Correlation and Simple Linear Regression Assignment

## Prince Agyabeng

Libraries needed

library(tidyverse)

## -- Attaching packages ----------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts -------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(GGally)

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

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

Task 1 Describing the Dataset There are 153 observations made up of 6 variables.Also the data is missing vaues for certain columns. I think Ozone will be the response variable

air = airquality  
str(airquality)

## 'data.frame': 153 obs. of 6 variables:  
## $ Ozone : int 41 36 12 18 NA 28 23 19 8 NA ...  
## $ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...  
## $ Wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...  
## $ Temp : int 67 72 74 62 56 66 65 59 61 69 ...  
## $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Day : int 1 2 3 4 5 6 7 8 9 10 ...

summary(airquality)

## Ozone Solar.R Wind Temp   
## Min. : 1.00 Min. : 7.0 Min. : 1.700 Min. :56.00   
## 1st Qu.: 18.00 1st Qu.:115.8 1st Qu.: 7.400 1st Qu.:72.00   
## Median : 31.50 Median :205.0 Median : 9.700 Median :79.00   
## Mean : 42.13 Mean :185.9 Mean : 9.958 Mean :77.88   
## 3rd Qu.: 63.25 3rd Qu.:258.8 3rd Qu.:11.500 3rd Qu.:85.00   
## Max. :168.00 Max. :334.0 Max. :20.700 Max. :97.00   
## NA's :37 NA's :7   
## Month Day   
## Min. :5.000 Min. : 1.0   
## 1st Qu.:6.000 1st Qu.: 8.0   
## Median :7.000 Median :16.0   
## Mean :6.993 Mean :15.8   
## 3rd Qu.:8.000 3rd Qu.:23.0   
## Max. :9.000 Max. :31.0   
##

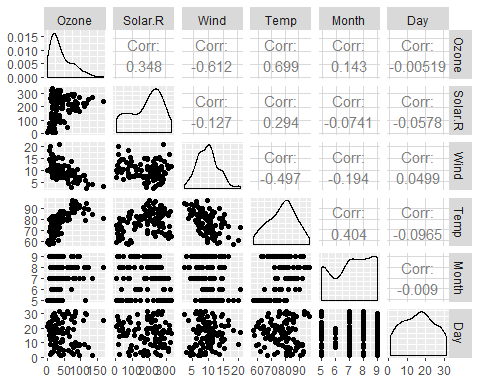
Task 2 Removing of missing data rows  
There are now 111 observation and 6 varibles of the new data air2

air2 = airquality %>% drop\_na()  
str(air2)

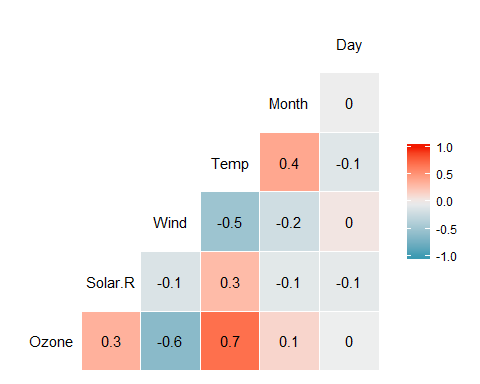
## 'data.frame': 111 obs. of 6 variables:  
## $ Ozone : int 41 36 12 18 23 19 8 16 11 14 ...  
## $ Solar.R: int 190 118 149 313 299 99 19 256 290 274 ...  
## $ Wind : num 7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...  
## $ Temp : int 67 72 74 62 65 59 61 69 66 68 ...  
## $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Day : int 1 2 3 4 7 8 9 12 13 14 ...

Task 3 A look at ggpairs plot for visualization and correlation. The best variable to predict Ozone appears to be Temp (correlation = 0.699

ggpairs(air2)

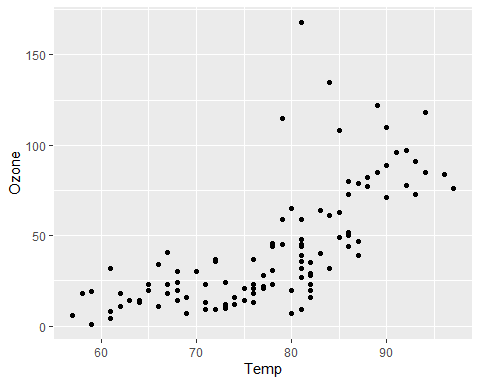


ggcorr(air2, label = TRUE)

 Which variable is most strongly correlated with the “Ozone” variable? Temp Which variable is least strongly correlated with the “Ozone” variable? Month

Task 4 Finding the relationship between Tempe and Ozone From the scatterplot, we can validate that there’s a strong positive relationship between Temp and Ozone.

ggplot(air2, aes(x=Temp, y = Ozone)) + geom\_point()



Task 5 Build a regression model1 using Temp to predict Ozone

model1 = lm(Ozone ~ Temp, data = air2)  
summary(model1)

##   
## Call:  
## lm(formula = Ozone ~ Temp, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.922 -17.459 -0.874 10.444 118.078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -147.6461 18.7553 -7.872 2.76e-12 \*\*\*  
## Temp 2.4391 0.2393 10.192 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23.92 on 109 degrees of freedom  
## Multiple R-squared: 0.488, Adjusted R-squared: 0.4833   
## F-statistic: 103.9 on 1 and 109 DF, p-value: < 2.2e-16

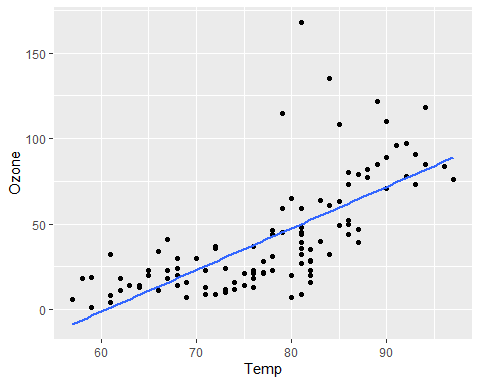
confint(model1)

## 2.5 % 97.5 %  
## (Intercept) -184.818372 -110.473773  
## Temp 1.964787 2.913433

The above model1 is significant as a unit change in Temp -147.64 change in the ozone layer. R2 is 0.488 which is good as as nearly half of the model can explain the data. The Temp variable is significant (p-value < 0.05) and it has explanatory power. The range of the slope falls between -184.818 and -110.473

Task 6 Plot the model It seems the model is good as the graph indicate a strong relationship between Temp an Ozone

ggplot(air2, aes(x=Temp, y = Ozone)) + geom\_point()+  
 geom\_smooth(method= "lm", se = FALSE)



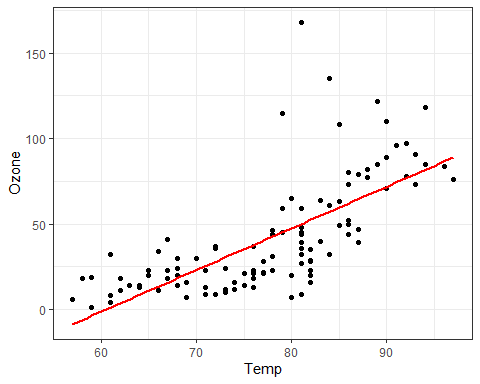
Task 7 Using the predict function

testdata = data.frame(Temp = c(80))  
predict(model1, newdata = testdata, interval = "predict")

## fit lwr upr  
## 1 47.48272 -0.1510188 95.11646

Task 8 **Assumption 1** The predictor and response variable mmust have a linear relationship. We can see the graph that there’s a linear relationship between Tem and Zone

ggplot(air2, aes(x=Temp,y=Ozone)) + geom\_point() + geom\_smooth(method="lm",se=FALSE, color="red") + theme\_bw()



**Assumption 2** Model errors (residuals) are independent

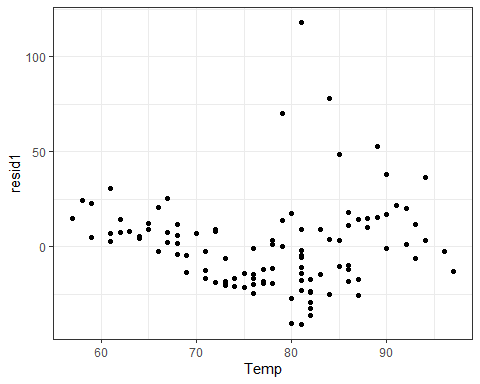
dwtest(model1)

##   
## Durbin-Watson test  
##   
## data: model1  
## DW = 1.8644, p-value = 0.2123  
## alternative hypothesis: true autocorrelation is greater than 0

We fail to reject the null hypothesis with a p-value greater than 0.05. This suggests that the residuals are likely independent.

**Assumption 3** Model residuals exhibit constant variance

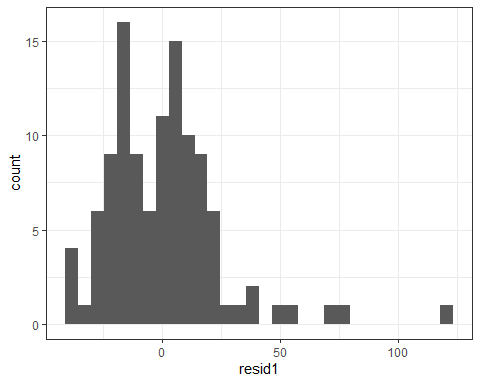
air2 = air2 %>% mutate(resid1 = model1$residuals)  
  
ggplot(air2,aes(x=Temp,y=resid1)) + geom\_point() + theme\_bw()

 From the graph the model residual does not exhibit constant variance thus violating this assumption.

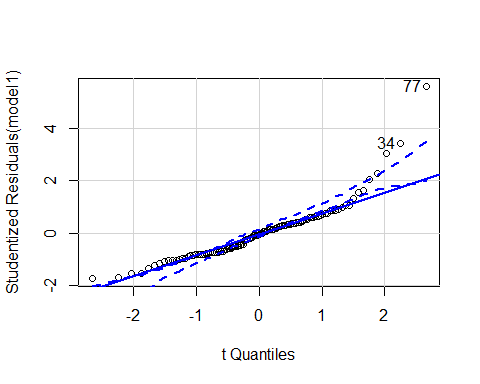
**Assumption 4** Model residuals are Normally-distributed

ggplot(air2,aes(x=resid1)) + geom\_histogram() + theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



##For a second check  
qqPlot(model1)



## [1] 34 77

From the histogram above, The residuals histogram is not normally distributed, thus violating the normal distribution assumption.

Task 9 The model in Task 5 can be used to predict the effects of Temp on the Ozone layer. Any unit of increase of Temp has a negative influence on the Ozone. when using this model, one must understand that the modle violates the model residual assumption and the normal distribution assumption.