Homework 5

Q1.Import “FAA-1.xls” into R

```
library("rJava")
library("xlsxjars")
library("xlsx")
faa1dataset<-read.xlsx("FAA1.xls",1,header=TRUE,stringsAsFactors =
default.stringsAsFactors())
```

Q2. Do data cleaning using the attached information.

```
# summary describes the data, the result shows that there are missing valus in speed_air
variable.
#replacing all missing values with mean of that variable
summary(faa1dataset)
```

```
 aircraft     duration        no_pasg      speed_ground          speed_air             height
airbus:400   Min.   : 14.76   Min.   :29.00   Min.   : 27.74       Min.   : 90.00          Min.   :-3.546
boeing:400   1st Qu.:119.49   1st Qu.:55.00   1st Qu.: 65.87   1st Qu.: 96.16         1st Qu.:23.338
             Median :153.95   Median :60.00   Median : 79.64       Median :100.99     Median :30.147
             Mean   :154.01   Mean   :60.13   Mean   : 79.54        Mean   :103.83       Mean   :30.122
             3rd Qu.:188.91   3rd Qu.:65.00   3rd Qu.: 92.33        3rd Qu.:109.48       3rd Qu.:36.981
             Max.   :305.62   Max.   :87.00   Max.   :141.22          Max.   :141.72        Max.   :59.946
                                                                      NA's   :600
    pitch         distance
 Min.   :2.284   Min.   :  34.08
 1st Qu.:3.658   1st Qu.: 900.95
 Median :4.020   Median :1267.44
 Mean   :4.018   Mean   :1544.52
 3rd Qu.:4.388   3rd Qu.:1960.44
 Max.   :5.927   Max.   :6533.05
```

There are 600 missing values in speed_air
Omitting na values

```
faa1dataset<-na.omit(faa1dataset)
#Data Preprocessing

resldataset<-faa1dataset[(faa1dataset$duration)>40,]
resldataset<-resldataset[resldataset$speed_ground>30,]
resldataset<-resldataset[resldataset$speed_ground<140,]
resldataset<-resldataset[resldataset$distance<6000,]
resldataset<-resldataset[resldataset$speed_air>30,]
resldataset<-resldataset[resldataset$speed_air<140,]
finaldataset<-resldataset[,2:dim(resldataset)[2]];
```

Q3) Do data visualization using R

```
summary(faa1dataset)
   aircraft      duration         no_pasg       speed_ground      speed_air          height
 airbus:400   Min.   : 14.76   Min.   :29.00   Min.   : 27.74   Min.   : 90.00   Min.   :-3.546
 boeing:400   1st Qu.:119.49   1st Qu.:55.00   1st Qu.: 65.87   1st Qu.: 96.16   1st Qu.:23.338
              Median :153.95   Median :60.00   Median : 79.64   Median :100.99   Median :30.147
              Mean   :154.01   Mean   :60.13   Mean   : 79.54   Mean   :103.83   Mean   :30.122
              3rd Qu.:188.91   3rd Qu.:65.00   3rd Qu.: 92.33   3rd Qu.:109.48   3rd Qu.:36.981
              Max.   :305.62   Max.   :87.00   Max.   :141.22   Max.   :141.72   Max.   :59.946
                                                                NA's   :600
     pitch          distance
 Min.   :2.284   Min.   :  34.08
 1st Qu.:3.658   1st Qu.: 900.95
 Median :4.020   Median :1267.44
 Mean   :4.018   Mean   :1544.52
 3rd Qu.:4.388   3rd Qu.:1960.44
 Max.   :5.927   Max.   :6533.05
```

```
cor(finaldataset)
                duration      no_pasg speed_ground    speed_air       height       pitch    distance
duration      1.00000000 -0.0403495644 -0.051917974  0.0233018467  0.010810083 -0.04289099 -0.05317204
no_pasg      -0.04034956  1.0000000000 -0.008571678  0.0007538724  0.001719813 -0.00651743 -0.02613177
speed_ground -0.05191797 -0.0085716781  1.000000000  0.2398451039 -0.022110859 -0.05350143  0.86797421
speed_air     0.02330185  0.0007538724  0.239845104  1.0000000000 -0.039769370 -0.02650357  0.40546086
height        0.01081008  0.0017198133 -0.022110859 -0.0397693697  1.000000000  0.02233026  0.13487894
pitch        -0.04289099 -0.0065174295 -0.053501426 -0.0265035744  0.022330263  1.00000000  0.06616551
distance     -0.05317204 -0.0261317665  0.867974207  0.4054608635  0.134878936  0.06616551  1.00000000
```

from the above correlation matrix, distance and speed_ground and speed_air has high correlation

lets examine the assumption for a model of the relationship between distance and speed_ground and speed_air

distance is the dependent variable or outcome.

Q4) Do model fitting and model diagnostics using R (what variable would you keep in the model?)

Lets fit the linear regression model to predict the distance using speed_ground and speed_air

But speed_ground and speed_air are highly correlated

```
cor(finaldataset$speed_ground,finaldataset$speed_air,method="pearson")
0.9883475
```

```
model1<-
lm(finaldataset$distance~finaldataset$speed_ground+finaldataset$speed_air+finaldataset$no_pa
sg+finaldataset$height,data=finaldataset)
summary(model1)

#Check normally distributed
hist(model1$residuals)

model2<-
lm(finaldataset$distance~finaldataset$speed_ground+finaldataset$speed_air,data=finaldataset)
summary(model2)
#Check normally distributed
hist(model2$residuals)
```

model1 not normal.
Model2 residuals normally distributed

```
model2<-lm(finaldataset$distance~finaldataset$speed_ground+finaldataset$speed_air, data=finaldataset)
summary(model2)
#Check normally distributed
hist(model2$residuals)
```

```
summary(model2)

Call:
lm(formula = finaldataset$distance ~ finaldataset$speed_ground + 
    finaldataset$speed_air, data = finaldataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-986.67 -307.64  -61.05  269.17 1444.69 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -5604.6299   310.3478  -18.06   <2e-16 ***
finaldataset$speed_ground    39.2329     0.7977   49.18   <2e-16 ***
finaldataset$speed_air       38.7124     3.0751   12.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 411 on 788 degrees of freedom
Multiple R-squared:  0.7947,	Adjusted R-squared:  0.7942 
F-statistic:  1525 on 2 and 788 DF,  p-value: < 2.2e-16
```

Residuals are very useful for checking the model assumptions

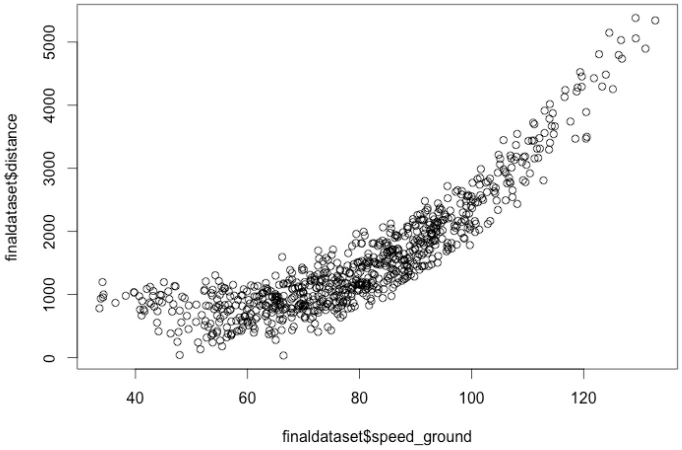
For fitting new regression model, the following assumptions will be considered
1.independet – check whether residuals are independent

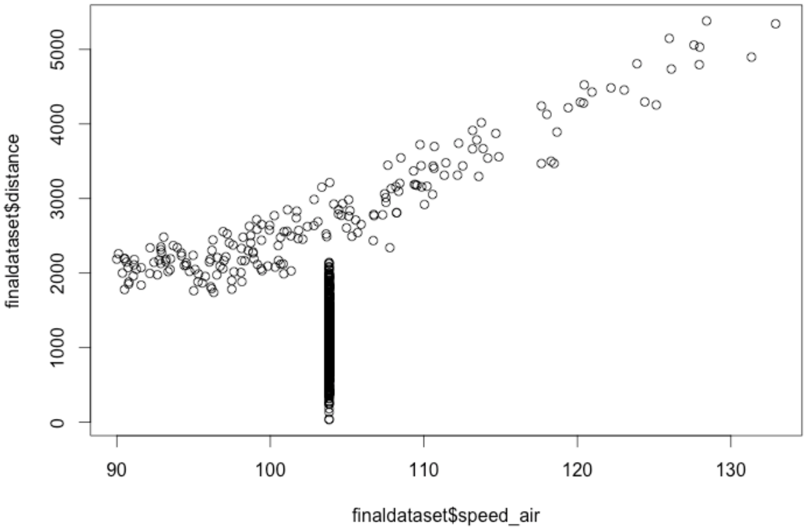
2.Normally distributed
3.Mean 0
4.Constant variance

To test the above assumptions, we use R diagnostic plots

1. Independent test

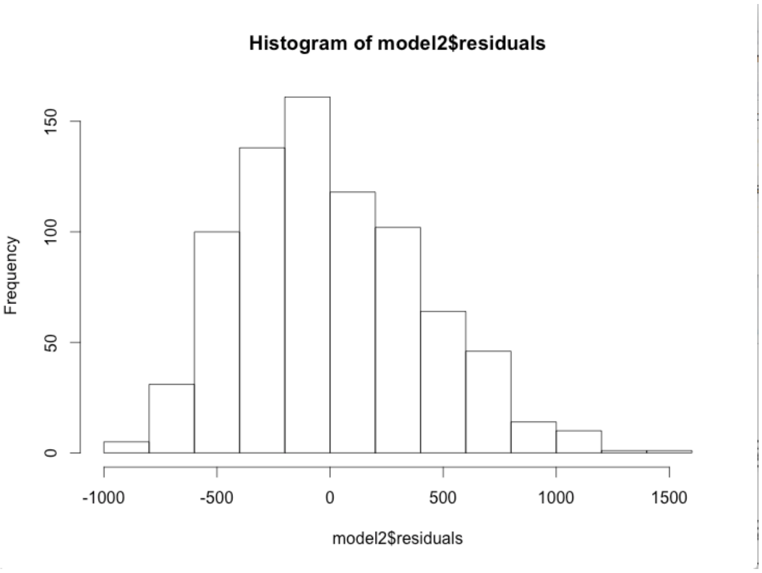
As the p-value is much less than 0.05, we reject the null hypothesis that $\beta = 0$. Hence there is a significant relationship between the variables in the linear regression model of the data set faithful





While speed_air value increasing, distance also increasing. Slope is positive

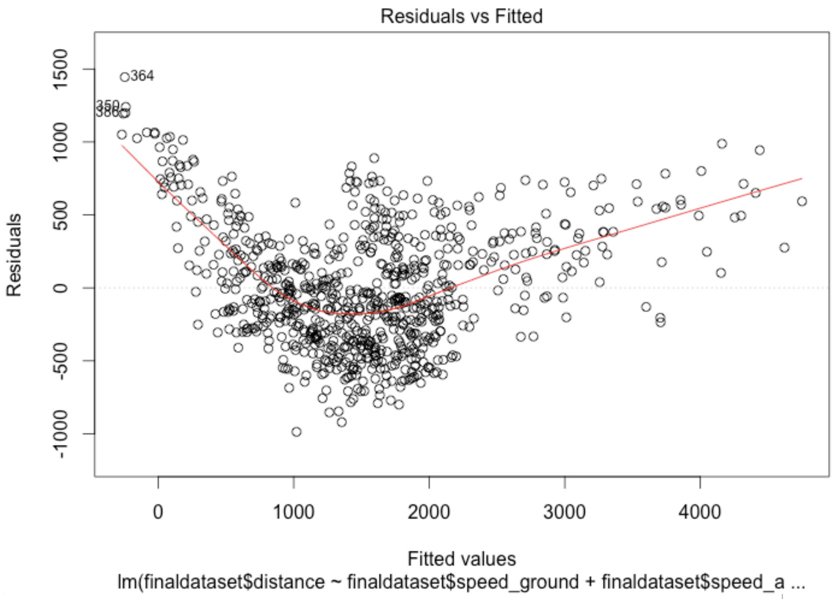
2.Normally distributed

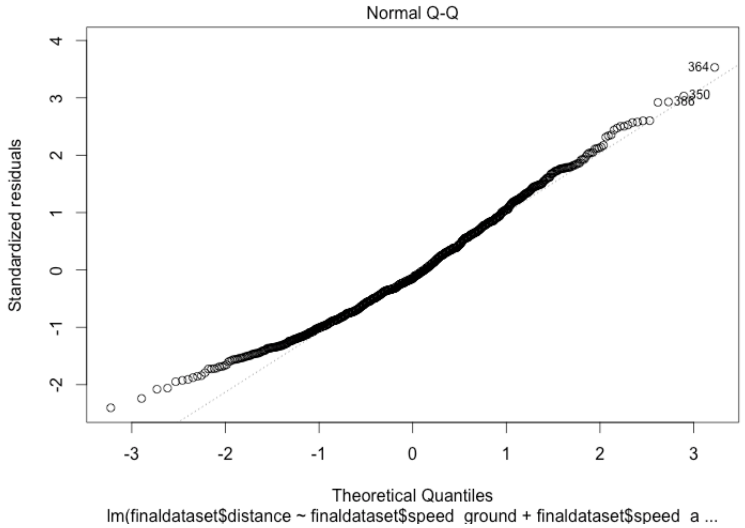


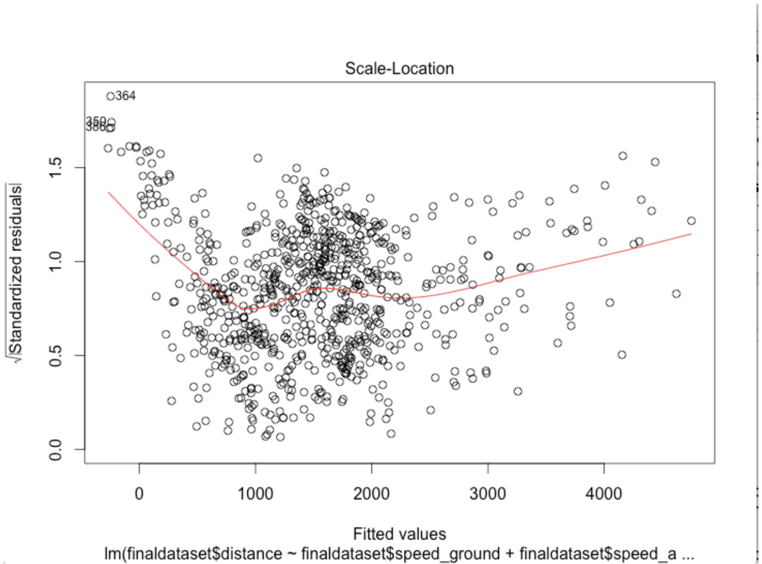
3. Mean =0 and constant variance

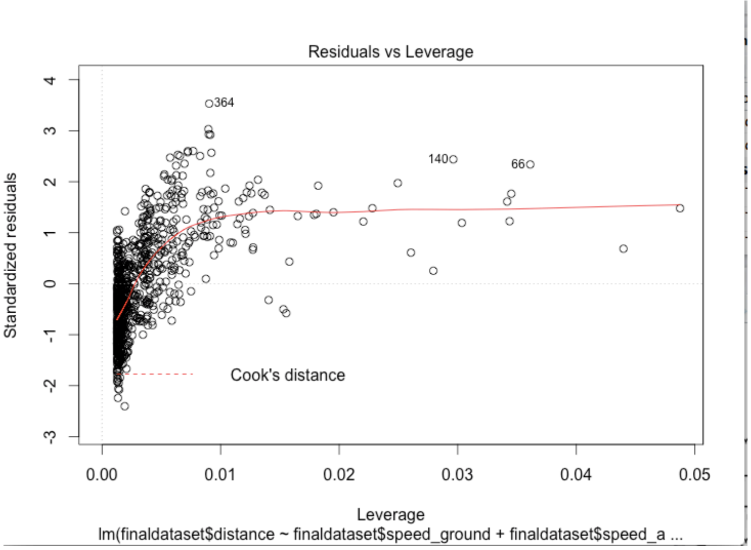
```
t.test(model$residuals,finaldataset$speed_ground)

                              Welch Two Sample t-test

data:  model$residuals and finaldataset$speed_ground
t = -5.8262, df = 793.84, p-value = 8.24e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -106.23931  -52.69204
sample estimates:
   mean of x     mean of y
-4.577228e-14  7.946567e+01

t.test(model$residuals,finaldataset$speed_air)

                              Welch Two Sample t-test

data:  model$residuals and finaldataset$speed_air
t = -7.6152, df = 790.26, p-value = 7.522e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -130.49297  -77.00588
sample estimates:
    mean of x     mean of y 
-4.577228e-14  1.037494e+02
```

4.Constant variance









5)

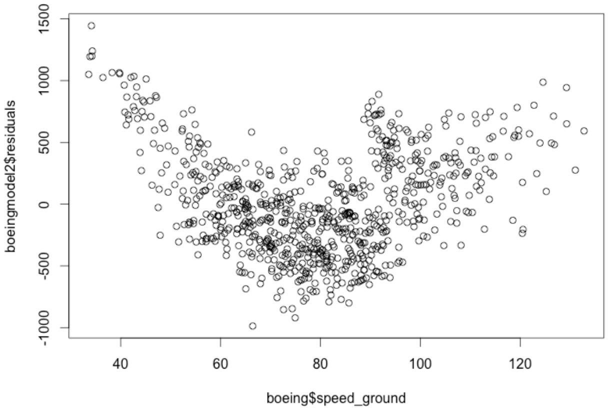
Model for aircraft make  boing

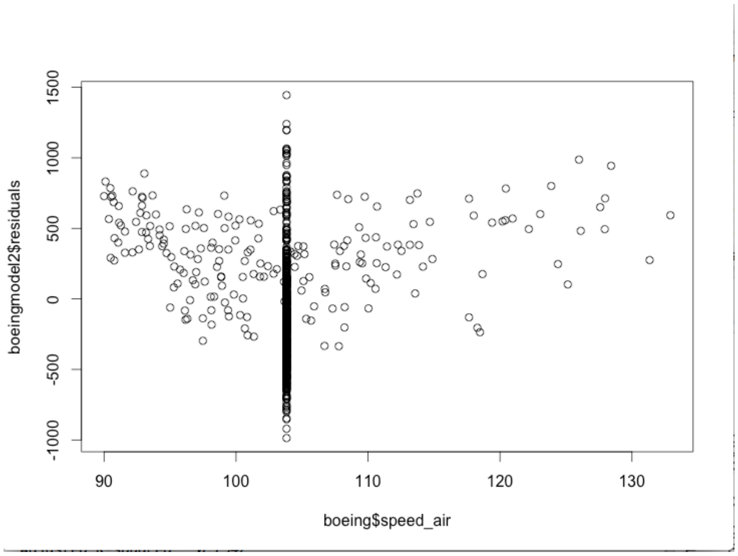
```
boeing<-resldataset[resldataset$aircraft=="boeing",]
boeing<-resldataset[,2:dim(resldataset)[2]];
boeingmodel2<-lm(boeing$distance~boeing$speed_ground+boeing$speed_air)
hist(boeingmodel2$residuals)
qqnorm(boeingmodel2$residuals)
qqline(boeingmodel2$residuals, col = "red")
plot(boeingmodel2)
```

```
summary(boeing)
   duration      no_pasg    speed_ground   speed_air      height       pitch       distance
 Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA
 1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA
 Median : NA   Median : NA   Median : NA   Median : NA   Median : NA   Median : NA   Median : NA
 Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN
 3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA
 Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA
```

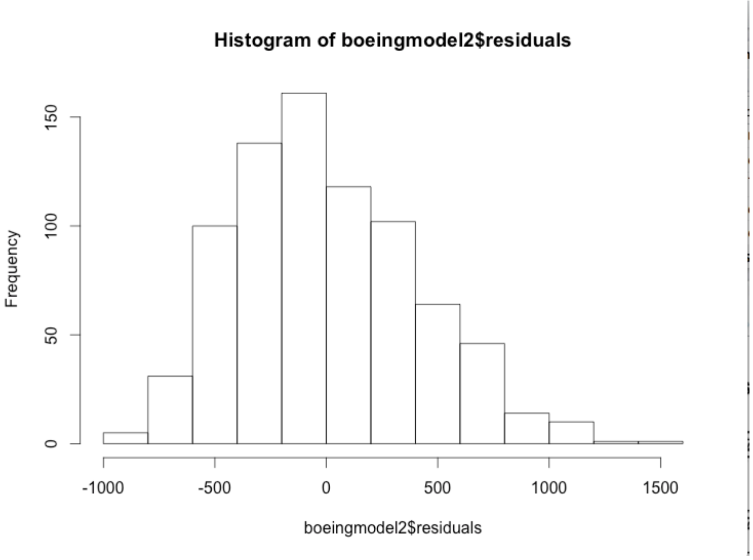
```
summary(boeingmodel2)

Call:
lm(formula = boeing$distance ~ boeing$speed_ground + boeing$speed_air)

Residuals:
   Min      1Q  Median      3Q     Max 
-986.67 -307.64  -61.05  269.17 1444.69 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -5604.6299   310.3478  -18.06   <2e-16 ***
boeing$speed_ground    39.2329     0.7977   49.18   <2e-16 ***
boeing$speed_air       38.7124     3.0751   12.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 411 on 788 degrees of freedom
Multiple R-squared:  0.7947,	Adjusted R-squared:  0.7942 
F-statistic:  1525 on 2 and 788 DF,  p-value: < 2.2e-16
```

For fitting new regression model, the following assumptions will be considered
1.independet – check whether residuals are independent

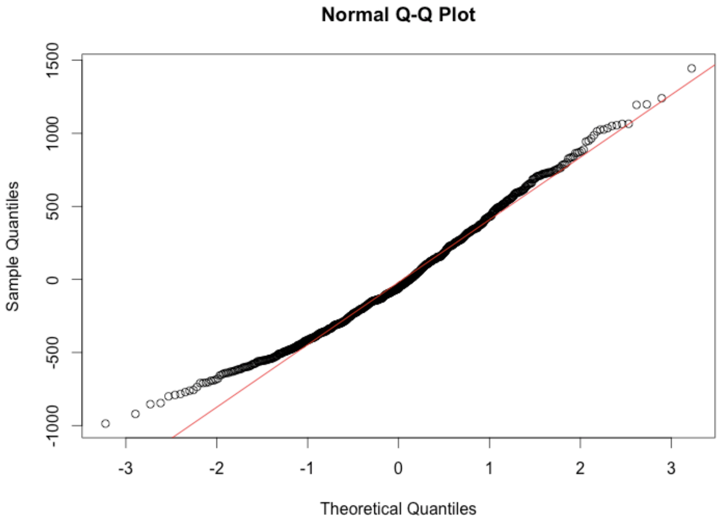


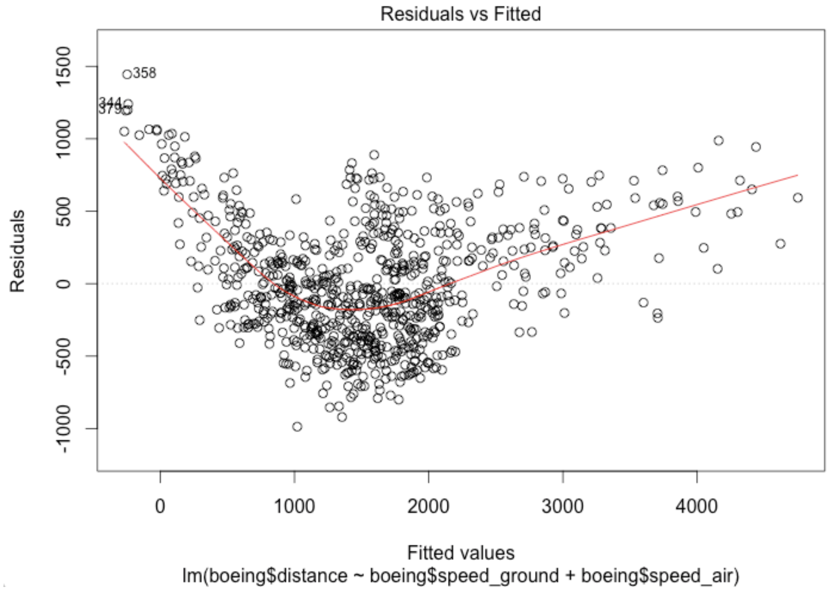


2.Normally distributed

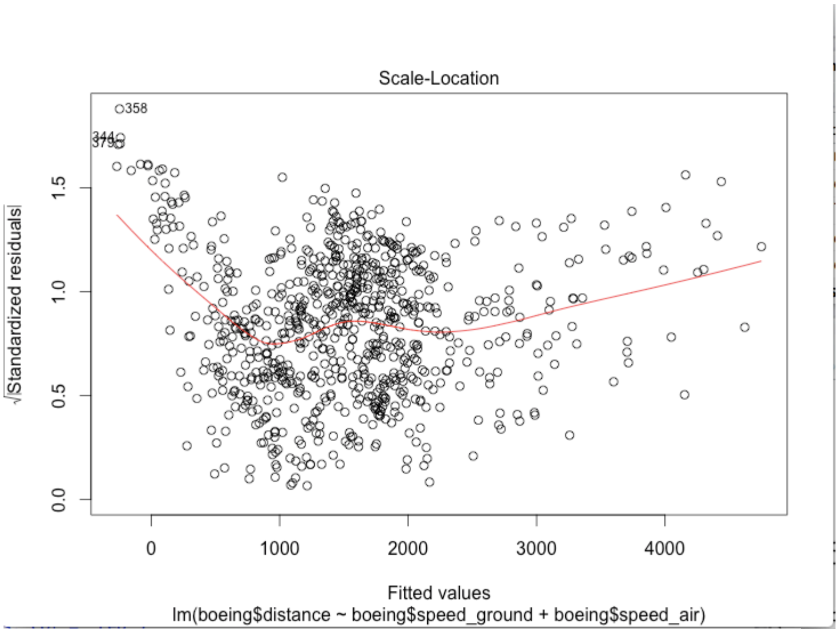


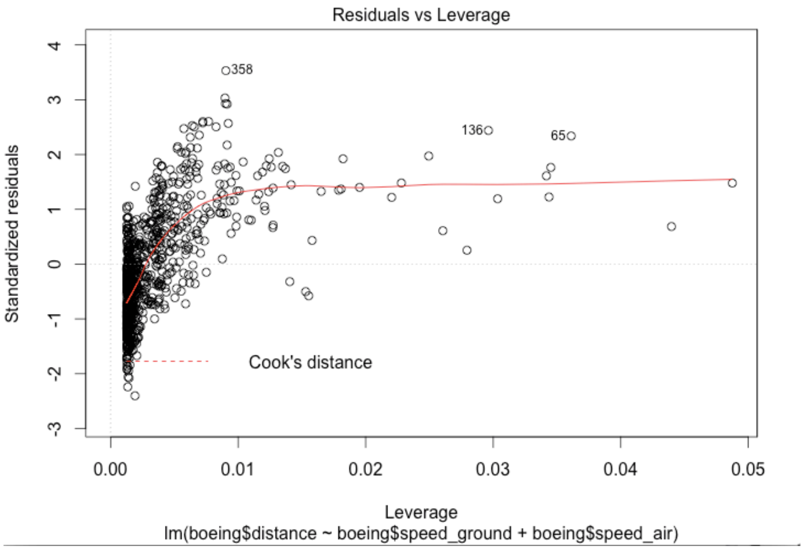
3.Mean 0





4.Constant variance



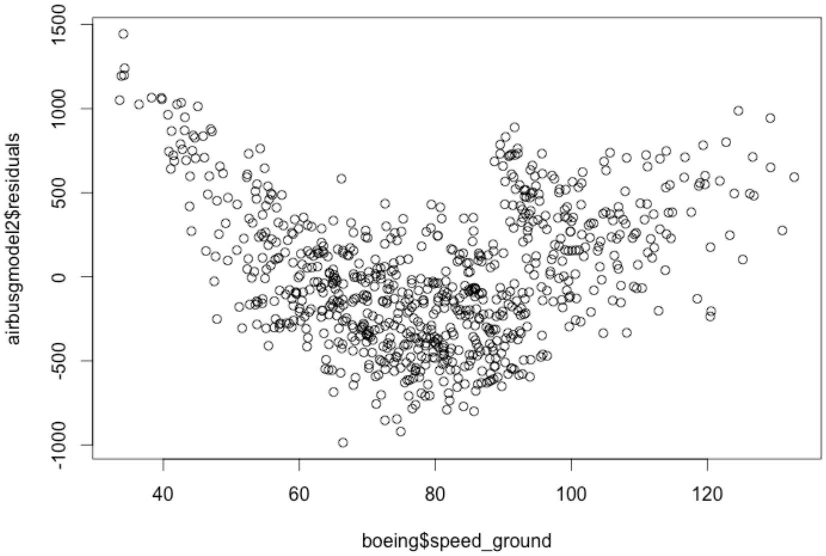


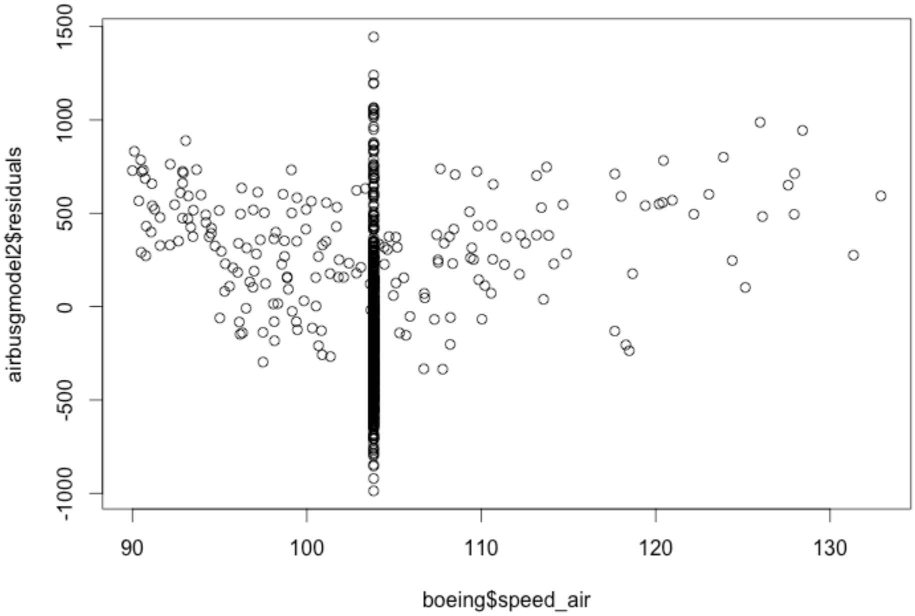
Airbus

```
airbus<-resldataset[resldataset$aircraft=="airbus",]
airbus<-resldataset[,2:dim(resldataset)[2]];
```

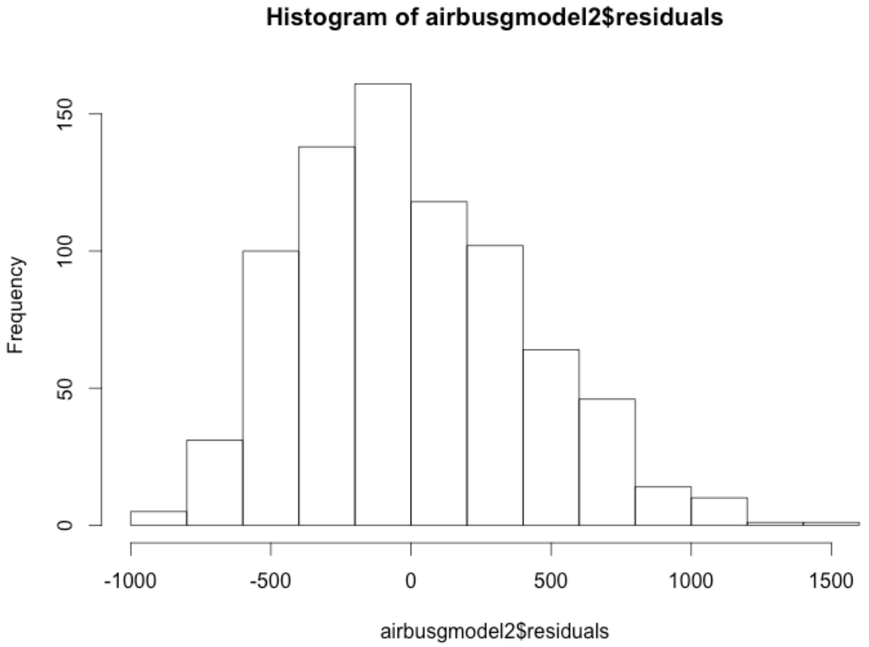
```
duration        no_pasg      speed_ground      speed_air       height          pitch
 Min.   : 41.95   Min.   :29.00   Min.   : 33.57   Min.   : 90.0   Min.   :-3.546   Min.   :2.284
 1st Qu.:119.68   1st Qu.:55.00   1st Qu.: 65.91   1st Qu.:103.8   1st Qu.:23.145   1st Qu.:3.654
 Median :154.24   Median :60.00   Median : 79.63   Median :103.8   Median :30.140   Median :4.017
 Mean   :154.79   Mean   :60.17   Mean   : 79.47   Mean   :103.7   Mean   :30.070   Mean   :4.016
 3rd Qu.:189.25   3rd Qu.:65.00   3rd Qu.: 92.13   3rd Qu.:103.8   3rd Qu.:36.896   3rd Qu.:4.388  
 Max.   :305.62   Max.   :87.00   Max.   :132.78   Max.   :132.9   Max.   :59.946   Max.   :5.927  
    distance      
 Min.   :  34.08  
 1st Qu.: 898.87  
 Median :1264.93  
 Mean   :1529.42  
 3rd Qu.:1949.22  
 Max.   :5381.96  

airbusgmodel2<-lm(airbus$distance~airbus$speed_ground+airbus$speed_air)
> summary(airbusgmodel2)

Call:
lm(formula = airbus$distance ~ airbus$speed_ground + airbus$speed_air)

Residuals:
    Min      1Q  Median      3Q     Max 
-986.67 -307.64  -61.05  269.17 1444.69 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -5604.6299   310.3478  -18.06   <2e-16 ***
airbus$speed_ground    39.2329     0.7977   49.18   <2e-16 ***
airbus$speed_air       38.7124     3.0751   12.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 411 on 788 degrees of freedom
Multiple R-squared:  0.7947,	Adjusted R-squared:  0.7942 
F-statistic:  1525 on 2 and 788 DF,  p-value: < 2.2e-16
```

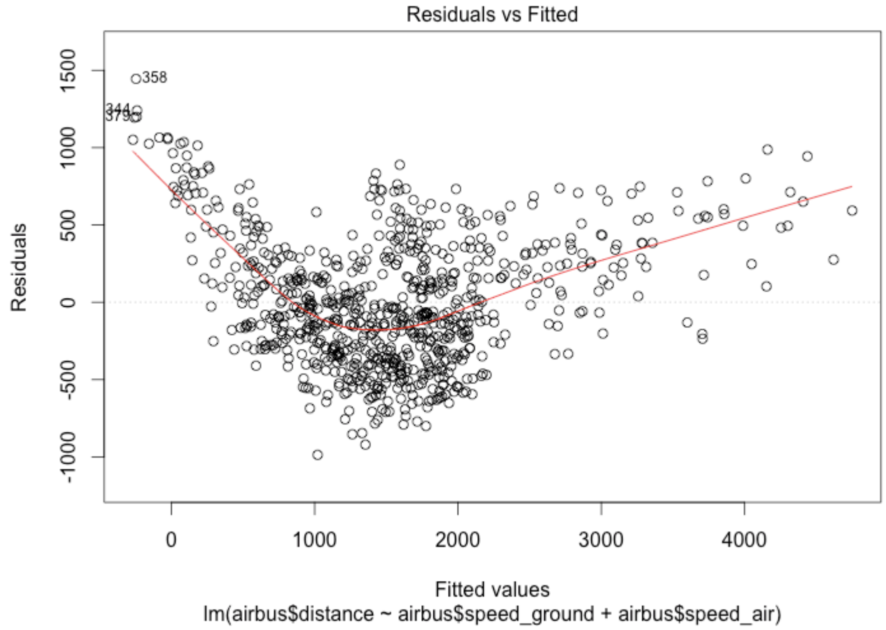
1.indepence

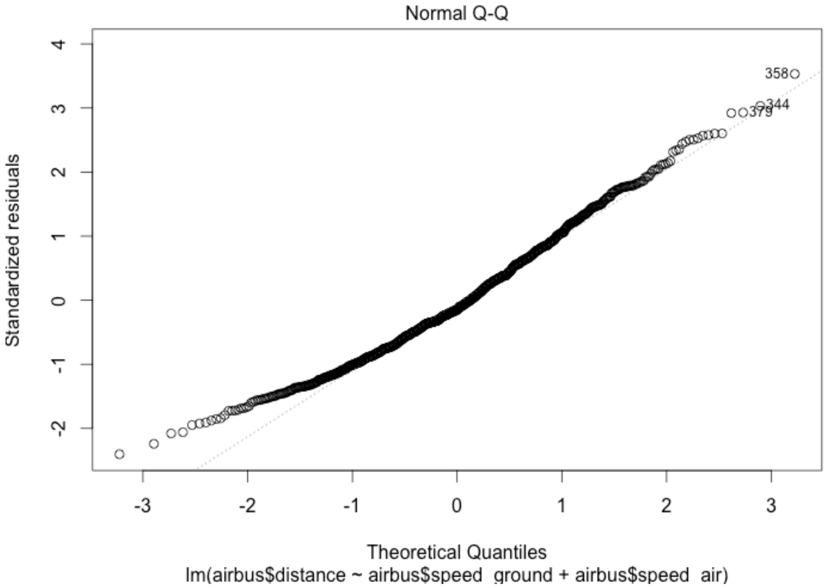

airbusgmodel2$residuals
-1000
-500
0
500
1000
1500
40
60
80
100
120
boeing$speed_ground


airbusgmodel2$residuals
-1000
-500
0
500
1000
1500
90
100
110
120
130
boeing$speed_air

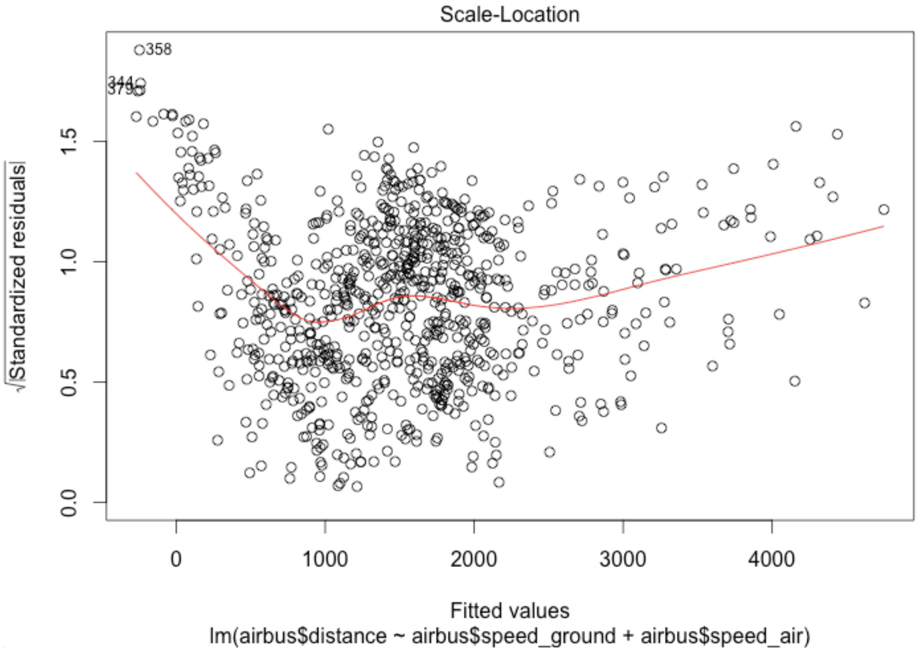
2.Normality

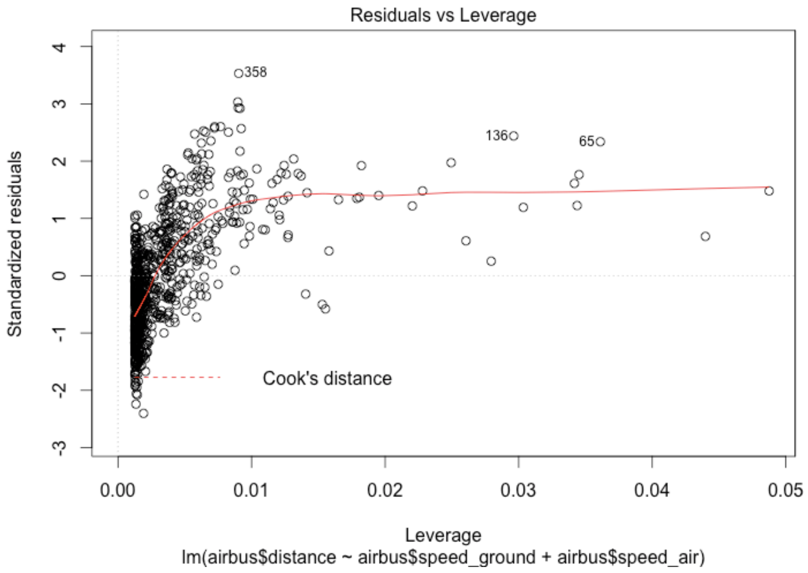

Histogram of airbusgmodel2$residuals
Frequency
0
50
100
150
-1000
-500
0
500
1000
1500
airbusgmodel2$residuals


Residuals vs Fitted
Residuals
-1000
-500
0
500
1000
1500
358
344
379
0
1000
2000
3000
4000
Fitted values
lm(airbus$distance ~ airbus$speed_ground + airbus$speed_air)



4. Constant variance





```
t.test(airbusgmodel2$residuals,finaldataset$speed_ground)

        Welch Two Sample t-test

data:  airbusgmodel2$residuals and finaldataset$speed_ground
t = -5.439, df = 793.34, p-value = 7.142e-08
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 -108.14536  -50.78598
sample estimates:
   mean of x    mean of y
1.102370e-13 7.946567e+01

t.test(airbusgmodel2$residuals,finaldataset$speed_air)

                Welch Two Sample t-test

data:  airbusgmodel2$residuals and finaldataset$speed_air
t = -7.1081, df = 790.22, p-value = 2.634e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -132.40103  -75.09782
sample estimates:
   mean of x    mean of y
1.102370e-13 1.037494e+02
```