# Correlation and Simple Linear Regression

## Dan Sanchez

### BAN 502

### Dr. Stephen Hill

### May 25th, 2020

library(tidyverse)  
library(GGally)  
library(car)  
library(lmtest)

### Task 1

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

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   
##

This data set is comprised of 6 variables and 153 observations. There are missing values in the Ozone and Solar.R columns. Ozone is likely to be the response variable.

### Task 2

Use the Tidyverse drop\_na function to remove any row with 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?

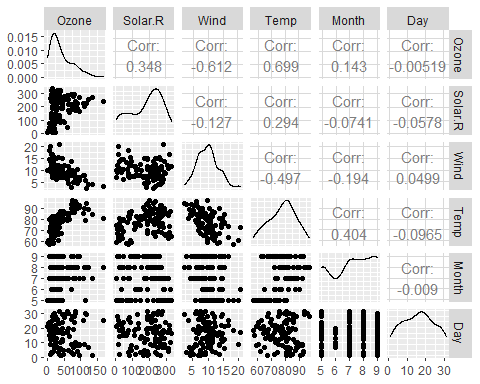
air2<- air%>%drop\_na(Solar.R,Ozone)

The air2 dataframe has 111 rows and 6 columns.

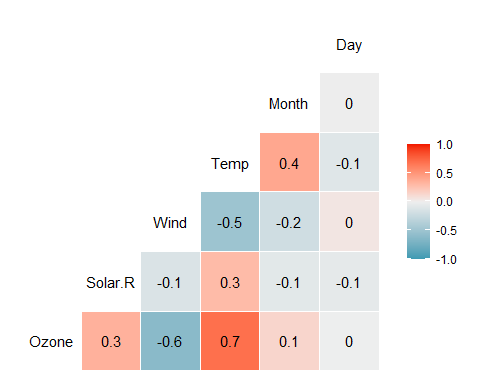
### Task 3

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.

ggpairs(`air2`)



ggcorr(air2,method = c("pairwise", "pearson"),cor\_matrix = NULL,label=TRUE)

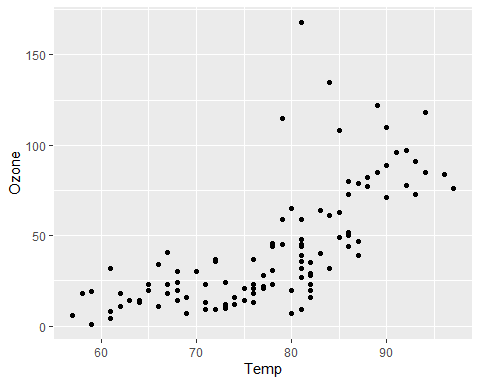


“Temp” is most strongly correlated with the “Ozone” variable. “Day” is least strongly correlated with the “Ozone” variable.

### 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”.

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



There is a postive linear relationship between “Temp” and “Ozone.” As Temp increases, Ozone tends to also increase. Between 80 and 90 degrees in Temp, there are a handulf of outliers that add some complexity to their otherwise linear relationship.

### Task 5

Create a linear regression model (called model1) using “Temp” to predict “Ozone”.

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

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

The p-value of Temp being less than .05 in this model indicates significance when predicting Ozone. The R square value of .488 is not bad. It would be helpful have a larger sample in this data set, perhaps spanning the entire year.

1. Use the code “confint(model1)” to generate 95% confidence intervals for the coefficients. In what range does the slope coefficient likely fall?

confint(model1)

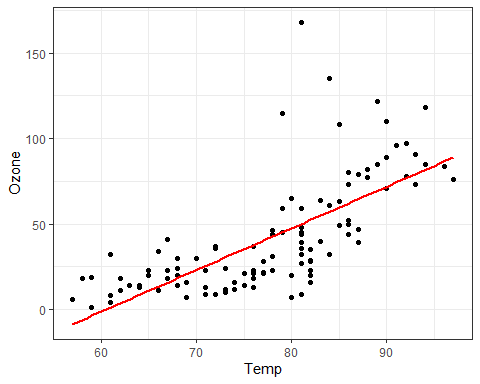
## 2.5 % 97.5 %  
## (Intercept) -184.818372 -110.473773  
## Temp 1.964787 2.913433

The slope coefficient likely falls between 1.964787 and 2.912333.

### 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)+  
 theme\_bw()



### Task 7

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

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

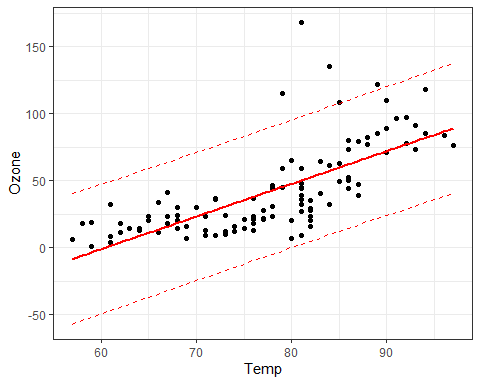
## fit lwr upr  
## 1 47.48272 -0.1510188 95.11646

### Task 8

Perform appropriate model diagnostics to verify whether or not the model appears to meet the four linear regression model assumptions.

temp\_var = predict(model1,interval = "prediction")  
new\_df = cbind(air2,temp\_var)

ggplot(new\_df, aes(x=`Temp`, y=`Ozone`))+  
 geom\_point()+  
 geom\_smooth(method="lm",color="red", se=FALSE)+  
 geom\_line(aes(y=lwr), color="red", linetype = "dashed")+  
 geom\_line(aes(y=upr), color="red", linetype = "dashed")+  
 theme\_bw()



*Assumption 1: The predictor and response variable have a linear relationship.*

The visualization above shows a liner relationship between Temp and Ozone.

*Assumption 2: Model errors (residuals) are independent*

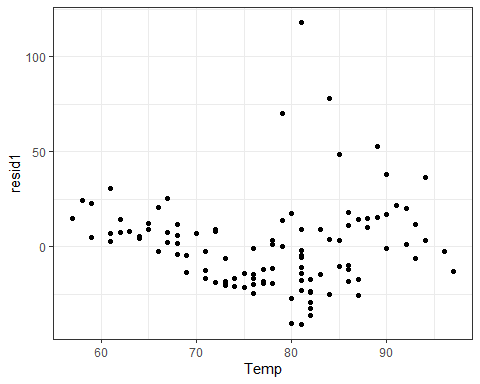
dwtest(model1)

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

Since the p-value from this Durbin-Watson test is greater than .05, it is likely that the model errors are independent.

*Assumption 3: Model residuals exhibit constant variance.*

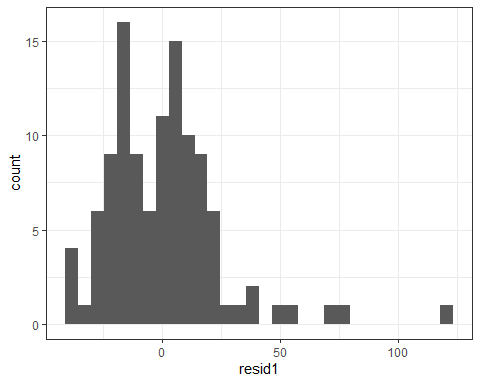
air2 = air2 %>% mutate(resid1 = model1$residuals)  
ggplot(air2,aes(x=Temp,y=resid1))+geom\_point()+theme\_bw()



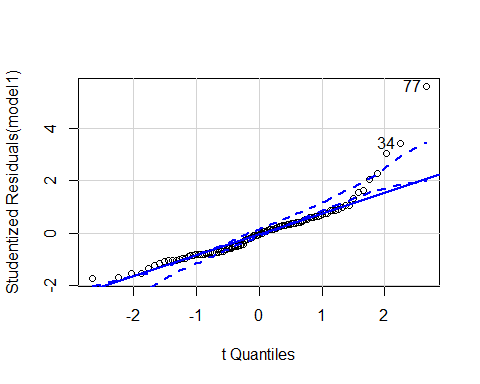
These models exhibit relatively constant variance, aside from the outliers between 79 and 85 degrees in Temp.

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

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



qqPlot(model1)



## [1] 34 77

The plots above show that the residuals are somewhat normally-distributed.

### 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 from Task 5 could be used to predict Ozone as temperatures rise and fall, perhaps in response to seasonal changes or climate change. This could help determine air quality in the geographic area, which could be used to recommend limits on outdoor activities and the promotion of carpooling and public transportation usage.

I would hesitate to use this model based upon the limited amount of data available from a single 5-month period. Another caution is that the model underestimates the Ozone in the 60s, overestimates Ozone in the 70s, and underestimates Ozone again for 87 degrees and higher.