A study on the factors affecting flight delays

MSBA 6210 Project Report

Anirudh Narayanan, Hamsika Venkataramanan, Hanyuan Chi, Zihao Jiang

08/19/2016

Introduction

Flight delays are not beneficial to any of the parties involved. For the passangers, it is a nuisance that disrupts their best-laid plans. For the airports, it sets a dominoes effect off where most subsequent flights are impacted to various extents. The airlines themselves suffer reputational damage in addition to financial damages. Hence, a comprehensive understanding of what drives flight delays would be a good start to addressing this issue.

Problem Scope

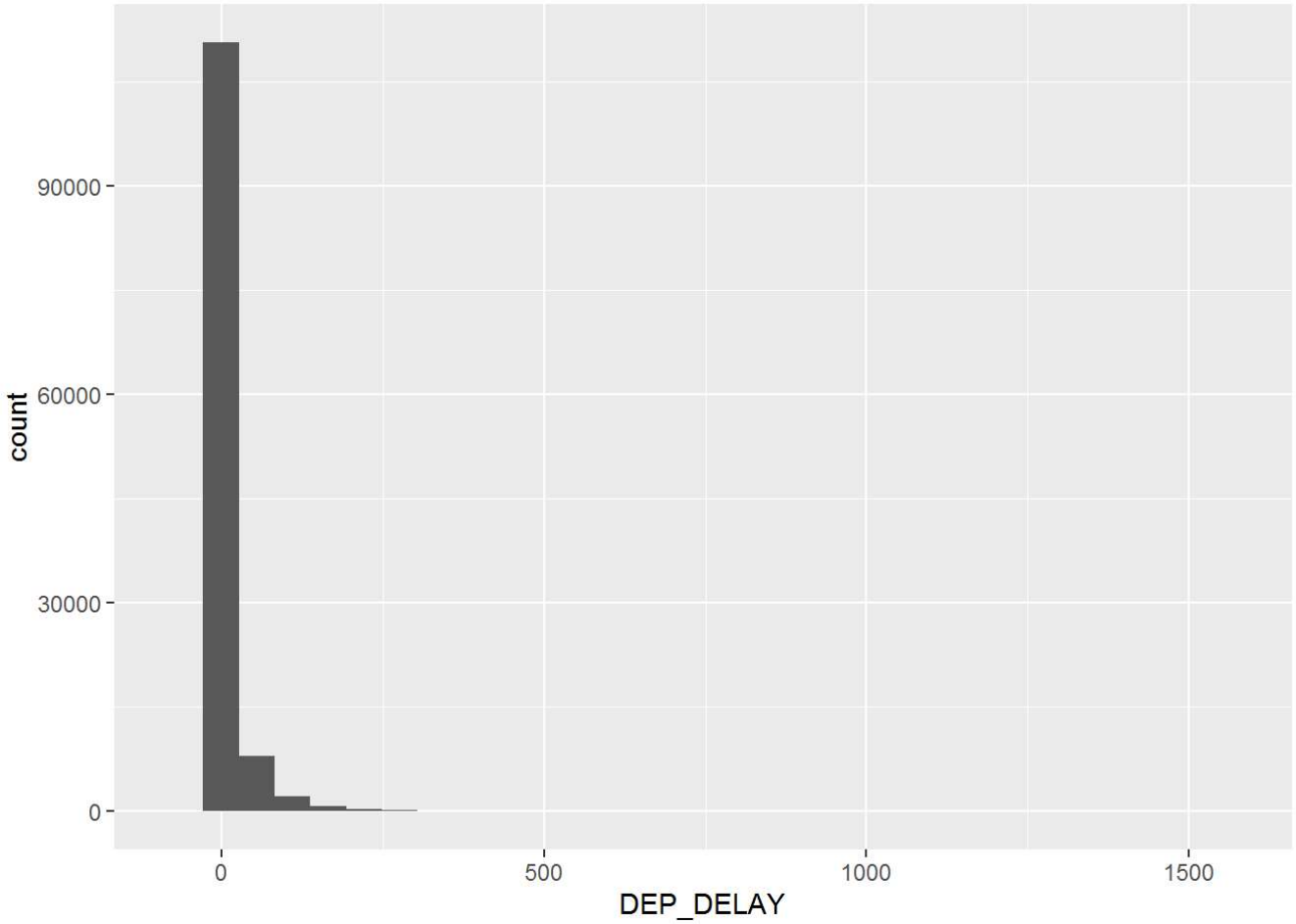
Our aim was to understand the major factors driving delays. We decided to consider delays at two major airports and contrast the results. This would allow us to check if factors at different airports have different impacts on flight delays. For this exercise, we chose Minneapolis - St. Paul (MSP) and Dallas/Fort Worth (DFW) International Airports. These airports were chosen because of their similarity in scale of operation and geographic and climatic differences. For example, snow was a common occurence in MSP and the staff there should be much better equipped at dealing with it as compared to DFW.

Data Sources and Description

The data was obtained from the United States Department of Transportation (Bureau of Transportation Statistics) for a period of 12 months from January 2015 to December 2015. Weather data for each day wass obtained from Wunderground.com.
Delay Data: http://www.transtats.bts.gov/ (http://www.transtats.bts.gov/)
Weather Data:https://www.wunderground.com/history/ (https://www.wunderground.com/history/)

Response Variable

Minutes of delay was the response variable used to build the model. Flights which departed earlier than scheduled were set to 0 values. A look at the distribution of the response showed us that it was right skewed. Hence, we considered *log of departure delays* as our response variable.



Predictors

Predictor variables were considered from two broad categories - weather conditions and flight logistics. For weather conditions, variables like precipitation, temperature, visibility and special events were considered. For flight logistics, variables like major vs minor airlines, time of day and season were considered.

- Friday through Sunday were defined as weekends
- Time of Departure of the flight were categorized into ‘Night’ (2300 - 0400), ‘Morning’ (0400 - 1100), ‘Afternoon’ (1100 - 1600) and ‘Evening’ (1600 - 2200)
- We classified all airlines into major/minor airlines (https://en.wikipedia.org/wiki/List_of_largest_airlines_in_North_America) (considering top 20 as major and the rest as minor)
- Weather factors are available at a day level. We extend those conditions to all the hours in the day as an approximation

Analysis

Hypotheses

We had a few intuitions going into the analysis regarding the impact of some of the variables.

- We expected snow and other forms of precipitation to have a negative impact in both airports. However, the impact of snow would be much higher in DFW than MSP. Extending this to seasons, we expected fall-winter to be worse than spring-summer
- We expected weekends to have higher delays as they would be busier. Extending this to time of day, we expected evening and afternoon flights to be delayed more
- Major airlines were expected to fare better due to the resources available as compared to minor airlines

- We expected long-distance flights are more susceptible to delays as there is higher amount of preparation invovled

Model Selection

Given the number of predictor variables, the best fitting model was decided by backward selection rather than by forward substitution. We started with all possible, logically intuitive predictors in our model which includes logistic, weather variables for the current as well as for the previous day. The best fit model was arrived at when the removal of extra predictors did not have a positive impact on R squared or failed to reduce the standard error.

Results and Findings

Model explainability

The models for MSP and DFW were able to explain ~6% and ~10% of the respective variation in flight delays. Thus, we aren't able to achieve the desired comprehensiveness in terms of explaining delays. However, the factors that we did include were significant ones.

Similarities

Although the scale might be different, weather conditions seem to have similar impacts on both airports

- As expected, weather wreaks havoc on flights in terms of delays. An inch increase in precipiation leads to a **~32%** increase in delays for MSP and a **~10%** increase in delays for DFW
- **June** was found to have the highest amount of delays in both airports. This is contrary to our expectation that delays should be highest in winter months
- **Flight distance** was also found to have similar impacts in both locations. A 10% increase flight distance causes a **~0.6%** and **~1%** increase in delays for MSP and DFW respectively. This is supportive of our initial intuitions
- Weather events occuring in conjunction cause more delays than individual events. The most disruptive combinations are **Rain-Snow-Thunderstorm** for DFW and **Fog-Snow** for MSP, causing **~130%** and **~52%** more delays than event-free days respectively.

Contrasts

A few conditions were observed to have contrasting impacts in the two airports.

- **Sea level pressure** turned out to be significant factor in DFW whereas it did not in MSP. This could be attributed to the susceptibility of DFW to thunderstorms which form in low-pressure zones
- Major airlines were found to do better than minor in DFW (as was expected). The trend in MSP was flipped. Major airlines had **~4%** more delays than minor ones in MSP. This is contrary to our initial intuitions
- With respect to time of the day, **evenings** are the worst in DFW whereas **nights** are the worst in MSP.
- Flight delays are worse on weekends for DFW but better for MSP. The latter is contrary, again, to our initial expectations on this matter.

Note: *The interpretation for each predictor's impact is over and above the impact of all the other predictors*

Limitations and next steps

As mentioned earlier, these factors are only able to account for a fraction of the variation present in flight delays. Thus, these models do not fit the bill when it comes to predicting flight delays. To improve explainability, more variables would be included. Some examples of such variables are

- the degree to which the airports were busy at the time of the flight
- hourly weather variables (as opposed to the current day-level ones)

Some of the assumptions were seen to be violated and thus requires further investigating into.

Appendix

Model Results

MSP

```
##
## Call:
## lm(formula = log_delay2 ~ min_temperaturef + cloudcover + min_visibilitymiles +
##     mean_wind_speedmph + max_gust_speedmph + min_humidity + events +
##     airline_class + MONTH + ln_dist + prec_new + time_of_day +
##     is_weekend, data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3358 -0.8832 -0.5655  0.5709  6.8719
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.2208309  0.0634968  19.227  < 2e-16 ***
## min_temperaturef             -0.0130534  0.0005283 -24.706  < 2e-16 ***
## cloudcover                    0.0073330  0.0026425   2.775  0.00552 **
## min_visibilitymiles           0.0045969  0.0020985   2.191  0.02848 *
## mean_wind_speedmph            0.0082265  0.0019284   4.266 1.99e-05 ***
## max_gust_speedmph             0.0036821  0.0009141   4.028 5.63e-05 ***
## min_humidity                  0.0021802  0.0004719   4.620 3.84e-06 ***
## eventsFog                     0.0973924  0.0439667   2.215  0.02675 *
## eventsFog-Rain-Thunderstorm   0.4042942  0.0391882  10.317  < 2e-16 ***
## eventsFog-Snow                0.5207792  0.0349507  14.900  < 2e-16 ***
## eventsRain                    0.0449959  0.0148457   3.031  0.00244 **
## eventsRain-Snow               0.3888364  0.0287277  13.535  < 2e-16 ***
## eventsRain-Snow-Thunderstorm  0.0804298  0.0772488   1.041  0.29779
## eventsRain-Thunderstorm       0.1248902  0.0183303   6.813 9.58e-12 ***
## eventsSnow                    0.1910539  0.0205339   9.304  < 2e-16 ***
## airline_class1                0.0366133  0.0092909   3.941 8.13e-05 ***
## MONTH1                       -0.6885630  0.0333261 -20.661  < 2e-16 ***
## MONTH2                       -0.7230168  0.0364237 -19.850  < 2e-16 ***
## MONTH3                       -0.6018235  0.0265755 -22.646  < 2e-16 ***
## MONTH4                       -0.5771428  0.0232810 -24.790  < 2e-16 ***
## MONTH5                       -0.3903140  0.0206987 -18.857  < 2e-16 ***
## MONTH7                       -0.0843472  0.0196767  -4.287 1.82e-05 ***
## MONTH8                       -0.1809464  0.0192561  -9.397  < 2e-16 ***
## MONTH9                       -0.3725031  0.0201011 -18.531  < 2e-16 ***
## MONTH10                      -0.5970179  0.0213689 -27.939  < 2e-16 ***
## MONTH11                      -0.6263724  0.0242576 -25.822  < 2e-16 ***
## MONTH12                      -0.6594159  0.0290776 -22.678  < 2e-16 ***
## ln_dist                       0.0629525  0.0065480   9.614  < 2e-16 ***
## prec_new                      0.3203090  0.0166131  19.280  < 2e-16 ***
## time_of_dayafternoon         -0.2897806  0.0094873 -30.544  < 2e-16 ***
## time_of_daymorning           -0.5361754  0.0097314 -55.097  < 2e-16 ***
## time_of_daynight              0.4826427  0.1127526   4.281 1.87e-05 ***
## is_weekend1                  -0.0495134  0.0083336  -5.941 2.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.376 on 121701 degrees of freedom
##   (207 observations deleted due to missingness)
## Multiple R-squared:  0.05955,    Adjusted R-squared:  0.05931
## F-statistic: 240.8 on 32 and 121701 DF,  p-value: < 2.2e-16
```

DFW

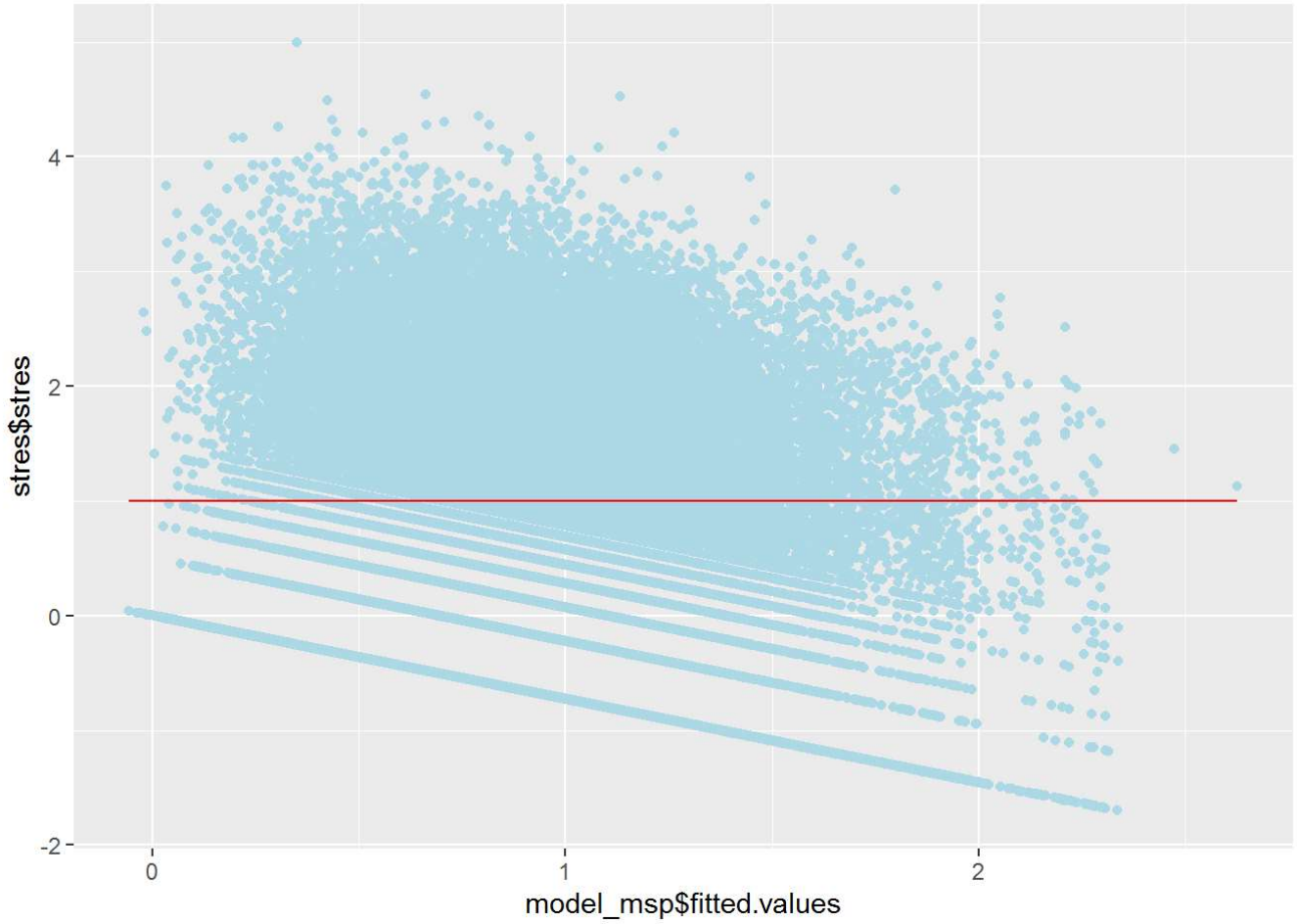
```
##
## Call:
## lm(formula = log_delay2 ~ min_temperaturef + min_visibilitymiles +
##     mean_wind_speedmph + max_gust_speedmph + min_humidity + min_humidity_prev +
##     min_sea_level_pressurein + events + airline_class + MONTH +
##     cloudcover + ln_dist + prec_new + time_of_day + is_weekend,
##     data = model_data_dfw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3657 -1.0867 -0.6429  1.1240  6.3392
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.9250844  0.7991883  12.419  < 2e-16
## min_temperaturef               -0.0238955  0.0005980 -39.956  < 2e-16
## min_visibilitymiles            -0.0084073  0.0016659  -5.047 4.50e-07
## mean_wind_speedmph             -0.0050318  0.0014676  -3.429 0.000607
## max_gust_speedmph               0.0100187  0.0007123  14.065  < 2e-16
## min_humidity                    0.0091093  0.0003386  26.903  < 2e-16
## min_humidity_prev               0.0042709  0.0002463  17.341  < 2e-16
## min_sea_level_pressurein       -0.2669640  0.0259255 -10.297  < 2e-16
## eventsFog                       0.4797021  0.0290576  16.509  < 2e-16
## eventsFog-Rain                  0.0878097  0.0371293   2.365 0.018032
## eventsFog-Rain-Snow-Thunderstorm -0.6116849  0.0741061  -8.254  < 2e-16
## eventsFog-Rain-Thunderstorm     0.6008501  0.0294588  20.396  < 2e-16
## eventsRain                      0.0579645  0.0124984   4.638 3.52e-06
## eventsRain-Snow                 0.7252737  0.0343114  21.138  < 2e-16
## eventsRain-Snow-Thunderstorm    1.3049121  0.0658185  19.826  < 2e-16
## eventsRain-Thunderstorm         0.3822418  0.0153092  24.968  < 2e-16
## eventsSnow                      0.5801600  0.0477130  12.159  < 2e-16
## eventsThunderstorm              0.1314281  0.0336099   3.910 9.22e-05
## airline_class1                 -0.0179983  0.0072325  -2.489 0.012827
## MONTH1                         -0.9400967  0.0250919 -37.466  < 2e-16
## MONTH2                         -1.2257198  0.0252925 -48.462  < 2e-16
## MONTH3                         -0.8044227  0.0197692 -40.691  < 2e-16
## MONTH4                         -0.8056211  0.0180907 -44.532  < 2e-16
## MONTH5                         -0.8387453  0.0166402 -50.405  < 2e-16
## MONTH7                         -0.0190739  0.0151625  -1.258 0.208407
## MONTH8                         -0.0331980  0.0151134  -2.197 0.028050
## MONTH9                         -0.4911842  0.0150785 -32.575  < 2e-16
## MONTH10                        -0.7896191  0.0164141 -48.106  < 2e-16
## MONTH11                        -1.0835773  0.0198728 -54.526  < 2e-16
## MONTH12                        -1.1494273  0.0230801 -49.802  < 2e-16
## cloudcover                     -0.0063510  0.0018808  -3.377 0.000734
## ln_dist                         0.1021666  0.0048557  21.040  < 2e-16
## prec_new                        0.1146818  0.0094411  12.147  < 2e-16
## time_of_dayafternoon           -0.1943504  0.0072494 -26.809  < 2e-16
## time_of_daymorning             -0.5174450  0.0073351 -70.544  < 2e-16
## time_of_daynight               -0.0516518  0.0488116  -1.058 0.289972
## is_weekend1                     0.0149209  0.0062451   2.389 0.016885
##
## (Intercept)                    ***
## min_temperaturef               ***
## min_visibilitymiles            ***
## mean_wind_speedmph             ***
## max_gust_speedmph              ***
```

```
## min_humidity                      ***
## min_humidity_prev                 ***
## min_sea_level_pressurein          ***
## eventsFog                         ***
## eventsFog-Rain                    *
## eventsFog-Rain-Snow-Thunderstorm ***
## eventsFog-Rain-Thunderstorm       ***
## eventsRain                        ***
## eventsRain-Snow                   ***
## eventsRain-Snow-Thunderstorm      ***
## eventsRain-Thunderstorm           ***
## eventsSnow                        ***
## eventsThunderstorm                ***
## airline_class1                    *
## MONTH1                            ***
## MONTH2                            ***
## MONTH3                            ***
## MONTH4                            ***
## MONTH5                            ***
## MONTH7
## MONTH8                            *
## MONTH9                            ***
## MONTH10                           ***
## MONTH11                           ***
## MONTH12                           ***
## cloudcover                        ***
## ln_dist                           ***
## prec_new                          ***
## time_of_dayafternoon              ***
## time_of_daymorning                ***
## time_of_daynight
## is_weekend1                       *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.513 on 254268 degrees of freedom
## Multiple R-squared:  0.0959, Adjusted R-squared:  0.09577
## F-statistic: 749.2 on 36 and 254268 DF,  p-value: < 2.2e-16
```

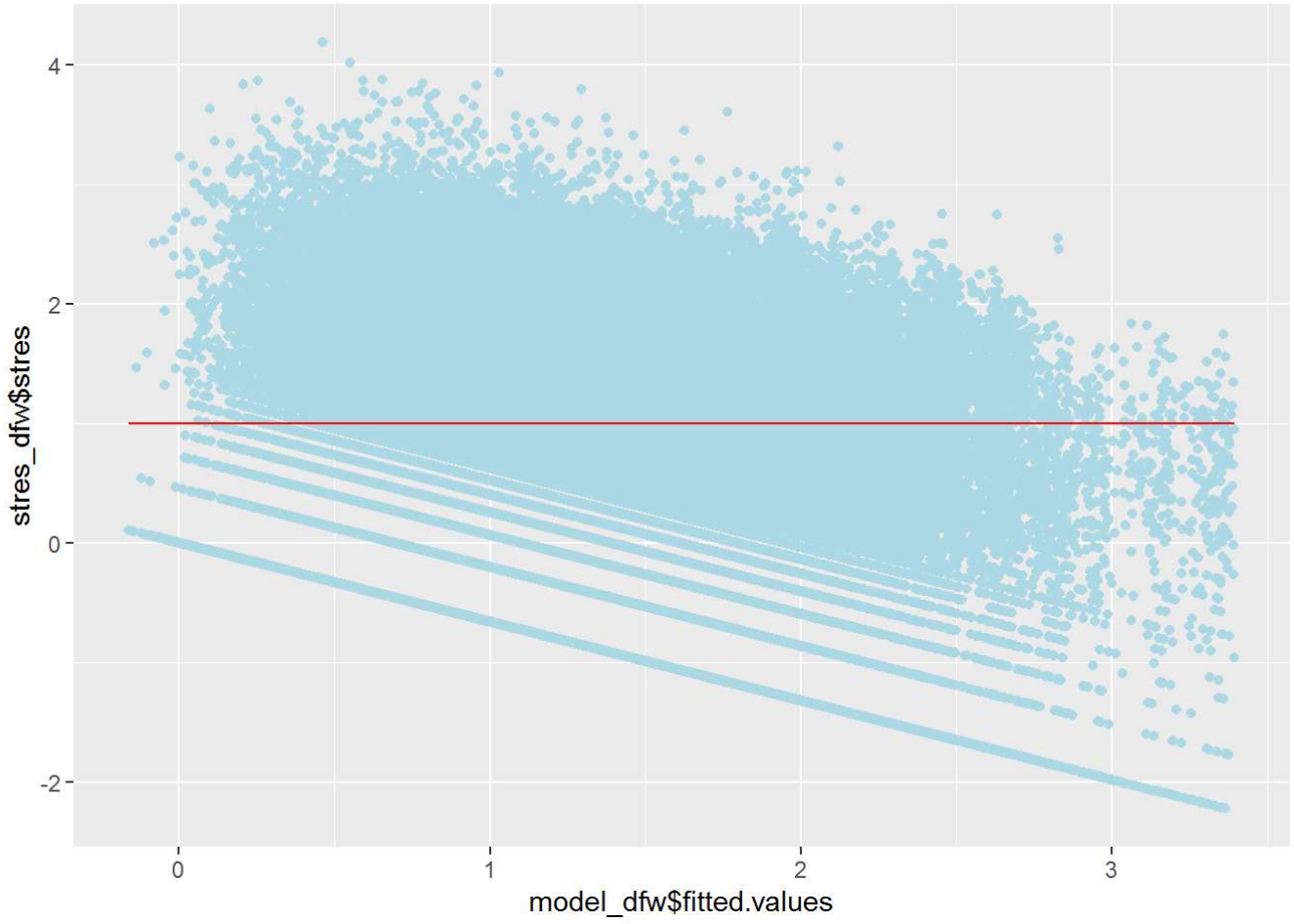
Residual plots

We can see from the residual plots that they are not normally distributed for either MSP or DFW.

MSP



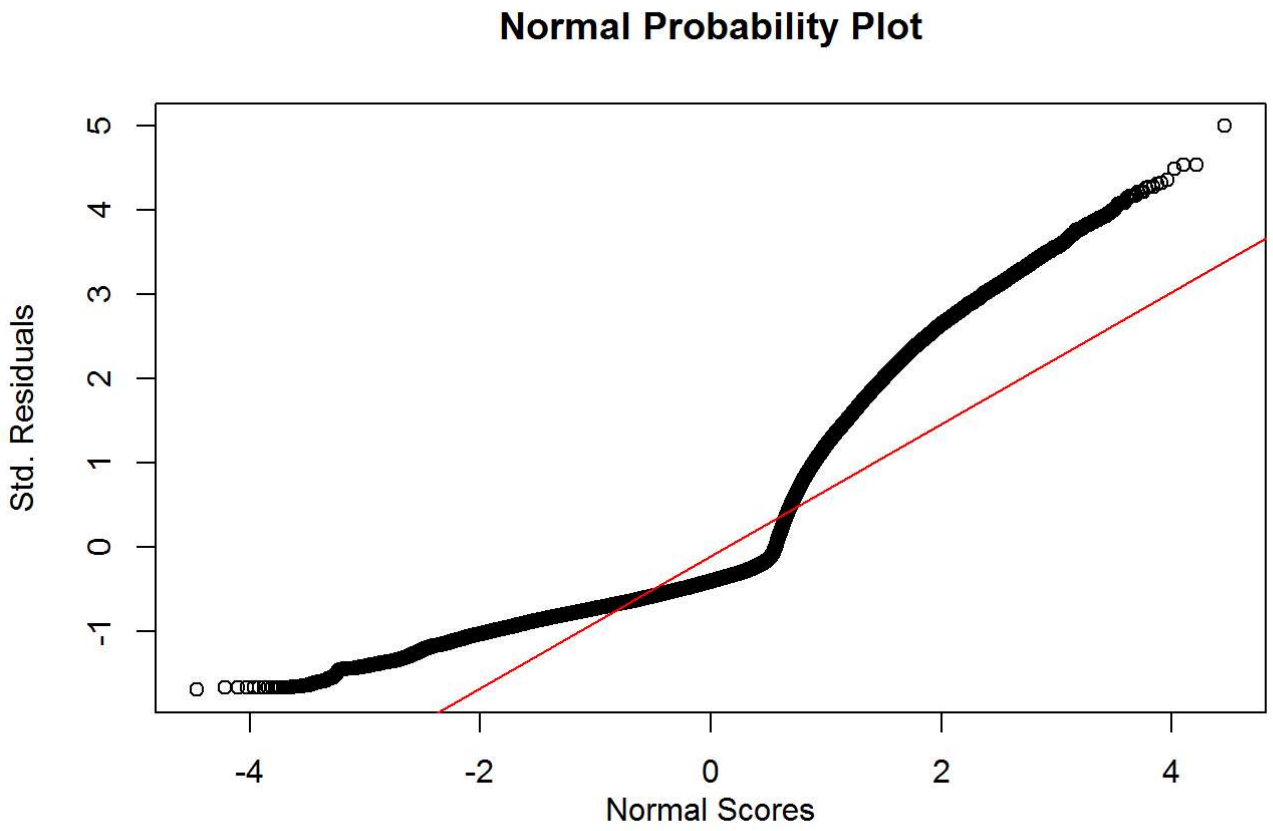
DFW



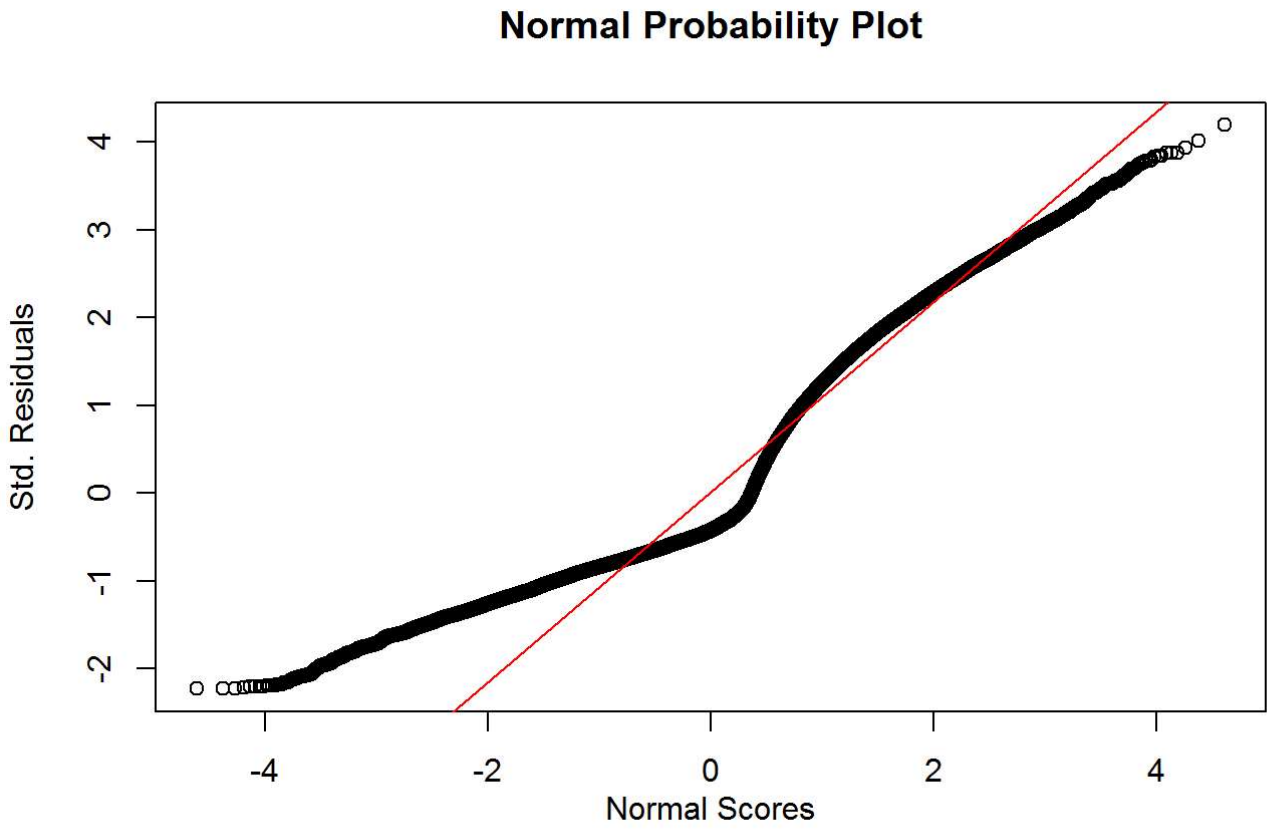
Residual Normality

The normality violation is further supported by the normal probability plots

MSP



DFW



Correlation matrices

There is no evidence for high degrees of correlation for either DFW or MSP.

MSP

```
##                     min_temperaturef   cloudcover min_visibilitymiles
## min_temperaturef          1.00000000 -0.022733986        0.0401953331
## cloudcover               -0.02273399  1.000000000       -0.6108297657
## min_visibilitymiles       0.04019533 -0.610829766        1.0000000000
## mean_wind_speedmph       -0.12018179  0.201245185       -0.2205635440
## min_humidity             -0.09013983  0.629847839       -0.5853277920
## ln_dist                  -0.01081712 -0.002279078        0.0001291282
## prec_new                  0.18548096  0.278765365       -0.4446175679
##                     mean_wind_speedmph min_humidity       ln_dist
## min_temperaturef          -0.120181786  -0.09013983 -0.0108171220
## cloudcover                 0.201245185   0.62984784 -0.0022790781
## min_visibilitymiles       -0.220563544  -0.58532779  0.0001291282
## mean_wind_speedmph         1.000000000   0.06140659  0.0034703349
## min_humidity               0.061406591   1.00000000 -0.0118932025
## ln_dist                    0.003470335  -0.01189320  1.0000000000
## prec_new                   0.158268370   0.27459575 -0.0034275144
##                         prec_new
## min_temperaturef     0.185480959
## cloudcover           0.278765365
## min_visibilitymiles -0.444617568
## mean_wind_speedmph   0.158268370
## min_humidity         0.274595747
## ln_dist             -0.003427514
## prec_new             1.000000000
```

DFW

```
##                          min_temperaturef min_visibilitymiles
## min_temperaturef              1.000000000          0.19581358
## min_visibilitymiles           0.195813582          1.00000000
## mean_wind_speedmph            0.067965543         -0.15142949
## min_humidity_prev            -0.195126870         -0.53505408
## min_sea_level_pressurein     -0.513807605          0.11007164
## cloudcover                   -0.145612452         -0.64009348
## ln_dist                       0.007958614         -0.00881235
## prec_new                     -0.020275006         -0.58387878
##                          mean_wind_speedmph min_humidity_prev
## min_temperaturef                0.067965543       -0.19512687
## min_visibilitymiles            -0.151429494       -0.53505408
## mean_wind_speedmph              1.000000000        0.14323333
## min_humidity_prev               0.143233332        1.00000000
## min_sea_level_pressurein       -0.304798592       -0.02679294
## cloudcover                      0.205520778        0.53687920
## ln_dist                         0.003172921        0.00407672
## prec_new                        0.157749781        0.26767798
##                          min_sea_level_pressurein   cloudcover
## min_temperaturef                     -0.513807605 -0.145612452
## min_visibilitymiles                   0.110071640 -0.640093476
## mean_wind_speedmph                   -0.304798592  0.205520778
## min_humidity_prev                    -0.026792939  0.536879198
## min_sea_level_pressurein              1.000000000 -0.024549437
## cloudcover                           -0.024549437  1.000000000
## ln_dist                              -0.005084242  0.005172326
## prec_new                             -0.134483844  0.395010320
##                              ln_dist    prec_new
## min_temperaturef          0.007958614 -0.02027501
## min_visibilitymiles      -0.008812350 -0.58387878
## mean_wind_speedmph        0.003172921  0.15774978
## min_humidity_prev         0.004076720  0.26767798
## min_sea_level_pressurein -0.005084242 -0.13448384
## cloudcover                0.005172326  0.39501032
## ln_dist                   1.000000000  0.01056443
## prec_new                  0.010564432  1.00000000
```