.097  0.00204 **
## SeasonWinter           2.084e-02  4.860e-02   0.429  0.66821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2967 on 643 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.9223, Adjusted R-squared:  0.9209
## F-statistic: 636.2 on 12 and 643 DF,  p-value: < 2.2e-16
```

Figure 4: Summary plot of duration LR model (Log)

```
##
## Call:
## lm(formula = log(Total_Duration) ~ avg_relative_humidity + avg_pressure_sea +
##     avg_visibility + avg_health_index + avg_cloud_cover_4 + avg_temperature +
##     precipitation + COVID + Season - 1, data = duration.data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.23061 -0.08093  0.08416  0.20816  0.80928
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## avg_relative_humidity -6.004e-04  2.355e-03  -0.255  0.79881
## avg_pressure_sea       3.790e-02  2.782e-02   1.362  0.17355
## avg_visibility        -5.809e-07  7.098e-06  -0.082  0.93480
## avg_health_index      -9.853e-04  3.729e-02  -0.026  0.97893
## avg_cloud_cover_4      2.169e-02  2.668e-02   0.813  0.41653
## avg_temperature        1.319e-03  2.959e-03   0.446  0.65606
## precipitation         -1.422e-03  4.899e-03  -0.290  0.77179
## COVIDFALSE            -3.986e+00  2.877e+00  -1.386  0.16634
## COVIDTRUE             -3.975e+00  2.877e+00  -1.382  0.16760
## SeasonSpring          -5.946e-03  6.740e-02  -0.088  0.92973
## SeasonSummer           1.733e-01  5.572e-02   3.111  0.00195 **
## SeasonWinter           3.892e-02  6.254e-02   0.622  0.53400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3819 on 643 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.06104,    Adjusted R-squared:  0.04352
## F-statistic: 3.483 on 12 and 643 DF,  p-value: 5.269e-05
```

Figure 5: Summary of cost LR model (Original)

```
##
## Call:
## lm(formula = `Average Cost Per Session` ~ avg_temperature + avg_relative_humidity +
##     avg_dew_point + avg_pressure_sea + avg_pressure_station +
##     avg_visibility + avg_health_index + precipitation + avg_cloud_cover_4 +
##     COVID + Season + n_exercises - 1, data = daily_average_cost_per_hour)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.564  -2.503  -0.191   2.142  33.262
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## avg_temperature        3.820e-02  8.709e-02   0.439 0.660988
## avg_relative_humidity  8.213e-03  2.496e-02   0.329 0.742163
## avg_dew_point         -9.363e-02  9.237e-02  -1.014 0.310830
## avg_pressure_sea      -3.109e+01  1.400e+01  -2.222 0.026386 *
## avg_pressure_station   3.217e+01  1.425e+01   2.258 0.024063 *
## avg_visibility        -1.613e-06  4.367e-05  -0.037 0.970537
## avg_health_index       7.364e-01  2.256e-01   3.264 0.001115 **
## precipitation         -7.053e-02  3.275e-02  -2.154 0.031360 *
## avg_cloud_cover_4      3.387e-01  1.660e-01   2.041 0.041393 *
```

```
## COVIDFALSE            -2.665e+01  1.839e+01  -1.449 0.147491
## COVIDTRUE             -3.012e+01  1.841e+01  -1.637 0.101853
## SeasonSpring           5.264e-01  3.944e-01   1.335 0.182103
## SeasonSummer           1.079e+00  3.218e-01   3.352 0.000814 ***
## SeasonWinter           2.262e+00  4.243e-01   5.331 1.06e-07 ***
## n_exercises            1.325e-01  4.256e-02   3.112 0.001878 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.602 on 2462 degrees of freedom
##   (153 observations deleted due to missingness)
## Multiple R-squared:  0.9779, Adjusted R-squared:  0.9777
## F-statistic:  7254 on 15 and 2462 DF,  p-value: < 2.2e-16
```

Figure 6: Summary of cost LR model (BoxCox)

```
##
## Call:
## lm(formula = `Average Cost Per Session`^0.1 ~ avg_temperature +
##     avg_relative_humidity + avg_dew_point + avg_pressure_sea +
##     avg_pressure_station + avg_visibility + avg_health_index +
##     precipitation + avg_cloud_cover_4 + COVID + Season + n_exercises -
##     1, data = daily_average_cost_per_hour)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.116731 -0.010880  0.000559  0.010912  0.108794
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## avg_temperature        2.345e-04  3.962e-04   0.592 0.554007
## avg_relative_humidity  7.096e-05  1.136e-04   0.625 0.532144
## avg_dew_point         -4.105e-04  4.203e-04  -0.977 0.328741
## avg_pressure_sea      -1.136e-01  6.367e-02  -1.784 0.074499 .
## avg_pressure_station   1.184e-01  6.483e-02   1.826 0.067898 .
## avg_visibility         5.813e-08  1.987e-07   0.293 0.769837
## avg_health_index       3.220e-03  1.027e-03   3.136 0.001731 **
## precipitation         -3.607e-04  1.490e-04  -2.421 0.015554 *
## avg_cloud_cover_4      1.210e-03  7.552e-04   1.603 0.109132
## COVIDFALSE             1.104e+00  8.367e-02  13.192  < 2e-16 ***
## COVIDTRUE              1.088e+00  8.374e-02  12.991  < 2e-16 ***
## SeasonSpring           1.804e-03  1.795e-03   1.005 0.314906
## SeasonSummer           5.263e-03  1.464e-03   3.595 0.000331 ***
## SeasonWinter           1.001e-02  1.930e-03   5.187 2.31e-07 ***
## n_exercises            6.126e-04  1.936e-04   3.164 0.001576 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02094 on 2462 degrees of freedom
```

```
##   (153 observations deleted due to missingness)
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 7.443e+05 on 15 and 2462 DF,  p-value: < 2.2e-16
```