HW 8

April 8, 2021

1 IST 387 HW 8

[1]:
```
# Enter your name here: Connor Hanan
```

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

[1]:
```
# 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**.

[2]:
```
air <- airquality
```

[3]:
```
str(air)
summary(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 …
```

```
     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
```

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**.

[4]:
```
?airquality
```

[ ]:
```
#Ozone is mean ozone in ppb at Roosevelt island
#Solar.R is solar radiation at central park
#wind is avg wind speed in mph at laguardia airport
#temp is max daily temp in deg F at laguardia airport
```

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

[15]:
```
table(is.na(air$Ozone))
table(is.na(air$Solar.R))
table(is.na(air$Wind))
table(is.na(air$Temp))
```

```
FALSE
  153
```

```
FALSE
  153
```

```
FALSE
  153
```

```
FALSE
  153
```

D. Use the **na_interpolation()** function from the **imputeTS package** from HW 6 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.

[8]:
```
#install.packages('imputeTS')
```

```
also installing the dependencies ‘png’, ‘jpeg’, ‘gridtext’, ‘ggtext’,
‘stinepack’


Updating HTML index of packages in '.Library'

Making 'packages.html' …
 done
```

[10]:
```
library(tidyverse)
library(MASS)
library(imputeTS)
```

[14]:
```
air %>%
mutate(Ozone = na_interpolation(Ozone),
       Solar.R = na_interpolation(Solar.R)) -> air
```

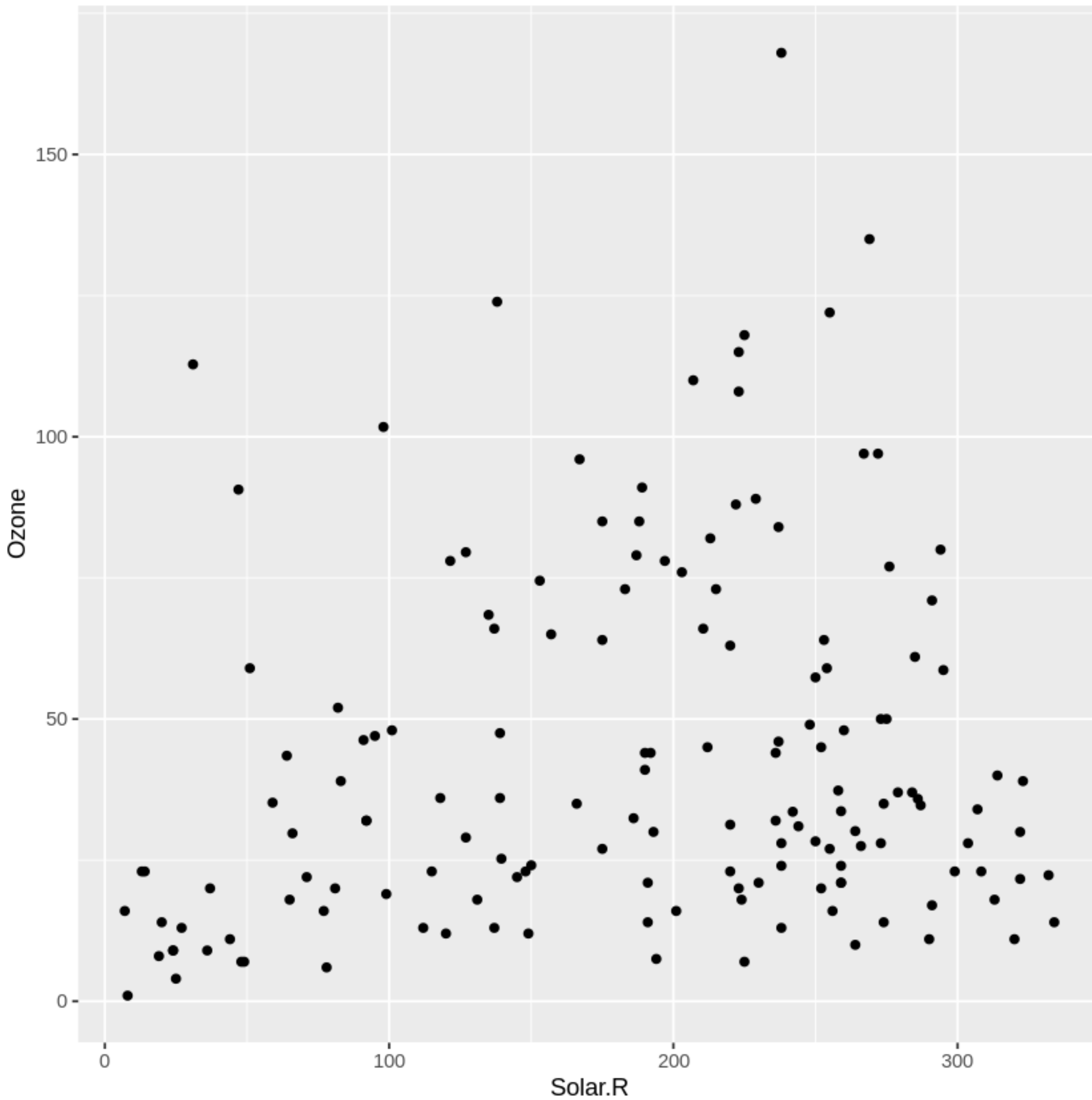
[16]:
```
table(is.na(air$Ozone))
table(is.na(air$Solar.R))
table(is.na(air$Wind))
table(is.na(air$Temp))
```

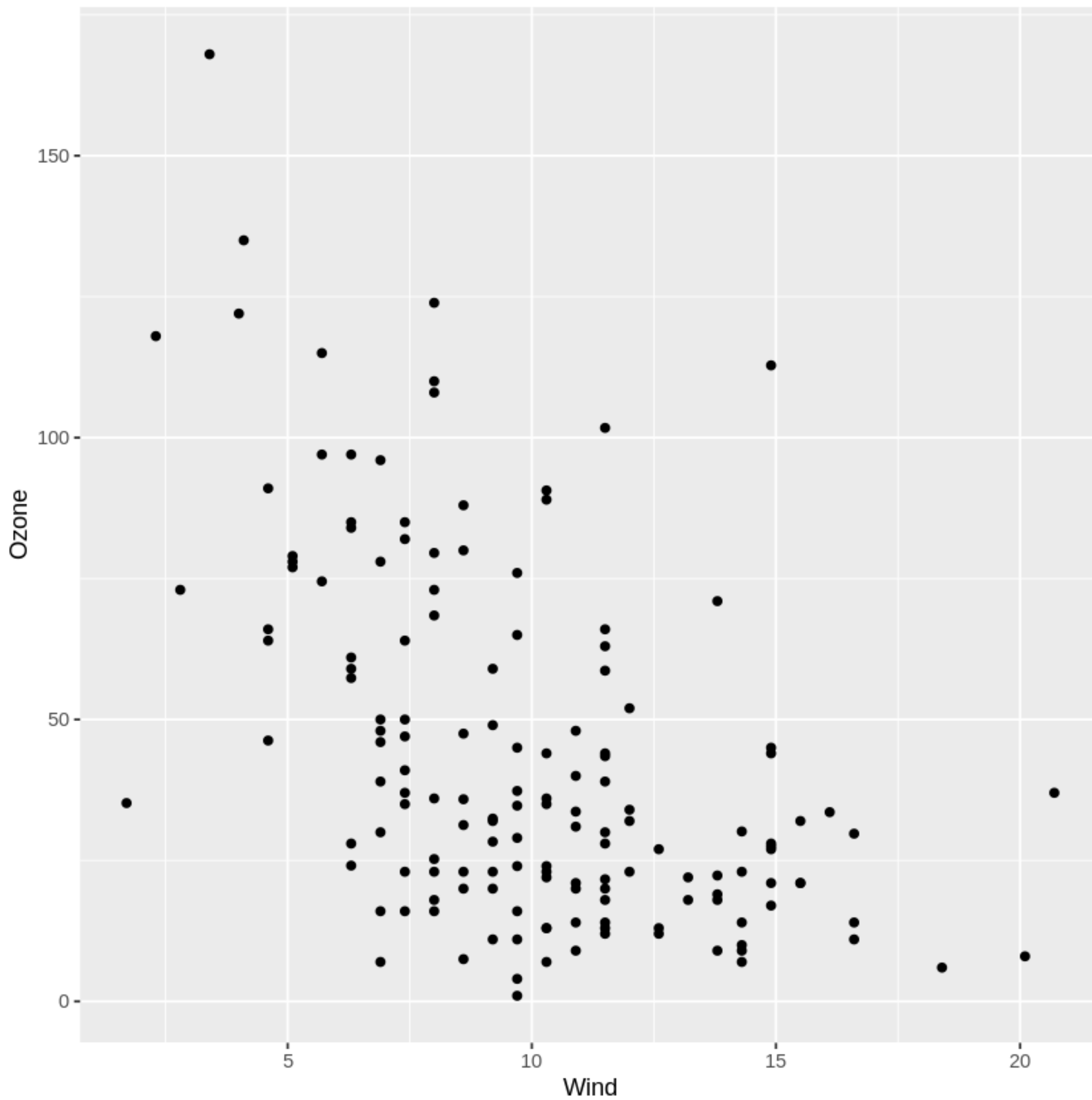
```
FALSE
  153


FALSE
  153


FALSE
  153


FALSE
  153
```

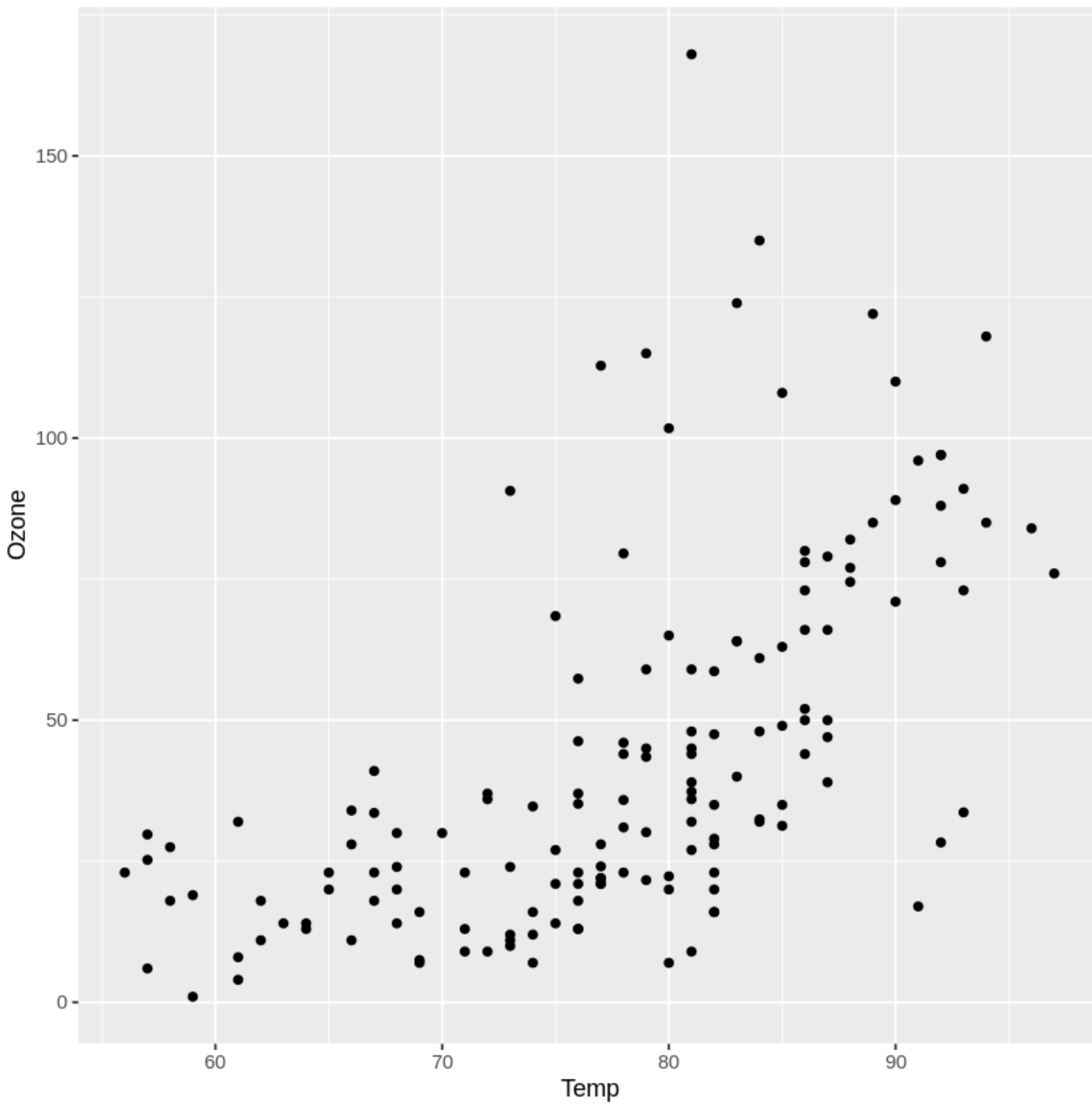
E. Create **3 bivariate scatterplots (X-Y) plots** for each of the predictors with the outcome. **Hint:** In each case, put **Ozone on the Y-axis**, and a **predictor on the X-axis**. Add a comment to each, describing the plot and explaining whether there appears to be a **linear relationship** between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

[29]:
```
ggplot(air)+
geom_point(aes(Solar.R,Ozone)) #wide distribution, no direct correlation

ggplot(air)+
geom_point(aes(Wind,Ozone)) #fairly linear distribution in the negative
↪direction
```

```
ggplot(air)+
geom_point(aes(Temp,Ozone)) #fairly linear distribution in the positive␣
↪direction
```




Ozone
150
100
50
0
5
10
15
20
Wind



F. Next, create a **simple regression model** predicting **Ozone based on Wind**. Refer to page 202 in the text for syntax and explanations of 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.

[20]:
```
lmout <- lm(Ozone~Wind,air)
summary(lmout) #coefficient of Wind is -4.5925, and it is statistically␣
↪significant
#adjusted Rsquared is 25.27%, meaning this variable is about a quarter of all␣
↪the factors that affect Ozone
```

```
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
```

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.**

[21]:
```
lmout <- lm(Ozone~Solar.R+Wind+Temp,air)
summary(lmout)
```

```
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.

[30]:
```
#adjusted r squared is now at 42.07%, meaning these three variables make up␣
↪almost half of all the factors which impact the Ozone level
#wind and temp are statistically significant, as their pr values are very␣
↪small, but Solar.R is not nearly as significant as it has a much higher pr␣
↪value
```

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

[23]:
```
predDF <- data.frame(Solar.R=290, Wind=13, Temp=61)
```

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

[24]:
```
predict(lmout,predDF)
```

1: 10.9463978698246

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**.

[28]:
```
lmout <- lm(Temp~Solar.R+Wind+Ozone,air)
summary(lmout) #the adjusted r squared is 40.3%, meaning that the impact of the␣
↪three variables on Temp is about two fifths of all factors on Temp
```

```
Call:
lm(formula = Temp ~ Solar.R + Wind + Ozone, 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 ***
Solar.R      0.015751   0.006737   2.338  0.02072 *
Wind        -0.580176   0.195774  -2.963  0.00354 **
Ozone        0.139055   0.021907   6.348 2.49e-09 ***
---
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
```