HW 8

August 1, 2021

1 IST 387 HW 8

[1]:
```
# Enter your name here: Ezra Cohen
```

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

[2]:
```
# 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<-data.frame(airquality)
str(air)
summary(air)
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
```

A data.frame: 153 × 6

Ozone	Solar.R	Wind	Temp	Month	Day
<int>	<int>	<dbl>	<int>	<int>	<int>
41	190	7.4	67	5	1
36	118	8.0	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
NA	NA	14.3	56	5	5
28	NA	14.9	66	5	6
23	299	8.6	65	5	7
19	99	13.8	59	5	8
8	19	20.1	61	5	9
NA	194	8.6	69	5	10
7	NA	6.9	74	5	11
16	256	9.7	69	5	12
11	290	9.2	66	5	13
14	274	10.9	68	5	14
18	65	13.2	58	5	15
14	334	11.5	64	5	16
34	307	12.0	66	5	17
6	78	18.4	57	5	18
30	322	11.5	68	5	19
11	44	9.7	62	5	20
1	8	9.7	59	5	21
11	320	16.6	73	5	22
4	25	9.7	61	5	23
32	92	12.0	61	5	24
NA	66	16.6	57	5	25
NA	266	14.9	58	5	26
NA	NA	8.0	57	5	27
23	13	12.0	67	5	28
45	252	14.9	81	5	29
115	223	5.7	79	5	30
96	167	6.9	91	9	1
78	197	5.1	92	9	2
73	183	2.8	93	9	3
91	189	4.6	93	9	4
47	95	7.4	87	9	5
32	92	15.5	84	9	6
20	252	10.9	80	9	7
23	220	10.3	78	9	8
21	230	10.9	75	9	9
24	259	9.7	73	9	10
44	236	14.9	81	9	11
21	259	15.5	76	9	12
28	238	6.3	77	9	13
9	24	10.9	71	9	14
13	112	11.5	71	9	15
46	237	6.9	78	9	16
18	224	13.8	67	9	17
13	27	10.3	76	9	18
24	238	10.3	68	9	19
16	201	8.0	82	9	20

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**.

[3]:
```
?airquality
#Ozone is the mean number of ozone in the air in parts per million from 1pm to␣
↪3pm(I assume this is what they mean by 1300 to 1500) at Roosevelt Park on␣
↪any given day,  Solar.r is the frequency band of 4000-7700 Angstroms in␣
↪Langleys  from 8am to 12am at Central Park on any given day, Wind is the␣
↪average wind speed in miles per hour from 7 a.m. to 10 a.m. at LaGuardia␣
↪Airport on any given day, And temp is the maximum degrees in Fahrenheit at␣
↪LaGuardia Airport on any given day
```

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

[8]:
```
match(TRUE,is.na(air$Ozone))
match(TRUE,is.na(air$Solar.R))
match(TRUE,is.na(air$Wind))
match(TRUE,is.na(air$Temp))
#There is at least one missing value in the first two columns but the second␣
↪two have no missing values
```

5

5

<NA>

<NA>

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.

[15]:
```
#install.packages("imputeTS")
#library(imputeTS)
air$Ozone<-na_interpolation(air$Ozone)
air$Solar.R<-na_interpolation(air$Solar.R)
```

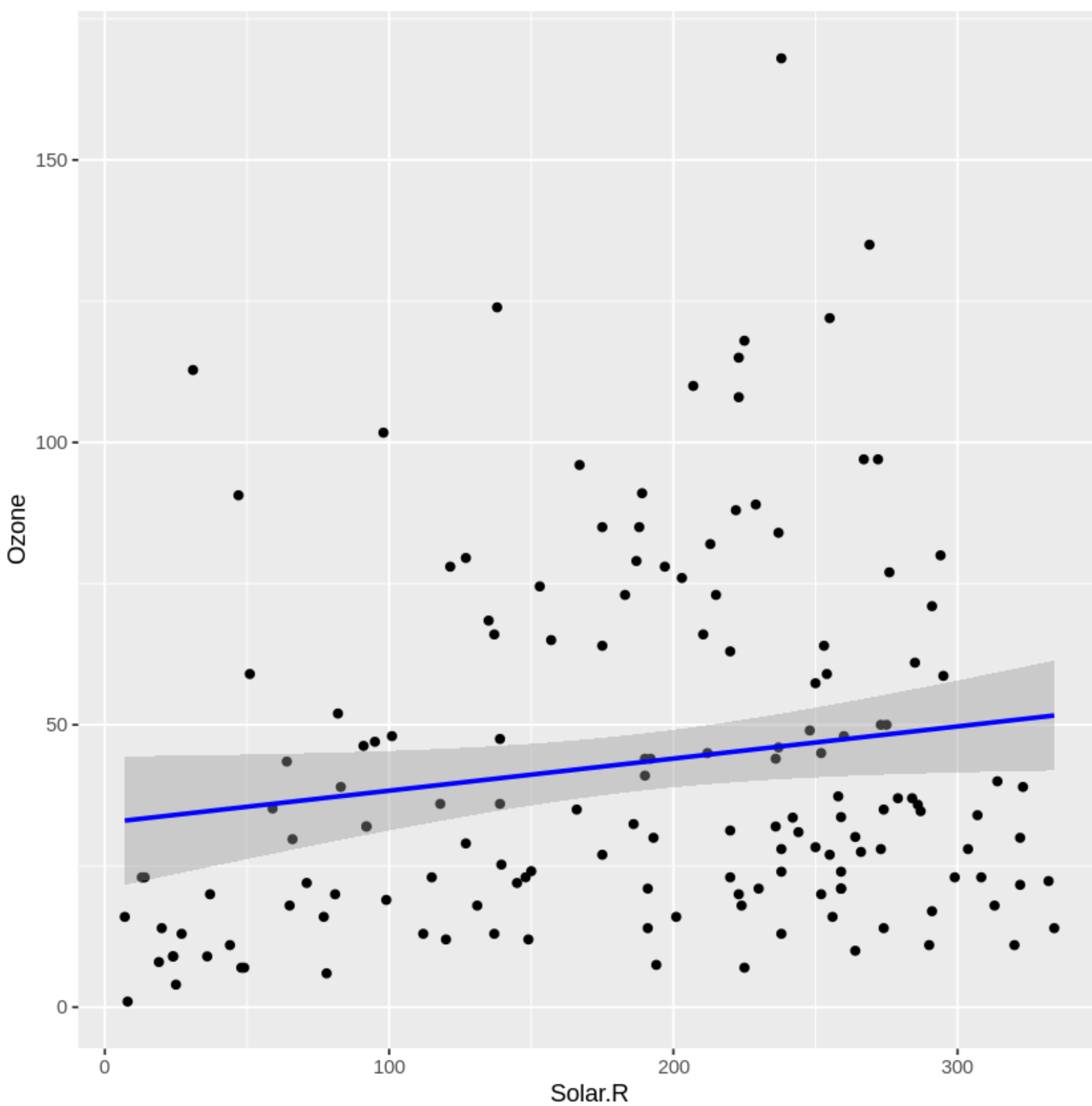
E. Create **3 bivariate scatterplots (X-Y) plots** for each of the predictors with the outcome. **Hint:** In each case, put **Ozone on the Y-axis**, and a **predictor on the X-axis**. Add a comment to each, describing the plot and explaining whether there appears to be a **linear relationship** between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

[18]:
```
library(ggplot2)
plot1<-ggplot(air,aes(x=Solar.R,y=Ozone))+geom_point()+geom_smooth(method =␣
↪"lm", color = "blue")#For the first graph there does not appear to be any␣
↪sort of relationship between the two
```

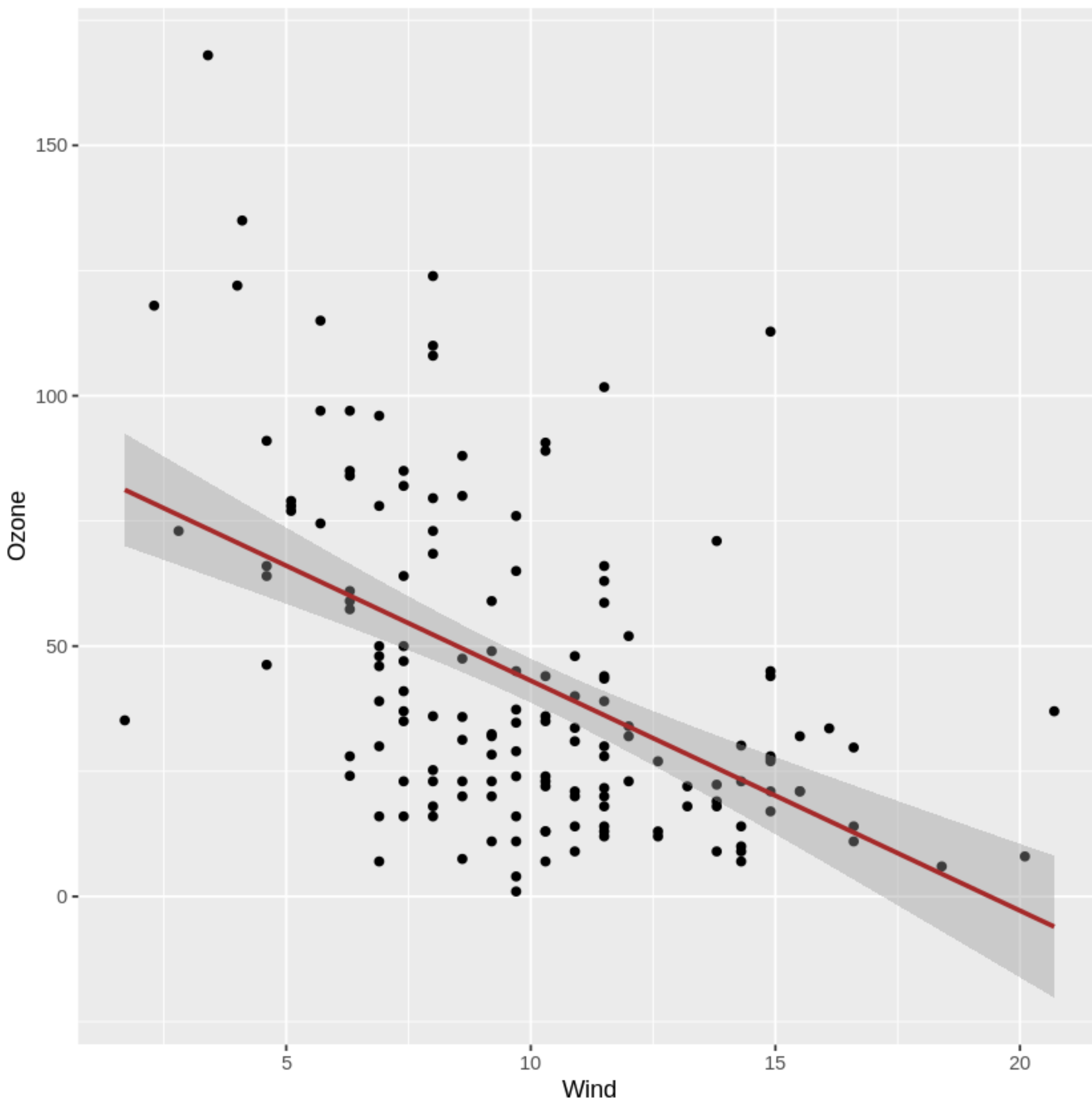
```
plot2<-ggplot(air,aes(x=Wind,y=Ozone))+geom_point()+geom_smooth(method = "lm",
  color = "brown")#For the second graph there seems to be an inverse
  relationship between the two and there is no overall downward trend of the
  line
plot3<-ggplot(air,aes(x=Temp,y=Ozone))+geom_point()+geom_smooth(method = "lm",
  color = "orange")#For the last graph there seems to be an upward trend of
  the line
plot1
plot2
plot3
```

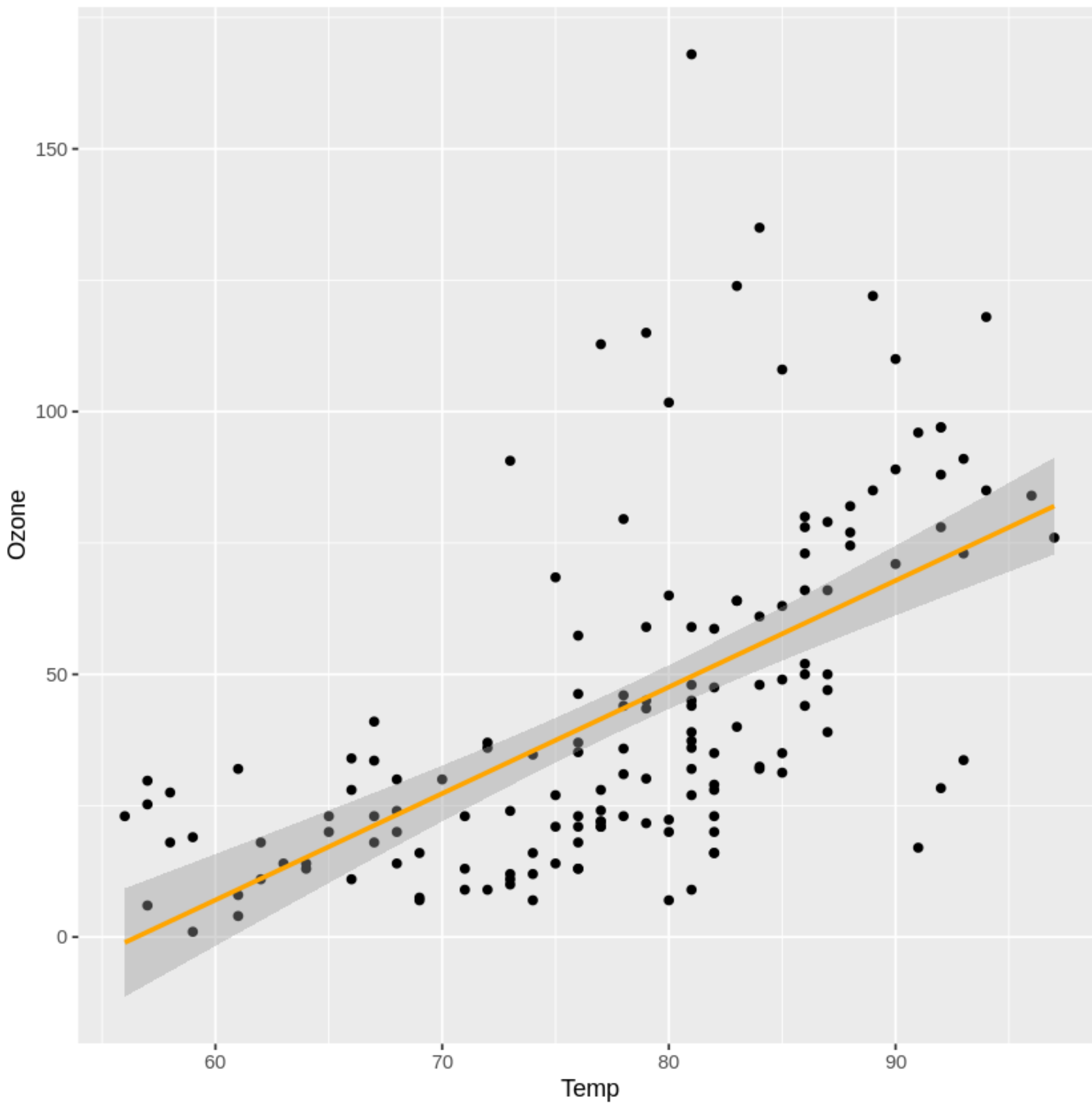
```
`geom_smooth()` using formula 'y ~ x'

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



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





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.

[19]:
```
lmair<-lm(formula=Ozone~Wind,data=air)
summary(lmair)
```

```
#The slope is -4.5925, it seems to be incredibly significant based on the P␣
↪value, but as indicated by the negative slope and I would also assume the␣
↪negative T value is also showing this, it has an inverse relationship as I␣
↪said earlier, the adjusted r-squared value is .2527 Which is really low, but␣
↪everything else is indicating that there is at the very least correlation,␣
↪and I think the reason the value is so low is because most of the points␣
↪don't fall on the line and some can be quite far from the line but the␣
↪points all still do follow a downward Trend none the less
```

```
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]:
```
lmair2<-lm(formula=Ozone~Wind+Solar.R+Temp,data=air)
summary(lmair2)
```

```
Call:
lm(formula = Ozone ~ Wind + Solar.R + 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 *
Wind         -2.69669    0.63085  -4.275 3.40e-05 ***
```

```
Solar.R       0.01654    0.02272   0.728   0.4678
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.

[ ]:
```
#The adjusted r-squared value is 0.4207 which is much better than the last one,␣
↪this is probably due to the inclusion of temp which the graphs also showed a␣
↪correlation between it and ozone, The statistically significant predictors␣
↪are wind and temp, their estimates are -2.69669 for wind and 1.53072 for␣
↪temp, the standard error is relatively low at 0.63085 for wind and 0.24115␣
↪for temp
```

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

[22]:
```
predDF <- data.frame(Solar.R=290, Wind=13, Temp=61)
```

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

[23]:
```
predict(lmair2,predDF)
```

1: 10.9463978698245

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**.

[24]:
```
lmair3<-lm(formula=Temp~Wind+Solar.R+Ozone,data=air)
summary(lmair3)
```

```
Call:
lm(formula = Temp ~ Wind + Solar.R + 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 ***
```

```
Wind        -0.580176   0.195774  -2.963  0.00354 **
Solar.R      0.015751   0.006737   2.338  0.02072 *
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
```

[ ]:
```
#The adjusted r-squared value is .403 which is slightly worse than for the
 ↪previous model but not by much from the P values that we can see temperature
 ↪is most significantly correlated to Ozone, then to wind and then the least
 ↪to solar radiation, but the fact that even solar has one asterisk means it
 ↪is at least slightly correlated
```