Multiple Imputation

Linxi Li T00622714

2022-10-11

This instruction is for imputation missing data

We are using mice() package to deal with the missing data

So we installed the package

```
#install.packages("mice")
```

```
library(mice)
```

```
##
## Attaching package: 'mice'
## The following object is masked from 'package:stats':
##
##     filter
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(VIM)
```

```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
```

R Markdown

I created a dataset and save as a csv document:

```
data <- read.csv("/Users/ccc/Desktop/create data.csv")
data
```

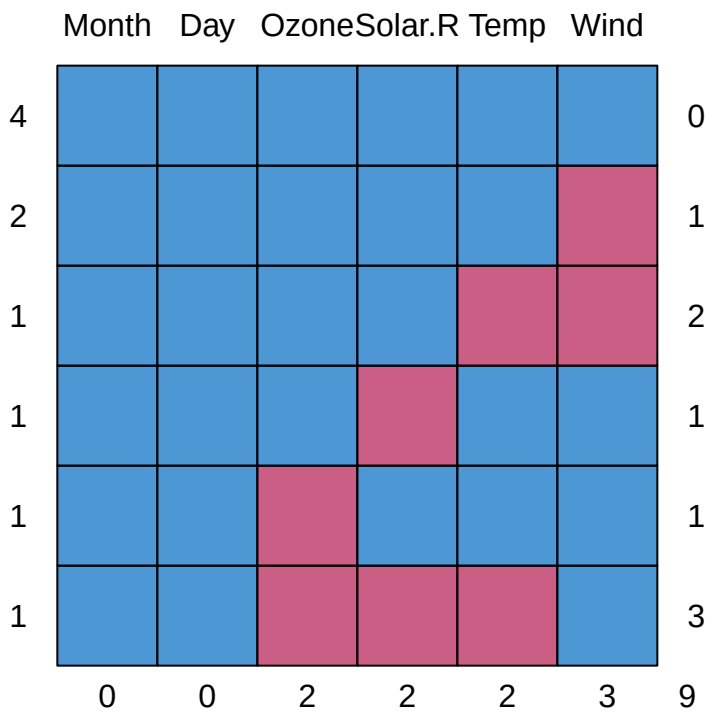
```
##    Ozone Solar.R Wind Temp Month Day
## 1     41     190  7.4   67     5   1
## 2     36     118  8.0   72     5   2
## 3     12     149 12.6   74     5   3
```

```
## 4     18     313 11.5   62     5   4
## 5     NA      NA 14.3   NA     5   5
## 6     28      NA 14.9   66     5   6
## 7     20     200   NA   70     5   7
## 8     NA     167 13.6   71     5   8
## 9     31     177   NA   NA     5   9
## 10    39     213   NA   73     5  10
```

Including Plots

I am going to show the plot of how the data missing, the red spot represents the missing data

```
md.pattern(data)
```




```
##   Month Day Ozone Solar.R Temp Wind
## 4     1   1     1       1    1    1 0
## 2     1   1     1       1    1    0 1
## 1     1   1     1       1    0    0 2
## 1     1   1     1       0    1    1 1
## 1     1   1     0       1    1    1 1
## 1     1   1     0       0    0    1 3
##       0   0     2       2    2    3 9
```

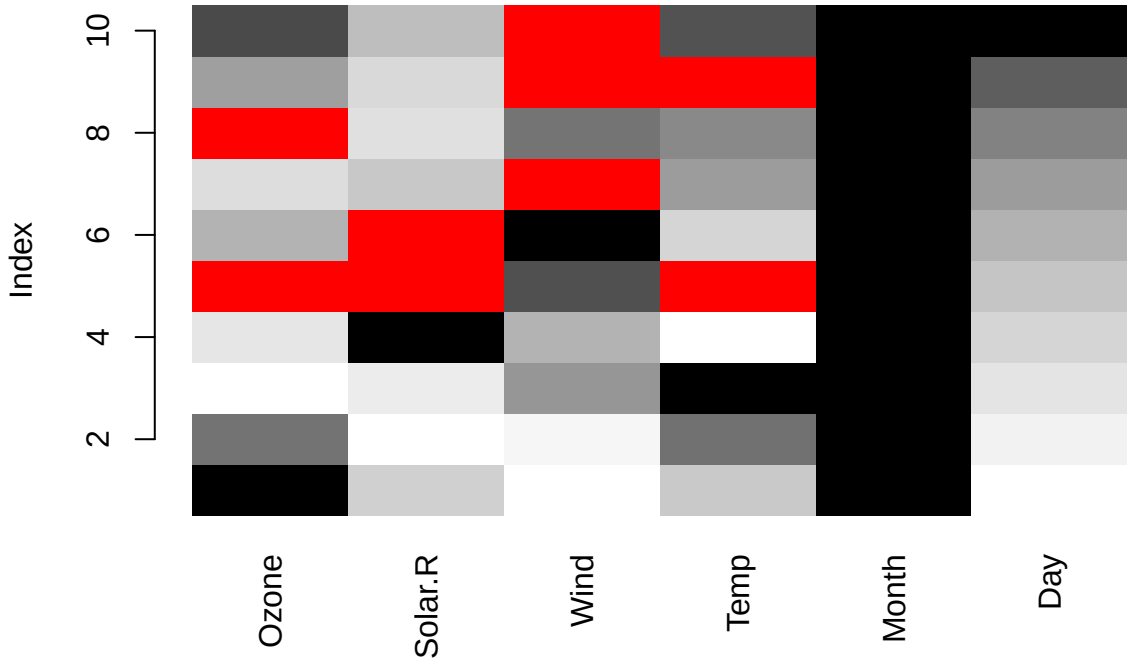
```
md.pairs(data)
```

```
## $rr
##         Ozone Solar.R Wind Temp Month Day
## Ozone       8       7    5    7     8   8
## Solar.R     7       8    5    7     8   8
## Wind        5       5    7    6     7   7
## Temp        7       7    6    8     8   8
```

```
## Month       8       8    7    8    10  10
## Day         8       8    7    8    10  10
##
## $rm
##         Ozone Solar.R Wind Temp Month Day
## Ozone       0       1    3    1     0   0
## Solar.R     1       0    3    1     0   0
## Wind        2       2    0    1     0   0
## Temp        1       1    2    0     0   0
## Month       2       2    3    2     0   0
## Day         2       2    3    2     0   0
##
## $mr
##         Ozone Solar.R Wind Temp Month Day
## Ozone       0       1    2    1     2   2
## Solar.R     1       0    2    1     2   2
## Wind        3       3    0    2     3   3
## Temp        1       1    1    0     2   2
## Month       0       0    0    0     0   0
## Day         0       0    0    0     0   0
##
## $mm
##         Ozone Solar.R Wind Temp Month Day
## Ozone       2       1    0    1     0   0
## Solar.R     1       2    0    1     0   0
## Wind        0       0    3    1     0   0
## Temp        1       1    1    2     0   0
## Month       0       0    0    0     0   0
## Day         0       0    0    0     0   0
```

we can also draw the data distribution of each column of the data set.
In the figure, the red color indicates the missing values, and the transition colors from black to gray to white indicate the different values, the more transition colors indicate the more scattered data values, and the less transition colors indicate the more concentrated data values.

```
matrixplot(data)
```



Next, we use the mice() function, which is used to populate the data

```
imputed_Data <- mice(data, m=10, maxit = 5, method = 'pmm', seed = 500)
```

```
##
##  iter imp variable
##   1   1  Ozone  Solar.R  Wind  Temp
##   1   2  Ozone  Solar.R  Wind  Temp
##   1   3  Ozone  Solar.R  Wind  Temp
##   1   4  Ozone  Solar.R  Wind  Temp
##   1   5  Ozone  Solar.R  Wind  Temp
##   1   6  Ozone  Solar.R  Wind  Temp
##   1   7  Ozone  Solar.R  Wind  Temp
##   1   8  Ozone  Solar.R  Wind  Temp
##   1   9  Ozone  Solar.R  Wind  Temp
##   1   10  Ozone  Solar.R  Wind  Temp
##   2   1  Ozone  Solar.R  Wind  Temp
##   2   2  Ozone  Solar.R  Wind  Temp
##   2   3  Ozone  Solar.R  Wind  Temp
##   2   4  Ozone  Solar.R  Wind  Temp
##   2   5  Ozone  Solar.R  Wind  Temp
##   2   6  Ozone  Solar.R  Wind  Temp
##   2   7  Ozone  Solar.R  Wind  Temp
##   2   8  Ozone  Solar.R  Wind  Temp
##   2   9  Ozone  Solar.R  Wind  Temp
##   2   10  Ozone  Solar.R  Wind  Temp
##   3   1  Ozone  Solar.R  Wind  Temp
##   3   2  Ozone  Solar.R  Wind  Temp
##   3   3  Ozone  Solar.R  Wind  Temp
##   3   4  Ozone  Solar.R  Wind  Temp
##   3   5  Ozone  Solar.R  Wind  Temp
##   3   6  Ozone  Solar.R  Wind  Temp
```

```
##   3   7  Ozone  Solar.R  Wind  Temp
##   3   8  Ozone  Solar.R  Wind  Temp
##   3   9  Ozone  Solar.R  Wind  Temp
##   3   10  Ozone  Solar.R  Wind  Temp
##   4   1  Ozone  Solar.R  Wind  Temp
##   4   2  Ozone  Solar.R  Wind  Temp
##   4   3  Ozone  Solar.R  Wind  Temp
##   4   4  Ozone  Solar.R  Wind  Temp
##   4   5  Ozone  Solar.R  Wind  Temp
##   4   6  Ozone  Solar.R  Wind  Temp
##   4   7  Ozone  Solar.R  Wind  Temp
##   4   8  Ozone  Solar.R  Wind  Temp
##   4   9  Ozone  Solar.R  Wind  Temp
##   4   10  Ozone  Solar.R  Wind  Temp
##   5   1  Ozone  Solar.R  Wind  Temp
##   5   2  Ozone  Solar.R  Wind  Temp
##   5   3  Ozone  Solar.R  Wind  Temp
##   5   4  Ozone  Solar.R  Wind  Temp
##   5   5  Ozone  Solar.R  Wind  Temp
##   5   6  Ozone  Solar.R  Wind  Temp
##   5   7  Ozone  Solar.R  Wind  Temp
##   5   8  Ozone  Solar.R  Wind  Temp
##   5   9  Ozone  Solar.R  Wind  Temp
##   5   10  Ozone  Solar.R  Wind  Temp
```

```
## Warning: Number of logged events: 1
```

```
#note that:
#m, the number of fill matrices for the multi-fill method, default is 5
#maxit, the maximum number of iterations, default is 5
#method, the method used to fill, and pmm is predictive mean matching.
#We can use methods(mice) to see what methods are available.
```

```
summary(imputed_Data)
```

```
## Class: mids
## Number of multiple imputations:  10
## Imputation methods:
##   Ozone Solar.R    Wind    Temp   Month     Day
##   "pmm"   "pmm"   "pmm"   "pmm"      ""      ""
## PredictorMatrix:
##         Ozone Solar.R Wind Temp Month Day
## Ozone       0       1    1    1     0   1
## Solar.R     1       0    1    1     0   1
## Wind        1       1    0    1     0   1
## Temp        1       1    1    0     0   1
## Month       1       1    1    1     0   1
## Day         1       1    1    1     0   0
## Number of logged events:  1
##   it im dep     meth   out
## 1  0  0     constant Month
```

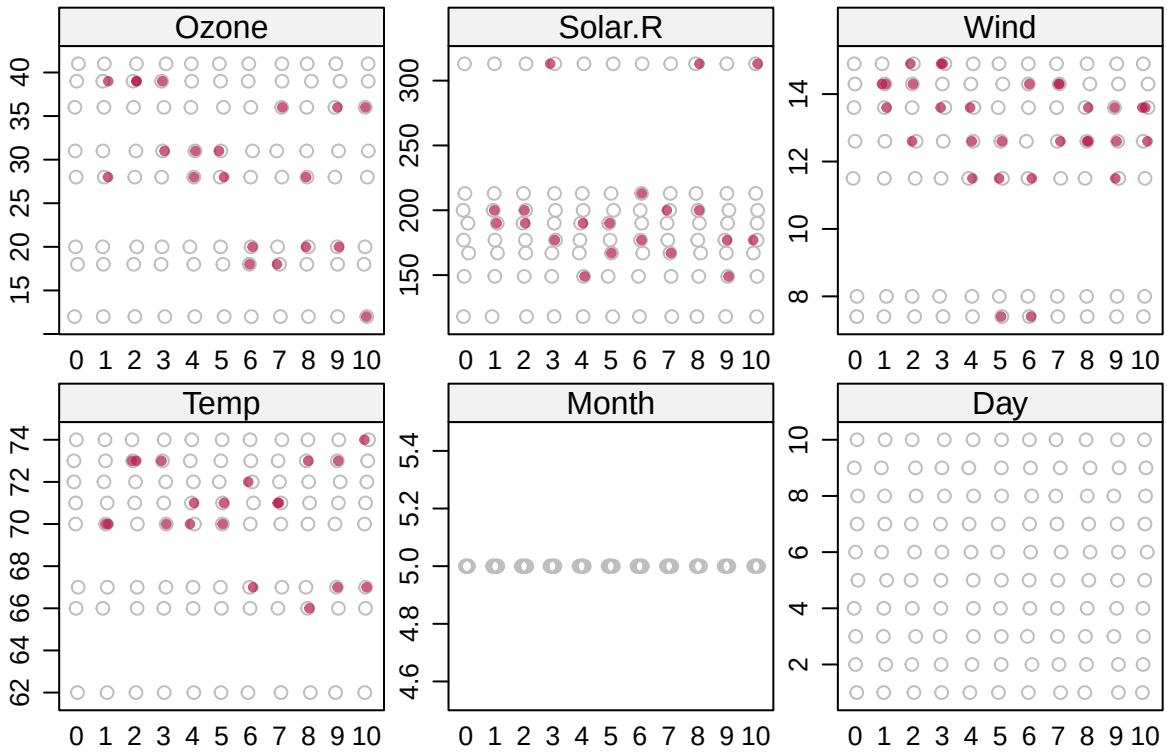
View the result of imputation:

```
imputed_Data$imp
```

```
## $Ozone
##    1  2  3  4  5  6  7  8  9 10
## 5 28 39 39 31 28 18 18 20 20 12
## 8 39 39 31 28 31 20 36 28 36 36
##
## $Solar.R
##     1   2   3   4   5   6   7   8   9  10
## 5 190 190 177 149 190 177 200 313 149 177
## 6 200 200 313 190 167 213 167 200 177 313
##
## $Wind
##       1    2    3    4    5    6    7    8    9   10
## 7  14.3 14.3 14.9 13.6 11.5 14.3 14.3 12.6 12.6 13.6
## 9  13.6 12.6 14.9 11.5  7.4  7.4 14.3 13.6 13.6 12.6
## 10 14.3 14.9 13.6 12.6 12.6 11.5 12.6 12.6 11.5 13.6
##
## $Temp
##    1  2  3  4  5  6  7  8  9 10
## 5 70 73 70 70 71 67 71 66 73 74
## 9 70 73 73 71 70 72 71 73 67 67
##
## $Month
##  [1] 1  2  3  4  5  6  7  8  9  10
## <0 rows> (or 0-length row.names)
##
## $Day
##  [1] 1  2  3  4  5  6  7  8  9  10
## <0 rows> (or 0-length row.names)
```

View the plot of Sub-panel observations, grouped by independent individual indicators, were populated for the 10 groups of data.

```
stripplot(imputed_Data, col=c("grey",mdc(2)),pch=c(1,20))
```



Analyze the results and optimize the model

```
fit=with(imputed_Data,lm(Ozone ~ Wind + Solar.R + Temp))
summary(fit)
```

```
## # A tibble: 40 x 6
##    term        estimate std.error statistic p.value  nobs
##    <chr>          <dbl>     <dbl>     <dbl>   <dbl> <int>
##  1 (Intercept)  58.3      142.        0.411    0.695    10
##  2 Wind         -0.900      1.60     -0.563    0.594    10
##  3 Solar.R      -0.0521     0.132    -0.394    0.707    10
##  4 Temp         -0.114      1.85     -0.0616   0.953    10
##  5 (Intercept)  10.7      141.        0.0761   0.942    10
##  6 Wind         -0.771      1.82     -0.425    0.686    10
##  7 Solar.R      -0.0203     0.141    -0.144    0.890    10
##  8 Temp          0.472      1.84      0.256    0.806    10
##  9 (Intercept)  22.1      178.        0.124    0.905    10
## 10 Wind         -1.04       2.02     -0.513    0.626    10
## # ... with 30 more rows
## # i Use `print(n = ...)` to see more rows
```

Using the with() function, a multiple linear regression analysis model was performed on the five interpolated data sets, and a t-test was performed to determine the validity of each variable in the data set.

```
pooled=pool(fit)
pool.r.squared(fit)
```

```
##           est      lo 95     hi 95      fmi
```

```
## R^2 0.2221467 0.07498096 0.7445742 0.129885
```

```
completeData1 <- complete(imputed_Data,1)
data1 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData1)
summary(data1)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.7746  -5.1361   0.7358   5.6927  12.9730
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.31005  141.88485   0.411    0.695
## Wind         -0.90002    1.59767  -0.563    0.594
## Solar.R      -0.05206    0.13199  -0.394    0.707
## Temp         -0.11404    1.84987  -0.062    0.953
##
## Residual standard error: 11.27 on 6 degrees of freedom
## Multiple R-squared:  0.1433, Adjusted R-squared:  -0.2851
## F-statistic: 0.3345 on 3 and 6 DF,  p-value: 0.8014
```

```
completeData2 <- complete(imputed_Data,2)
data2 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData2)
summary(data2)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.8597  -5.2442   0.8145   8.5848   9.6851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.7013   140.5720   0.076    0.942
## Wind         -0.7713     1.8167  -0.425    0.686
## Solar.R      -0.0203     0.1413  -0.144    0.890
## Temp          0.4716     1.8410   0.256    0.806
##
## Residual standard error: 12.06 on 6 degrees of freedom
## Multiple R-squared:  0.1023, Adjusted R-squared:  -0.3465
## F-statistic: 0.228 on 3 and 6 DF,  p-value: 0.8737
```

```
completeData3 <- complete(imputed_Data,3)
data3 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData3)
summary(data3)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData3)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.342  -5.320   2.237   6.008  10.980
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.131228 178.301558   0.124    0.905
## Wind         -1.035191   2.018303  -0.513    0.626
## Solar.R      -0.009977   0.145069  -0.069    0.947
## Temp          0.320823   2.417798   0.133    0.899
##
## Residual standard error: 11.47 on 6 degrees of freedom
## Multiple R-squared:  0.1139, Adjusted R-squared:  -0.3291
## F-statistic: 0.2571 on 3 and 6 DF,  p-value: 0.8539
```

```
completeData4 <- complete(imputed_Data,4)
data4 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData4)
summary(data4)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.300  -4.227   1.570   4.221  13.696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.45448  107.84811   0.681    0.521
## Wind         -1.59014    1.33010  -1.196    0.277
## Solar.R      -0.05053    0.09493  -0.532    0.614
## Temp         -0.23769    1.37203  -0.173    0.868
##
## Residual standard error: 9.885 on 6 degrees of freedom
## Multiple R-squared:  0.2583, Adjusted R-squared:  -0.1126
## F-statistic: 0.6965 on 3 and 6 DF,  p-value: 0.5873
```

```
completeData5 <- complete(imputed_Data,5)
data5 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData5)
summary(data5)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.101  -4.981   1.576   4.175  13.586
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.58663  104.23476   0.572    0.588
## Wind         -1.17950    1.24334  -0.949    0.379
## Solar.R      -0.04411    0.09444  -0.467    0.657
```

```
## Temp         -0.13582    1.33105  -0.102    0.922
##
## Residual standard error: 10.31 on 6 degrees of freedom
## Multiple R-squared:  0.1927, Adjusted R-squared:  -0.2109
## F-statistic: 0.4774 on 3 and 6 DF,  p-value: 0.7096
```

```
completeData6 <- complete(imputed_Data,6)
data6 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData6)
summary(data6)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0449  -4.0377  -0.9255   4.4515  12.5549
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.145e+01  1.006e+02   0.512   0.6272
## Wind        -2.249e+00  1.070e+00  -2.101   0.0803 .
## Solar.R     -9.284e-04  9.104e-02  -0.010   0.9922
## Temp         1.455e-02  1.238e+00   0.012   0.9910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.287 on 6 degrees of freedom
## Multiple R-squared:  0.4364, Adjusted R-squared:  0.1545
## F-statistic: 1.548 on 3 and 6 DF,  p-value: 0.2964
```

```
completeData7 <- complete(imputed_Data,7)
data7 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData7)
summary(data7)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData7)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9513  -5.4772   0.0259   5.5218  13.3096
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.40201  108.18809   0.641    0.545
## Wind         -1.60921    1.40836  -1.143    0.297
## Solar.R      -0.05353    0.09891  -0.541    0.608
## Temp         -0.16485    1.39744  -0.118    0.910
##
## Residual standard error: 10.82 on 6 degrees of freedom
## Multiple R-squared:  0.2578, Adjusted R-squared:  -0.1133
## F-statistic: 0.6946 on 3 and 6 DF,  p-value: 0.5882
```

```
completeData8 <- complete(imputed_Data,8)
data8 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData8)
```

```
summary(data8)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData8)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.4994  -5.1225   0.7875   5.6020  12.1010
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.23805  109.21177   0.368    0.725
## Wind         -1.70001    1.53209  -1.110    0.310
## Solar.R      -0.02226    0.09398  -0.237    0.821
## Temp          0.17566    1.43819   0.122    0.907
##
## Residual standard error: 10.05 on 6 degrees of freedom
## Multiple R-squared:  0.2803, Adjusted R-squared:  -0.07961
## F-statistic: 0.7788 on 3 and 6 DF,  p-value: 0.5473
```

```
completeData9 <- complete(imputed_Data,9)
data9 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData9)
summary(data9)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData9)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7532  -5.3736   0.1835   3.7458  13.7425
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.64984  103.00884   0.987    0.362
## Wind         -1.84390    1.38655  -1.330    0.232
## Solar.R      -0.06396    0.09429  -0.678    0.523
## Temp         -0.56937    1.28488  -0.443    0.673
##
## Residual standard error: 10.47 on 6 degrees of freedom
## Multiple R-squared:  0.2776, Adjusted R-squared:  -0.08365
## F-statistic: 0.7684 on 3 and 6 DF,  p-value: 0.5522
```

```
completeData10 <- complete(imputed_Data,10)
data10 <- lm(Ozone ~ Wind + Solar.R + Temp, data=completeData10)
summary(data10)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Solar.R + Temp, data = completeData10)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.149  -9.352   2.198   5.499  15.120
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.90329  147.22848   0.468    0.656
## Wind         -1.72637    2.25908  -0.764    0.474
## Solar.R      -0.01252    0.13188  -0.095    0.927
## Temp         -0.25861    2.02990  -0.127    0.903
##
## Residual standard error: 12.12 on 6 degrees of freedom
## Multiple R-squared:  0.1968, Adjusted R-squared:  -0.2048
## F-statistic: 0.4901 on 3 and 6 DF,  p-value: 0.7019
```

By checking the R-square, I am going to use the imputation 8, as our missing data.

So our model will be: Ozone = 40.23805-1.70001Wind-0.02226Solar.R+0.17566Temp