Correlation & Simple Linear Regression

# MIS 502

### Danielle Harris

library(tidyverse)

## ── Attaching packages ────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.7  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(GGally)

##   
## Attaching package: 'GGally'

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

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 ...

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   
##

ggpairs(air)

## Warning: Removed 37 rows containing non-finite values (stat\_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =  
## "pearson", : Removed 42 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =  
## "pearson", : Removed 37 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =  
## "pearson", : Removed 37 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =  
## "pearson", : Removed 37 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =  
## "pearson", : Removed 37 rows containing missing values

## Warning: Removed 42 rows containing missing values (geom\_point).

## Warning: Removed 7 rows containing non-finite values (stat\_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =  
## "pearson", : Removed 7 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =  
## "pearson", : Removed 7 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =  
## "pearson", : Removed 7 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =  
## "pearson", : Removed 7 rows containing missing values

## Warning: Removed 37 rows containing missing values (geom\_point).

## Warning: Removed 7 rows containing missing values (geom\_point).

## Warning: Removed 37 rows containing missing values (geom\_point).

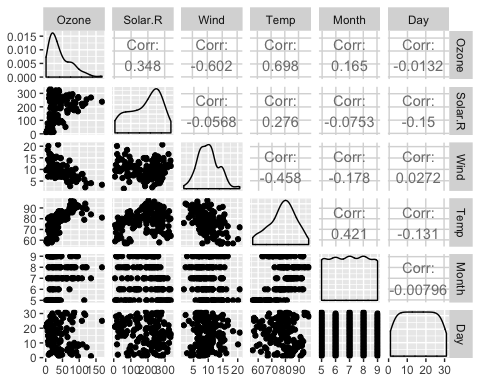
## Warning: Removed 7 rows containing missing values (geom\_point).

## Warning: Removed 37 rows containing missing values (geom\_point).

## Warning: Removed 7 rows containing missing values (geom\_point).

## Warning: Removed 37 rows containing missing values (geom\_point).

## Warning: Removed 7 rows containing missing values (geom\_point).



# a) The dataset “air” displays air quality between the months May- October.

# b) There are 153 observations and 6 variables.

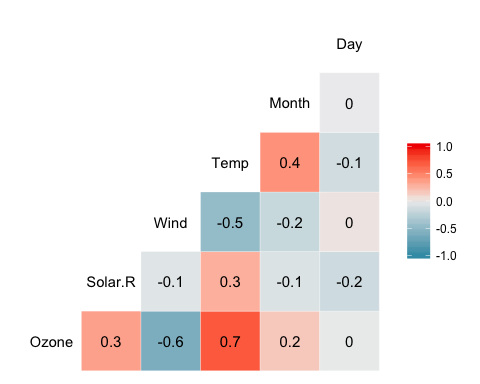
# c) There are 37 missing values.

# d) Temperature is most likely to be the respose varaible (Y).

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

# There are now 111 rows and 6 columns in the new data frame “air2”.

ggcorr(air, label=TRUE)



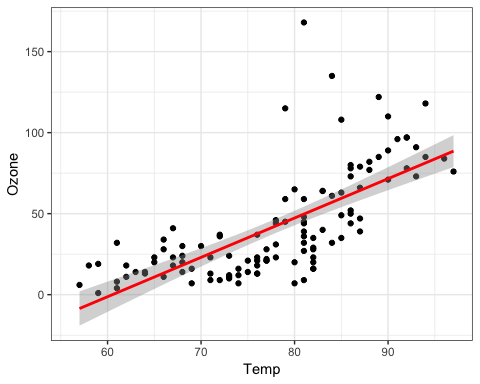
# a) The variable “Temperature”" is most strongly correlated with the variable “Ozone”.

# b) The varaiable “Day” is least strongly correlated with the variable “Ozone”.

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

## Warning: Removed 37 rows containing non-finite values (stat\_smooth).

## Warning: Removed 37 rows containing missing values (geom\_point).



# According to the graph, Temperature tends to increase as Ozone increases.

model1= lm(Ozone ~ Temp, air)  
summary(model1)

##   
## Call:  
## lm(formula = Ozone ~ Temp, data = air)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.729 -17.409 -0.587 11.306 118.271   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -146.9955 18.2872 -8.038 9.37e-13 \*\*\*  
## Temp 2.4287 0.2331 10.418 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23.71 on 114 degrees of freedom  
## (37 observations deleted due to missingness)  
## Multiple R-squared: 0.4877, Adjusted R-squared: 0.4832   
## F-statistic: 108.5 on 1 and 114 DF, p-value: < 2.2e-16

# a) The quality of this model is strong with a p-value of <2e-16 (<0.05), and an R Square value of 0.4832.

confint(model1)

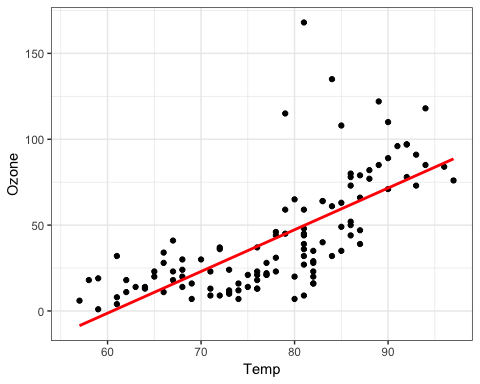
## 2.5 % 97.5 %  
## (Intercept) -183.222241 -110.768741  
## Temp 1.966871 2.890536

# b) The slope coefficient likely falls between 1.9 and 2.9.

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

## Warning: Removed 37 rows containing non-finite values (stat\_smooth).

## Warning: Removed 37 rows containing missing values (geom\_point).



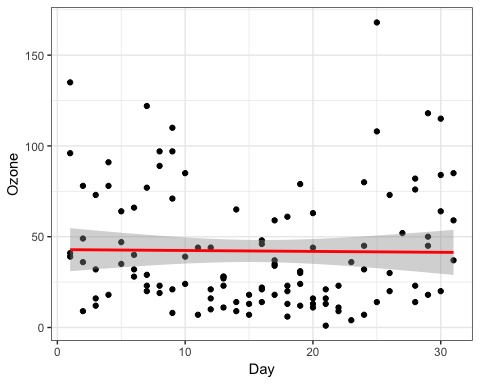
testdata= data.frame(Temp=80)  
predict(model1, newdata= testdata, interval= "predict")

## fit lwr upr  
## 1 47.30077 0.110647 94.4909

ggplot(air, aes(x= Day, y= Ozone))+  
 geom\_point()+  
 geom\_smooth(method="lm", color="red")+  
 theme\_bw()

## Warning: Removed 37 rows containing non-finite values (stat\_smooth).

## Warning: Removed 37 rows containing missing values (geom\_point).



# The graph displays a fairly neutral correlation between the variables “Day” and “Ozone”, meaning that one variable is not likely to be a predictor of the other variable.

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

##   
## Call:  
## lm(formula = Ozone ~ Day, data = air)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.86 -24.29 -10.86 21.16 126.34   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.90387 6.28832 6.823 4.53e-10 \*\*\*  
## Day -0.04986 0.35306 -0.141 0.888   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33.13 on 114 degrees of freedom  
## (37 observations deleted due to missingness)  
## Multiple R-squared: 0.0001749, Adjusted R-squared: -0.008595   
## F-statistic: 0.01994 on 1 and 114 DF, p-value: 0.8879

# a) There is little relationship between the variables “Day” and “Ozone” as the p-value is greater than 0.05 and the R Squared value is significantly low.

confint(model2)

## 2.5 % 97.5 %  
## (Intercept) 30.446753 55.3609902  
## Day -0.749275 0.6495535

# b) The slope coefficient likely falls between -0.7 and 0.7.

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

## Warning: Removed 37 rows containing non-finite values (stat\_smooth).

## Warning: Removed 37 rows containing missing values (geom\_point).

