HW 8

Intro to Data Science HW 8

```
# Enter your name here: Chaithra Kopparam Cheluvaiah
```


Attribution statement: (choose only one and delete the rest)

```
# 1. I did this homework by myself, with help from the book and the professor.
```

The chapter on **linear models** ("Lining Up Our Models") introduces **linear predictive modeling** using the tool known as **multiple regression**. The term "multiple regression" has an odd history, dating back to an early scientific observation of a phenomenon called **"regression to the mean."** These days, multiple regression is just an interesting name for using **linear modeling** to assess the **connection between one or more predictor variables and an outcome variable**.

In this exercise, you will **predict Ozone air levels from three predictors**.

A. We will be using the **airquality** data set available in R. Copy it into a dataframe called **air** and use the appropriate functions to **summarize the data**.

```
air <- airquality
help("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  
## 
```

```r
head(air)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

B. In the analysis that follows, **Ozone** will be considered as the **outcome variable**, and **Solar.R**, **Wind**, and **Temp** as the **predictors**. Add a comment to briefly explain the outcome and predictor variables in the dataframe using **?airquality**.

```r
# description of variables:
# Ozone is mean ozone measured in parts per billion
# Sloar.R is solar radiation measured in langleys
# Wind is the average wind speed measured in miles per hour
# Temp is max daily temperature measured in Fahrenheit

# trying to predict how ozone concentration or ozone levels are affected by
# solar radiation,wind speed and the temperature by taking the sample data
# from New York Air Quality Measurements
```

C. Inspect the outcome and predictor variables – are there any missing values? Show the code you used to check for that.

```r
table(is.na(air$Ozone)) # 37 missing values in the outcome variable
```

```
##
## FALSE  TRUE
##   116    37
```

```r
table(is.na(air$Solar.R)) # 7 missing values in the Solar Radition attribute
```

```
##
## FALSE  TRUE
##   146     7
```

```r
table(is.na(air$Wind)) # no missig value
```

```
##
## FALSE
##   153
```

```r
table(is.na(air$Temp)) # no missing value
```

```
##
## FALSE
##   153
```

D. Use the **na__interpolation()** function from the **imputeTS package** (remember this was used in a previous HW) to fill in the missing values in each of the 4 columns. Make sure there are no more missing values using the commands from Step C.

```
library(imputeTS)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
air$Ozone <- na_interpolation(air$Ozone)
air$Solar.R <- na_interpolation(air$Solar.R)

# ensuring there are no NA
table(is.na(air$Ozone))
```

```
##
## FALSE
##   153
```

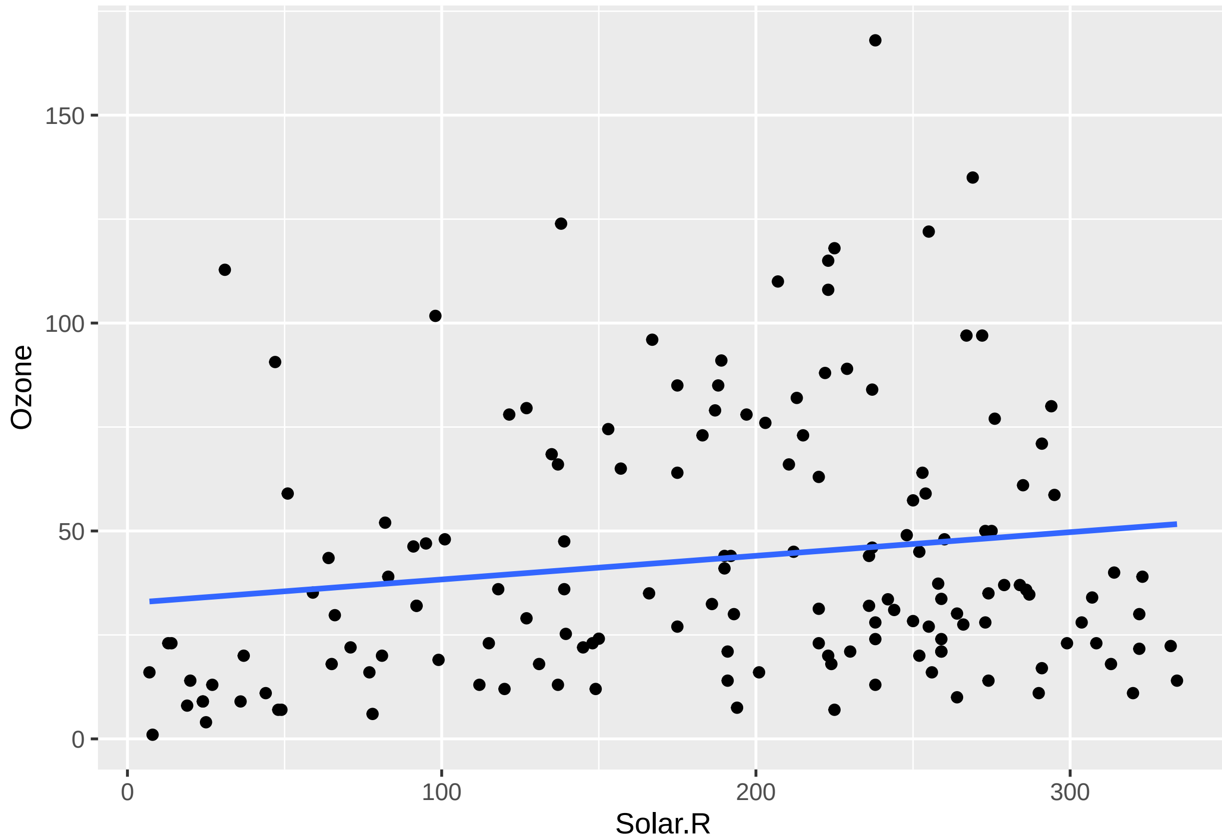
```
table(is.na(air$Solar.R))
```

```
##
## FALSE
##   153
```

E. Create **3 bivariate scatterplots (X-Y) plots** (using ggplot), for each of the predictors with the outcome. **Hint:** In each case, put **Ozone on the Y-axis**, and a **predictor on the X-axis**. Add a comment to each, describing the plot and explaining whether there appears to be a **linear relationship** between the outcome variable and the respective predictor.

```
library(ggplot2)
ggplot(air, aes(x=Solar.R, y=Ozone)) +
geom_point() +
geom_smooth(method="lm", se=FALSE)
```

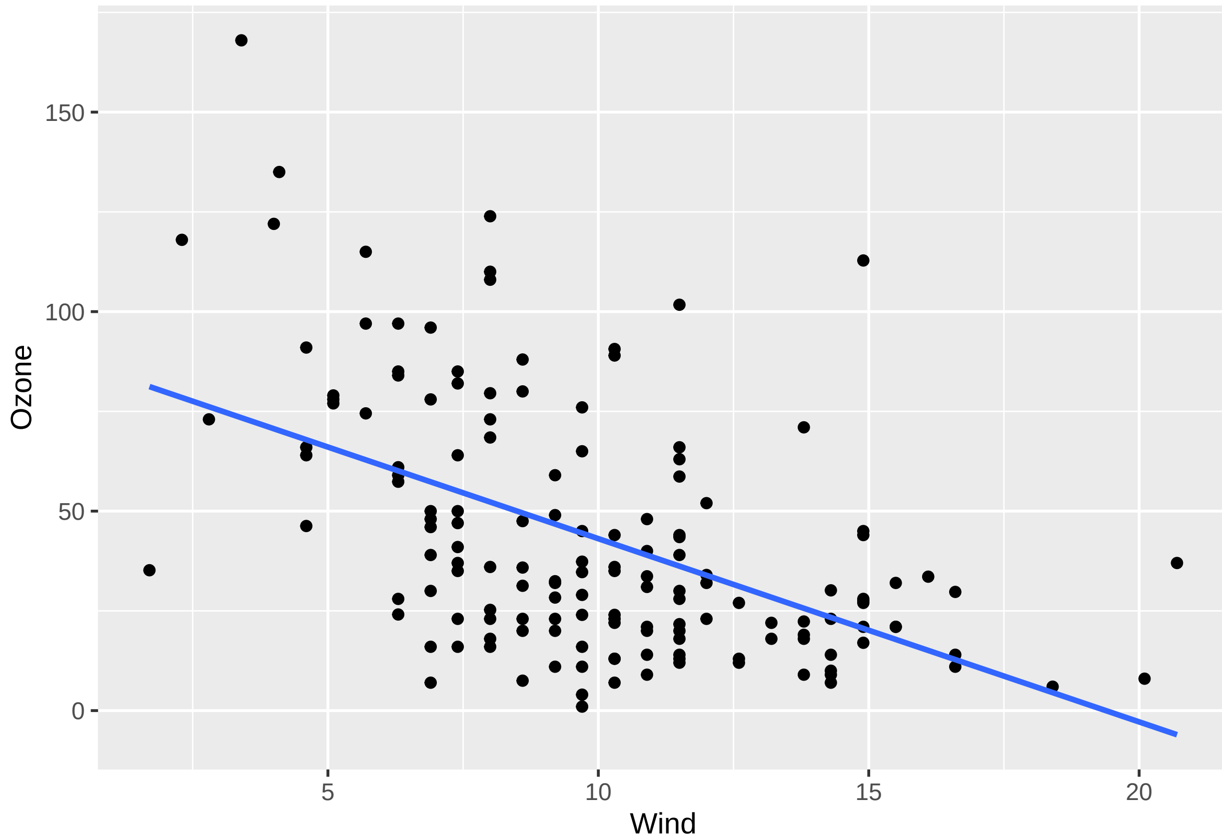
```
## `geom_smooth()` using formula 'y ~ x'
```




```
#from the plot, It appears, there is a very weak positive relation between Ozone and Solar.R
```

```
ggplot(air, aes(x=Wind, y=Ozone)) +
geom_point() +
geom_smooth(method="lm", se=FALSE)
```

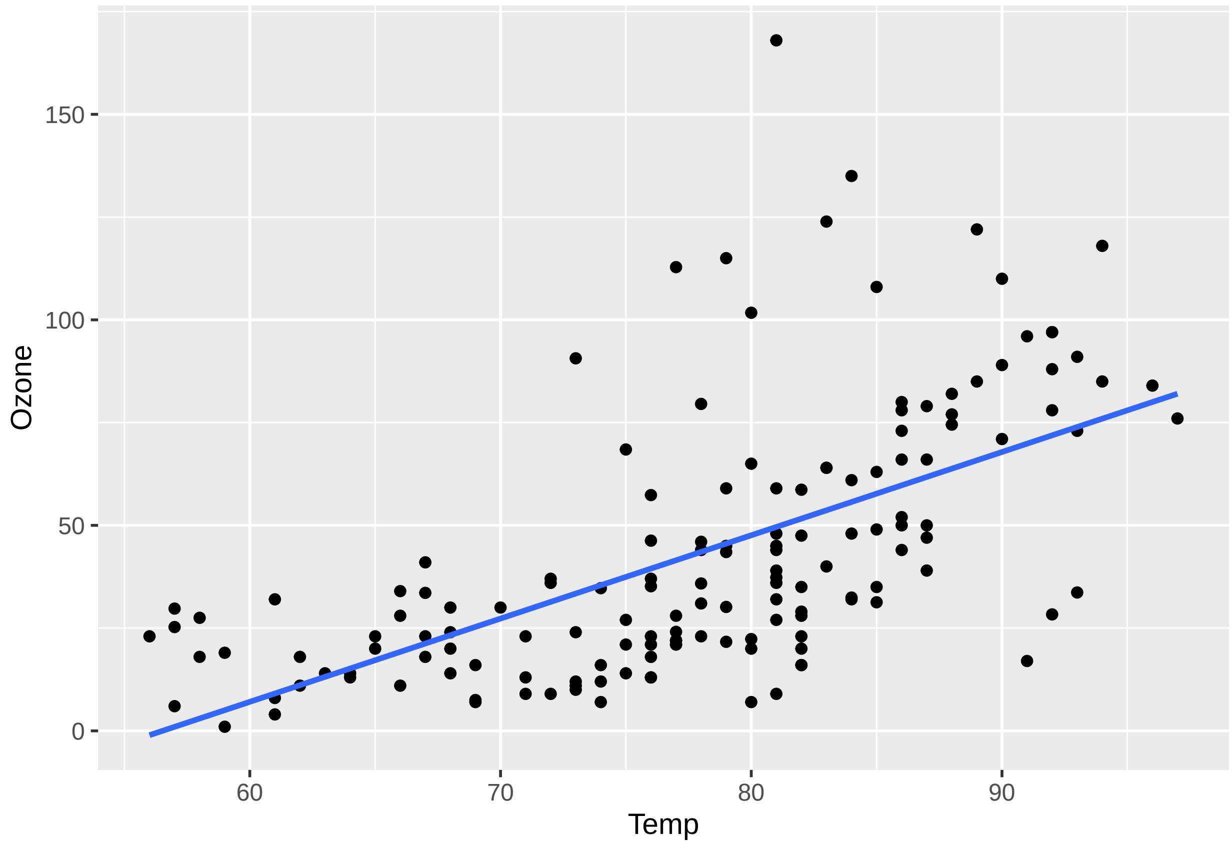
```
## `geom_smooth()` using formula 'y ~ x'
```




```
#from the plot, It appears, there is a strong negative relation between Ozone and Wind
```

```
ggplot(air, aes(x=Temp, y=Ozone)) +
geom_point() +
geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#from the plot, It appears, there is a strong positive relation between Ozone and Temp
```

F. Next, create a **simple regression model** predicting **Ozone based on Wind**, using the **lm( )** command. In a comment, report the **coefficient** (aka **slope** or **beta weight**) of **Wind** in the regression output and, **if it is statistically significant**, **interpret it** with respect to **Ozone**. Report the **adjusted R-squared** of the model and try to explain what it means.

```
windModel <- lm(Ozone~Wind,data=air)
summary(windModel)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.332 -18.332  -4.155  14.163  94.594
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.0205     6.6991  13.288  < 2e-16 ***
## Wind         -4.5925     0.6345  -7.238 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 27.56 on 151 degrees of freedom
## Multiple R-squared:  0.2576, Adjusted R-squared:  0.2527
## F-statistic: 52.39 on 1 and 151 DF,  p-value: 2.148e-11
```

```
# co-efficient of Wind is -4.59:
# for every additional increase in Wind speed, Ozone concentration is predicted to
# decreases by 4.59 ppb

# p-value 2.148e-11 is under 0.05 cutoff: it is statistically significant.
# This shows that Wind and Ozone are related and we should be considering in the
# regression model for predicting Ozone level

# Adjusted R-Square 0.2527: Wind speed accounts for about 25.27% of the Ozone level
```

G. Create a **multiple regression model** predicting **Ozone** based on **Solar.R**, **Wind**, and **Temp**. **Make sure to include all three predictors in one model – NOT three different models each with one predictor.**

```
multipleReg <- lm(Ozone~Solar.R+Wind+Temp, data=air) # multiple linear regression
summary(multipleReg)
```

```
##
## Call:
## lm(formula = Ozone ~ Solar.R + Wind + Temp, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.651 -15.622  -4.981  12.422 101.411
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -52.16596   21.90933  -2.381   0.0185 *
## Solar.R       0.01654    0.02272   0.728   0.4678
## Wind         -2.69669    0.63085  -4.275 3.40e-05 ***
## Temp          1.53072    0.24115   6.348 2.49e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.26 on 149 degrees of freedom
## Multiple R-squared:  0.4321, Adjusted R-squared:  0.4207
## F-statistic: 37.79 on 3 and 149 DF,  p-value: < 2.2e-16
```

H. Report the **adjusted R-Squared** in a comment – how does it compare to the adjusted R-squared from Step F? Is this better or worse? Which of the predictors are **statistically significant** in the model? In a comment, report the coefficient of each predictor that is statistically significant. Do not report the coefficients for predictors that are not significant.

```
# Adjusted R-squared: 0.4207 (multiple regression) is greater than
# Adusted R-squared of linear regression: 0.2527 from Step F

# Multiple Regression is better model than Linear regression with Wind speed
```

```r
# Wind speed and daily temperature are statistically significant

# co-efficients of predictors that are significant
#Wind          -2.69669
#Temp           1.53072
```

I. Create a one-row data frame like this:

```r
predDF <- data.frame(Solar.R=300, Wind=15, Temp=70) # test data
```

and use it with the **predict( )** function to predict the **expected value of Ozone**:

```r
predict(multipleReg, predDF) # predicting ozone concentration of test data
```

```
##        1
## 19.49483
```

J. Create an additional **multiple regression model**, with **Temp** as the **outcome variable**, and the other **3 variables** as the **predictors**.

Review the quality of the model by commenting on its **adjusted R-Squared**.

```r
multipleRegTemp <- lm(Temp~Ozone+Solar.R+Wind, data=air)
summary(multipleRegTemp)
```

```
##
## Call:
## lm(formula = Temp ~ Ozone + Solar.R + Wind, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.831  -4.802   1.174   4.880  18.004
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.693222   2.796787  26.707  < 2e-16 ***
## Ozone        0.139055   0.021907   6.348 2.49e-09 ***
## Solar.R      0.015751   0.006737   2.338  0.02072 *
## Wind        -0.580176   0.195774  -2.963  0.00354 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.313 on 149 degrees of freedom
## Multiple R-squared:  0.4148, Adjusted R-squared:  0.403
## F-statistic: 35.21 on 3 and 149 DF,  p-value: < 2.2e-16
```

```r
# Adjusted R-squared:  0.403
# It means Ozone concentration, Solar Radiation, and Wind Speed accounts for
# about 40.3% of daily temperature
```

```
# Quality of the model appears to be good as the p-value of the model 2.2e-16 is under 0.05
# also, all the predictors have p-values under 0.05 that imples the predictors are
# significant in predicting daily temperature

# To decide if this a better model, we need to perform regression with other prdictors and
# compare the Adjusted R-Square values with other models. However, 40.3% seems to be pretty
# good number
# Since, we do not have other model to compare with, we can decide if Adjusted R-sqaure
# is closer to 1
# adjusted R-sqaure 0.403 is not very close to 0 so, its comparitively a better model
```