Correlation and Simple Linear Regression Assignment

2/2/2020

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages ----------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 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(ggplot2)

library(GGally)

## Warning: package 'GGally' was built under R version 3.6.2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

##   
## Attaching package: 'GGally'

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

library(ggplot2)

air = airquality

#Task 1: There are 6 variables and 153 observations. There is some missing data. Temperature is most likely to be the Y variable.

air2 = air %>% drop\_na()  
str(air2)

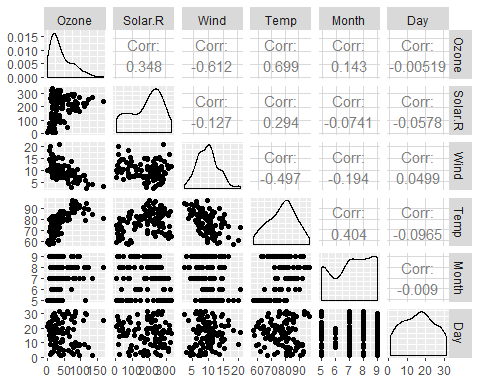
## 'data.frame': 111 obs. of 6 variables:  
## $ Ozone : int 41 36 12 18 23 19 8 16 11 14 ...  
## $ Solar.R: int 190 118 149 313 299 99 19 256 290 274 ...  
## $ Wind : num 7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...  
## $ Temp : int 67 72 74 62 65 59 61 69 66 68 ...  
## $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Day : int 1 2 3 4 7 8 9 12 13 14 ...

#Task 2: There are now 111 observations after removing the missing data.

library(ggcorrplot)

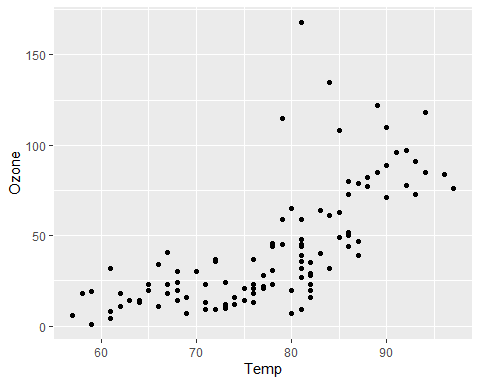
## Warning: package 'ggcorrplot' was built under R version 3.6.2

ggpairs(air2)



#Task 3: The variable “Temperature” is most strongly correlated with the “Ozone” variable, while “Wind” is least strongly correlated with the “Ozone” variable?

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



#Task4: Temperature and Ozone appear to have a high correlation.

model1 = lm(Temp ~ Ozone, air2)  
summary(model1)

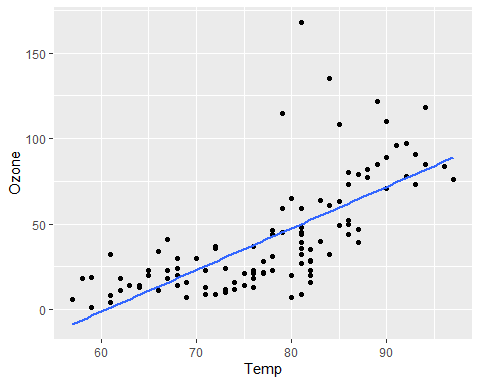
##   
## Call:  
## lm(formula = Temp ~ Ozone, data = air2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.980 -4.775 1.825 4.228 12.425   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.37059 1.05151 65.97 <2e-16 \*\*\*  
## Ozone 0.20006 0.01963 10.19 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.851 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) 67.2865285 71.4546496  
## Ozone 0.1611525 0.2389608

#Task 5: # a. This is a good model because the p-value is less than .05, and the R squared value is almost .5 suggesting significance. # b. The slope coefficient likely falls between 67 and 71.

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



#Task7: Prediction for ozone when temp is 80

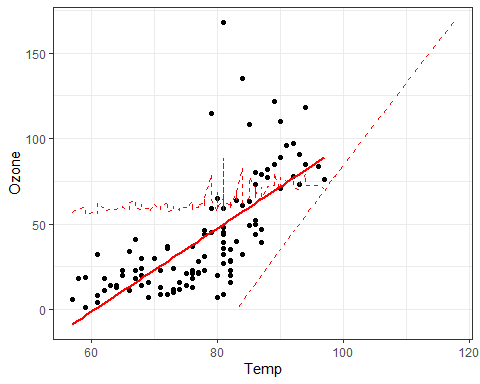
69.37059 + 0.20006 \* 80

## [1] 85.37539

temp\_var = predict(model1, interval = "prediction")

## Warning in predict.lm(model1, interval = "prediction"): predictions on current data refer to \_future\_ responses

new\_df = cbind(air2, temp\_var)  
ggplot(new\_df, aes(x=Temp, y=Ozone))+  
 geom\_point()+  
 geom\_smooth(method = "lm", se = FALSE, color = "red")+  
 geom\_line(aes(y=lwr), color = "red", linetype = "dashed")+  
 geom\_line(aes(x=upr), color = "red", linetype = "dashed")+  
 theme\_bw()



#Task 8: To ensure the linear regression model is valid, we look for: # 1. The predictor and response variable have a linear relationship - TRUE. # 2. Model errors (residuals) are independent - TRUE # 3. Model residuals exhibit constant variance - TRUE # 4. Model residuals are Normally-distributed

#Task9: The model we constructed in Task 5 would be used to identify the lower and upper bounds. I have no concerns against recommending the model for use.