## Simple Linear Regression and Correlation

### Anthony Sumter

Libraries

#install.packages("tidyverse","GGally","car","lmtest")  
library(tidyverse)

## -- Attaching packages ----------------------------------------------------- tidyverse 1.2.1 --

## 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)

## Warning: package 'GGally' was built under R version 3.6.2

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

##   
## Attaching package: 'GGally'

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

library(car)

## Warning: package 'car' was built under R version 3.6.2

## 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)

## Warning: package 'lmtest' was built under R version 3.6.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.6.2

##   
## Attaching package: 'zoo'

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

#### Task 1

Read-in the airqualtiy data set (a default R dataset) as a data frame called “air”. Describe this dataset. How many variables and observations are there? Is there any missing data? Which variable is likely to be the response (Y) variable?

**The air data frame decribes the daily air quality within the months of May to September and which varibles impact it. In this data set there are 6 varibles and a total of 153 observations. Some data does happen to be missing in the ozone and solar.r variables. Out of all the variables it appears that the Ozone varible is likely to be the response (Y) variable.**

air = airquality

#### Task 2

In Task 1 you would have discovered that there is missing data in two of the variables: Ozone and Solar.R. Here we’ll choose deletion of rows with any missing data. Save your new data frame (with missing data removed) as a data frame named “air2”. How many rows and columns remain in this new (air2) data frame?

**In the new data frame of air2 there are 111 observations (rows) and 6 variables (columns) that remain.**

air2 <- air %>% drop\_na()

#### Task 3

Use the ggpairs function to develop a visualization of and to calculate correlation for the combinations of variables in this dataset. Then use the “ggcorr” function to develop a correlation matrix for the variables.

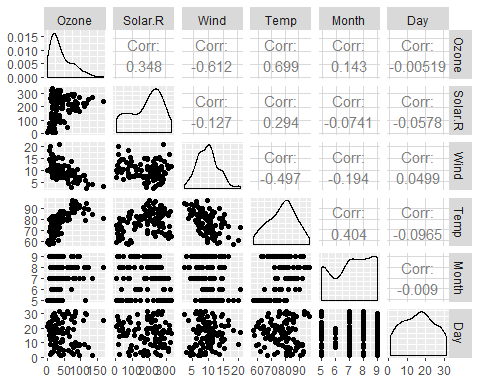
Which variable is most strongly correlated with the “Ozone” variable?

**The variable that is most strongly correlated with the Ozone Variable is the Temp Variable.**

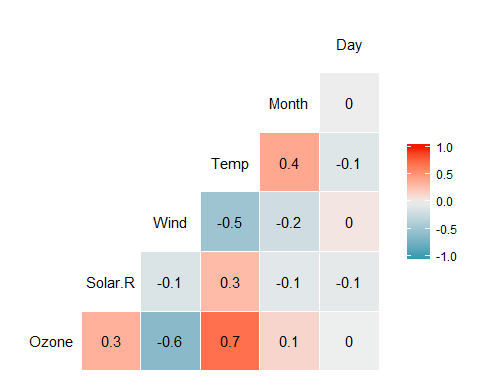
Which variable is least strongly correlated with the “Ozone” variable?

**The variable taht is least strongly correlated with the Ozone variable is the Wind variable**

ggpairs(air2)



ggcorr(air2, label = TRUE)

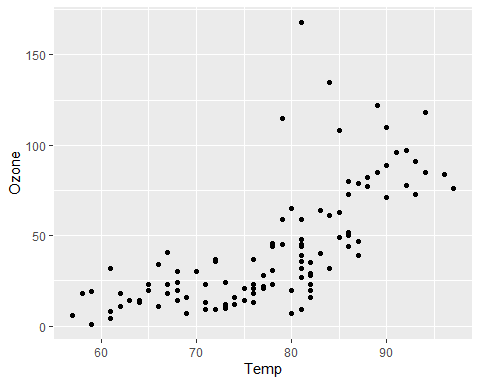


#### Task 4

Plot “Temp” (x axis) versus “Ozone” (y axis) using the “ggplot” function. Choose an appropriate chart type. Describe the relationship between “Temp” and “Ozone”.

**Based on the chart below the relationship that the two variables share is that when Temp goes up so does Ozone.**

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



#### Task 5

Create a linear regression model (called model1) using “Temp” to predict “Ozone”. a. Discuss the quality of this model (mention the R square value and significance of the predictor variable). b. Use the code “confint(model1)” to generate 95% confidence intervals for the coefficients. In what range does the slope coefficient likely fall?

**a.When evaluating the quality of the model the R square value is ok at .48 and the Temp variable is significant with a p value less than .05 and has an intuitive sign. Thus this would be a good model**

model1 <- lm(Ozone ~ Temp, 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

**b.Based on the below data the range that that the slope coefficient likely falls in is from 1.965 to 2.913.**

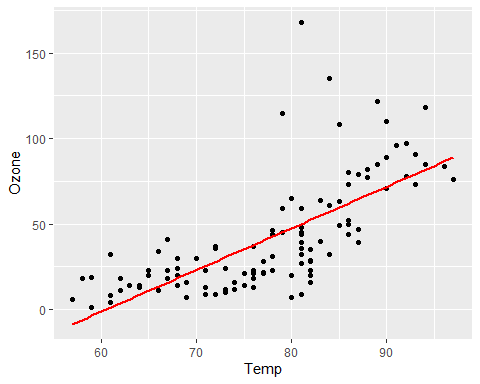
confint(model1)

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

#### Task 6

Re-do Task 4 to include the regression line.

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



#### Task 7

Develop a prediction for “Ozone” when “Temp” is 80.

**Based on the below data it is predicted that when Temp reaches 80 that the Ozone will be 47 (rounded).**

Predict80 = data.frame(Temp = c(80))  
predict(model1, newdata = Predict80, interval = "predict")

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

#### Task 8

Perform appropriate model diagnostics to verify whether or not the model appears to meet the four linear regression model assumptions. Provide a brief comment on each assumptions validity for this model.

**Assumption 1: The predictor variable and response variable have a linear relationship based upon the plot visualization and that the Temp variable is a significant one.**

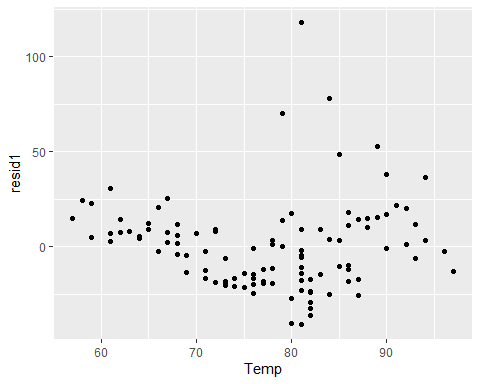
**Assumption 2: Per the below data We fail to reject the null hypothesis with a p-value greater than 0.05 at .21 and this suggests that the residuals are likely independent.**

dwtest(model1)

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

**Assumption 3: It appears just like in the first plot between the predctor variable and response variable that there is a relationship between the residuals and Temp Variable.**

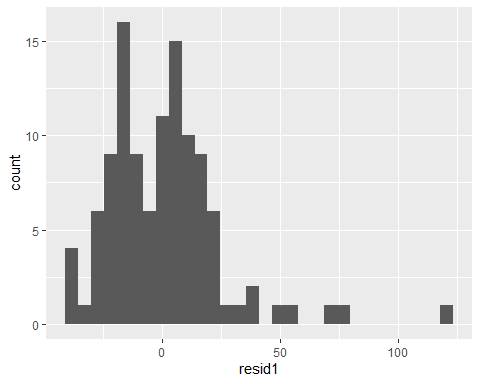
air2 = air2 %>% mutate(resid1 = model1$residuals)  
ggplot(air2,aes(x=Temp,y=resid1)) + geom\_point()



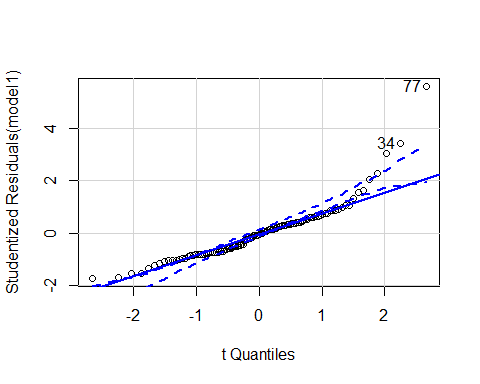
**Assumption 4: The Normal Probability Plot confirms what we saw in the histogram that there is normal distribution based on the points being aligned in a straight line.**

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

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



qqPlot(model1)



## [1] 34 77

#### Task 9

How might the model that you constructed in Task 5 be used? Are there any cautions or concerns that you would have when recommending the model for use?

**The model in task 5 might be used to predict the onzone quality based upon the temp for a particular day. There may be caution or concern as it is not a 100% indicator that the Ozone will good or bad based upon temp for that day.Even with that being the case it can be used to get an get a decent idea of what to expect in terms of air quality for the day.**