Airline Stock Price and Delays

Student

2022-11-10

Libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##      filter, lag

## The following objects are masked from 'package:base':
##
##      intersect, setdiff, setequal, union
```

Initial Data

There are 3 initial data files. There is a data file for each of 3 chosen airlines with all BTS monthly delay data

```
dal_data <- read.csv("delta_lbc.csv")
sw_data <- read.csv("southwest_lbc.csv")
ual_data <- read.csv("united_lbc.csv")
```

Delta Model1

This is a model of stock price regressed only on individual summary delay values (not the Arr_del15 agregated value) reported in flight count increments. There are additional delay values reported in minutes that are not used in this model. The full data set of pre and post COVID months are used.

```
dal_model1 <- lm(Open.Stock.Price ~ Sum.of.carrier_ct
                 + Sum.of.weather_ct + Sum.of.nas_ct + Sum.of.security_ct
                 + Sum.of.late_aircraft_ct + Sum.of.arr_cancelled
                 , data = dal_data)

par(mfrow = c(2,2))
plot(dal_model1)
```

```
par(mfrow = c(1,1))
```
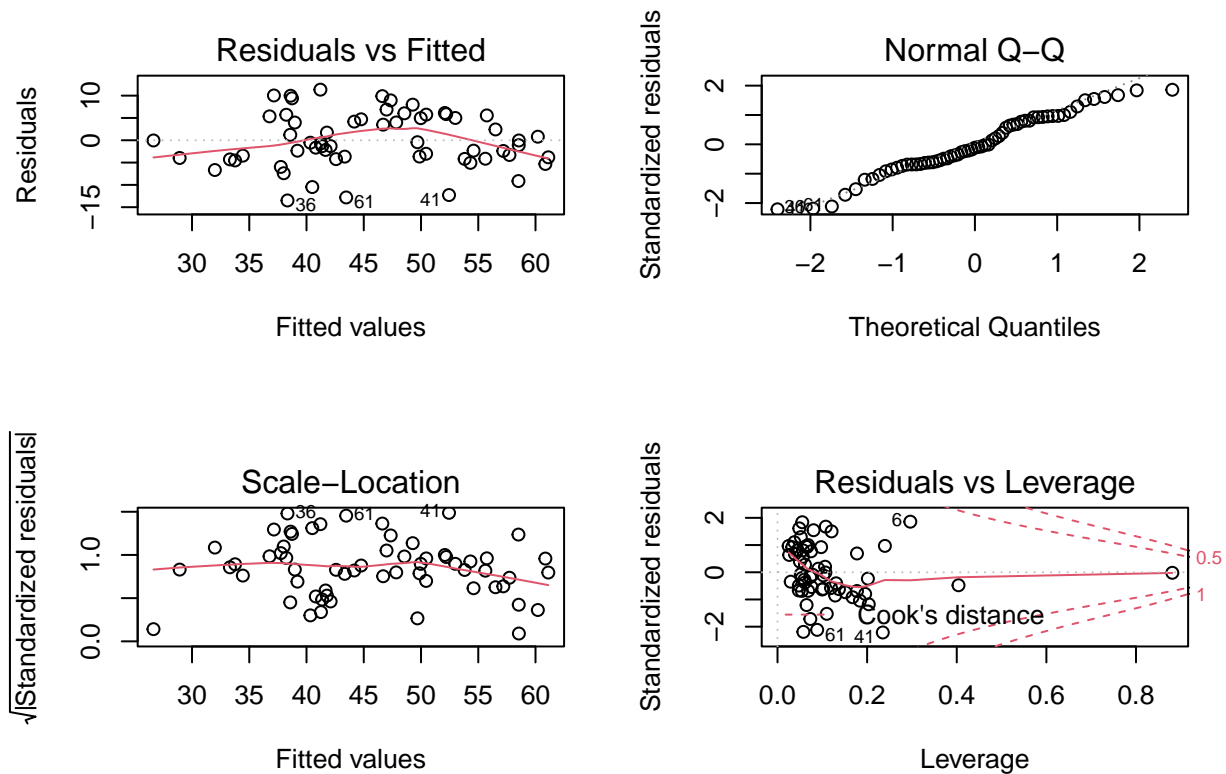
Review of Diagnostic Plots for Delta Regression Model1

The Residuals vs Fitted plot indicates a linear model absent of predominate non-linear shapes. The Normal Q-Q is ok with a bit of drift on the top right data values straying from the diagonal line. The Spread-Location plot is ok also with residuals spread fairly equally.. The Residuals vs Leverage plot calls is pretty good.

```
summary(dal_model1)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
##     Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
##     Sum.of.arr_cancelled, data = dal_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4882  -3.9687  -0.6903   4.9931  11.3364
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             30.7215903  3.4254225   8.969 2.78e-12 ***
## Sum.of.carrier_ct       -0.0016266  0.0023818  -0.683   0.4976
## Sum.of.weather_ct        0.0054750  0.0084110   0.651   0.5178
## Sum.of.nas_ct            0.0102267  0.0019186   5.330 1.97e-06 ***
## Sum.of.security_ct       0.0181690  0.1033754   0.176   0.8611
```

```
## Sum.of.late_aircraft_ct  -0.0044426   0.0038072   -1.167    0.2484
## Sum.of.arr_cancelled      -0.0010160   0.0004345   -2.338    0.0231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.342 on 54 degrees of freedom
## Multiple R-squared:  0.6683, Adjusted R-squared:  0.6315
## F-statistic: 18.13 on 6 and 54 DF,  p-value: 2.156e-11
```

Review of the Fit and Significance for Delta Regression Model1

The Sum.of.arr_cancelled coefficient is significant at a greater than 95% level. The null hypothisis that Sum.of.arr_cancelled is not significant can be rejected. The R-sqared value is a bit belowthe financial industry standard of above 0.7 for a high level of significance, although fairly close at 0.6683. The Sum.of.nas_ct high significance may be correlation and is a positive coefficient. Field knowledge would put more of an emphasis on Sum.of.arr_cancelled which is significant at 97% plus.

Look at the Flight Delay Linar Regression Model Coefficients That Are in Minutes

```
dal_model1b <- lm(Open.Stock.Price ~ Sum.of.carrier_delay
                  + Sum.of.weather_delay + Sum.of.nas_delay + Sum.of.security_delay
                  + Sum.of.late_aircraft_delay
                  , data = dal_data)
par(mfrow = c(2,2))
plot(dal_model1b)
```
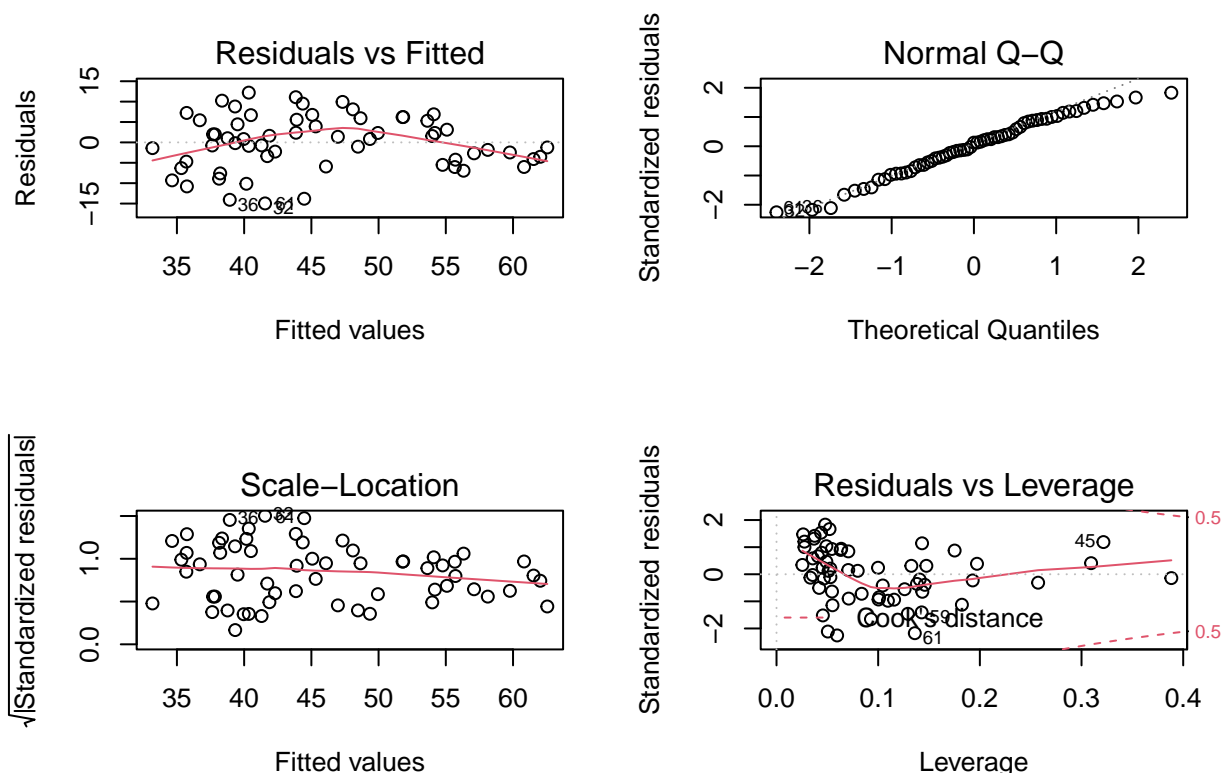
```
par(mfrow = c(1,1))
```

```
summary(dal_model1b)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_delay + Sum.of.weather_delay +
##     Sum.of.nas_delay + Sum.of.security_delay + Sum.of.late_aircraft_delay,
##     data = dal_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9652  -4.2385   0.8247   5.4296  12.2049
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.390e+01  2.903e+00  11.680  < 2e-16 ***
## Sum.of.carrier_delay      -4.567e-07  2.442e-05  -0.019   0.9851
## Sum.of.weather_delay       8.494e-05  6.935e-05   1.225   0.2258
## Sum.of.nas_delay           1.842e-04  3.608e-05   5.104 4.27e-06 ***
## Sum.of.security_delay      6.777e-04  1.301e-03   0.521   0.6046
## Sum.of.late_aircraft_delay -1.023e-04  4.718e-05  -2.168   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.836 on 55 degrees of freedom
## Multiple R-squared:  0.6076, Adjusted R-squared:  0.5719
## F-statistic: 17.03 on 5 and 55 DF,  p-value: 3.829e-10
```

Review of the Fit and Significance of Model1b

R-squared is too low at 0.6076. Sum.of.late_aircraft_delay with a correct negative value is good but Sum.of.nas_delay is positive and not in line with domain or field knowledge.

Go Back to Model1 For Southwest and United

Southwest Model1

This is a model of stock price regressed only on summary delay values reported in flight count increments. There are additional delay values reported in minutes that are not used in this model. The full data set of pre and post COVID months are used.

```
sw_model1 <- lm(Open.Stock.Price ~ Sum.of.carrier_ct
             + Sum.of.weather_ct + Sum.of.nas_ct + Sum.of.security_ct
             + Sum.of.late_aircraft_ct + Sum.of.arr_cancelled
             , data = sw_data)
```

```
par(mfrow = c(2,2))
plot(sw_model1)
```

```
par(mfrow = c(1,1))
```
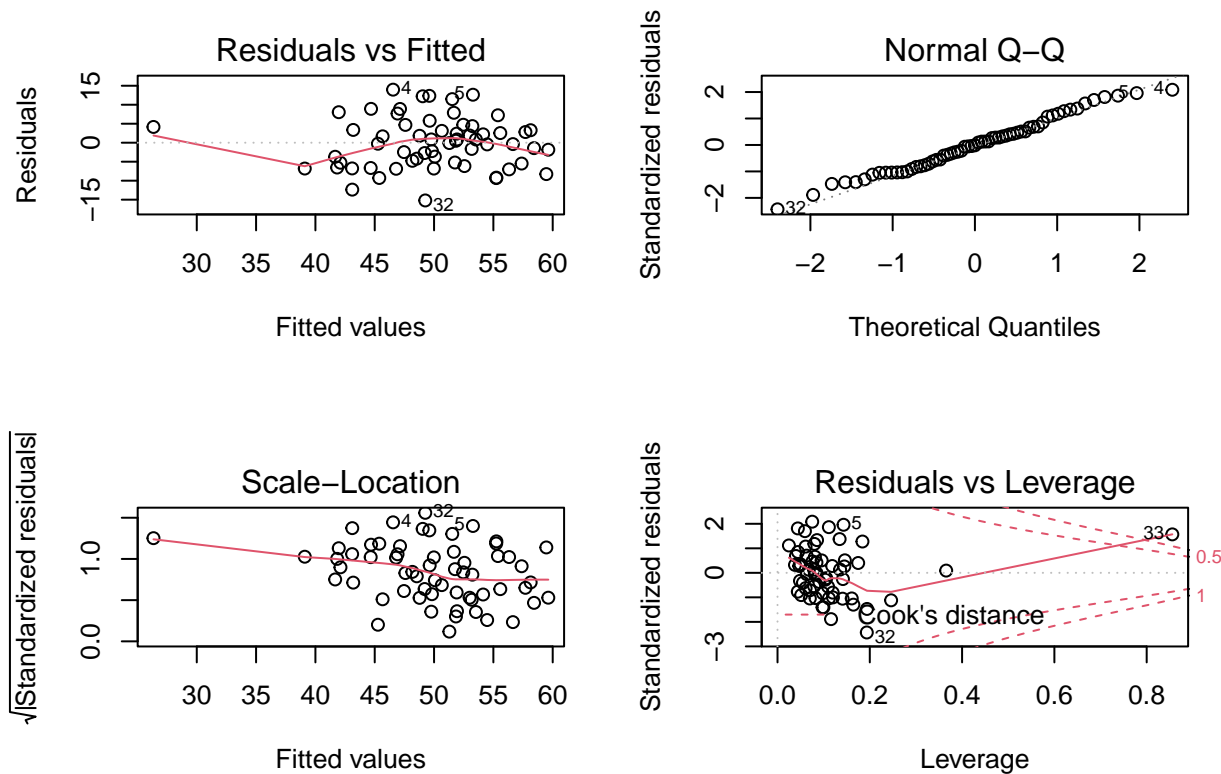
Review of Diagnostic Plots for Southwest Regression Model1

The Residuals vs Fitted plot indicates a linear model absent of non-linear shapes. The Normal Q-Q is ok with a bit of drift on the top right data values straying from the diagonal line. The Spread-Location plot is ok also with residuals spread fairly equally except on the left side of the plot. The Residuals vs Leverage plot calls out data point 33 as a potential influential case. This data point is approximately 2 months after the start of COVID. While data point 33 can arguably be removed from the data set prior work with Delta data and other work not in this R Markdown indicate it is a not feasible approach.

```
summary(sw_model1)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
##     Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
##     Sum.of.arr_cancelled, data = sw_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.208  -5.272  -0.093   4.149  13.965
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          38.5768082  3.2175153  11.990  < 2e-16 ***
## Sum.of.carrier_ct    -0.0005726  0.0011239  -0.509   0.6125
```

```
## Sum.of.weather_ct         -0.0088185  0.0065027  -1.356   0.1807
## Sum.of.nas_ct              0.0092238  0.0018555   4.971  7.1e-06 ***
## Sum.of.security_ct         0.0602469  0.0445704   1.352   0.1821
## Sum.of.late_aircraft_ct   -0.0020403  0.0011107  -1.837   0.0717 .
## Sum.of.arr_cancelled       -0.0003392  0.0001418  -2.393   0.0202 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.962 on 54 degrees of freedom
## Multiple R-squared:  0.4353, Adjusted R-squared:  0.3726
## F-statistic: 6.938 on 6 and 54 DF,  p-value: 1.678e-05
```

## Review of the Fit and Significance Southwest Model1

The R-sqared value is a below the financial industry standard of above 0.7 for a high level of significance at 0.4353. The Sum.of.arr_cancelled and Sum.of.nas_ct coefficients again come up similar to the Delta Airline models.

United Model1

This is a model of stock price regressed only on summary delay values reported in flight count increments. There are additional delay values reported in minutes that are not used in this model. The full data set of pre and post COVID months are used.

```
ual_model1 <- lm(Open.Stock.Price ~ Sum.of.carrier_ct
                 + Sum.of.weather_ct + Sum.of.nas_ct + Sum.of.security_ct
                 + Sum.of.late_aircraft_ct + Sum.of.arr_cancelled
                 , data = ual_data)

par(mfrow = c(2,2))
plot(ual_model1)
```

```
par(mfrow = c(1,1))
```
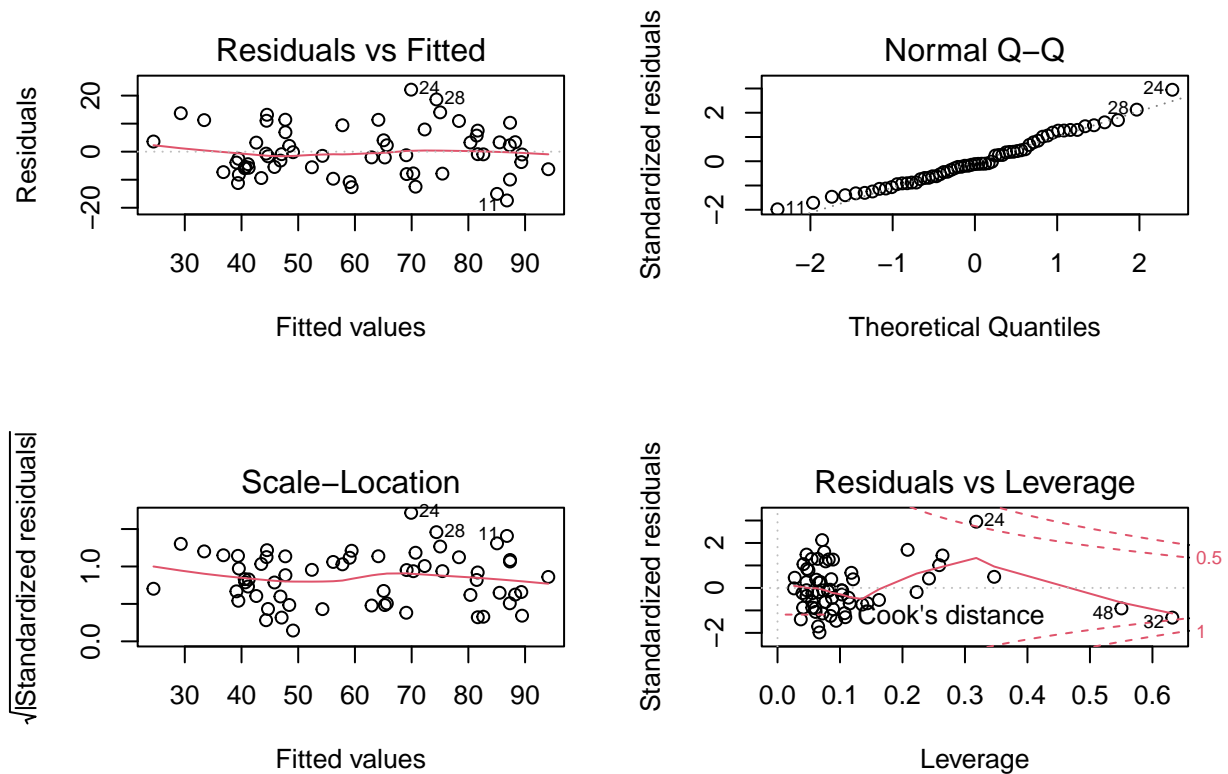
Review of Diagnostic Plots for United Regression Model1

The Residuals vs Fitted plot indicates a linear model absent of non-linear shapes. The Normal Q-Q is good. The Spread-Location plot is good with residuals spread fairly equally. The Residuals vs Leverage plot is a scattered somewhat but not top right or bottom right possible high-influence variables.

```
summary(ual_model1)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
##     Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
##     Sum.of.arr_cancelled, data = ual_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.400  -5.901  -1.010   5.740  22.119
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          40.2938833  3.2669519  12.334  < 2e-16 ***
## Sum.of.carrier_ct    -0.0097907  0.0021966  -4.457 4.23e-05 ***
## Sum.of.weather_ct    -0.0168770  0.0082410  -2.048  0.04544 *
## Sum.of.nas_ct         0.0154198  0.0020535   7.509 6.15e-10 ***
## Sum.of.security_ct   -0.5181368  0.3495411  -1.482  0.14406
```

7

```
## Sum.of.late_aircraft_ct   0.0009860   0.0030146    0.327   0.74487
## Sum.of.arr_cancelled      -0.0016752   0.0006269   -2.672   0.00995 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.089 on 54 degrees of freedom
## Multiple R-squared:  0.8269, Adjusted R-squared:  0.8076
## F-statistic: 42.99 on 6 and 54 DF,  p-value: < 2.2e-16
```

Review of the Fit and Significance United Model1

Sum.of.carrier_ct, Sum.of.weather_ct, and Sum.of.arr_cancelled are all significant at a greater than 95% level and have a correct negative value coefficient. The Sum.of.nas_ct coefficient is positive and not in line with field knowledge, this is identical to the Delta Airlines and SW Airlines models that had low R-squared values. The R-sqared value is above the financial industry standard of above 0.7 for a high level of significance at 0.8269. The null hypothesis that the coefficients are not significant can be rejected.

Overview Conclusion

Delay counts for Southwest do not model for fit or significance with the data. Delay counts for Delta are are closer for fit and significance for the Sum.of.arr_cancelled coefficient but the R-squared value is still too low. United models well on Sum.of.carrier_ct, Sum.of.weather_ct, and Sum.of.arr_cancelled for fit and signifcance of the model and correct sign of the coefficients.

## Pre and Post COVID Data Modeling - Post COVID Meaning After Outbreak

Use new yyyymm field for better COVID delineation of data points 03 March 2020 and 04 April are left off on purpose as transition months

```
dal_pre_data <- filter(dal_data, yyyymm < 202003)
dal_post_data <- filter(dal_data, yyyymm > 202004)
sw_pre_data <- filter(sw_data, yyyymm < 202003)
sw_post_data <- filter(sw_data, yyyymm > 202004)
ual_pre_data <- filter(ual_data, yyyymm < 202003)
ual_post_data <- filter(ual_data, yyyymm > 202004)
```

Delta pre-COVID model with flight count data

```
dal_pre_model <- lm(Open.Stock.Price ~ Sum.of.carrier_ct
                + Sum.of.weather_ct + Sum.of.nas_ct + Sum.of.security_ct
                + Sum.of.late_aircraft_ct + Sum.of.arr_cancelled
                    , data = dal_pre_data)

par(mfrow = c(2,2))
plot(dal_pre_model)
```

```r
par(mfrow = c(1,1))
```
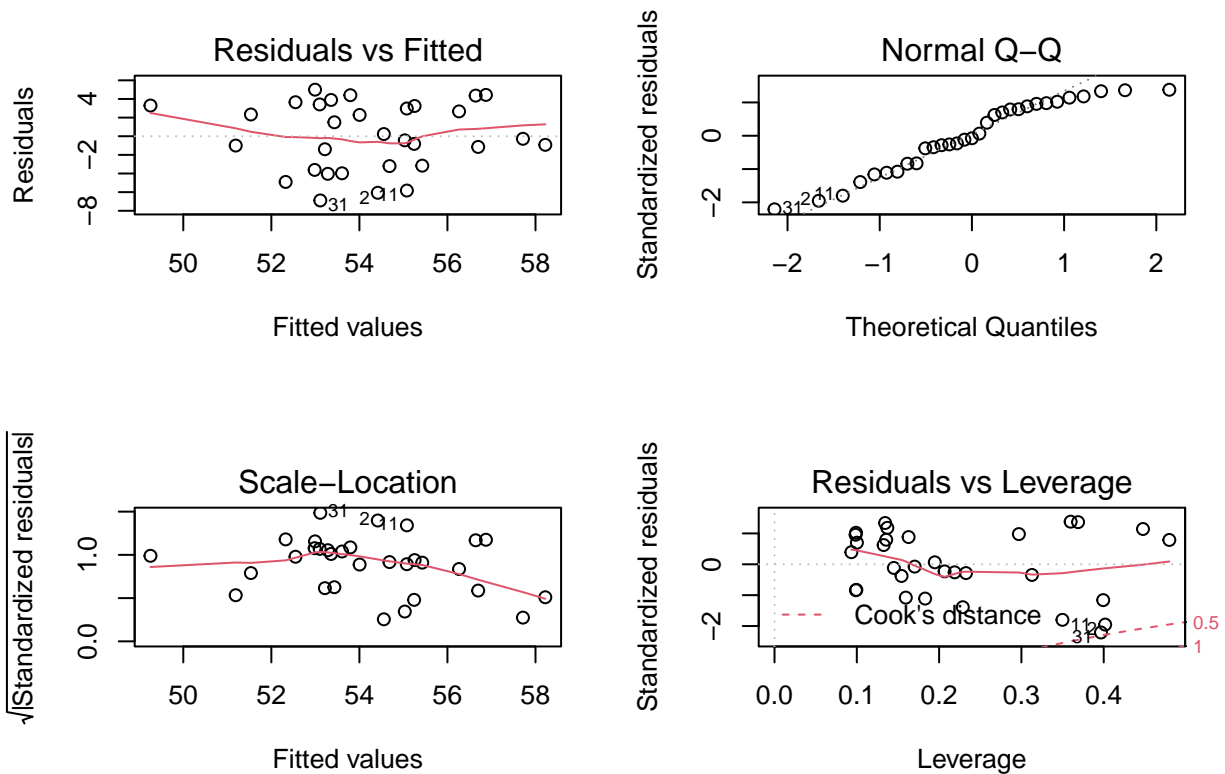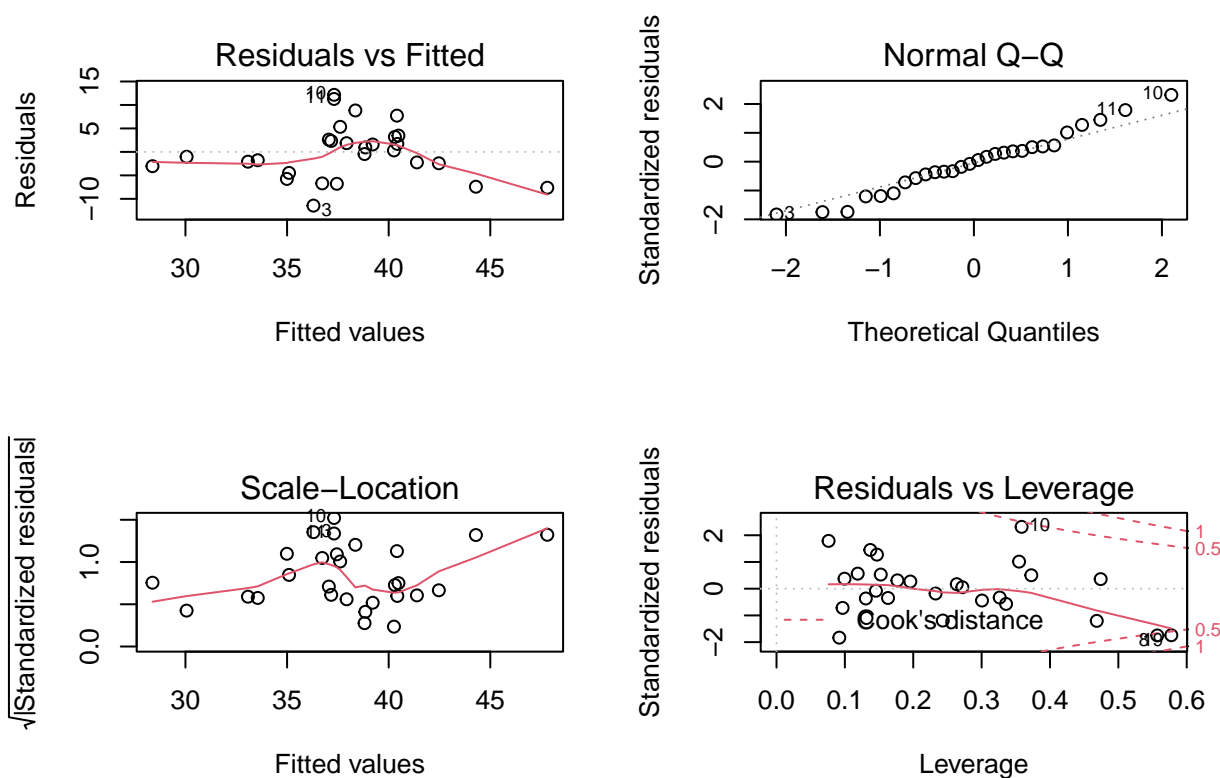
```r
summary(dal_pre_model)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
##     Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
##     Sum.of.arr_cancelled, data = dal_pre_data)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -6.8920 -3.1810 -0.2753  3.2707  4.9971
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             44.273587   4.369541  10.132  3.8e-10 ***
## Sum.of.carrier_ct        0.002423   0.003685   0.657    0.517
## Sum.of.weather_ct       -0.005033   0.008558  -0.588    0.562
## Sum.of.nas_ct            0.003145   0.002511   1.252    0.223
## Sum.of.security_ct      -0.205759   0.242805  -0.847    0.405
## Sum.of.late_aircraft_ct -0.001517   0.004364  -0.348    0.731
## Sum.of.arr_cancelled     0.002542   0.002577   0.986    0.334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.017 on 24 degrees of freedom
## Multiple R-squared:  0.2279, Adjusted R-squared:  0.0349
## F-statistic: 1.181 on 6 and 24 DF,  p-value: 0.3494
```

9

Conclusion - poor R-Squared value

Delta post-COVID model with flight count data

```
dal_post_model <- lm(Open.Stock.Price ~ Sum.of.carrier_ct
                + Sum.of.weather_ct + Sum.of.nas_ct + Sum.of.security_ct
                + Sum.of.late_aircraft_ct + Sum.of.arr_cancelled
                  , data = dal_post_data)

par(mfrow = c(2,2))
plot(dal_post_model)
```



```
par(mfrow = c(1,1))
```

```
summary(dal_post_model)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
##     Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
##     Sum.of.arr_cancelled, data = dal_post_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4481  -3.4044  -0.0765   2.7593  12.1562
##
## Coefficients:
```

```
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              25.652207   4.396232   5.835 8.6e-06 ***
## Sum.of.carrier_ct         0.006485   0.005978   1.085   0.2903
## Sum.of.weather_ct        -0.014110   0.018704  -0.754   0.4590
## Sum.of.nas_ct             0.007054   0.003361   2.099   0.0481 *
## Sum.of.security_ct        0.126935   0.145386   0.873   0.3925
## Sum.of.late_aircraft_ct  -0.009865   0.006440  -1.532   0.1405
## Sum.of.arr_cancelled     -0.000451   0.003054  -0.148   0.8840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.559 on 21 degrees of freedom
## Multiple R-squared:  0.3189, Adjusted R-squared:  0.1244
## F-statistic: 1.639 on 6 and 21 DF,  p-value: 0.1859
```

Conclusion - poor R-Squared value

Southwest pre-COVID model with flight count data

```
sw_pre_model <- lm(Open.Stock.Price ~ Sum.of.carrier_ct
            + Sum.of.weather_ct + Sum.of.nas_ct + Sum.of.security_ct
            + Sum.of.late_aircraft_ct + Sum.of.arr_cancelled
                , data = sw_pre_data)

par(mfrow = c(2,2))
plot(sw_pre_model)
```


11
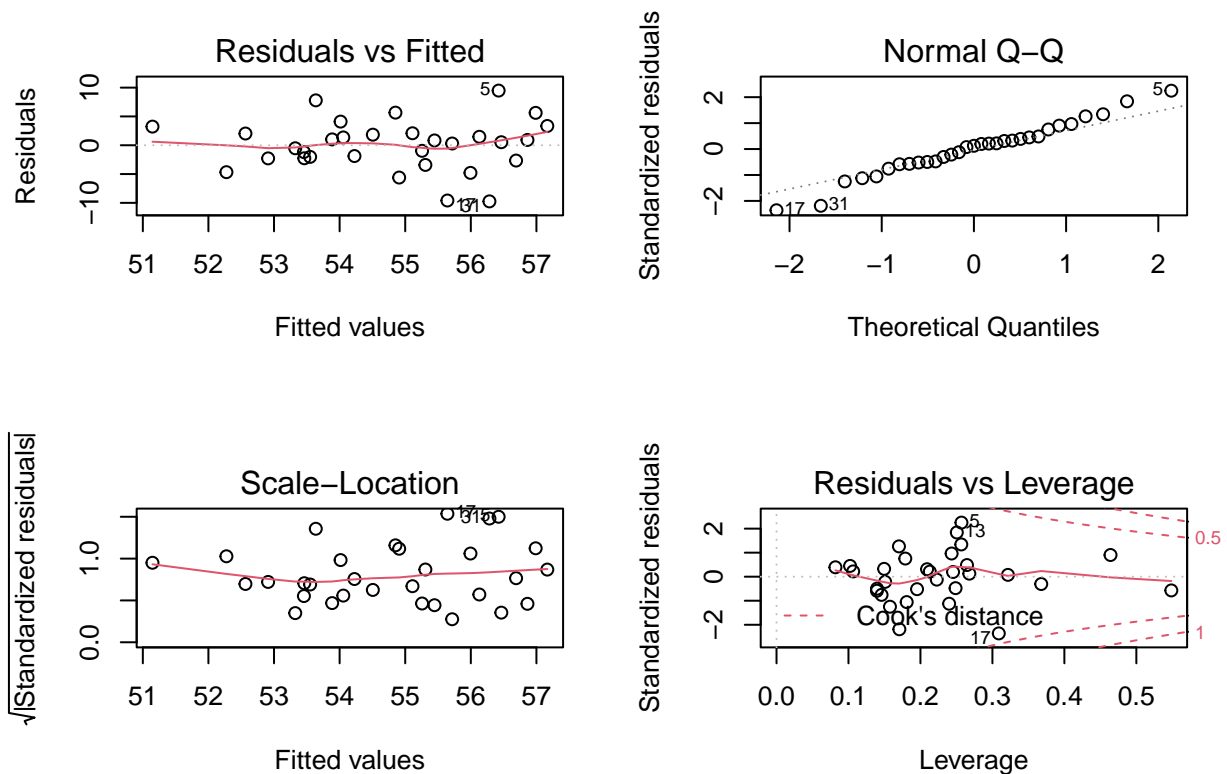
```
par(mfrow = c(1,1))
```
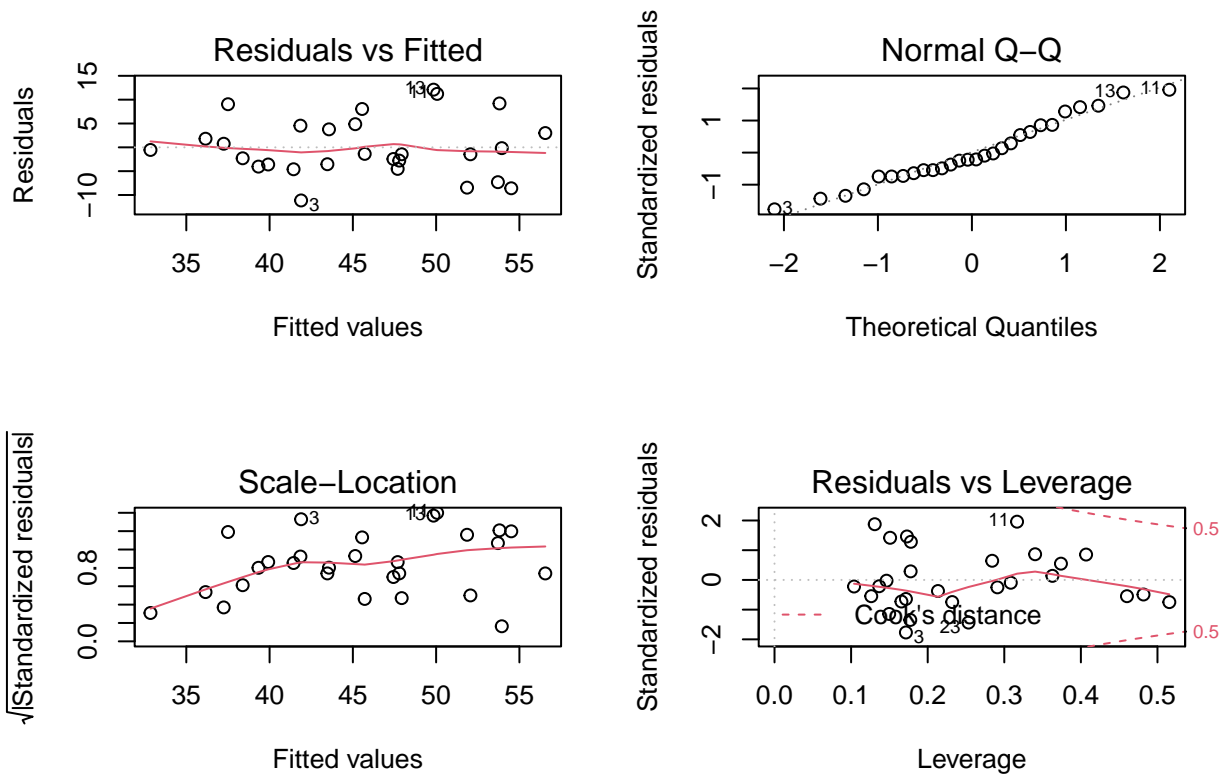
```
summary(sw_pre_model)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
##     Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
##     Sum.of.arr_cancelled, data = sw_pre_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7439 -2.2709  0.5248  2.0574  9.4861
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             57.1496561  5.6869250  10.049 4.46e-10 ***
## Sum.of.carrier_ct       -0.0030296  0.0022826  -1.327    0.197
## Sum.of.weather_ct       -0.0026779  0.0065958  -0.406    0.688
## Sum.of.nas_ct            0.0015373  0.0024260   0.634    0.532
## Sum.of.security_ct       0.0582456  0.0501551   1.161    0.257
## Sum.of.late_aircraft_ct  0.0007863  0.0013545   0.580    0.567
## Sum.of.arr_cancelled     0.0005434  0.0008653   0.628    0.536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.891 on 24 degrees of freedom
## Multiple R-squared:   0.11,  Adjusted R-squared:  -0.1125
## F-statistic: 0.4944 on 6 and 24 DF,  p-value: 0.8061
```

Conclusion - poor R-Squared value

Southwest post-COVID model with flight count data

```
sw_post_model <- lm(Open.Stock.Price ~ Sum.of.carrier_ct
              + Sum.of.weather_ct + Sum.of.nas_ct + Sum.of.security_ct
              + Sum.of.late_aircraft_ct + Sum.of.arr_cancelled
                , data = sw_post_data)
```

```
par(mfrow = c(2,2))
plot(sw_post_model)
```

```r
par(mfrow = c(1,1))
```
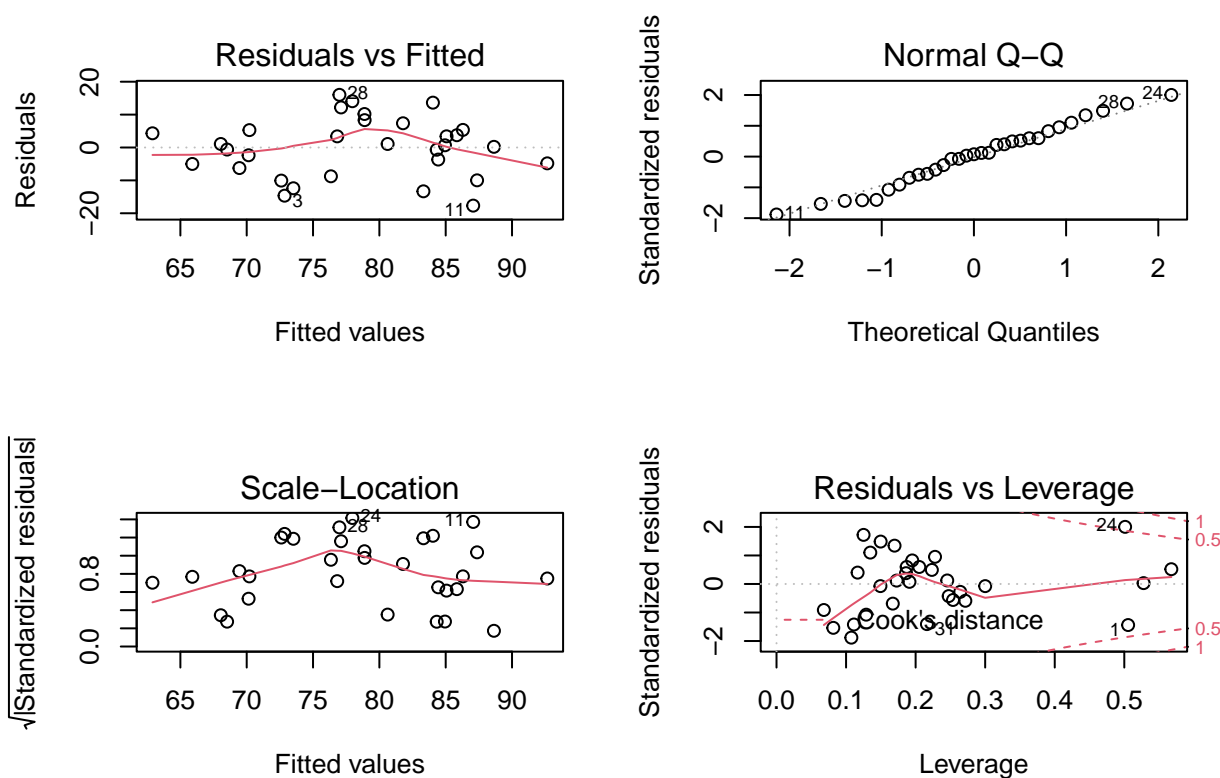
```r
summary(sw_post_model)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
##     Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
##     Sum.of.arr_cancelled, data = sw_post_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.143  -3.709  -1.412   3.966  12.096
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             27.1899012  4.5889400   5.925 7.01e-06 ***
## Sum.of.carrier_ct        0.0036812  0.0027160   1.355  0.18970
## Sum.of.weather_ct        0.0096749  0.0122297   0.791  0.43773
## Sum.of.nas_ct            0.0127968  0.0042448   3.015  0.00660 **
## Sum.of.security_ct       0.1406426  0.0835034   1.684  0.10694
## Sum.of.late_aircraft_ct -0.0072218  0.0021084  -3.425  0.00254 **
## Sum.of.arr_cancelled     0.0001229  0.0010111   0.122  0.90439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.922 on 21 degrees of freedom
## Multiple R-squared:  0.523,  Adjusted R-squared:  0.3868
## F-statistic: 3.838 on 6 and 21 DF,  p-value: 0.009684
```

13

Conclusion - poor R-Squared value

United pre-COVID model with flight count data

```
ual_pre_model <- lm(Open.Stock.Price ~ Sum.of.carrier_ct
                + Sum.of.weather_ct + Sum.of.nas_ct + Sum.of.security_ct
                + Sum.of.late_aircraft_ct + Sum.of.arr_cancelled
                    , data = ual_pre_data)

par(mfrow = c(2,2))
plot(ual_pre_model)
```



```
par(mfrow = c(1,1))
```

```
summary(ual_pre_model)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
##     Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
##     Sum.of.arr_cancelled, data = ual_pre_data)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -17.6730  -5.6269   0.6749   5.3186  16.0112
##
## Coefficients:
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             52.915002  11.143843   4.748 7.87e-05 ***
## Sum.of.carrier_ct       -0.009331   0.014289  -0.653   0.5200
## Sum.of.weather_ct       -0.011549   0.015072  -0.766   0.4510
## Sum.of.nas_ct            0.009123   0.004456   2.047   0.0517 .
## Sum.of.security_ct      -0.352202   0.649058  -0.543   0.5924
## Sum.of.late_aircraft_ct  0.004484   0.005600   0.801   0.4312
## Sum.of.arr_cancelled    -0.001871   0.004234  -0.442   0.6626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.945 on 24 degrees of freedom
## Multiple R-squared:  0.4213, Adjusted R-squared:  0.2766
## F-statistic: 2.912 on 6 and 24 DF,  p-value: 0.02807
```

Conclusion: poor R-Squared value

United post-COVID model with flight count data

```
ual_post_model <- lm(Open.Stock.Price ~ Sum.of.carrier_ct
                + Sum.of.weather_ct + Sum.of.nas_ct + Sum.of.security_ct
                + Sum.of.late_aircraft_ct + Sum.of.arr_cancelled
                    , data = ual_post_data)

par(mfrow = c(2,2))
plot(ual_post_model)
```
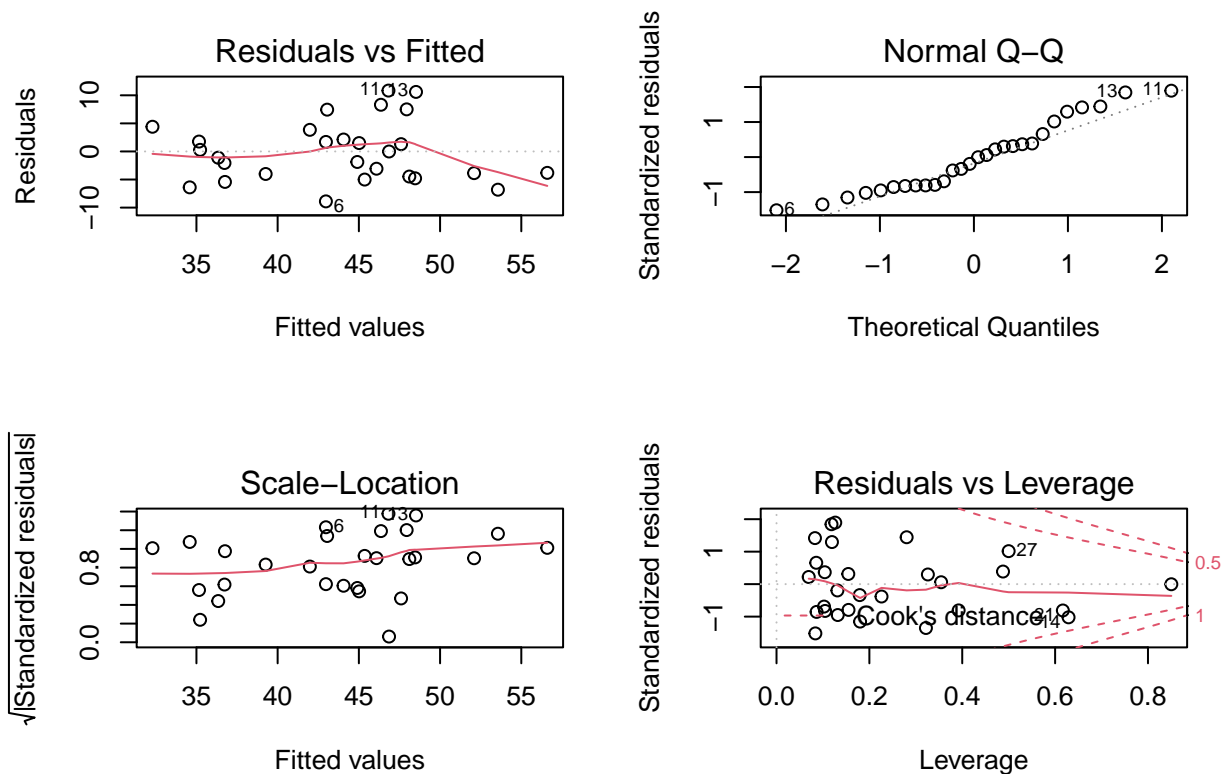
```
par(mfrow = c(1,1))
```

```
summary(ual_post_model)
```

```
##
## Call:
## lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
##      Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
##      Sum.of.arr_cancelled, data = ual_post_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8906 -4.1261 -0.5602  2.5539 10.8374
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             33.436533   2.901514  11.524 1.53e-10 ***
## Sum.of.carrier_ct        0.001813   0.004750   0.382 0.706572
## Sum.of.weather_ct       -0.012349   0.010969  -1.126 0.272969
## Sum.of.nas_ct            0.019903   0.004213   4.725 0.000115 ***
## Sum.of.security_ct      -0.466950   0.329385  -1.418 0.170963
## Sum.of.late_aircraft_ct -0.015424   0.004764  -3.237 0.003946 **
## Sum.of.arr_cancelled     0.000709   0.003174   0.223 0.825409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.121 on 21 degrees of freedom
## Multiple R-squared:  0.5607, Adjusted R-squared:  0.4351
## F-statistic: 4.466 on 6 and 21 DF,  p-value: 0.004594
```

Conclusion: poor R-Squared value too

Pre and Post COVID Data and Models

Inconclusive - none of the Pre and Post COVID Data models pass a R-squared value test.