Multiple Regression Model to Predict Water Consumption

Ruth Kristianingsih

October 17, 2017

Introduction

A person spent a three-hour period outside, he recorded the temperature outside, the time he spent mowing the grass, and their water consumption. In this analysis, the goal is to analyze the dependence of water to the temperature and time of mowing the grass. And to predict or model the data using the linear regression and analyze the significance.

Data

The data below is the data of experiment in data frame, where the data of temperature is in F, water is in ounces, dan time is in hours period.

```
##   Temperature Water Time
## 1          75    16 1.85
## 2          83    20 1.25
## 3          85    25 1.50
## 4          85    27 1.75
## 5          92    32 1.15
## 6          97    48 1.75
## 7          99    48 1.60
```

General Statistical Analysis

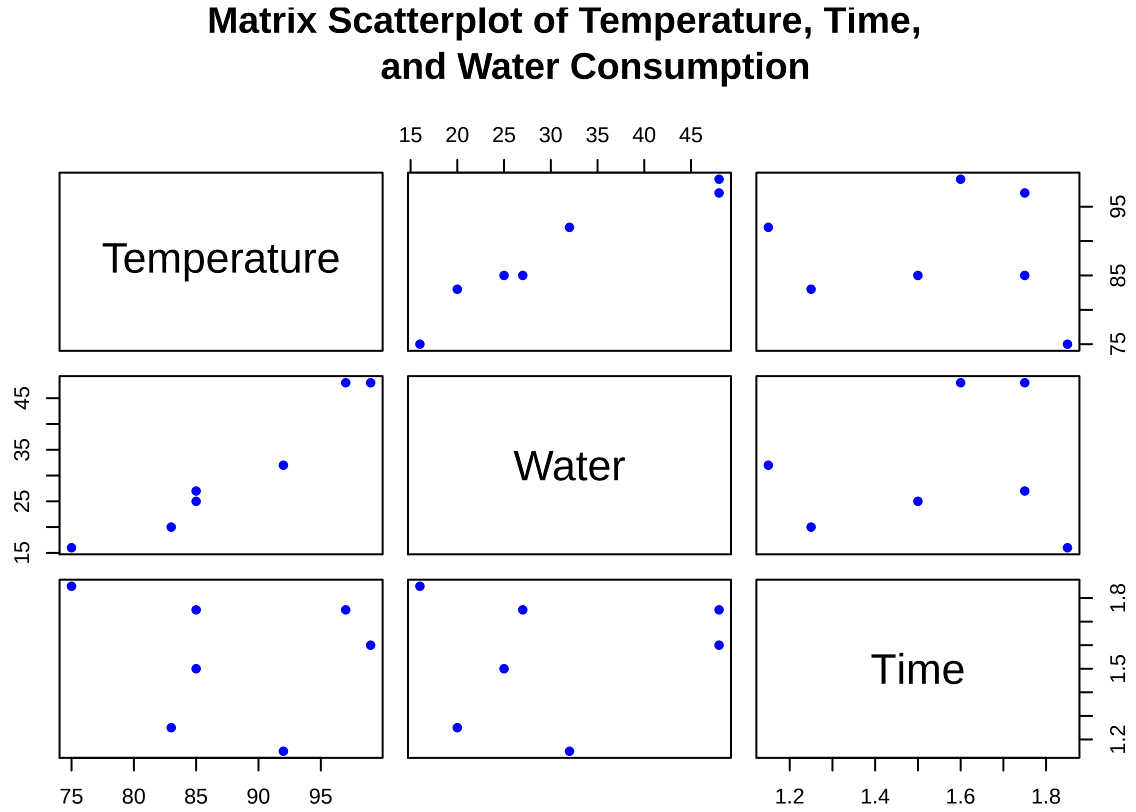
The general statistical analysis will be conducted, including the summary below. It is shown that the maximum value of temperature, water consumption, and time mowing grass are 99 F, 48 ounces, and 1.85 hours respectively. It can be seen also the minimum value, average value, and median of each variable of data.

```
summary(task2dat)
```

```
##   Temperature       Water            Time      
##  Min.   :75.0   Min.   :16.00   Min.   :1.150  
##  1st Qu.:84.0   1st Qu.:22.50   1st Qu.:1.375  
##  Median :85.0   Median :27.00   Median :1.600  
##  Mean   :88.0   Mean   :30.86   Mean   :1.550  
##  3rd Qu.:94.5   3rd Qu.:40.00   3rd Qu.:1.750  
##  Max.   :99.0   Max.   :48.00   Max.   :1.850
```

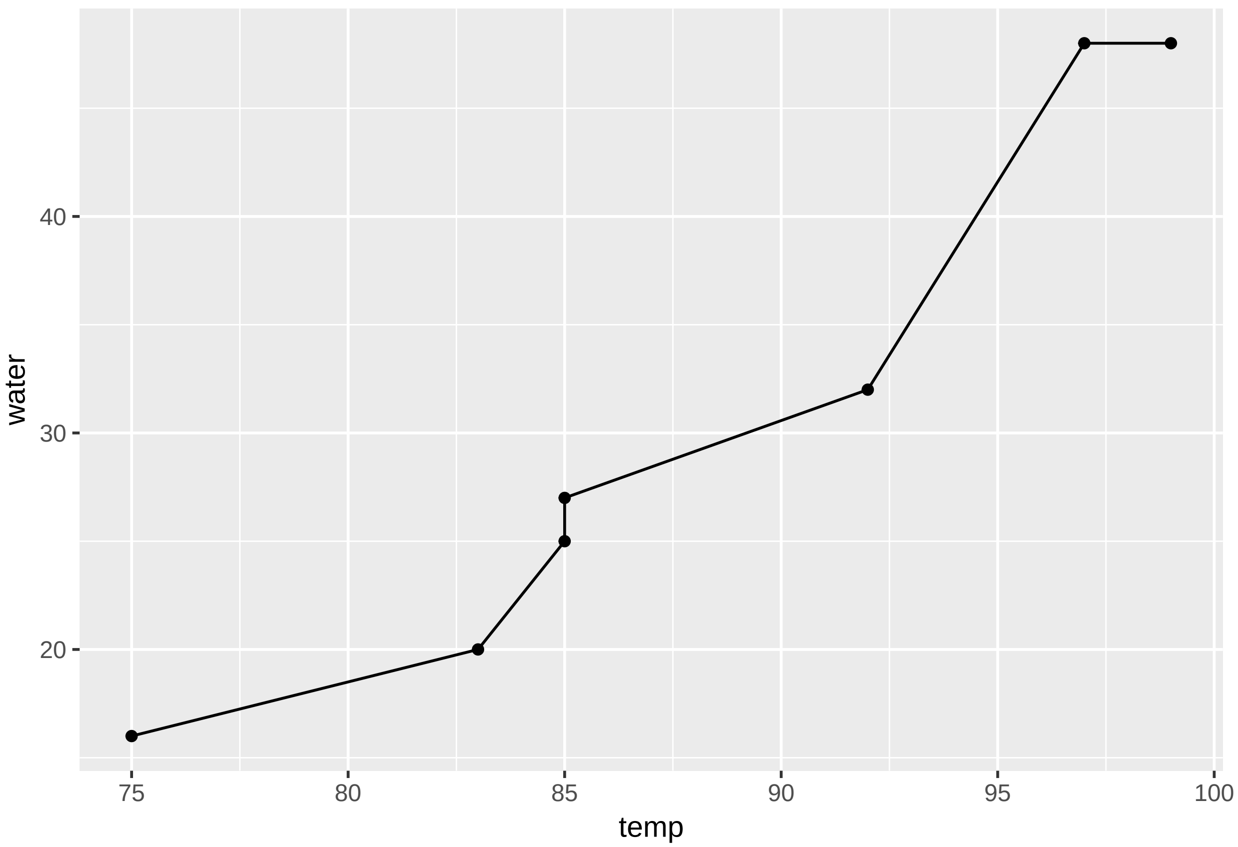
The following graph is the visualization of the relationship among the variables in single image. For instance, it can be seen how time mowing grass is related to water consumption. Another relationship that can be seen is the relationship between temperature and water, if the the temperature increase, the water consumption is getting higher.

```r
plot(task2dat, pch=16, col="blue", main="Matrix Scatterplot of Temperature, Time,
     and Water Consumption")
```



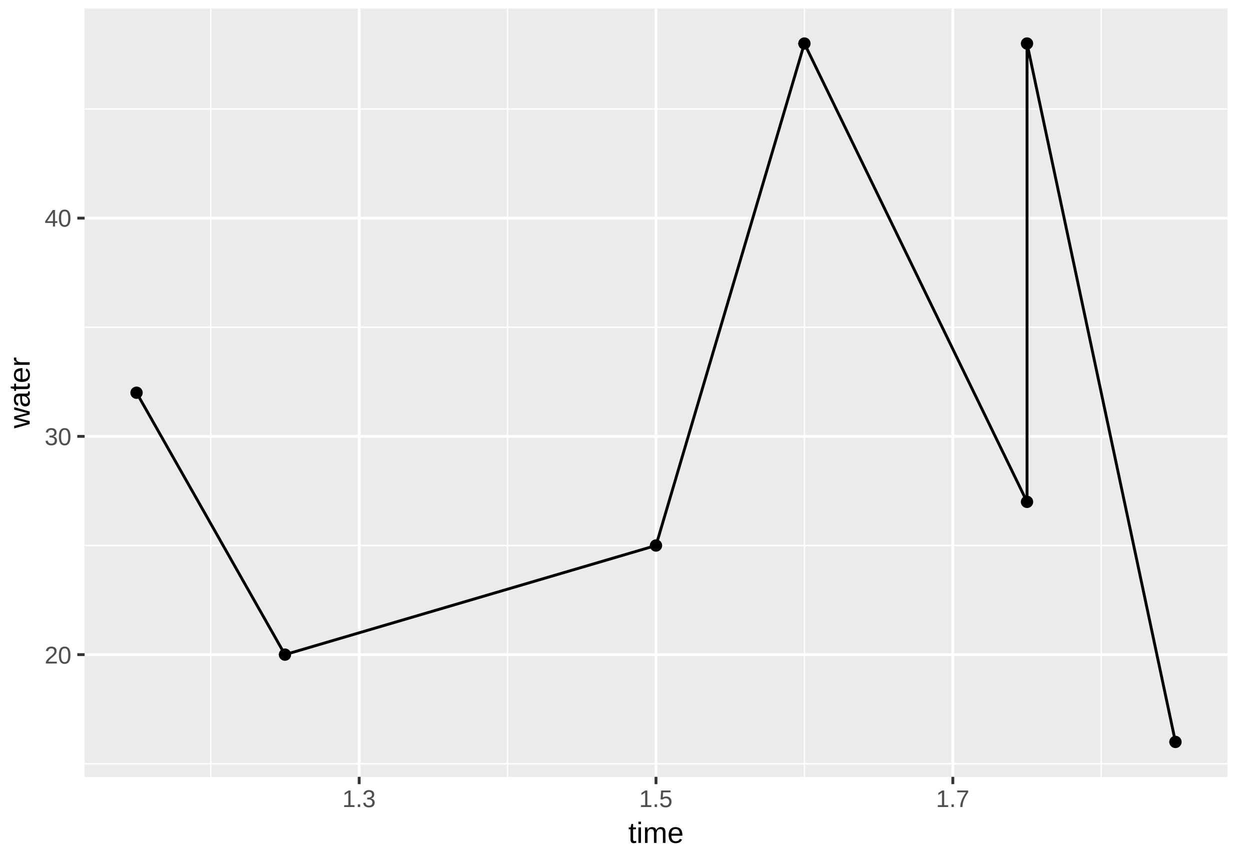
Now, using the graph below, it is clearly shown that temperature and water consumption has linear relationship.

```r
library(ggplot2)
ggplot(task2dat, aes(x=temp, y=water)) + geom_point() +geom_line()
```



However, the relationship between time and water does not seem to be a linear.

```
library(ggplot2)
ggplot(task2dat, aes(x=time, y=water)) + geom_point() +geom_line()
```



Model

Two model will be constructed using function lm (), the first model will consider the interaction between predictors, and the intercation will be removed in the second model.

1. Multiple Linear Regression with Interaction Between Time and Temperature

The first model is multiple linear regression with considering the interaction between temperature and time. The summarry shown below states that p-value for both time and the interaction between time and temperature are not less then alpha 0.5, as well as the p-value for temperature. Then analysis of significance using ANOVA is needed in order to analyze whether the interaction is significance, if it is not, then there is no interaction between time and temperature.

```
fit2 <- lm( task2dat$Water ~ task2dat$Temperature + task2dat$Time
            + task2dat$Temperature:task2dat$Time)
summary(fit2)
```

```
##
## Call:
## lm(formula = task2dat$Water ~ task2dat$Temperature + task2dat$Time +
##     task2dat$Temperature:task2dat$Time)
##
## Residuals:
##         1         2         3         4         5         6         7
##  0.96174  0.58085 -0.65018 -1.80965 -0.01499  1.08362 -0.15139
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -129.56029   56.02595  -2.313   0.1038
## task2dat$Temperature                  1.60298    0.64025   2.504   0.0874
## task2dat$Time                        17.20797   32.91635   0.523   0.6373
## task2dat$Temperature:task2dat$Time   -0.05377    0.37759  -0.142   0.8958
##
## (Intercept)
## task2dat$Temperature               .
## task2dat$Time
## task2dat$Temperature:task2dat$Time
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.433 on 3 degrees of freedom
## Multiple R-squared:  0.9937, Adjusted R-squared:  0.9874
## F-statistic: 157.7 on 3 and 3 DF,  p-value: 0.0008479
```

Analysis of Variance

Now, analysis of variance using ANOVA will be done. Below is the result of the test. With the results below, it can be seen that both predictor temperature and time are significance. Despite that, the interaction of time and temperature is not significance since the p-value of it is greater then 0.05. To verify the results of ANOVA which assumes the normality of the data, the exact permutation test using F-statistics will be done.

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: task2dat$Water
##                                    Df Sum Sq Mean Sq  F value   Pr(>F)
## task2dat$Temperature                1 905.53  905.53 441.1987 0.000236 ***
## task2dat$Time                       1  65.13   65.13  31.7346 0.011066 *
## task2dat$Temperature:task2dat$Time  1   0.04    0.04   0.0203 0.895796
## Residuals                           3   6.16    2.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Below is the results of exact permutation test using F-statistics.

```
library(combinat)
```

```
##
## Attaching package: 'combinat'

## The following object is masked from 'package:utils':
##
##     combn
```

```
fit2 <- lm( task2dat$Water ~ task2dat$Temperature + task2dat$Time
            + task2dat$Temperature:task2dat$Time)
b <- anova(fit2)
Ftimetrue <- b$"F value"[1]
Ftemptrue <- b$"F value"[2]
```

```
Fttinterac <- b$"F value"[3]
n=length(task2dat$Water)
nr = fact(n) #number of arrangement to be examined
Ftime=numeric(Ftimetrue)
Ftemp=numeric(Ftemptrue)
Ftt= numeric(Fttinterac)
for (i in 1:nr){
  newater <- sample(task2dat$Water,7)
  fit2 <- lm( newater ~ task2dat$Temperature + task2dat$Time
              + task2dat$Temperature:task2dat$Time)
  b <- anova(fit2)
  Ftime[i] <- b$"F value"[1]
  Ftemp[i] <- b$"F value"[2]
  Ftt[i] <- b$"F value"[3]
}
length(Ftime[Ftime>=Ftimetrue])/nr
```

```
## [1] 0.001587302
```

```
length(Ftemp[Ftemp>=Ftemptrue])/nr
```

```
## [1] 0.01071429
```

```
length(Ftt[Ftt>=Fttinterac])/nr
```

```
## [1] 0.8896825
```

Based on the results above, it is stated that the p-value for predictor time and temperature are consistent with the results obtained from ANOVA before which is the p-value for both predictors are less than 0.05. However, the p-value for the interaction between time and temperature is greater than 0.05, which means that it is not consistent and there is no interaction between time and temperature.

2. Multiple Linear Regression Without Interaction Between Time and Temperature

Now, an attempt to model or fit the data using multiple linear regression without considering the interaction between time and temperature will be done below.

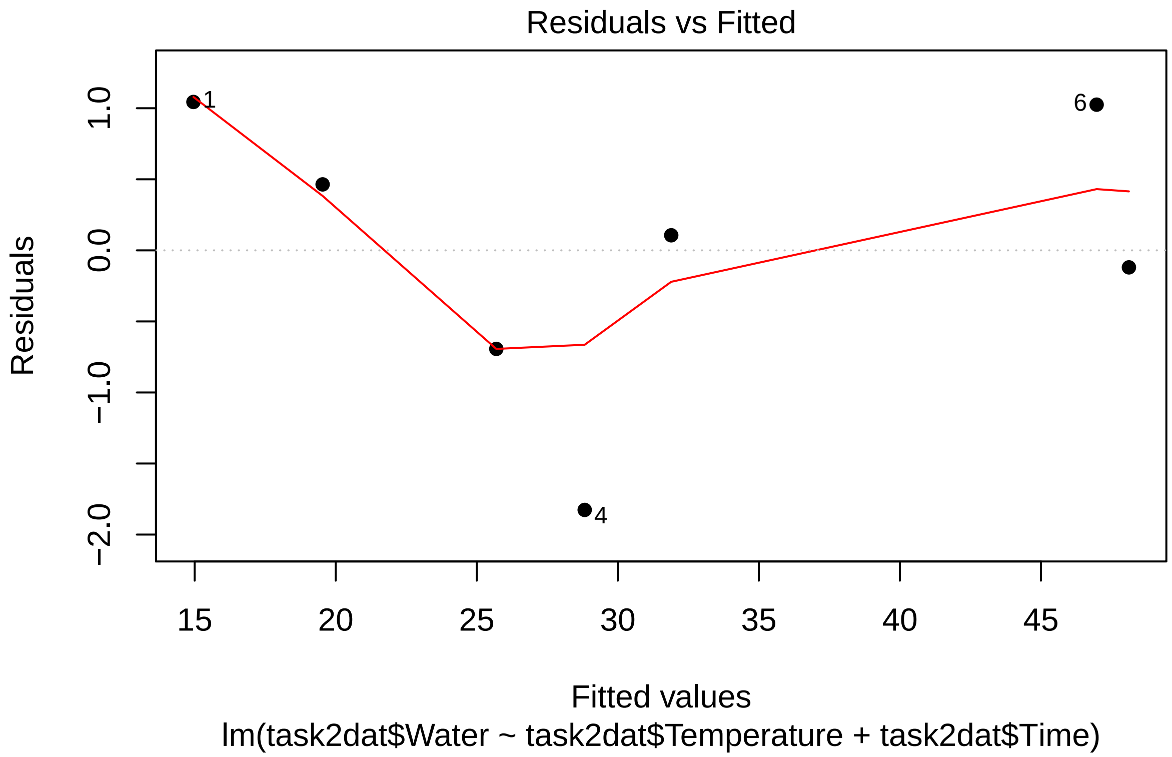
```
fit1 <- lm( task2dat$Water ~ task2dat$Temperature + task2dat$Time)
summary(fit1)
```

```
##
## Call:
## lm(formula = task2dat$Water ~ task2dat$Temperature + task2dat$Time)
##
## Residuals:
##       1       2       3       4       5       6       7
##  1.0441  0.4642 -0.6935 -1.8264  0.1061  1.0252 -0.1197
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -121.65500    6.54035 -18.601 4.92e-05 ***
## task2dat$Temperature    1.51236    0.06077  24.886 1.55e-05 ***
## task2dat$Time          12.53168    1.93302   6.483  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.245 on 4 degrees of freedom
## Multiple R-squared:  0.9937, Adjusted R-squared:  0.9905
## F-statistic: 313.2 on 2 and 4 DF,  p-value: 4.027e-05
```

From the results above, it is shown that both time and temperature are significance predictor for water consumption, as it is written in the summary that the p-value for both variables are less than 0.05.

Now, the residuals will be plotted in the following graph.

```r
plot(fit1, pch=16, which=1)
```



Analysis of Variance

Analysis of variance will be shown below, and from the results, it can be seen that the p-value for both temperature and time are significance (less than 0.05).

```r
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: task2dat$Water
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## task2dat$Temperature  1 905.53  905.53 584.316 1.737e-05 ***
## task2dat$Time         1  65.13   65.13  42.029  0.002918 **
## Residuals             4   6.20    1.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of exact permutation below also shows that both predictors significantly contribute to the model.

```r
library(combinat)
fit1 <- lm( task2dat$Water ~ task2dat$Temperature + task2dat$Time)
b <- anova(fit1)
Ftimetrue <- b$"F value"[1]
Ftemptrue <- b$"F value"[2]
n=length(task2dat$Water)
nr = fact(n) #number of arrangement to be examined
Ftime=numeric(Ftimetrue)
Ftemp=numeric(Ftemptrue)
for (i in 1:nr){
  newater <- sample(task2dat$Water,7)
  fit1 <- lm( newater ~ task2dat$Temperature + task2dat$Time)
  b <- anova(fit1)
  Ftime[i] <- b$"F value"[1]
  Ftemp[i] <- b$"F value"[2]
}
length(Ftime[Ftime>=Ftimetrue])/nr
```

```
## [1] 0.001984127
```

```r
length(Ftemp[Ftemp>=Ftemptrue])/nr
```

```
## [1] 0.01190476
```

And the model obtained above can be written as:

$$Water = -121.65 + 1.51 Temperature + 12.53 Time$$

Here below is the picture of the data and the prediction in 3D.

```r
library(rgl)
newdata <- expand.grid(temp=seq(70,100,by=5),time=seq(1.25,1.85,by=0.05))
newdata$pp <- round(predict(fit1, task2dat=newdata),2)

with(task2dat,plot3d(time,temp,water, col="blue", size=1, type="s", main="3D Linear Model Fit"))
with(newdata,surface3d(unique(time),unique(temp),pp,col=3,alpha=0.5))
rglwidget()
```

Conclusion

Since it is shown that the interaction between temperature and time should not be considered. Then the best model for the data is multiple linear regression without interaction between predictors. (The second model)