Stats

```r
library(tidyr)
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v dplyr   1.0.10
v tibble  3.1.8      v stringr 1.4.1
v readr   2.1.3      v forcats 0.5.2
v purrr   0.3.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(ggplot2)
library(GGally)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```r
ozone = read_csv("ozone.csv")
```

```
Rows: 111 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (4): radiation, temperature, wind, ozone

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Question 5

You can add options to executable code like this

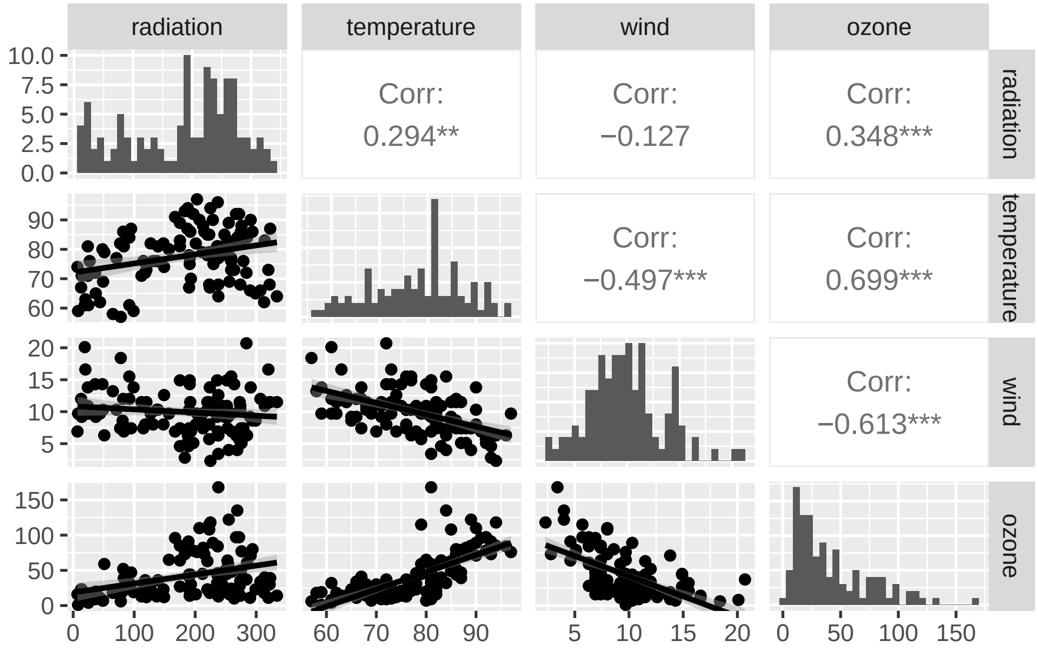
```r
summary(ozone)
```

```
   radiation      temperature         wind             ozone      
 Min.   :  7.0   Min.   :57.00   Min.   : 2.300   Min.   :  1.0  
 1st Qu.:113.5   1st Qu.:71.00   1st Qu.: 7.400   1st Qu.: 18.0  
 Median :207.0   Median :79.00   Median : 9.700   Median : 31.0  
 Mean   :184.8   Mean   :77.79   Mean   : 9.939   Mean   : 42.1  
 3rd Qu.:255.5   3rd Qu.:84.50   3rd Qu.:11.500   3rd Qu.: 62.0  
 Max.   :334.0   Max.   :97.00   Max.   :20.700   Max.   :168.0  
```

Here we can see that wind and ozone have some pretty extremely high max values compared to both the median

```r
ggpairs(ozone, lower = list(continuous = "smooth"), diag = list(continuous = "barDiag"),
    axisLabels = "show")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Firstly, as it can be observed in the graph that ozone has a very significant positive skewness and is possibly normally distributed. It also noticeable from the ozone histogram that it resembles a normal distribution with positive skewness.

We can also observe that ozone have a good correlation with temperature with only a small amount of variance overall with the exception of a few points between the 3rd quartile and the maximum, we can also see that there is a a positive slope meaning that as temperature increases the amount of ozone detected increases as well.

Furthermore, radiation has a correlation with a positive slope with ozone so radiation has a positive effect on ozone. The variance is more extreme between the 2nd quartile and maximum but maintaining a relatively low variance between the minimum and the 2nd quartile.

Lastly, Wind's correlation with ozone has a negative slope, meaning that has wind increases the less ozone is detected. Most of the variance below the line of best fit, is between the first and third quartile while the values that are more on the extreme, between the the minimum and 1st quartile and the 3rd quartile and the maximum are almost all above the line of best fit. The wind histogram also displays what looks to be a normal distribution with close to zero skewness with some irregularities near the 15 bin.

```
model = lm(ozone ~ radiation + temperature + wind, data = ozone)

summary(model)
```

```
Call:
lm(formula = ozone ~ radiation + temperature + wind, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.485 -14.210  -3.556  10.124  95.600 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -64.23208   23.04204  -2.788  0.00628 ** 
radiation     0.05980    0.02318   2.580  0.01124 *  
temperature   1.65121    0.25341   6.516 2.43e-09 ***
wind         -3.33760    0.65384  -5.105 1.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.17 on 107 degrees of freedom
Multiple R-squared:  0.6062,	Adjusted R-squared:  0.5952 
F-statistic: 54.91 on 3 and 107 DF,  p-value: < 2.2e-16
```

```
## we can observe that the intercept so when \t

## residuals are the values of the differences between the line we made
## and the observations

## the coefficients are the point estimations intercept is the beta0
## and the wt is the beta1
```

First thing we can observe is the confidence intervals of the 3 variables, both the temperature and wind have confidence intervals of 99,9% as it can be see by the 3 stars next to their respective p-values, radiation is in the 95% confidence interval but is close to the 99% confidence interval.

From the estimates we can than wind is the variable with the biggest impact per unit however when comparing it is also important to compare using the minimum and maximum values so we can determine how much each of the independent variables have been recorded to affect the ozone readings so we will be using the minimum and maximum to determine the maximum and minimum variance.

About the Coefficients, we can observe that radiation is the least impactful of the 3 independent variables, the changes in ozone radiation detected vary between [0.4186,19.9732] from the minimum and maximum values, which compared to the [94.11897,160.16737] minimum and maximum variance from the temperature readings which is by far the most impactful variable or the [-7.67648,-66.752] variance from the minimum and maximum values from the wind readings.

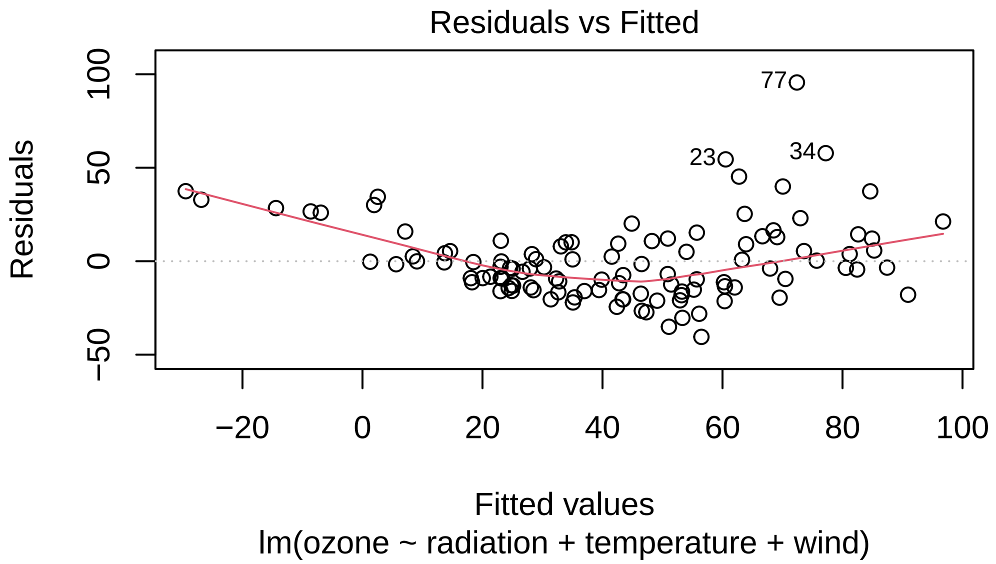
the intercept value is impossible so lets grpah what we have and have a look

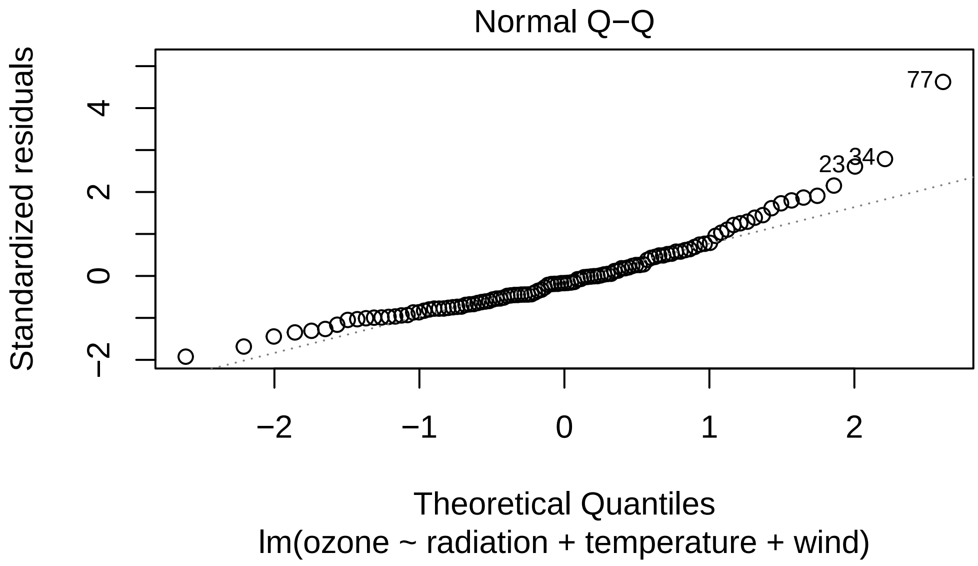
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.23208 23.04204 -2.788 0.00628 ** radiation 0.05980 0.02318 2.580 0.01124 *
temperature 1.65121 0.25341 6.516 2.43e-09 ***wind -3.33760 0.65384 -5.105 1.45e-06***

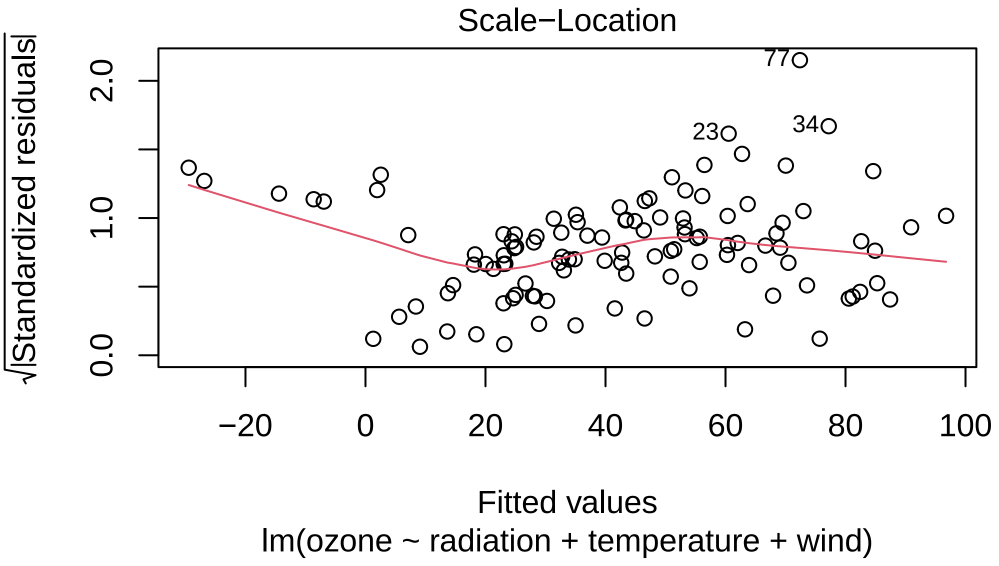
radiation temperature wind ozone
Min. : 7.0 Min. :57.00 Min. : 2.300 Min. : 1.0
1st Qu.:113.5 1st Qu.:71.00 1st Qu.: 7.400 1st Qu.: 18.0
Median :207.0 Median :79.00 Median : 9.700 Median : 31.0
Mean :184.8 Mean :77.79 Mean : 9.939 Mean : 42.1
3rd Qu.:255.5 3rd Qu.:84.50 3rd Qu.:11.500 3rd Qu.: 62.0
Max. :334.0 Max. :97.00 Max. :20.700 Max. :168.0

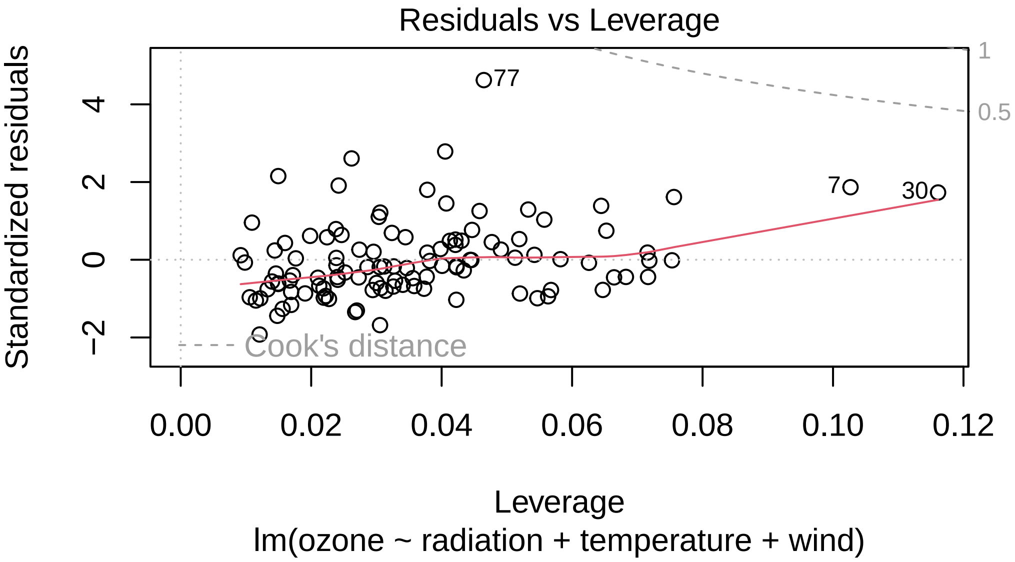
$\beta 0 = \beta 1$

```r
plot(model)
```




Normal Q−Q
Standardized residuals
−2
0
2
4
77
23
34
−2
−1
0
1
2
Theoretical Quantiles
lm(ozone ~ radiation + temperature + wind)


Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
77
23
34
−20
0
20
40
60
80
100
Fitted values
lm(ozone ~ radiation + temperature + wind)



As observed here on Residuals vs Fitted graph