M440B HW 15

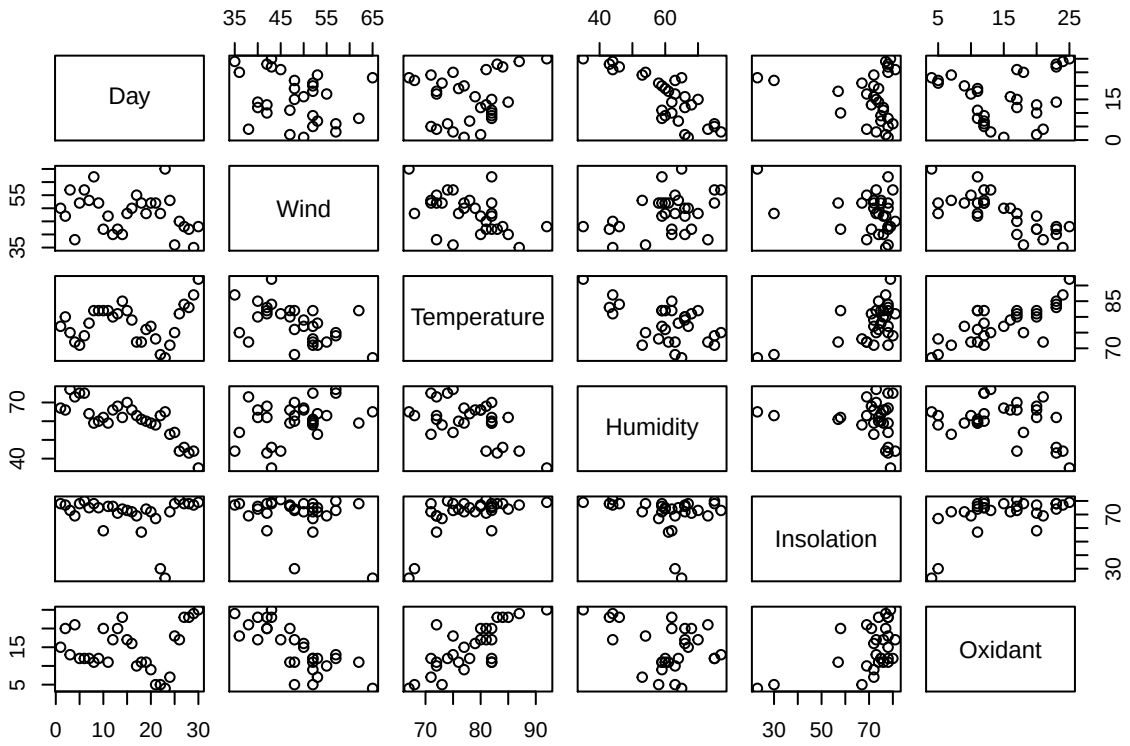
Erick Castillo

4/25/2021

Problem 56A. This problem is broken down into three major sections.

(i.) Type the commands given in the pdf and find the most interesting information from the scatterplots and the correlation matrix.

```
pairs(air[1:6])
```




```
cor(air)
```

```
##                   Day       Wind Temperature   Humidity Insolation    Oxidant
## Day         1.0000000 -0.2771054   0.1827177 -0.8129835 -0.1643671  0.1045661
## Wind       -0.2771054  1.0000000  -0.4953908  0.3743735 -0.3209849 -0.7657126
```

```
## Temperature  0.1827177 -0.4953908   1.0000000 -0.5435176  0.5668331  0.7589575
## Humidity    -0.8129835  0.3743735  -0.5435176  1.0000000 -0.1837282 -0.3521647
## Insolation  -0.1643671 -0.3209849   0.5668331 -0.1837282  1.0000000  0.5051419
## Oxidant      0.1045661 -0.7657126   0.7589575 -0.3521647  0.5051419  1.0000000
```

Beginning with the scatterplots, there are a few positive correlations present being Temperature with Oxidant; Insolation with Oxidant; and Wind with Humidity. There are a few negatively correlated variables, those being Day with Humidity; Temperature with Wind; Temperature with Humidity; and Wind with Oxidant.

The correlation matrix agrees with the variables I listed above; however, some of the values were not as strongly correlated as I thought. For example, Wind and Humidity have a correlation coefficient of 0.37.

Notice that as the days go by, the humidity levels decrease. As the wind levels rise, the oxidant levels decrease along. As the temperature rises, both insolation and oxidant levels rise. As the temperature increases, the humidity levels tend to decrease. All these conclusions were derived from looking at the correlation matrix. Only values ≥ 0.5.

(ii.) Run a multiple regression using all 5 predictors.

```
fullmod <- lm(Oxidant~., data = air)
summary(fullmod)
```

```
##
## Call:
## lm(formula = Oxidant ~ ., data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6920 -1.1675  0.2582  1.8289  4.0773
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.04010   21.20961  -0.568  0.57553
## Day          -0.02997    0.13995  -0.214  0.83227
## Wind         -0.44749    0.09103  -4.916 5.14e-05 ***
## Temperature   0.55714    0.15347   3.630  0.00133 **
## Humidity      0.06818    0.13336   0.511  0.61384
## Insolation    0.01822    0.05583   0.326  0.74694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.977 on 24 degrees of freedom
## Multiple R-squared:  0.7984, Adjusted R-squared:  0.7564
## F-statistic: 19.01 on 5 and 24 DF,  p-value: 1.203e-07
```

The above output has a lot of information to interpret. To begin, notice that the Wind and Temperature variables are significant at a 1% level. The other variables: Day, Humidity, and Insolation are not significant because of their huge p-values. This can also be seen in their t-values as they are small in magnitude when compared to the significant predictors.

The $R^2_{Adj} \approx 0.756$. This means that 75.6% of the variability in Oxidant is explained by the model.

This is a relatively good fit for the data. This model could however be reduced to fit in line with the principle of parsimony.

(iii.) This part of the question asks to use the regsubsets function.

```
mod <- summary(regsubsets(Oxidant~., data = air, method = c('exhaustive','seqrep')))
mod$outmat
```

```
##          Day Wind Temperature Humidity Insolation
## 1  ( 1 ) " " "*"  " "         " "      " "
## 2  ( 1 ) " " "*"  "*"         " "      " "
## 3  ( 1 ) " " "*"  "*"         "*"      " "
## 4  ( 1 ) " " "*"  "*"         "*"      "*"
## 5  ( 1 ) "*" "*"  "*"         "*"      "*"
```

From the above output, it is clear that the best model is given by Wind being the only predictor. The second best model is given by Wind and Temperature included. The third best model is given by Humidity, Wind, and Temperature present.

(iv.) Use the R^2 criterion to determine the best subset model. That is, which model uses the least amount of predictors while still yielding a reasonably high R^2?

```
mod1 <- lm(Oxidant~Wind, data = air)
mod2 <- lm(Oxidant~Wind+Temperature, data = air)
mod3 <- lm(Oxidant~Wind+Temperature+Humidity, data = air)
```

Now that the models were created, I will generate their respective summaries:

```
summary(mod1)
```

```
##
## Call:
## lm(formula = Oxidant ~ Wind, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9266 -2.5923  0.2065  2.6636  6.9077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.3171     4.8976   9.253 5.19e-10 ***
## Wind         -0.6331     0.1005  -6.300 8.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.948 on 28 degrees of freedom
## Multiple R-squared:  0.5863, Adjusted R-squared:  0.5715
## F-statistic: 39.68 on 1 and 28 DF,  p-value: 8.205e-07
```

The above $R^2_{Adj} \approx 0.57$. This is pretty low. Let's see what the following models generate.

```
summary(mod2)
```

```
##
## Call:
## lm(formula = Oxidant ~ Wind + Temperature, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3939 -1.8608  0.5826  1.9461  4.9661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.20334   11.11810  -0.468    0.644
## Wind        -0.42706    0.08645  -4.940 3.58e-05 ***
## Temperature  0.52035    0.10813   4.812 5.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.95 on 27 degrees of freedom
## Multiple R-squared:  0.7773, Adjusted R-squared:  0.7608
## F-statistic: 47.12 on 2 and 27 DF,  p-value: 1.563e-09
```

The above $R^2_{Adj} \approx 0.76$. This is better than the first model. So far this is the best of the top 3 models.

```
summary(mod3)
```

```
##
## Call:
## lm(formula = Oxidant ~ Wind + Temperature + Humidity, data = air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5887 -1.1686  0.1978  1.9004  4.1544
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.60697   13.07154  -1.270    0.215
## Wind         -0.44620    0.08513  -5.241 1.78e-05 ***
## Temperature   0.60190    0.11764   5.117 2.47e-05 ***
## Humidity      0.09850    0.06316   1.559    0.131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.874 on 26 degrees of freedom
## Multiple R-squared:  0.7964, Adjusted R-squared:  0.7729
## F-statistic: 33.89 on 3 and 26 DF,  p-value: 3.904e-09
```

The above $R^2_{Adj} \approx 0.77$. This does not differ much from the second model. The addition of humidity does not appear to be significant.

This means that the best of the three models from the regsubsets function is the second model. with Wind and Temperature included as predictor variables.