Module 2: Assignment 1

Chris Collins

January 31, 2019

#install.packages("tidyverse")  
#install.packages("GGally")  
library(tidyverse)

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

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.8  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.3.1 v forcats 0.3.0

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

library(GGally)

##   
## Attaching package: 'GGally'

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

air=airquality

# 1a. This dataset uses 6 different variables to best predict/test the air quality using a Regression Line.

# 1b. 153 observations of 6 variables

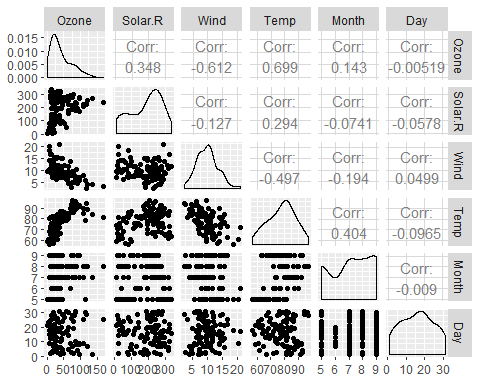
# 1c. Missing data found in the Ozone and Solar. R attributes

# 1d. The response variable (Y) is likely to be Ozone.

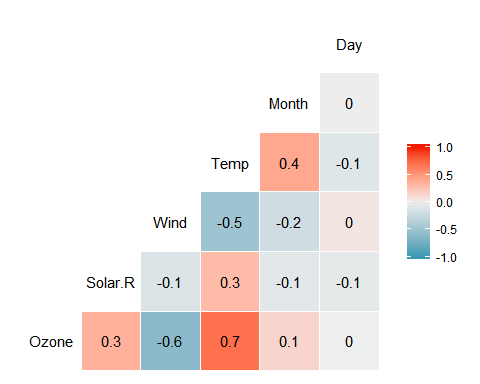
air2=air%>%filter(!is.na(Ozone))%>%filter(!is.na(Solar.R))

# 2 There are 111 observations of 6 variables in the new data frame “air2”.

ggpairs(air2)



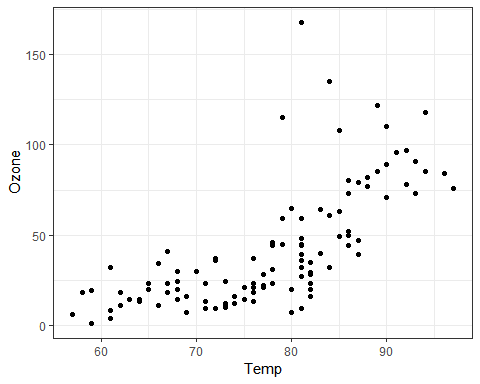
ggcorr(air2,label = TRUE)



# 3a. ozone and temp at 0.699

# 3b. ozone and day at -0.00519

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



mod1=lm(Ozone~Temp,air2)  
summary(mod1)

##   
## 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

# 5a. The p-value shows that it is a significant variable as it is less than .001. The Multiple R-squared is 0.488 and the Adjusted R-squared is 0.4833; therefore, it is an OK R-square value, not too high or low.

confint(mod1,"Temp",level=0.95)

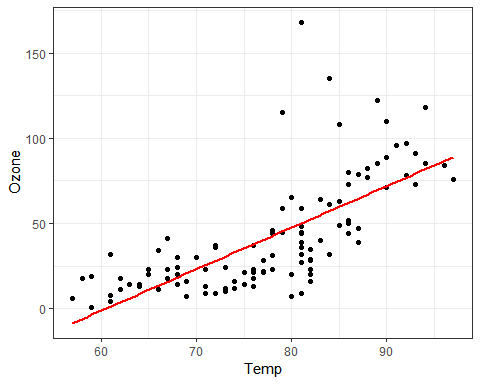
## 2.5 % 97.5 %  
## Temp 1.964787 2.913433

summary(mod1)

##   
## 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

# 5b. Given that the slope is 2.4391 that would mean that it would fall in between the median and 3rd quartile with 95% confidence level.

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



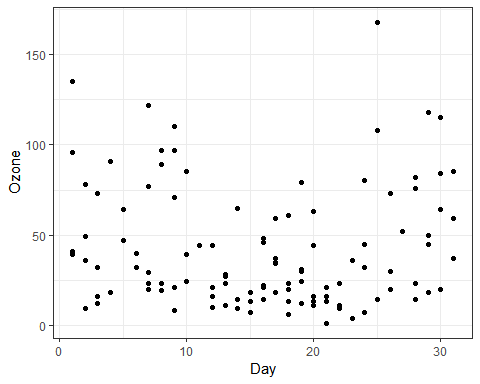
summary(air2)

## Ozone Solar.R Wind Temp   
## Min. : 1.0 Min. : 7.0 Min. : 2.30 Min. :57.00   
## 1st Qu.: 18.0 1st Qu.:113.5 1st Qu.: 7.40 1st Qu.:71.00   
## Median : 31.0 Median :207.0 Median : 9.70 Median :79.00   
## Mean : 42.1 Mean :184.8 Mean : 9.94 Mean :77.79   
## 3rd Qu.: 62.0 3rd Qu.:255.5 3rd Qu.:11.50 3rd Qu.:84.50   
## Max. :168.0 Max. :334.0 Max. :20.70 Max. :97.00   
## Month Day   
## Min. :5.000 Min. : 1.00   
## 1st Qu.:6.000 1st Qu.: 9.00   
## Median :7.000 Median :16.00   
## Mean :7.216 Mean :15.95   
## 3rd Qu.:9.000 3rd Qu.:22.50   
## Max. :9.000 Max. :31.00

air=airquality  
air2=air%>% filter(!is.na(Ozone))%>%filter(!is.na(Solar.R))  
  
x=air2$Temp  
y=air2$Ozone  
  
model1=lm(y~x)  
predict(model1,data.frame("x"=80), interval = 'p')

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

ggplot(air2,aes(x=Day,y=Ozone))+geom\_point()+theme\_bw()



# 8.The plot is scattered and there seems to be no correlation between day and ozone levels. With that being said, we can predict a high p-value meaning no significance.

mod2=lm(Ozone~Day,air2)  
summary(mod2)

##   
## Call:  
## lm(formula = Ozone ~ Day, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.00 -24.23 -11.04 19.96 126.08   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.41536 6.64353 6.384 4.32e-09 \*\*\*  
## Day -0.01983 0.36604 -0.054 0.957   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33.43 on 109 degrees of freedom  
## Multiple R-squared: 2.693e-05, Adjusted R-squared: -0.009147   
## F-statistic: 0.002936 on 1 and 109 DF, p-value: 0.9569

# 9a. The day and ozone variables are not significant at all and have (essentially) 0 as Multiple R-squared and an Adjusted R-squared. This means they are not related at all and have no correlation.

confint(mod2,"Day",level=0.95)

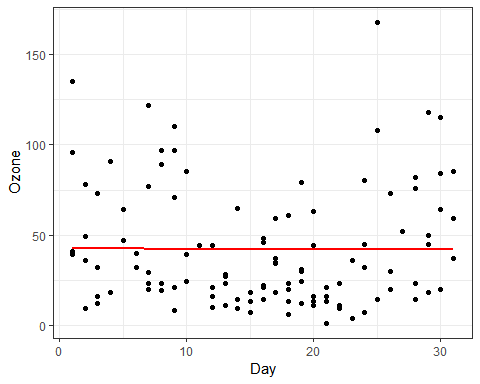
## 2.5 % 97.5 %  
## Day -0.745321 0.7056539

summary(mod2)

##   
## Call:  
## lm(formula = Ozone ~ Day, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.00 -24.23 -11.04 19.96 126.08   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.41536 6.64353 6.384 4.32e-09 \*\*\*  
## Day -0.01983 0.36604 -0.054 0.957   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33.43 on 109 degrees of freedom  
## Multiple R-squared: 2.693e-05, Adjusted R-squared: -0.009147   
## F-statistic: 0.002936 on 1 and 109 DF, p-value: 0.9569

# 9b. Given that the slope is -0.01983, it would fall in between the Median and 3rd quartile with 95% confidence level.

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



summary(air2)

## Ozone Solar.R Wind Temp   
## Min. : 1.0 Min. : 7.0 Min. : 2.30 Min. :57.00   
## 1st Qu.: 18.0 1st Qu.:113.5 1st Qu.: 7.40 1st Qu.:71.00   
## Median : 31.0 Median :207.0 Median : 9.70 Median :79.00   
## Mean : 42.1 Mean :184.8 Mean : 9.94 Mean :77.79   
## 3rd Qu.: 62.0 3rd Qu.:255.5 3rd Qu.:11.50 3rd Qu.:84.50   
## Max. :168.0 Max. :334.0 Max. :20.70 Max. :97.00   
## Month Day   
## Min. :5.000 Min. : 1.00   
## 1st Qu.:6.000 1st Qu.: 9.00   
## Median :7.000 Median :16.00   
## Mean :7.216 Mean :15.95   
## 3rd Qu.:9.000 3rd Qu.:22.50   
## Max. :9.000 Max. :31.00