# Simple Linear Regression and Correlation

## Miguel Fernandez

### BAN-502

# Import libraries  
library(tidyverse)  
library(knitr)  
library(GGally)  
library(lmtest)  
  
# Set global theme  
theme\_set(theme\_bw())

#### Task 1

# Read in data set  
air = airquality

The data set air contains daily air quality measures in New York in 1973. Data was collected from May 1 to September 30. In the five month period, 153 observations were made across six variables. These include Ozone, Solar.R, Wind, Temp, Month and Day. The summary below reveals 37 missing values from the Ozone variable and 7 missing values from the Solar.R variable. For the purposes of this exercise, I believe the response variable to be Ozone This variable is a measure of the mean ozone in parts per billion.

# Use str() to get number of rows and columns  
# str(air)  
# Generate summary of the data frame  
kable(summary(air))

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Ozone | Solar.R | Wind | Temp | Month | Day |
|  | Min. : 1.00 | Min. : 7.0 | Min. : 1.700 | Min. :56.00 | Min. :5.000 | Min. : 1.0 |
|  | 1st Qu.: 18.00 | 1st Qu.:115.8 | 1st Qu.: 7.400 | 1st Qu.:72.00 | 1st Qu.:6.000 | 1st Qu.: 8.0 |
|  | Median : 31.50 | Median :205.0 | Median : 9.700 | Median :79.00 | Median :7.000 | Median :16.0 |
|  | Mean : 42.13 | Mean :185.9 | Mean : 9.958 | Mean :77.88 | Mean :6.993 | Mean :15.8 |
|  | 3rd Qu.: 63.25 | 3rd Qu.:258.8 | 3rd Qu.:11.500 | 3rd Qu.:85.00 | 3rd Qu.:8.000 | 3rd Qu.:23.0 |
|  | Max. :168.00 | Max. :334.0 | Max. :20.700 | Max. :97.00 | Max. :9.000 | Max. :31.0 |
|  | NA’s :37 | NA’s :7 | NA | NA | NA | NA |

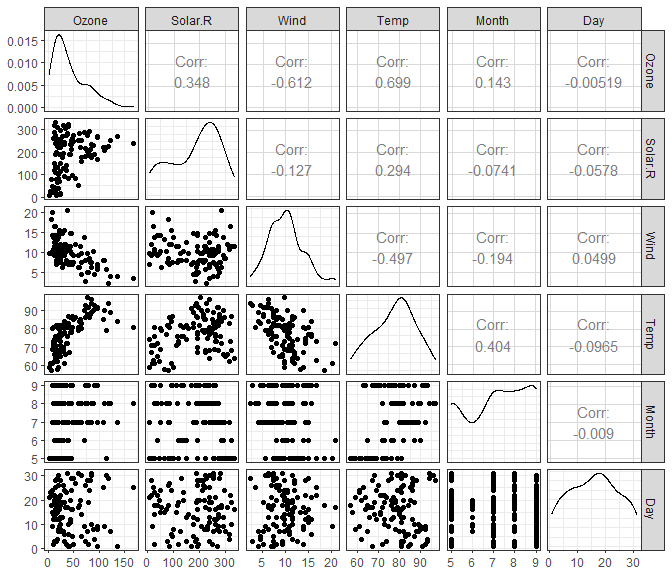
#### Task 2

# Remove null values from data set  
air2 <- air %>%  
 drop\_na()  
  
# Use str() to get number of rows and columns  
# str(air2)

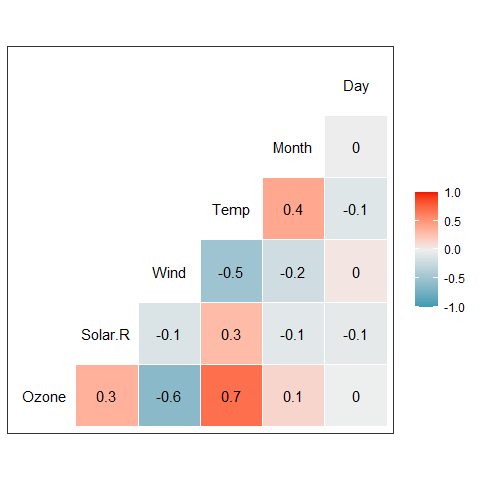
After removing the missing values, there are now 111 observations in the data set as well as the same six variables.

#### Task 3

# Create matrix of plots  
ggpairs(air2)



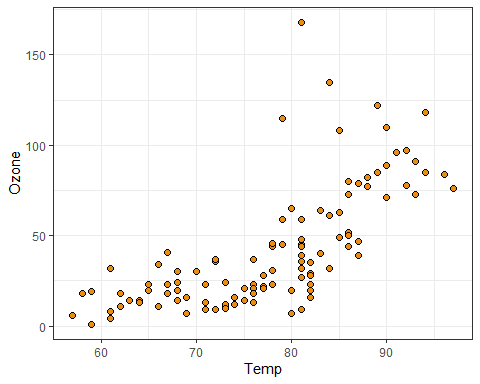
ggcorr(air2, label = "TRUE")



In the correlation matrix, Temp is the most strongly correlated to Ozone while Wind is the least strongly correlated.

#### Task 4

# Plot Temp vs Ozone  
ggplot(air2, aes(x = Temp, y = Ozone)) +  
 geom\_point(shape = 21,  
 size = 2,  
 fill = '#eb8d13')



The plot Temp versus Ozone visualizes the positive correlation from the matrix plot above. As Temp increases, so too, does Ozone. A non-linear method might describe the data more accurately but further analysis is needed to confirm that. There are also outliers present in the data set for Ozone. For example, the point representing 150 parts per billion is unusually high. This could be an accurate recording or another possible explanation is that an extra zero was input when entering the data. The reading might have been 15, which would be closer to other recordings given the temperature.

#### Task 5

# Create linear model  
model1 <- lm(Ozone ~ Temp, air2)  
# Generate model statistics  
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

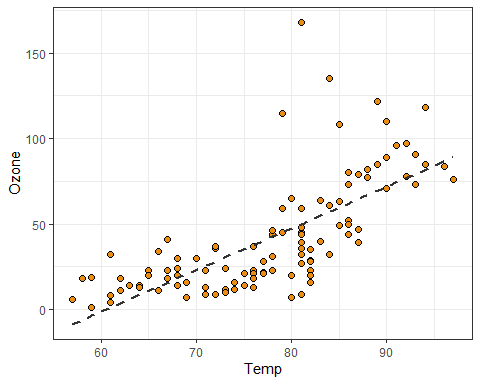
# Generate confidence interval  
confint(model1)

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

The model statistics generated above provide useful measures for interpreting the strength and validity of the model. Temp has a coefficient of 2.44, meaning that for an increase in one degree Fahrenheit, the mean Ozone in ppb increases by 2.44. The p-value for Temp is near-zero, indicating that it is a significant predictor in the model. A confidence interval for Temp identifies a range that the true slope likely falls. We are 95 percent confident that the slope will fall between 1.96 and 2.91.

#### Task 6

# Plot Temp vs Ozone with regression line  
ggplot(air2, aes(x = Temp, y = Ozone)) +  
 geom\_point(shape = 21,  
 size = 2,  
 fill = '#eb8d13') +  
 geom\_smooth(method = 'lm',  
 se = FALSE,  
 color = 'gray20',  
 linetype = 'dashed')



#### Task 7

# Generate prediction using model  
test.data <- data.frame(Temp = 80)  
kable(predict(model1, newdata = test.data, interval = "predict"))

|  |  |  |
| --- | --- | --- |
| fit | lwr | upr |
| 47.48272 | -0.1510188 | 95.11646 |

The model predicted 47.48 ppb for Ozone when Temp is 80°F. However, the prediction interval is large and spans a range of nearly 100 parts per billion.

#### Task 8

**An appropriate model must meet the following criteria:**  
1. *The predictor and response variable must have a linear relationship.*  
By examining the plot from Task 4 and Task 6, we can confirm that this criteria has been met.

1. *Residuals are independent*  
   We can use the Durbin-Watson statistic to test for independence.

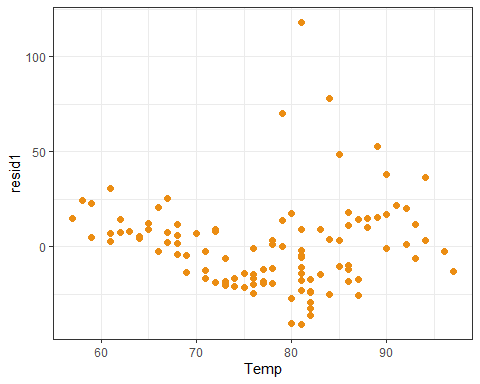
dwtest(model1)

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

The test returns a p-value greater than 0.05 which indicates that the residuals are independent.

1. *Residuals exhibit constant variance*

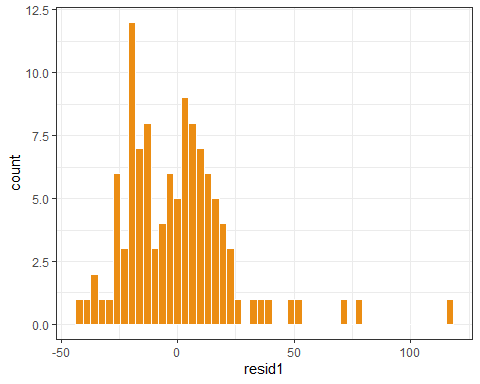
# Add residuals from model to data frame  
air2 <- air2 %>%  
 mutate(resid1 = model1$residuals)  
  
# Plot residuals to confirm constant variance  
ggplot(air2, aes(x = Temp, y = resid1)) +  
 geom\_point(size = 2, color = '#eb8d13')



Overall, the plot confirms constant variance. There is no distinct cone or funnel shape in the plot. They are a few outliers present in the data.

1. *Residuals are normally distributed.*

# Create histogram to confirm normality  
ggplot(air2, aes(x = resid1)) +  
 geom\_histogram(bins = 50,  
 color = 'white',  
 fill = '#eb8d13')



The residuals appear to be somewhat normally distributed. There are outliers present causing a positive skew in the data set.

#### Task 9

The model that was constructed above could be used to predict measures of ozone in the atmosphere and used to issue air quality warnings. The model has an R-squared value of 0.49 meaning that approximately 50 percent of ozone measures can be explained by temperature. While this is a somewhat strong model, adding more predictors could help increase the R-squared value giving analysts and meteorologists more confidence in the model. If air quality warnings attempt to restrict movement and activities, a stronger model is needed. This model was also built using data from New York City only. According to the [EPA](https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics), ozone is formed when sunlight reacts to particles of nitrogen oxides (NOx) emitted by cars, factories and other sources. New York City is a densely populated island and a model that uses data from New York City may not be appropriate to predict measures of ozone in a smaller city.