Stats

Table of contents

── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
✔ ggplot2 3.3.6 ✔ dplyr 1.0.10  
✔ tibble 3.1.8 ✔ stringr 1.4.1   
✔ readr 2.1.3 ✔ forcats 0.5.2   
✔ purrr 0.3.4   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
Registered S3 method overwritten by 'GGally':  
 method from   
 +.gg ggplot2  
  
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.

## 1 Question 1 a)

|  |
| --- |
| Exercise 1 a) |

## 2 Question 1 b)

|  |
| --- |
| Exercise 1 b) |

## 3 Question 1 c)

Part 1 of exercise 1 c)

|  |
| --- |
| Exercise 1 c) part 1 |

Part 2 of exercise 1 c)

|  |
| --- |
| Exercise 1 c) part 2 |

## 4 Question 2 a)

Part 1 of exercise 2 a)

|  |
| --- |
| Exercise 2 a) part 1 |

Part 2 of exercise 2 a)

|  |
| --- |
| Exercise 2 a) part 2 |

## 5 Question 2 C)

# set.seed(26041999)  
  
candidateBWins = 0  
  
## Here we set the number of times we will estimate  
  
## largeNumberOfExtimations = 1000000 reducing the number so it doesn't  
## take an hour to compile  
largeNumberOfExtimations = 1e+05  
  
for (i in 1:largeNumberOfExtimations) {  
  
 candidateB = 175 ## candidate B initial votes/support  
 candidateM = 184 ## candidate M initial votes/support  
  
 ## increment for each day from the initial day to the final day,  
 ## the 14th  
 for (i in 1:14) {  
  
 ## will be used to hold the new value of votes that continue to  
 ## support the same candidate  
 candidateBAuxCalc = 0  
 candidateMAuxCalc = 0  
  
 ## here we will estimate how many candidates still support the  
 ## same candidate by generating a vector where 1 means they  
 ## still support the same candidate after the end of each day  
 ## the 0 represent them changing the support and vote from one  
 ## candidate to another  
 candidateBAuxCalc = sum(sample(c(1, 0), candidateB, replace = TRUE,  
 prob = c(0.996, 0.004)))  
 candidateMAuxCalc = sum(sample(c(1, 0), candidateM, replace = TRUE,  
 prob = c(0.995, 0.005)))  
  
 ## here we calculate the number of votes that will be exchanged  
 ## by the candidates, so the votes that are subtracted from  
 ## what they had previously  
 votesMovedFromCandidaBToCandidateM = candidateB - candidateBAuxCalc  
 votesMovedFromCandidaMToCandidateB = candidateM - candidateMAuxCalc  
  
 ## here calculate that the new current votes for each candidate  
 ## by adding the current votes from the mps that still support  
 ## the same candidate plus the number of votes of the mps that  
 ## exchanged support for their candidate  
 candidateB = candidateBAuxCalc + votesMovedFromCandidaMToCandidateB  
 candidateM = candidateMAuxCalc + votesMovedFromCandidaBToCandidateM  
  
 }  
 ## check if the candidate B did win the election by holding the  
 ## majority of the votes  
 if (candidateB > candidateM) {  
 candidateBWins = candidateBWins + 1  
 }  
  
}  
  
## probability estimation of candidate B winning after 14 days of  
## campaign by holding the majority of the votes  
  
probabilityB = candidateBWins/largeNumberOfExtimations  
  
probabilityB

[1] 0.35812

As we can see the probability of candidate B winning after 14 days is approximately 0.35

## 6 Question 2 d)

# set.seed(26041999)  
  
candidateBWins = 0  
  
## Here we set the number of times we will estimate  
## largeNumberOfExtimations = 1000000 reducing the number so it doesn't  
## take an hour to compile  
largeNumberOfExtimations = 1e+05  
  
for (i in 1:largeNumberOfExtimations) {  
  
 candidateB = 175 ## candidate B initial votes/support  
 candidateM = 184 ## candidate M initial votes/support  
  
 ## increment for each day from the initial day to the final day,  
 ## the 60th  
 for (i in 1:60) {  
  
 ## will be used to hold the new value of votes that continue to  
 ## support the same candidate  
 candidateBAuxCalc = 0  
 candidateMAuxCalc = 0  
  
 ## here we will estimate how many candidates still support the  
 ## same candidate by generating a vector where 1 means they  
 ## still support the same candidate after the end of each day  
 ## the 0 represent them changing the support and vote from one  
 ## candidate to another  
 candidateBAuxCalc = sum(sample(c(1, 0), candidateB, replace = TRUE,  
 prob = c(0.996, 0.004)))  
 candidateMAuxCalc = sum(sample(c(1, 0), candidateM, replace = TRUE,  
 prob = c(0.995, 0.005)))  
  
 ## here we calculate the number of votes that will be exchanged  
 ## by the candidates, so the votes that are subtracted from  
 ## what they had previously  
 votesMovedFromCandidaBToCandidateM = candidateB - candidateBAuxCalc  
 votesMovedFromCandidaMToCandidateB = candidateM - candidateMAuxCalc  
  
 ## here calculate that the new current votes for each candidate  
 ## by adding the current votes from the mps that still support  
 ## the same candidate plus the number of votes of the mps that  
 ## exchanged support for their candidate  
 candidateB = candidateBAuxCalc + votesMovedFromCandidaMToCandidateB  
 candidateM = candidateMAuxCalc + votesMovedFromCandidaBToCandidateM  
  
 }  
 ## check if the candidate B did win the election by holding the  
 ## majority of the votes  
 if (candidateB > candidateM) {  
 candidateBWins = candidateBWins + 1  
  
 }  
  
}  
  
## probability estimation of candidate B winning after 14 days of  
## campaign by holding the majority of the votes  
  
probabilityB = candidateBWins/largeNumberOfExtimations  
  
probabilityB

[1] 0.77216

As we can see the probability of candidate B winning after 60 days is approximately 0.77

## 7 Question 3 a)

|  |
| --- |
| Exercise 3 a) |

## 8 Question 3 b)

|  |
| --- |
| Exercise 3 b) |

## 9 Question 3 c)

|  |
| --- |
| Exercise 3 c) |

## 10 Question 3 d)

|  |
| --- |
| Exercise 3 d) |

## 11 Question 3 e)

|  |
| --- |
| Exercise 3 e) |

## 12 Question 4 a)

|  |
| --- |
| Exercise 4 a) |

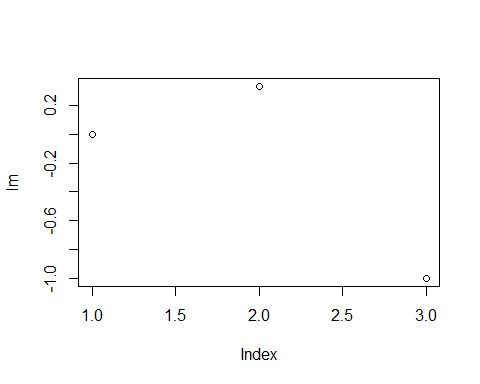
## 13 Question 4 c)

vectorY = c(0.573, 0.77, 0.652, 0.827, 0.821, 0.789, 0.898, 0.718, 0.382,  
 0.668, 0.647, 0.477, 0.661, 0.38, 0.87, 0.794, 0.783, 0.732, 0.629, 0.777,  
 0.6, 0.724, 0.553, 0.693, 0.687, 0.935, 0.494, 0.411, 0.53, 0.478)  
  
# rt <- polyroot(t)  
  
# replacing theta with x because it is easier to type on QWERTY  
# keyboards  
  
# k = ((x+2)/y^(x+2) - (x+3)/y^(x+3)) fy = ((x+2)/y^(x+2) -  
# (x+3)/y^(x+3)) \* (1 - y)\*y^(x+1)  
  
fy = polyroot(c(0, 1, -2, -3))  
  
fyReal = Re(fy)  
  
fyReal

[1] 0.0000000 0.3333333 -1.0000000

## 14 Question 4 d)

# lm = lm(fyReal)  
  
lm = fyReal  
  
plot(lm)



## 15 Question 5 a)

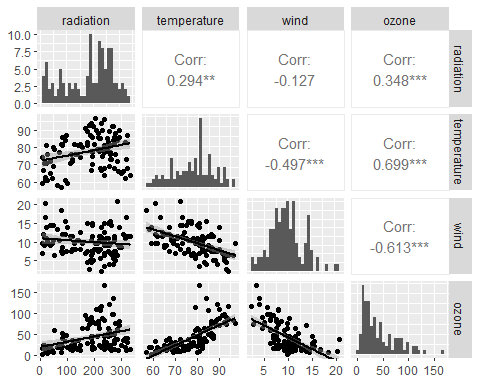
summary(ozone)

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

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 temperature have a positive and strong linear correlation with ozone 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 positive correlation 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 has a negative correlation with ozone, 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.

## 16 Question 5 b)

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

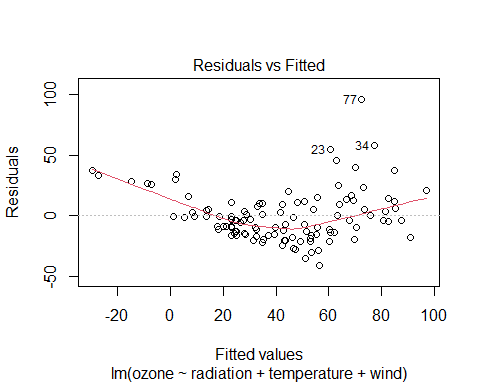
First thing we can observe is the confidence intervals of the 3 variables, both the temperature and wind have confidence intervals of 99,9% as it can be see by the 3 stars next to their respective p-values, radiation is in the 95% confidence interval but is close to the 99% confidence interval, meaning all 3 variables have a significant association and and are a meaningful addiction to our model. Significance being p-values < 0.05.

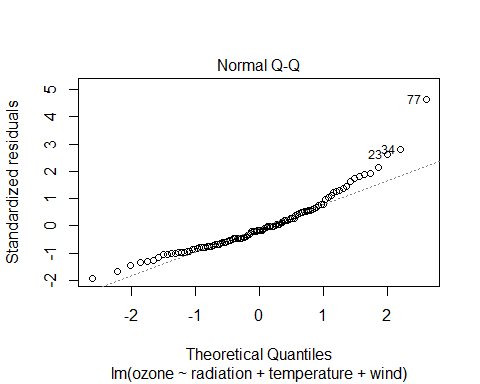
We can also see by the value of the R-squared and adjusted R-squared that this model around 60% of the variation in ozone levels.

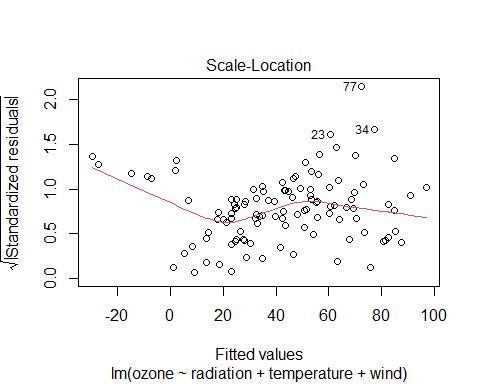
From the estimates we can determine than wind is the variable with the biggest impact per unit however when comparing it is also important to compare using the minimum and maximum values so we can determine how much each of the independent variables have been recorded to affect the ozone readings so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

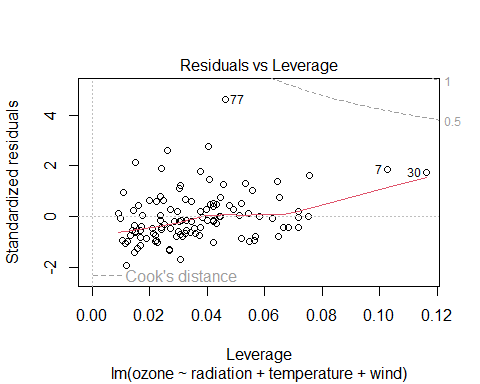
About the Coefficients, we can observe that radiation is the least impactful of the 3 independent variables, the changes in ozone radiation detected vary between [0.4186,19.9732] from the minimum and maximum values, which compared to the [94.11897,160.16737] minimum and maximum variance from the temperature 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.

plot(model)









## a residual is the distance between our data points and our  
## regression line

A good Residuals vs Fitted graph main characteristic should be that the residuals bounce randomly around the 0 line, that would suggest the linear is reasonable, in this case we can observe that the residuals follow a clear trend of an almost quadratic function.

Another Residuals vs Fitted graph main characteristic is the formation of an horizontal trend around 0, that would suggest the variance of errors would be equal. Since our graph does not exactly follow this trend we know that is noise introduced systematically to induce these changes in errors

The third characteristic of a Residuals vs Fitted graph should be no residuals stands out from the graph, meaning there are no outliers. As we can see this is not the case with 3 specific points, number 23, 34 and 77 even being labelled as outliers.

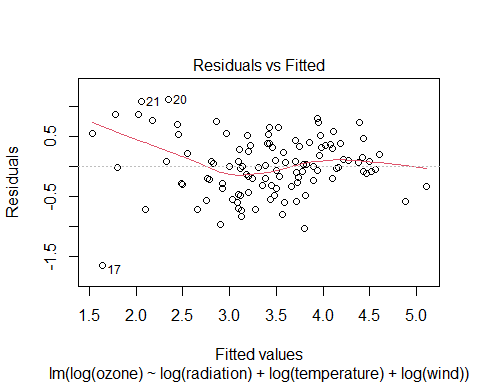
This means that the linear model is not exactly the best fit for this data and that maybe we should try a log transform to convert this data into a more linear fit.

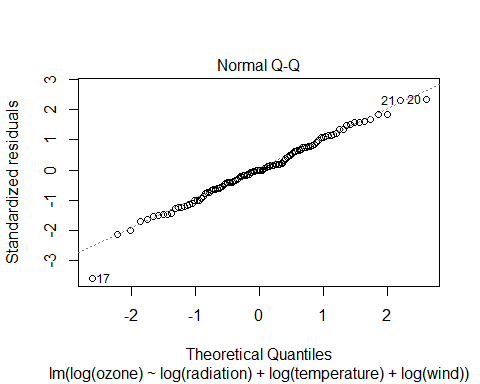
The Q-Q plot or Quantile-Quantile plot compares 2 probability distributions by comparing them against each other. Varshney, P. (2020) Q-Q plots explained, Medium. Towards Data Science. Available at: https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0 (Accessed: December 2, 2022).

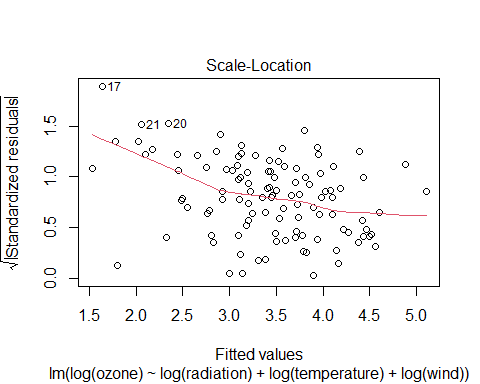
When the plotted points create a straight line then we can identify this distribution as a normally distribution because it is aligned with the standard normal, that is what we can identify on a Q-Q plot. As we can see our values follow very closely the fitted line with the exception of the extreme where it has a slight but systematic deviation from the standard line and a few outliers again labeled 23, 34 and 77.

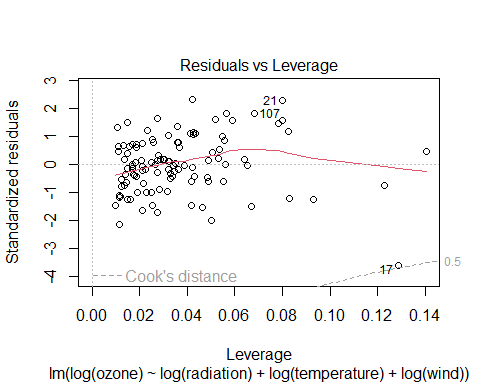
## 17 Question 5 c)

## log(ozone) = β0+β1 log(radiation)+β2 log(temperature)+β3  
## log(wind)+εi where εi ∼ N (0, σ2) β0 = intercept  
  
modelLogFriend = lm(log(ozone) ~ log(radiation) + log(temperature) + log(wind),  
 data = ozone)  
  
  
plot(modelLogFriend)









A good Residuals vs Fitted graph main characteristic should be that the residuals bounce randomly around the 0 line, that would suggest the linear is reasonable, in this case we can observe that the residuals don’t follow any clear trend and look mostly randomly scattered.

Another Residuals vs Fitted graph main characteristic is the formation of an horizontal trend around 0, that would suggest the variance of errors would be equal. Since our graph does not exactly follow this trend we know that is noise introduced systematically to induce these changes in errors

The third characteristic of a Residuals vs Fitted graph should be no residuals stands out from the graph, meaning there are no outliers. As we can see this is not the case with especially number 17 even being labelled as outliers with many others scattered very far away from the main concentration of points.

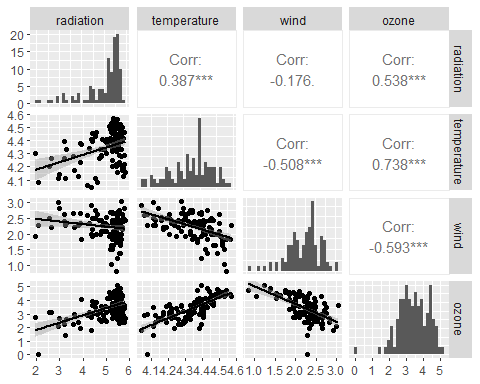
This means that the linear model is not exactly the best fit for this data and that maybe we should try a log transform to convert this data into a more linear fit.

The Q-Q plot or Quantile-Quantile plot compares 2 probability distributions by comparing them against each other. Varshney, P. (2020) Q-Q plots explained, Medium. Towards Data Science. Available at: https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0 (Accessed: December 2, 2022).

When the plotted points create a straight line then we can identify this distribution as a normally distribution because it is aligned with the standard normal, that is what we can identify on a Q-Q plot. As we can see our values follow with extreme precision the fitted line meaning we have normalized the distribution by log transforming the parameters.

## so we can better see the new log transformed data I will make a new  
## log transformed data set and then apply ggpairs to it  
  
ozoneLogged = ozone  
  
ozoneLogged$radiation = log(ozoneLogged$radiation)  
ozoneLogged$wind = log(ozoneLogged$wind)  
ozoneLogged$temperature = log(ozoneLogged$temperature)  
ozoneLogged$ozone = log(ozoneLogged$ozone)  
  
  
ggpairs(ozoneLogged, 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`.



summary(ozoneLogged)

radiation temperature wind ozone   
 Min. :1.946 Min. :4.043 Min. :0.8329 Min. :0.000   
 1st Qu.:4.732 1st Qu.:4.263 1st Qu.:2.0015 1st Qu.:2.890   
 Median :5.333 Median :4.369 Median :2.2721 Median :3.434   
 Mean :4.979 Mean :4.346 Mean :2.2261 Mean :3.416   
 3rd Qu.:5.543 3rd Qu.:4.437 3rd Qu.:2.4423 3rd Qu.:4.127   
 Max. :5.811 Max. :4.575 Max. :3.0301 Max. :5.124

In the histograms we can immediately spot that ozone, wind and temperature seem to have a normal distribution without any immediately significant looking skewness, while bith temperature and wind were already with a normal looking distribution this was not the case for the ozone before the log transform. Radiation skewdness seem to have increased even more barely having any values on the skewed left area due to such an high skewness level.

summary(modelLogFriend)

Call:  
lm(formula = log(ozone) ~ log(radiation) + log(temperature) +   
 log(wind), data = ozone)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1.63961 -0.30073 -0.00097 0.34414 1.11545   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -10.55570 2.08818 -5.055 1.79e-06 \*\*\*  
log(radiation) 0.30500 0.05868 5.198 9.73e-07 \*\*\*  
log(temperature) 3.20478 0.46019 6.964 2.79e-10 \*\*\*  
log(wind) -0.66305 0.13751 -4.822 4.74e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.4907 on 107 degrees of freedom  
Multiple R-squared: 0.6876, Adjusted R-squared: 0.6788   
F-statistic: 78.49 on 3 and 107 DF, p-value: < 2.2e-16

# summary(model)

First thing we can observe is the confidence intervals of the 3 variables, all the 3 variables have confidence intervals of 99,9% as it can be see by the 3 stars next to their respective p-values, meaning all 3 variables have a significant association and and are a meaningful addiction to our model. Significance being p-values < 0.05. As such there is an increase in significance of the variables compared to the previous model although the changes in Confidence Interval are not significant enough to be very impactful.

We can also see by the value of the R-squared and adjusted R-squared that this model around 68% of the variation in ozone levels, being able to explain 8% more than the original model.

From the estimates we can observe than wind is no longer the variable with the biggest impact per unit. This is now in fact the temperature that becomes by far the most impactful per unit. however when comparing it is also important to compare using the minimum and maximum values plus how quickly the increase or decrease the level of ozone we can check that temperature becomes incredibly more capable of altering ozone levels.

We will be using the minimum and maximum with the correlation to determine how much each of the independent variables have been recorded to affect the ozone readings.

About the Coefficients, we can observe that radiation is still the least impactful of the 3 independent variables even thought it has significantly increased in ratio for ozone reading compared to the previous linear model. We can also determine by the ggpairs graphs with the log transformed variables that most of the radiation are close to increasing the most possible inside our scale with only a few points in the middle of the scale and some outliers farther away.

The changes in ozone by radiation detected vary between [0.59353,1.772355] from the minimum and maximum values, which increases slightly the minimum and actually increase the relevance of the radiation in comparison with the other explanatory variables.

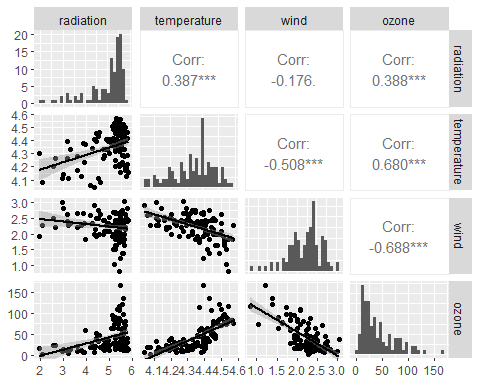
which compared to the [12.95692554,14.6618685] minimum and maximum variance from the temperature readings which is still by far the most impactful variable has decreased greatly in the interval of captured values meaning that while the explanatory variable is still incredibly impactful it has now a much lesser degree to how much its values impact the level of ozone read.

Finally, the wind, the only variable here that reduces the read ozone level from [0,-3.3974682] according to the minimum and maximum values continues to be relevant with its level of changing negatively the number of ozone ppb we have detected.

A very obvious change with the new model is how much closer, comparatively to the other variables, has the variance between minimum and maximum became, making each of the variables have similar weights with how much they can change ozone readings in comparison to the other variables.

## for us to see the difference between the log explanatory variables  
## and the unchanged ozone level  
  
ozoneLogged = ozone  
  
ozoneLogged$radiation = log(ozoneLogged$radiation)  
ozoneLogged$wind = log(ozoneLogged$wind)  
ozoneLogged$temperature = log(ozoneLogged$temperature)  
ozoneLogged$ozone = ozoneLogged$ozone  
  
  
ggpairs(ozoneLogged, 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`.



summary(ozoneLogged)

radiation temperature wind ozone   
 Min. :1.946 Min. :4.043 Min. :0.8329 Min. : 1.0   
 1st Qu.:4.732 1st Qu.:4.263 1st Qu.:2.0015 1st Qu.: 18.0   
 Median :5.333 Median :4.369 Median :2.2721 Median : 31.0   
 Mean :4.979 Mean :4.346 Mean :2.2261 Mean : 42.1   
 3rd Qu.:5.543 3rd Qu.:4.437 3rd Qu.:2.4423 3rd Qu.: 62.0   
 Max. :5.811 Max. :4.575 Max. :3.0301 Max. :168.0

Initially we can see that when compared to the log transform model the histograms have not had a significant change with the exception of ozone who has returned to have its initial model distribution.

Each of the individual log transformed values seem to follow a very similar patern when comparing the log transformed and untransformed ozone variable. The most significant variance would how radiation has a much less step slope when using the unstransformed scaled compared temperature and wind who slopes seem to remaind almost identicals