# Simple Linear Regression and Correlation

## Module 2 Assignment 1

### Brian Fortier

Libraries:

#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 0.7.7  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.1.1 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

Create Data Frame:

air = airquality

This data set contains 153 observations of 6 variables. It has variables Ozone, Solar.R, Wind, Temp, Month and Day. Some of the data is missing mainly for the Ozone variable and a couple entries for the Solar variable. Temperature is likely to be the response variable because it will seemingly change based on the other entries.

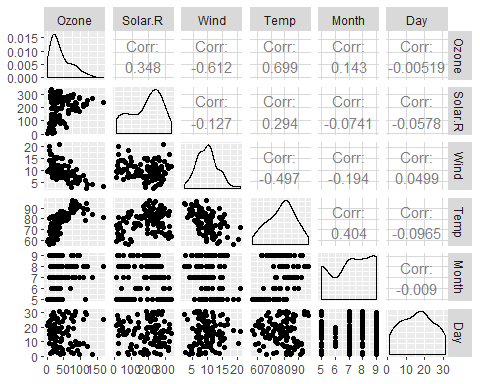
Deleting Rows:

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

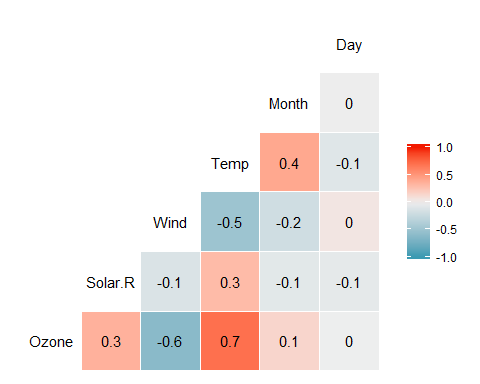
After this command, we are left with 111 observations of 6 variables.

Visualization:

ggpairs(air2)



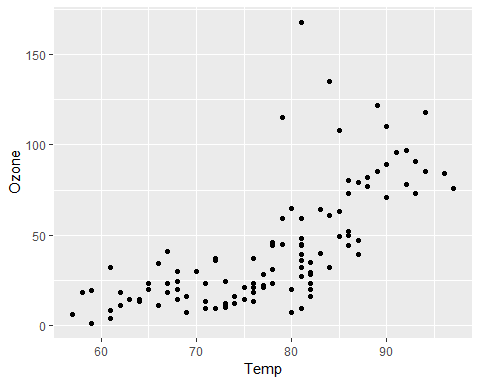
ggcorr(air2, label = TRUE)



Temperature is most strongly correlated with the Ozone variable and Wind is the least strongly correlated.

Plot (Temp vs. Ozone):

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



Using this scatterplot graph, we see a slight positive correlation showing the ozone increases as Temp increases. However we see basically a neutral correlation until about when Temp = 80.

Linear Regression Model 1:

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)

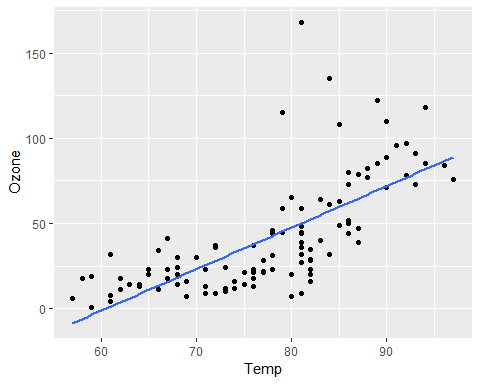
## 2.5 % 97.5 %  
## (Intercept) -184.818372 -110.473773  
## Temp 1.964787 2.913433

Based on this model, we see our slope of 2.4391 which shows us a positve increase in Ozone as we increase Temp. We have a small p value (<0.05) which shows a significant relationship between the two variables. And our multiple R-squared value at 0.488 adds to our conclusion that our model is a good model.

Looking at our confidence interval, our slope coefficient would likely fall below the 97.5% marker. Our coefficient being 2.439 and the marker at 97.5% being 2.91.

Re-Plot with Regression Line:

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



We can see adding in the regression line to our graph shows how it follows the points and relationship between Temp and Ozone.

Ozone Prediction with Temp = 80:

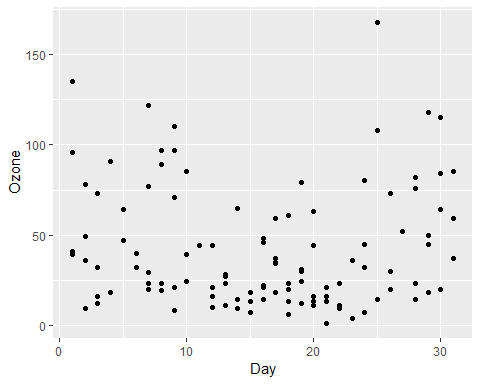
testdata = data.frame(Temp = 80)  
predict(model1, newdata = testdata, interval = "predict")

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

This prediction gives us an Ozone value of 47.48 when we have a Temp = 80. Our prediction inverval runs between (-0.151, 95.116) which is a very wide interval for our data.

Plot (Day vs. Ozone):

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



We do not see much of a relationship between Day and Ozone. Describing it as a “U” shape would be the best description of this scatterplot graph.

Linear Regression Model 2:

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)

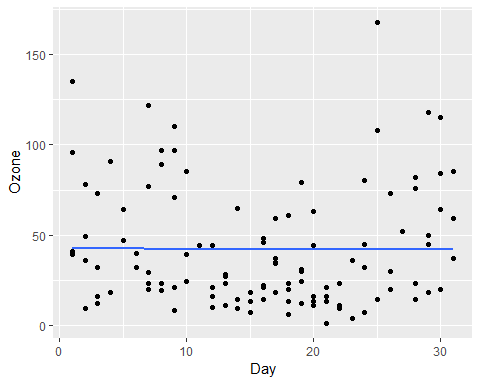
## 2.5 % 97.5 %  
## (Intercept) 29.248109 55.5826192  
## Day -0.745321 0.7056539

After reviewing our regession model information, we see we have basically a 0 slope with y-intercept of 42. Our p-value is almost at 1 which shows a very insignificant relationship between our variables and again a 0 for our multiple R-squared value.  
This is not a good model for these variables.

For this model, we would see our slope coefficient to fall somewhere around the 50% marker giving us little confidence in our model with the 97.5% marker being at 0.71 and our slope equaling -0.02.

Re-Plot with Regression Line:

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



Viwing our scatterplot with the regression line justifies our conclusion of a poor model. The regression line is basically a straight line through our data points with no correlation showing.