Applying linear regression to study flights delay

Salma Elshahawy

10/12/2019

Contents

Part 1 - Introduction

I decided to use the dataset of nycflights13::flights: package included with R. This package contains information about all flights that departed from NYC (i.e., EWR, JFK and LGA) in 2013: 336,776 flights with 16 variables. To help understand what causes delays, it also includes a number of other useful datasets: weather, planes, airports, airlines. Source: Bureau of transportation statistics

H0(null hypotithes) -> No associations between departure delay and arrival delay

HA(alternative hypotithes) -> There are associations between departure delay and arrival delay.

- Research Questions:
  - Are the actual departure delay associated with the arrival delay?

Part 2 - Data

- Variables:
  - variable_1 -> dep_delay - independent variable, numerical - discrete
  - outcome -> arr_delay, numerical - numerical - discrete
- There are about 336,776 observation in the given dataset. Each observation represent flight full details.
- This is an observational study. I will draw my conclusions based on analyzing the existing data.

Part 3 - Exploratory data analysis

```
library(RCurl)
library(dplyr)
library(ggplot2)
library(ggExtra)

library(nycflights13)
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515         2      830
##  2  2013     1     1      533            529         4      850
##  3  2013     1     1      542            540         2      923
##  4  2013     1     1      544            545        -1     1004
##  5  2013     1     1      554            600        -6      812
##  6  2013     1     1      554            558        -4      740
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

```
summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median :2013   Median : 7.000   Median :16.00   Median :1401
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##                                                  NA's   :8255
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 106   Min.   : -43.00   Min.   :   1   Min.   :   1
##  1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124
##  Median :1359   Median :  -2.00   Median :1535   Median :1556
##  Mean   :1344   Mean   :  12.64   Mean   :1502   Mean   :1536
##  3rd Qu.:1729   3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945
##  Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##                 NA's   :8255      NA's   :8713
##    arr_delay          carrier              flight       tailnum
##  Min.   : -86.000   Length:336776      Min.   :   1   Length:336776
##  1st Qu.: -17.000   Class :character   1st Qu.: 553   Class :character
##  Median :  -5.000   Mode  :character   Median :1496   Mode  :character
##  Mean   :   6.895                      Mean   :1972
##  3rd Qu.:  14.000                      3rd Qu.:3465
##  Max.   :1272.000                      Max.   :8500
##  NA's   :9430
##     origin              dest              air_time        distance
##  Length:336776      Length:336776      Min.   : 20.0   Min.   :  17
##  Class :character   Class :character   1st Qu.: 82.0   1st Qu.: 502
##  Mode  :character   Mode  :character   Median :129.0   Median : 872
##                                        Mean   :150.7   Mean   :1040
##                                        3rd Qu.:192.0   3rd Qu.:1389
##                                        Max.   :695.0   Max.   :4983
##                                        NA's   :9430
##       hour           minute        time_hour
##  Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
```

```
##  1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :13.00   Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :13.18   Mean   :26.23   Mean   :2013-07-03 05:22:54
##  3rd Qu.:17.00   3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :23.00   Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

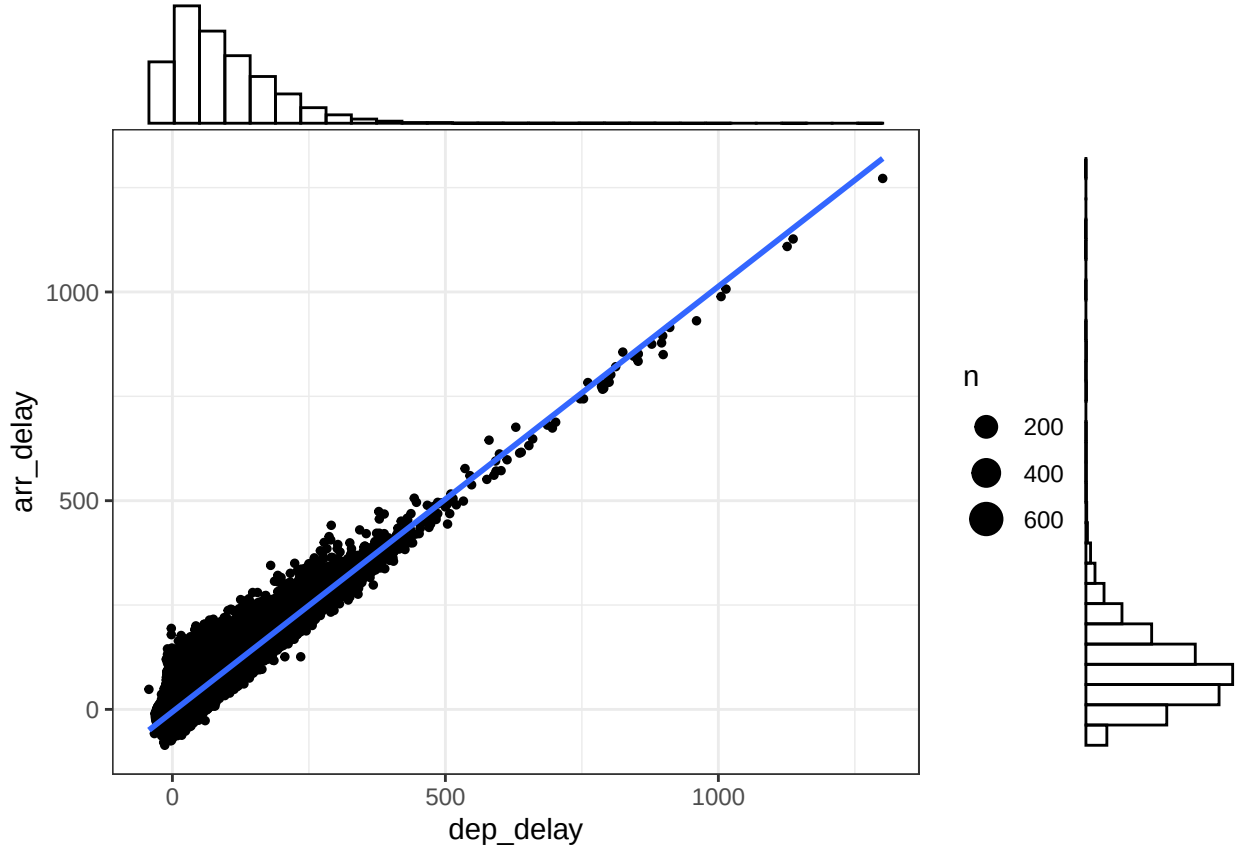
```
# taking a subset
sub_set <- flights[c(6,9,10,16)]
sub_set
```

```
## # A tibble: 336,776 x 4
##    dep_delay arr_delay carrier distance
##        <dbl>     <dbl> <chr>      <dbl>
##  1         2        11 UA          1400
##  2         4        20 UA          1416
##  3         2        33 AA          1089
##  4        -1       -18 B6          1576
##  5        -6       -25 DL           762
##  6        -4        12 UA           719
##  7        -5        19 B6          1065
##  8        -3       -14 EV           229
##  9        -3        -8 B6           944
## 10        -2         8 AA           733
## # ... with 336,766 more rows
```

```
summary(sub_set)
```

```
##    dep_delay          arr_delay          carrier             distance
##  Min.   : -43.00   Min.   : -86.000   Length:336776      Min.   :  17
##  1st Qu.:  -5.00   1st Qu.: -17.000   Class :character   1st Qu.: 502
##  Median :  -2.00   Median :  -5.000   Mode  :character   Median : 872
##  Mean   :  12.64   Mean   :   6.895                      Mean   :1040
##  3rd Qu.:  11.00   3rd Qu.:  14.000                      3rd Qu.:1389
##  Max.   :1301.00   Max.   :1272.000                      Max.   :4983
##  NA's   :8255      NA's   :9430
```

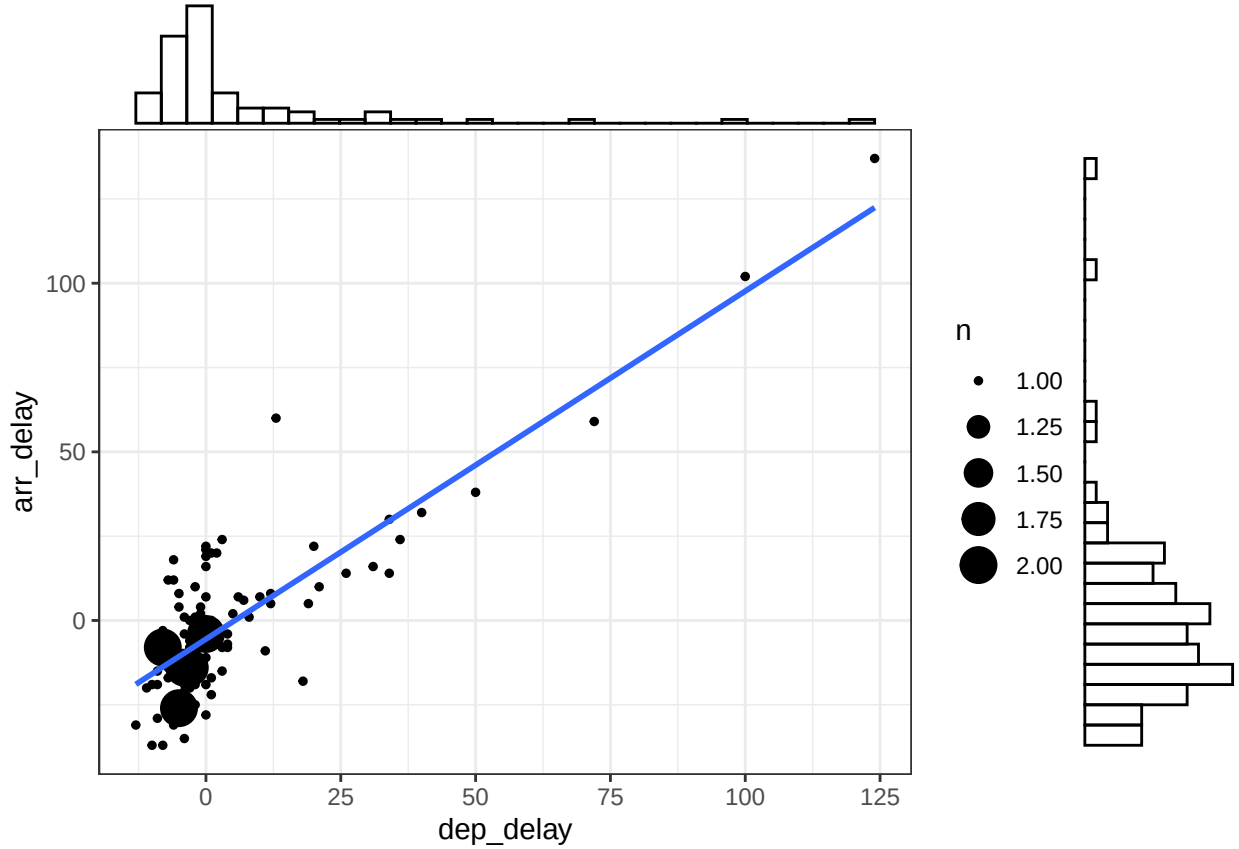
```
## get statistical analysis for the whole population
theme_set(theme_bw())  # pre-set the bw theme.
g <- ggplot(sub_set, aes(dep_delay, arr_delay)) +
  geom_count() +
  geom_smooth(method="lm", se=F)

ggMarginal(g, type = "histogram", fill="transparent")
```




```
## sampling, get summary plots --> statistics for only sample of 100

sample_100 <- sample_n(sub_set, 100)

summary(sample_100)
```

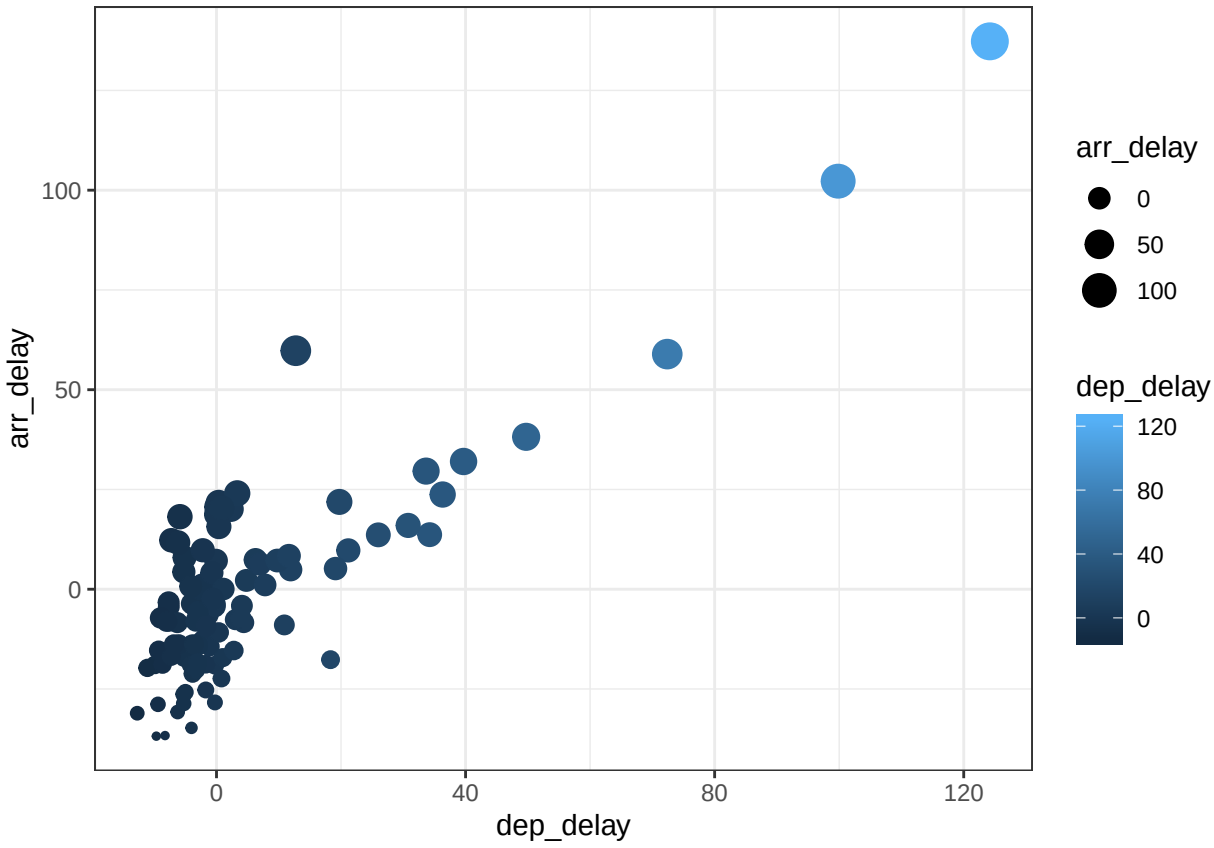
```
##    dep_delay          arr_delay          carrier             distance
##  Min.   :-13.000   Min.   :-37.0000   Length:100         Min.   : 187.0
##  1st Qu.: -5.000   1st Qu.:-17.0000   Class :character   1st Qu.: 531.2
##  Median : -1.000   Median : -6.0000   Mode  :character   Median : 948.0
##  Mean   :  4.814   Mean   : -0.5567                      Mean   :1038.6
##  3rd Qu.:  4.000   3rd Qu.:  8.0000                      3rd Qu.:1389.0
##  Max.   :124.000   Max.   :137.0000                      Max.   :2586.0
##  NA's   :3         NA's   :3
```

```
theme_set(theme_bw())  # pre-set the bw theme.
g <- ggplot(sample_100, aes(dep_delay, arr_delay)) +
  geom_count() +
  geom_smooth(method="lm", se=F)

ggMarginal(g, type = "histogram", fill="transparent")
```



Both dep_delay and arr_delap are right skewed distribution.

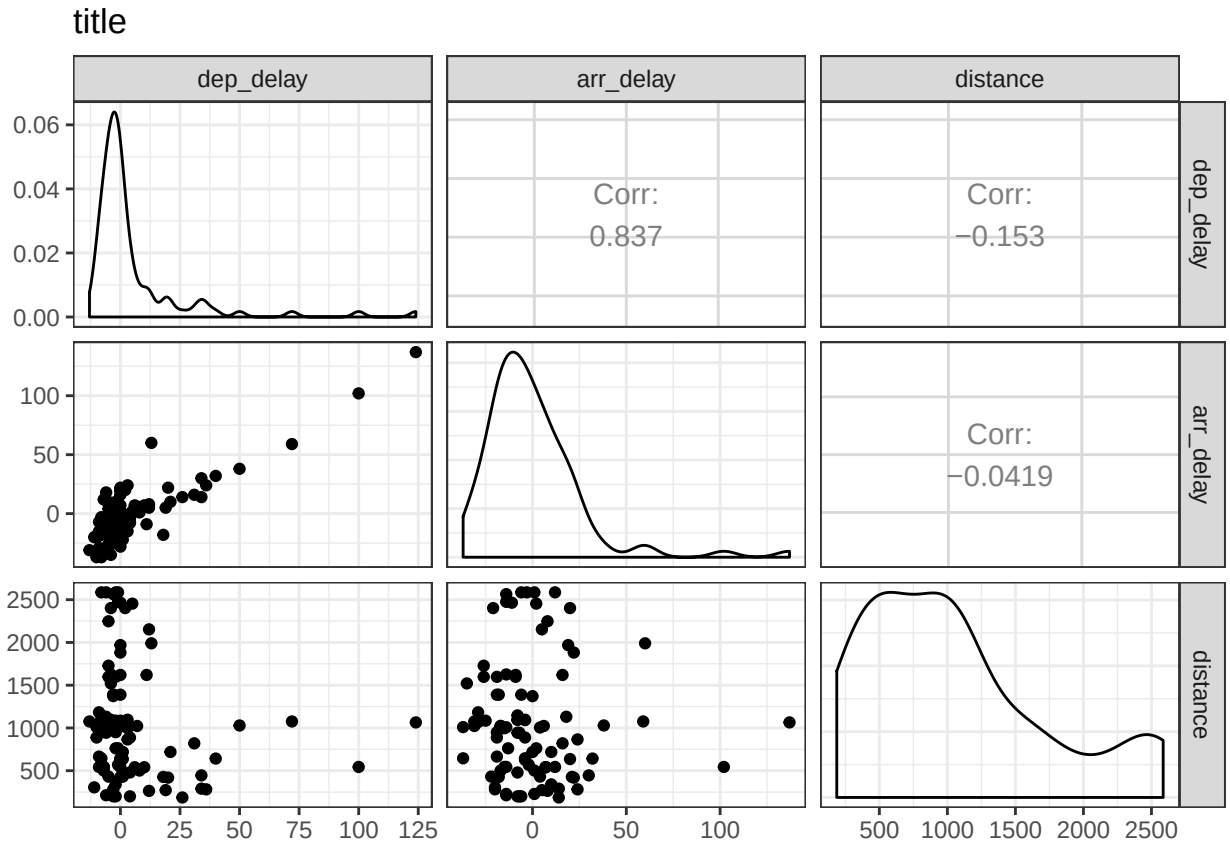
```
theme_set(theme_bw())  # pre-set the bw theme.
ggplot(sample_100, aes(dep_delay, arr_delay)) +
  geom_jitter(aes(colour = dep_delay, size = arr_delay), na.rm = TRUE)
```



Part 4 - Inference

This dataset doesn't follow the normal distribution. Since n = 100 => which is more than 25 we can do a linear regression model. Let's begin with the correlation which is a statistical tool to measure the level of linear dependence between two variables, that occur in pair

```
library(GGally)
sample_100 <- sample_100 %>%
  na.omit() %>%
  select(dep_delay, arr_delay, distance)

ggpairs(data = sample_100, title = "title")
```




```
# cor(sub_set$arr_delay, sub_set$dep_delay, use = "complete.obs")
```

The correlation between arr_delay and dep_delay is very strong as it close to 1 - strong correlation. However, relation doesn't mean causation. Now, let's build the linear regression model.

```
linearMod <- lm(arr_delay ~ dep_delay, data=sample_100)  # build linear regression model on full data
summary(linearMod)
```

```
##
## Call:
## lm(formula = arr_delay ~ dep_delay, data = sample_100)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.048  -9.349  -1.857   6.888  52.111
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.52420    1.47848  -3.736 0.000319 ***
## dep_delay    1.03179    0.06927  14.895  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.19 on 95 degrees of freedom
## Multiple R-squared:  0.7002, Adjusted R-squared:  0.697
## F-statistic: 221.8 on 1 and 95 DF,  p-value: < 2.2e-16
```

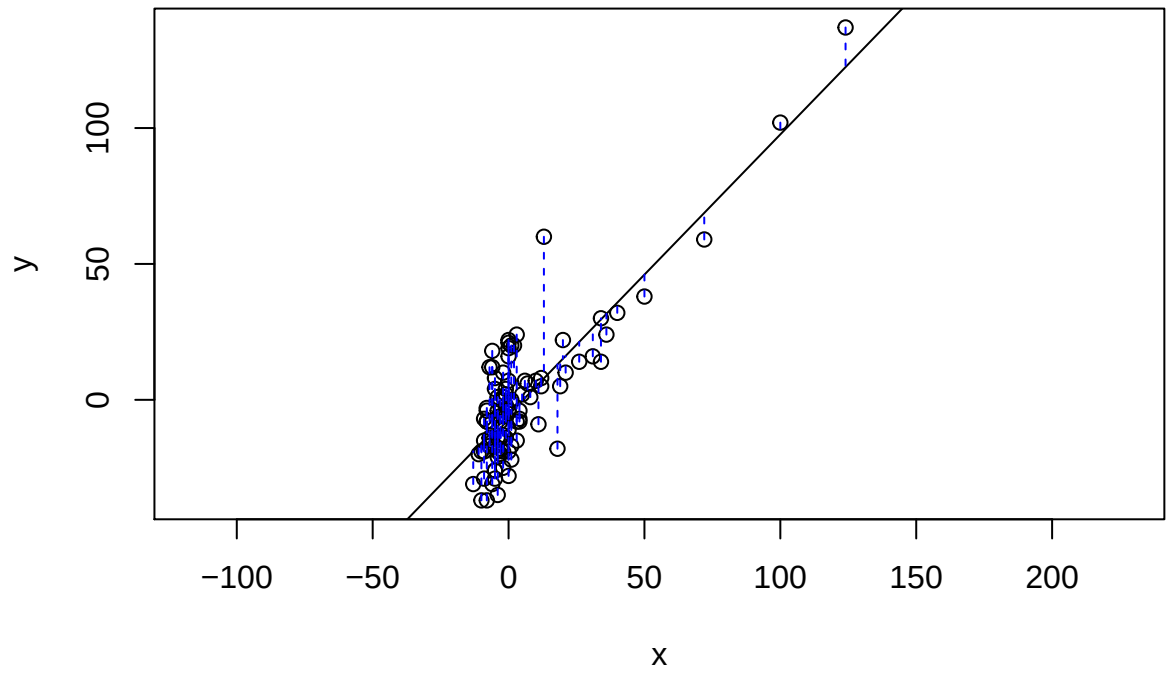
Check residuals distribution

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.

```
library(DATA606)
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

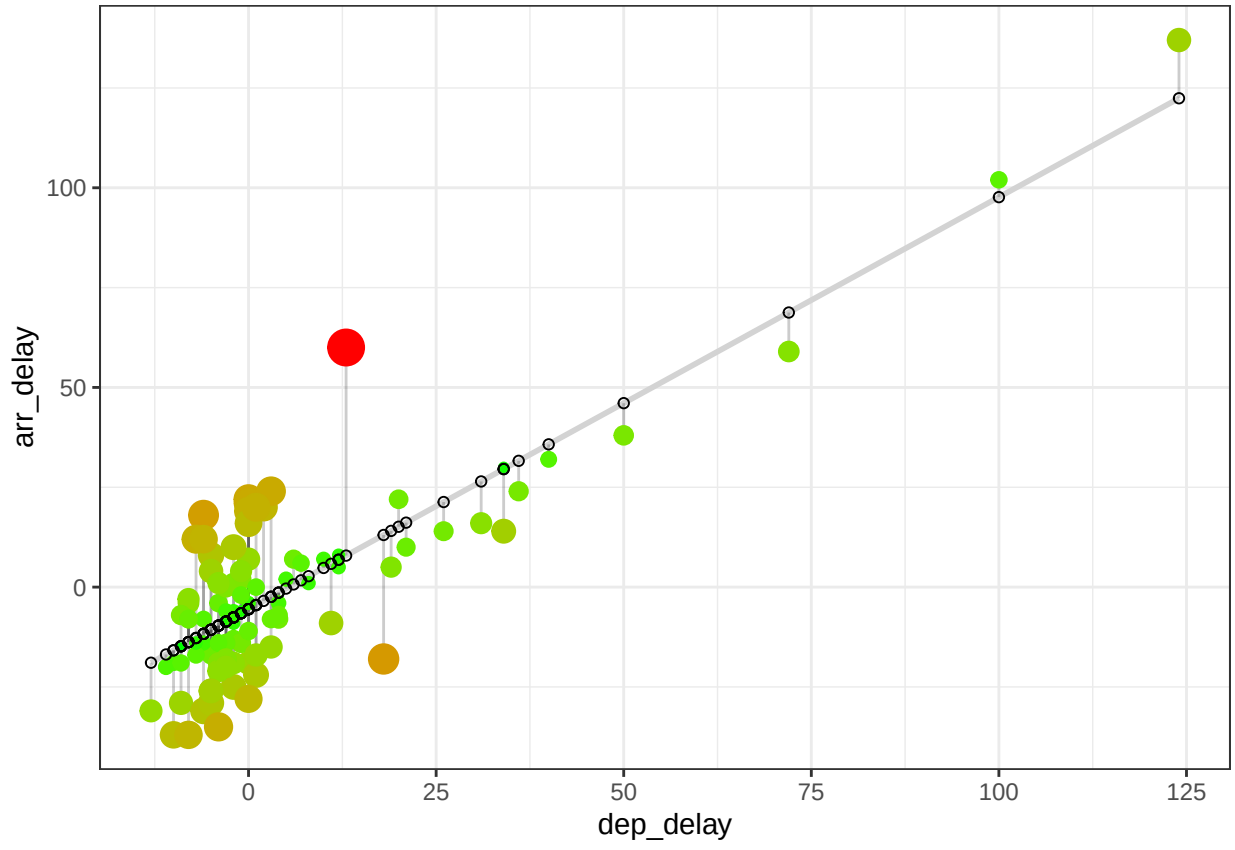
```
plot_ss(sample_100$dep_delay, sample_100$arr_delay)
```



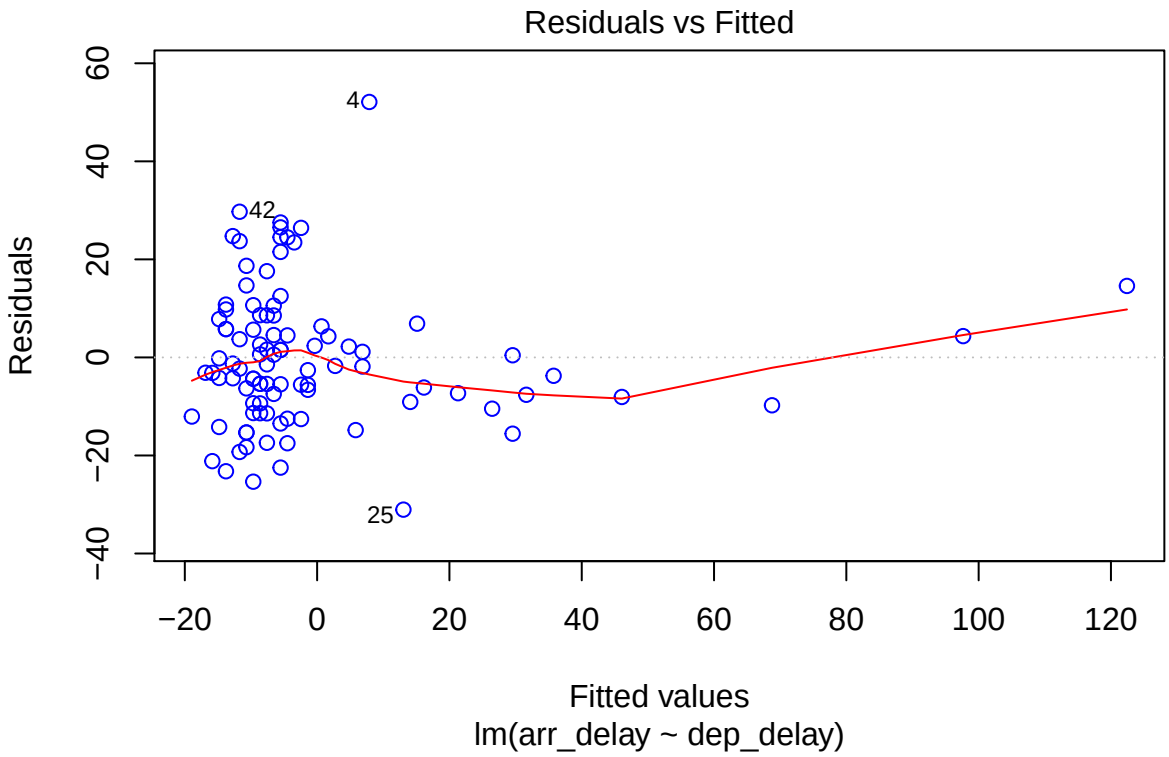

```
## Click two points to make a line.

## Call:
## lm(formula = y ~ x, data = pts)
##
```

```
## Coefficients:
## (Intercept)            x
##      -5.524        1.032
##
## Sum of Squares:  19118.25
```

```
fit <- linearMod
d <- sample_100
d$predicted <- predict(fit)   # Save the predicted values
d$residuals <- residuals(fit) # Save the residual values
ggplot(d, aes(x = dep_delay, y = arr_delay)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +     # regression line
  geom_segment(aes(xend = dep_delay, yend = predicted), alpha = .2) +      # draw line from point to li
  geom_point(aes(color = abs(residuals), size = abs(residuals))) +  # size of the points
  scale_color_continuous(low = "green", high = "red") +             # colour of the points mapped to re
  guides(color = FALSE, size = FALSE) +                             # Size legend removed
  geom_point(aes(y = predicted), shape = 1) +
  theme_bw()
```



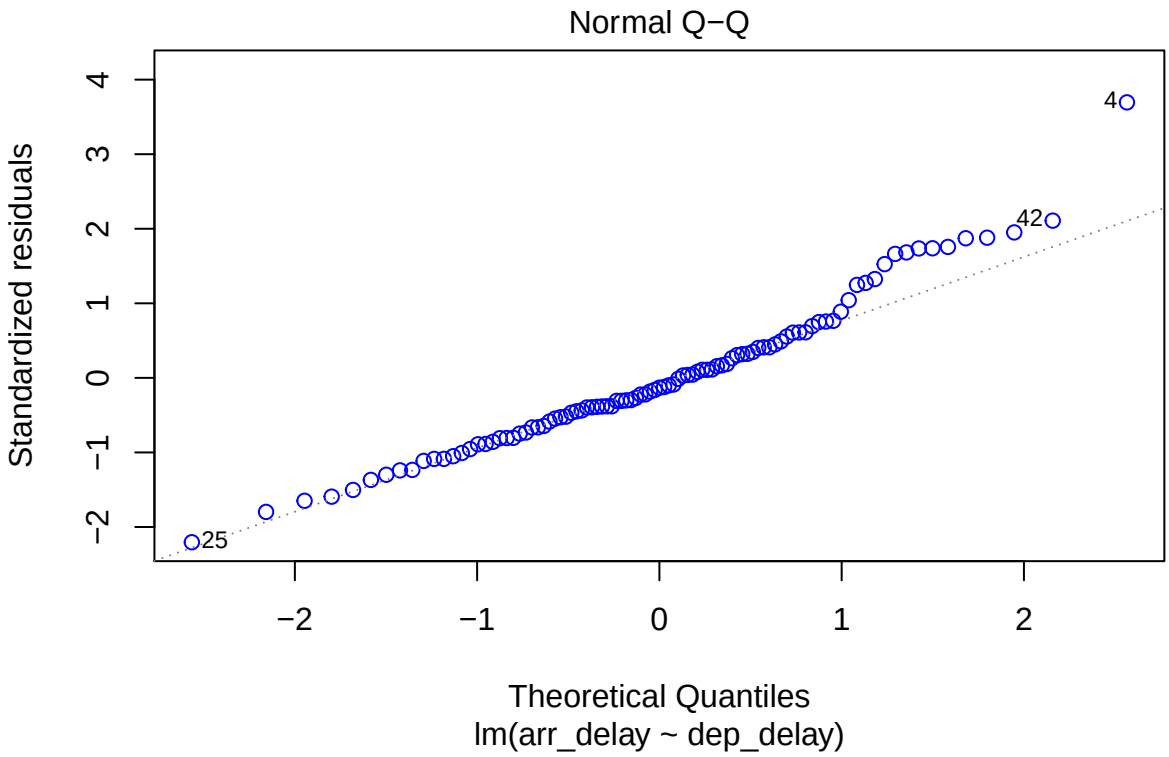

```
plot(fit, which=1, col=c("blue"))
```



When we look at the plot above, we see that the data does not have any obvious distinct pattern. While it is slightly curved, it has equally spread residuals around the horizontal line without a distinct pattern.

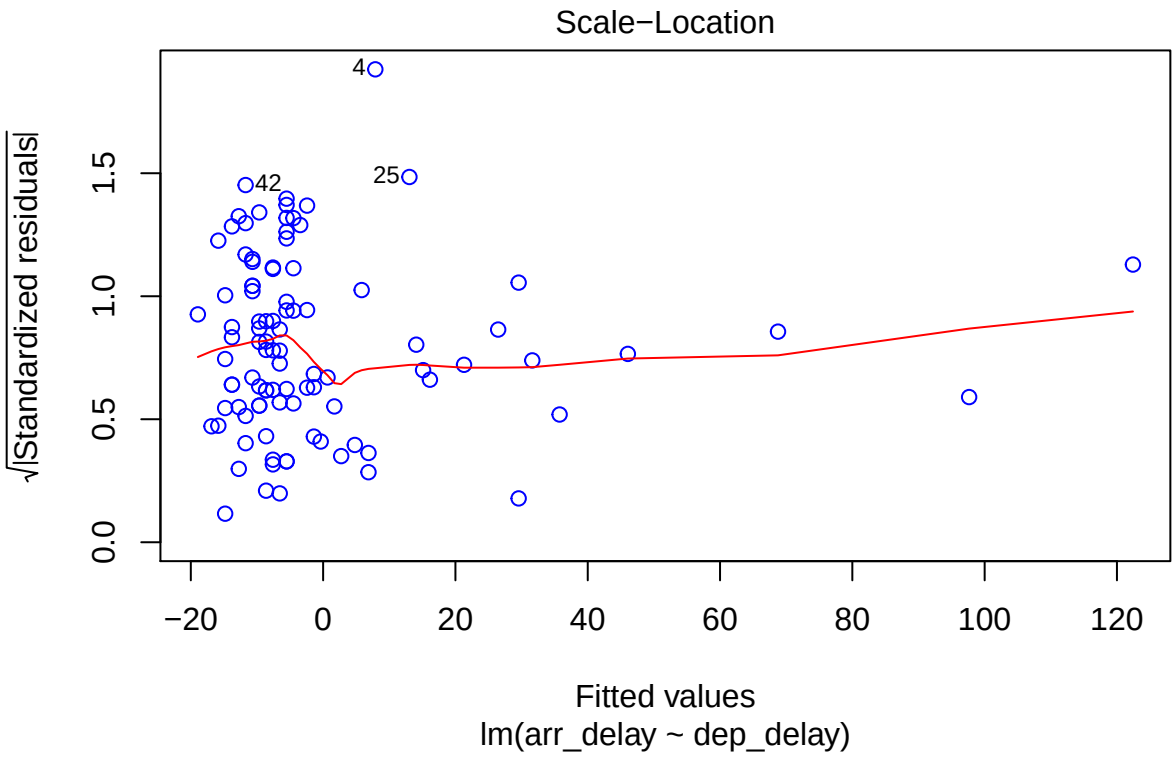
This is a good indication it is not a non-linear relationship.

```
plot(fit, which=2, col=c("blue")) # qqplot
```



For our model, the Q-Q plot shows pretty good alignment to the the line with a few points at the top slightly offset. Probably not significant and a reasonable alignment.

```
plot(fit, which=3, col=c("blue"))  # Scale-Location Plot
```



The residuals are reasonably well spread above and below a pretty horizontal line however the beginning of the line does have more points so has less variance there.

Residual analysis plots are a very useful tool for assessing aspects of veracity of a linear regression model on a particular dataset and testing that the attributes of a dataset meet the requirements for linear regression.

Now that we have built the linear model, we also have established the relationship between the predictor and response in the form of a mathematical formula for arrival delay (arr_delay) as a function for departure delay. For the above output, we can notice the 'Coefficients' part having two components: Intercept: -6.94, distance: 1.019 These are also called the beta coefficients. In other words,

arr_delay = Intercept + (beta ∗ dep_delay)

arr_delay = -5.899 + 1.02*dep_delay

Part 5 - Conclusion

as a conclusion, I would go with refusing the Null hypotethis that there is no associations between arrival delay and departure delay. However, We need to consider other attributes that has a confounding effects on the arrival times.

References

- Flights database

- University of Virginia Library