# Intro to Data Science HW 8

##### Copyright Jeffrey Stanton, Jeffrey Saltz, and Jasmina Tacheva

#Rutwik Ghag

### 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**.

1. 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**.

library(ggplot2)  
library(MASS)  
air <- data.frame(airquality)  
#View(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   
##

1. 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**.

#the ozone variable has 37 na values. Solar.R, wind and Temp are variables that are explaining current weather condition that lead to the ozone value  
#there are no NA values in wind and temp. Ozone and Solar.R have 37 and 7 NA values respectively.

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

air[is.na(air$Ozone),]

## Ozone Solar.R Wind Temp Month Day  
## 5 NA NA 14.3 56 5 5  
## 10 NA 194 8.6 69 5 10  
## 25 NA 66 16.6 57 5 25  
## 26 NA 266 14.9 58 5 26  
## 27 NA NA 8.0 57 5 27  
## 32 NA 286 8.6 78 6 1  
## 33 NA 287 9.7 74 6 2  
## 34 NA 242 16.1 67 6 3  
## 35 NA 186 9.2 84 6 4  
## 36 NA 220 8.6 85 6 5  
## 37 NA 264 14.3 79 6 6  
## 39 NA 273 6.9 87 6 8  
## 42 NA 259 10.9 93 6 11  
## 43 NA 250 9.2 92 6 12  
## 45 NA 332 13.8 80 6 14  
## 46 NA 322 11.5 79 6 15  
## 52 NA 150 6.3 77 6 21  
## 53 NA 59 1.7 76 6 22  
## 54 NA 91 4.6 76 6 23  
## 55 NA 250 6.3 76 6 24  
## 56 NA 135 8.0 75 6 25  
## 57 NA 127 8.0 78 6 26  
## 58 NA 47 10.3 73 6 27  
## 59 NA 98 11.5 80 6 28  
## 60 NA 31 14.9 77 6 29  
## 61 NA 138 8.0 83 6 30  
## 65 NA 101 10.9 84 7 4  
## 72 NA 139 8.6 82 7 11  
## 75 NA 291 14.9 91 7 14  
## 83 NA 258 9.7 81 7 22  
## 84 NA 295 11.5 82 7 23  
## 102 NA 222 8.6 92 8 10  
## 103 NA 137 11.5 86 8 11  
## 107 NA 64 11.5 79 8 15  
## 115 NA 255 12.6 75 8 23  
## 119 NA 153 5.7 88 8 27  
## 150 NA 145 13.2 77 9 27

air[is.na(air$Solar),]

## Ozone Solar.R Wind Temp Month Day  
## 5 NA NA 14.3 56 5 5  
## 6 28 NA 14.9 66 5 6  
## 11 7 NA 6.9 74 5 11  
## 27 NA NA 8.0 57 5 27  
## 96 78 NA 6.9 86 8 4  
## 97 35 NA 7.4 85 8 5  
## 98 66 NA 4.6 87 8 6

air[is.na(air$Wind),]

## [1] Ozone Solar.R Wind Temp Month Day   
## <0 rows> (or 0-length row.names)

air[is.na(air$Temp),]

## [1] Ozone Solar.R Wind Temp Month Day   
## <0 rows> (or 0-length row.names)

#ozone and solar have missing values while wind and temp don't.

1. 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)

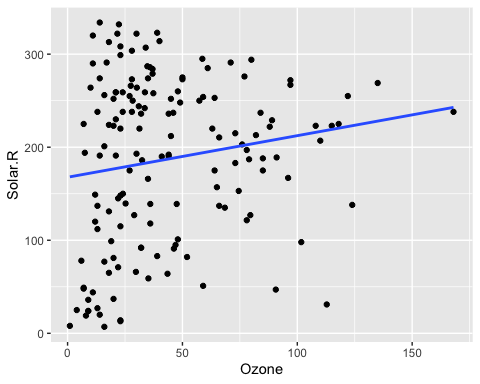
## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

air$Ozone <- na\_interpolation(air$Ozone)  
air$Solar.R <- na\_interpolation(air$Solar.R)  
#View(air)

1. 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.

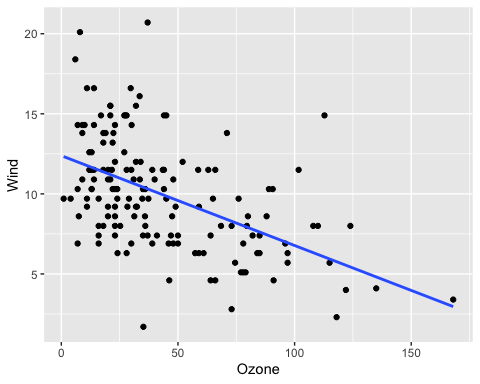
ggplot(data = air) + aes(x=Ozone,y=Solar.R) + geom\_point() +geom\_smooth(method = "lm", se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'



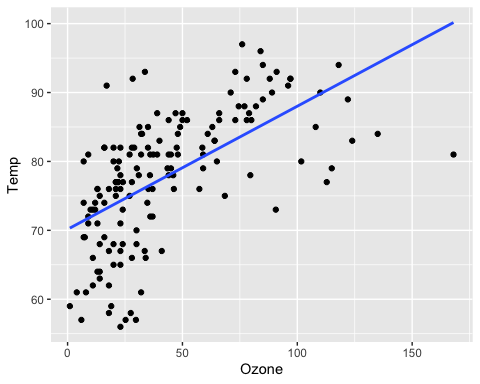
ggplot(data = air) + aes(x=Ozone,y=Wind) + geom\_point() +geom\_smooth(method = "lm", se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'



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

## `geom\_smooth()` using formula 'y ~ x'



#all three plots have a linear relationship.

1. 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.

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

##   
## 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

#it is statistically significant since the p value of wind predictor variable is negative i.e. it is less that 0.05 and furthermore there are threee star. Hence Wind is a good prediction variable for variable Ozone.

1. 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.**

lmout10 <- lm(Ozone~Solar.R + Wind + Temp ,data=air)  
summary(lmout10)

##   
## 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

1. 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 values of F is 0.2527 and G is 0.4207. Out of the two, G is a better prediction model because here the three variables are used to create a single prediction model instead of creating separate models and combining them.

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

predDF <- data.frame(Solar.R=290, Wind=13, Temp=61)

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

predict(lmout10,predDF)

## 1   
## 10.9464

1. 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**.

lmout100 <- lm(Temp~Ozone+Solar.R + Wind ,data=air)  
summary(lmout100)

##   
## 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

#the adjusted R-Squared value of this prediction model is 0.403  
#the quality of the model is good because the value of adjusted r-squared is high.