# Simple Linear Regression and Correlation Assignment

## BA502 Predictive Analytics

### Susan Wiggins

Library files needed for assignment, tidyverse,GGally.

library(tidyverse)

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

## v ggplot2 3.1.1 v purrr 0.3.2  
## v tibble 2.1.1 v dplyr 0.8.1  
## v tidyr 0.8.3 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)

##   
## Attaching package: 'GGally'

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

Reading in the airquality data set.

air = airquality  
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 ...

a.Describe the dataset. The dataset consists of variable data for Ozone, Solar.R, Wind, Temperature, Month, and Day.The dataset consists of 153 observations and 6 variables. The data is numeric for all variables. The measurement of data is from the range of May 1, 1973 to September 30, 1973.The data was obtained from the New York State Department of Conservation and the National Weather Service.

Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island

Solar.R: Solar radiation in Langleys in the frequency band 4000-7700 Angstroms from 0800 to 1200 hours at Central Park

Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport

Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

WorkCited:<http://rpubs.com/Nitika/linearRegression_Airquality>

b.How many variables and observations are there?

The dataset consists of 153 observations and 6 variables.

c.Is there any missing data?

Yes, missing data is indicated by NA’s for Ozone and Solar.R

d.Which variable is likely to be the response (Y) variable?

Solar.R is likely to be the response (Y) variable.

The methods for dealing with missing data:

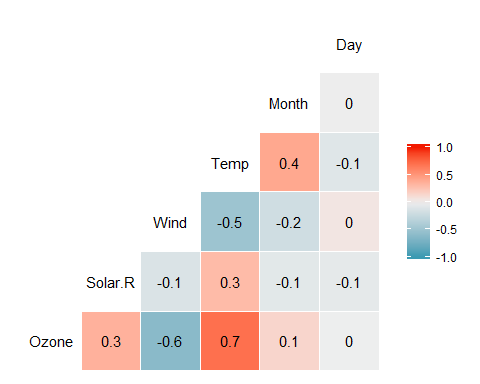
1. Delete the rows with missing data
2. Delete the columns with missing data.
3. Impute(i.e.,estimate or guess) values to replace the missing values.

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

How many rows and columns remain in this new (air2) data frame?

The air2 data frame consists of 111 observations and 6 variables.

ggcorr(air2, label = TRUE)



corr = round(cor(air2),2)  
corr

## Ozone Solar.R Wind Temp Month Day  
## Ozone 1.00 0.35 -0.61 0.70 0.14 -0.01  
## Solar.R 0.35 1.00 -0.13 0.29 -0.07 -0.06  
## Wind -0.61 -0.13 1.00 -0.50 -0.19 0.05  
## Temp 0.70 0.29 -0.50 1.00 0.40 -0.10  
## Month 0.14 -0.07 -0.19 0.40 1.00 -0.01  
## Day -0.01 -0.06 0.05 -0.10 -0.01 1.00

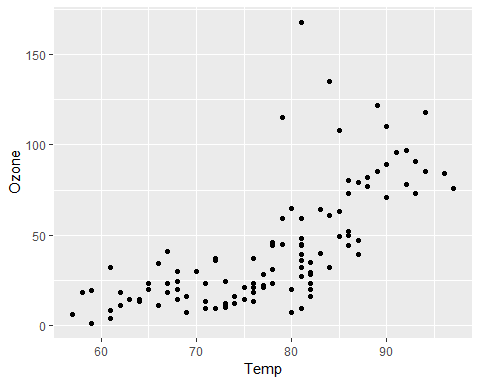
1. Which variable is most strongly correlated with the Ozone variable?

The variable which is most strongly correlated to “Ozone” is Temperature.

1. Which variable is least strongly correlated with the Ozone variable?

The variable least strongly correlated with “Ozone” is Day.

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



Describe the relationship between Temp and Ozone.

As Temperture increase at Fahrenheit(80) there is a gradual increase in Ozone between 50 and 100 parts per billion. Between the Fahrenheit temperature of 60 to 80 there is no increase due to Ozone the data remains on a continuous flow around 25 to 30 parts per billion.

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

confint(model1)

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

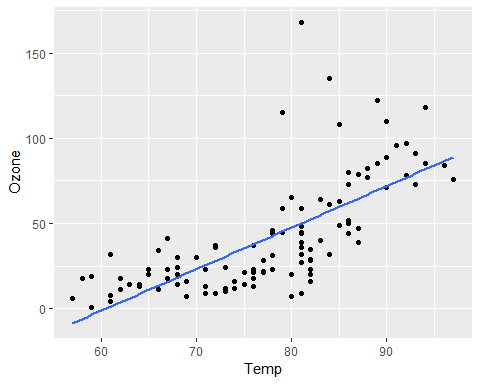
1. Discuss the quality of this model (mention the R square value and significance of the predictor variable).

R-squared value is OK at .488, the value is descently close to 1. The p-value is good less than .05 significantly.

1. Use the code confÃ¯nt(model1) to generate 95% conÃ¯fidence intervals for the coeffÃ¯cients. In what range does the slope coeffÃ¯cient likely fall?

The slope coefficient which is the intercept falls between the lower of -184.818372 and upper of -110.473773.

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



Develop a Prediction for Ozone when Temp is 80. The prediction for ozone when temperature is 80 is 47.48272.

predict(model1, data.frame("Temp" = 80))

## 1   
## 47.48272

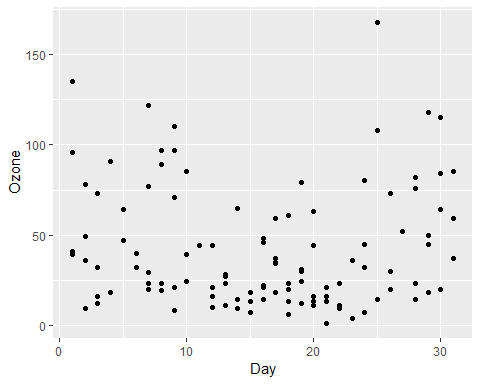
confint(model1)

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

Describe the relationship between Day and Ozone.

Based on the scatter plot there does not seem to be any linear relationship between Day and Ozone. The data consists in a varied dispersion of data throughout the graphical parameters.

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



model2 = lm(Ozone ~ Day, air2)  
summary(model2)

##   
## 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

confint(model2)

## 2.5 % 97.5 %  
## (Intercept) 29.248109 55.5826192  
## Day -0.745321 0.7056539

1. Discuss the quality of this model (mention the R square value and significance of the predictor variable).

The p value indicates very low significance due to a value of .957, a p value needs to be less than .05. The R-Squared value is at .00002693 also very low which does not show relevance of a good regression analysis.

1. Use the code confÃ¯nt(model1) to generate 95% conÃ¯fidence intervals for the coeffÃ¯cients. In what range does the slope coeffÃ¯cient likely fall?

The slope coefficient falls between 29.248109 lower limit and the upper limit for slope is 55.5826192.

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

