## Week 2 Assignment 1

### DEY, Sankha

#### Correlation and Simple Linear Regression Assignment

#### Answers Summary:

Task -1:

There are total 153 observations and 6 variables in the dataset. Five variables are integer and one is number (with decimal). Ozone and Solar.R variables have 37 and 7 null observations. Ozone is likely to be response (Y) variable.

Task-2:

111 rows and 6 colums are there in air2. Using drop\_na() we have removed 42 rows.

Task-3:

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

Task-4:

Temp and Ozone have a strong positive correlationship (~0.7). With Temp increased, Ozone also getting increased. There are few outliers though in the relationship.

Task-5.a:

p-value < 0.05, slope is positive 2.4 and R-squared value is 0.48. One unit of Temp change will result it 2.4 unit of Ozone change. All three suggest that Temp is a significant predictor of Ozone.

Task-5.b:

Task 5.b: 95% Confidence Intervals will be between -184.818 and -110.473. Slope coefficient likely to fall between 1.964 and 2.913.

Task -7:

When Temp = 80, then predicted Ozone is 47.48 with range varying from -0.151 (lower range) to 95.116 (upper range).

Task-8:

**Assumption 1:** The predictor and response variable have a linear relationship  
It seems reasonable to say that there is something of a linear relationship between these two variables (Temp and Ozone).  
**Assumption 2:** Model errors (residuals) are independent  
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

Seems there are not many changes in the variances of residuals. They represent constant variance.

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

The residuals histogram is normal. An alternative to the histogram is the Normal Probability Plot. The black points are aligned with the blue line from the bottom left to the upper right.

Task-9:

This model can be used to predict the Ozone volume with Temp change. Y Intercept is -147.6 and one unit of Temp change will result it 2.4 unit of Ozone change.  
Talking on cautions or concerns, this model has a wide range of differences between lower and upper range. The expected result may vary. Also, Temp is not the sole factor in determining the Ozone. There are other factors like Wind and Solar.R have also strong correlation with Ozone. Those need to be considered as well.

#install.packages("GGally")  
#install.packages("car")  
#install.packages("lmtest")  
# message = FALSE in final knit.  
library(tidyverse)  
library(car)  
library(GGally)  
library(lmtest)

#### Task 1

air = airquality  
summary(air)

## 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   
##

str(air)

## '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 ...

There are total 153 observations and 6 variables in the dataset. Five variables are integer and one is number (with decimal). Ozone and Solar.R variables have 37 and 7 null observations. Ozone is likely to be response (Y) variable.

#### Task 2

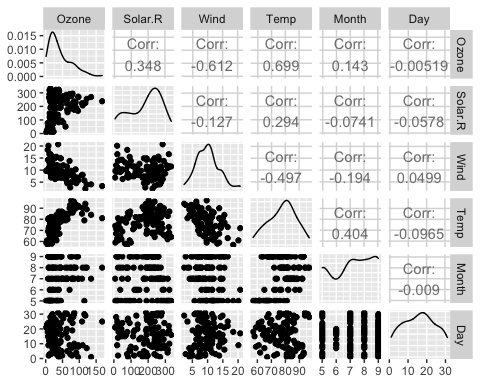
air2 <- air %>% drop\_na()  
str(air2)

## '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 ...

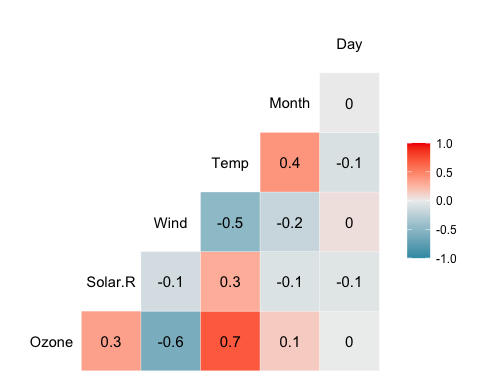
111 rows and 6 colums are there in air2. Using drop\_na() we have removed 42 rows.

#### Task 3

ggpairs(air2)



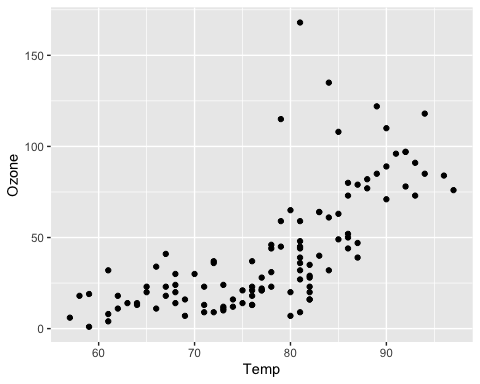
ggcorr(air2,label = TRUE)



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

#### Task 4

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



Temp and Ozone have a strong positive correlationship (~0.7). With Temp increased, Ozone also getting increased. There are few outliers though in the relationship.

#### Task 5.a

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

p-value < 0.05, slope is positive 2.4 and R-squared value is 0.48. One unit of Temp change will result it 2.4 unit of Ozone change. All three suggest that Temp is a significant predictor of Ozone.

#### Task 5.b

confint(model1)

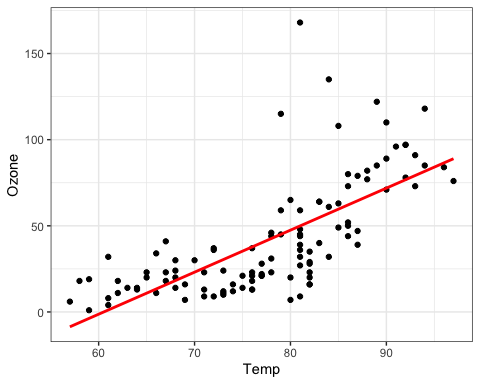
## 2.5 % 97.5 %  
## (Intercept) -184.818372 -110.473773  
## Temp 1.964787 2.913433

Task 5.b: 95% Confidence Intervals will be between -184.818 and -110.473. Slope coefficient likely to fall between 1.964 and 2.913.

#### Task 6

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

## `geom\_smooth()` using formula 'y ~ x'



#### Task 7

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

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

When Temp = 80, then predicted Ozone is 47.48 with range varying from -0.151 (lower range) to 95.116 (upper range).

#### Task 8

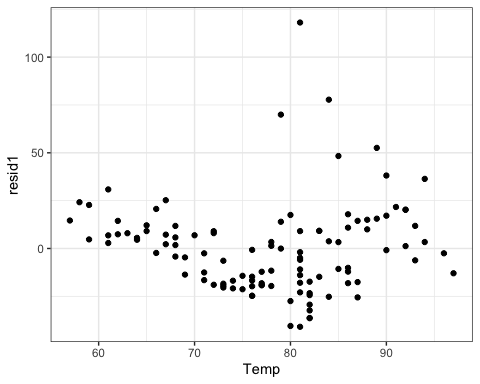
**Assumption 1:** The predictor and response variable have a linear relationship  
It seems reasonable to say that there is something of a linear relationship between these two variables (Temp and Ozone).  
**Assumption 2:** Model errors (residuals) are independent  
Let’s use the Durbin-Watson Test to examine independence of residuals. The dwtest function is from the lmtest package.

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  
Let’s examine a plot of residuals.

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

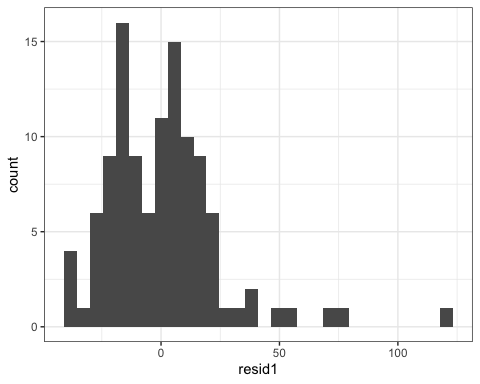


Seems there are not many changes in the variances of residuals. They represent constant variance.

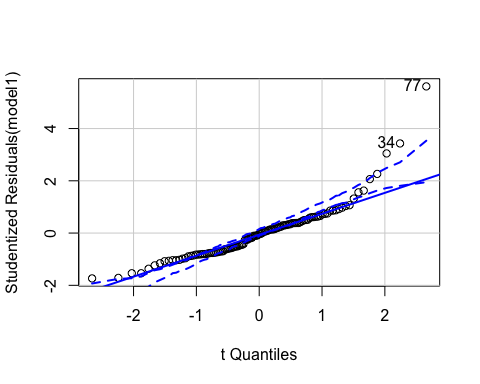
**Assumption 4:** Model residuals are Normally-distributed  
Let’s examine a histogram of the residuals.

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

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



qqPlot(model1)



## [1] 34 77

The residuals histogram is normal. An alternative to the histogram is the Normal Probability Plot. The black points are aligned with the blue line from the bottom left to the upper right.

#### Task 9

This model can be used to predict the Ozone volume with Temp change. Y Intercept is -147.6 and one unit of Temp change will result it 2.4 unit of Ozone change.  
Talking on cautions or concerns, this model has a wide range of differences between lower and upper range. The expected result may vary. Also, Temp is not the sole factor in determining the Ozone. There are other factors like Wind and Solar.R have also strong correlation with Ozone. Those need to be considered as well.