Assignment 2

79546 - Stephen K. Ng'etich

Contents

1 Pre-requisite

1.1 Load packages

```
# Clear variables
rm(list=ls())

library(readxl)
library(dplyr)
library(tidyverse)
library(lattice)
library(leaps)
library(MASS)
```

1.2 Load dataset

```
# Load Dataset
dataset <- read_excel("dataset/Dataset2.xlsx")
```

2 Exploratory Data Analysis

Convert `Region` column data type to `factor`

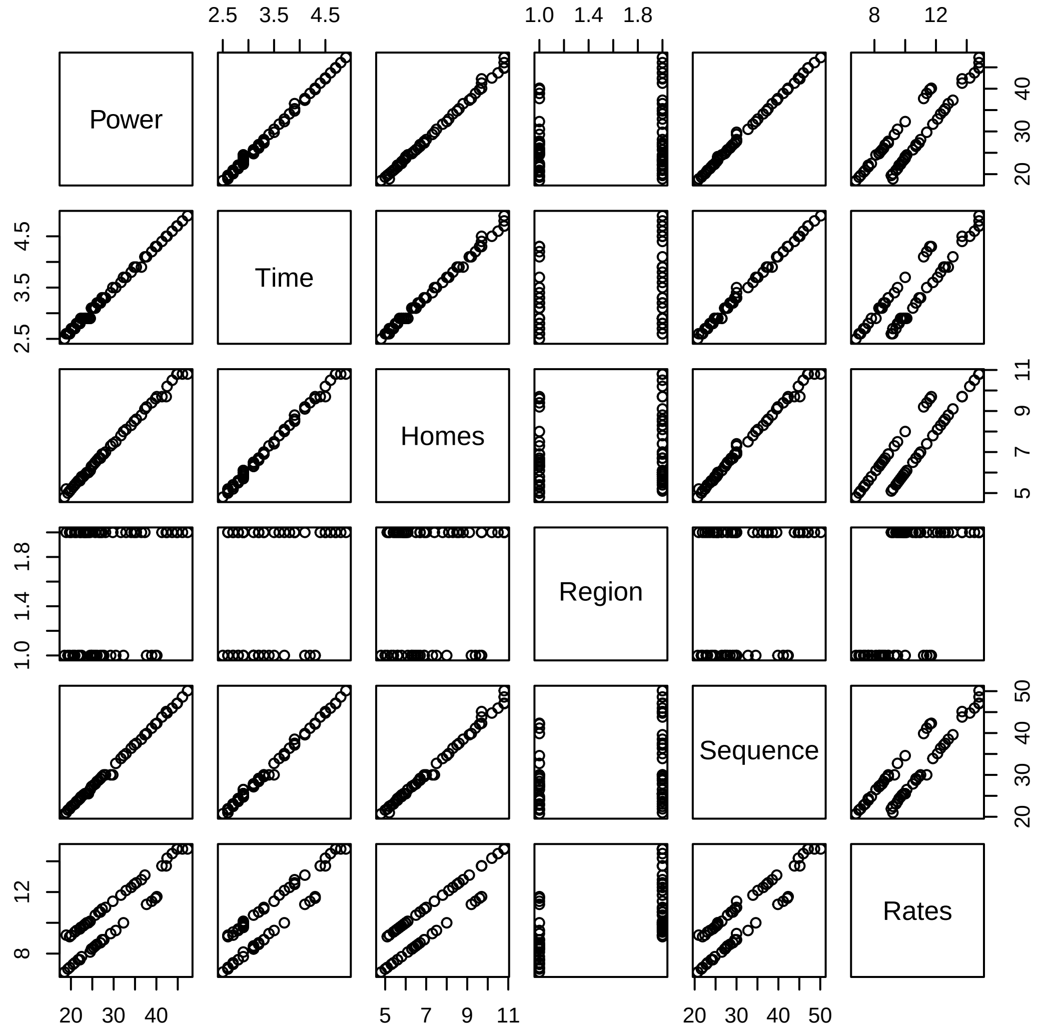
```
dataset$Region <- as.factor(dataset$Region)
```

2.1 Summary statistic

```
summary(dataset)
```

```
##      Power            Time           Homes         Region    Sequence    
##  Min.   :18.50   Min.   :2.500   Min.   : 4.800   1:24   Min.   :20.72  
##  1st Qu.:22.60   1st Qu.:2.900   1st Qu.: 5.800   2:34   1st Qu.:24.92  
##  Median :26.70   Median :3.200   Median : 6.700          Median :28.93  
##  Mean   :29.21   Mean   :3.405   Mean   : 7.226          Mean   :31.39  
##  3rd Qu.:35.17   3rd Qu.:3.900   3rd Qu.: 8.575          3rd Qu.:37.44  
##  Max.   :47.30   Max.   :4.900   Max.   :10.800          Max.   :50.12  
##      Rates      
##  Min.   : 6.80  
##  1st Qu.: 8.75  
##  Median : 9.95  
##  Mean   :10.40  
##  3rd Qu.:11.78  
##  Max.   :14.80  
```

```
plot(dataset)
```



3 Model Selection

3.1 Build the linear model

```
power_lm_model = lm(Power~ .,data=dataset)
summary(power_lm_model)
```

```
##
## Call:
## lm(formula = Power ~ ., data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47230 -0.17587 -0.05152  0.08181  0.91553
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.46083    0.75232  -4.600 2.66e-05 ***
## Time         0.98145    0.96223   1.020    0.312
## Homes        1.70436    0.29075   5.862 3.00e-07 ***
## Region2      0.08236    0.07156   1.151    0.255
## Sequence     0.54034    0.06563   8.233 4.76e-11 ***
## Rates             NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2609 on 53 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.999
## F-statistic: 1.424e+04 on 4 and 53 DF,  p-value: < 2.2e-16
```

From the summary, Homes show coefficient of Rates are NA this means that it does not add any information to the model.

```
# Remove Rates from the predictor variables
dataset1 = subset(dataset,select = -c(Rates))

power_lm_model =  lm(Power~ ., data=dataset1)
summary(power_lm_model)
```

```
##
## Call:
## lm(formula = Power ~ ., data = dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47230 -0.17587 -0.05152  0.08181  0.91553
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.46083    0.75232  -4.600 2.66e-05 ***
## Time         0.98145    0.96223   1.020    0.312
## Homes        1.70436    0.29075   5.862 3.00e-07 ***
## Region2      0.08236    0.07156   1.151    0.255
## Sequence     0.54034    0.06563   8.233 4.76e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2609 on 53 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.999
## F-statistic: 1.424e+04 on 4 and 53 DF,  p-value: < 2.2e-16
```

The following stepwise regression methods are used for model selection

1. Forward
2. Backward
3. Both

3.2 Forward Selection

```
modelfoward =  stepAIC(power_lm_model,direction="forward",trace = TRUE)
```

```
## Start:  AIC=-151.08
## Power ~ Time + Homes + Region + Sequence
```

```
step(modelfoward)
```

```
## Start:  AIC=-151.08
## Power ~ Time + Homes + Region + Sequence
##
##            Df Sum of Sq    RSS     AIC
## - Time      1    0.0708 3.6790 -151.95
## - Region    1    0.0902 3.6983 -151.65
## <none>                  3.6082 -151.08
## - Homes     1    2.3394 5.9475 -124.09
## - Sequence  1    4.6141 8.2223 -105.31
##
## Step:  AIC=-151.95
## Power ~ Homes + Region + Sequence
##
##            Df Sum of Sq     RSS      AIC
## - Region    1    0.0847  3.7637 -152.632
## <none>                   3.6790 -151.953
## - Homes     1    4.3768  8.0558 -108.495
## - Sequence  1    8.9427 12.6217  -82.451
##
## Step:  AIC=-152.63
## Power ~ Homes + Sequence
##
##            Df Sum of Sq     RSS      AIC
## <none>                   3.7637 -152.632
## - Homes     1    4.3297  8.0934 -110.225
## - Sequence  1    9.0959 12.8596  -83.368
```

```
##
## Call:
## lm(formula = Power ~ Homes + Sequence, data = dataset1)
##
## Coefficients:
## (Intercept)        Homes     Sequence
##     -2.6955       1.8676       0.5864
```

From forward selection of the optimal model is

$$Power = Homes + Sequence$$

with the lowest *AIC of -152.63*

3.3 Backward Selection

```
modelfoward =  stepAIC(power_lm_model,direction="backward",trace = TRUE)
```

```
## Start:  AIC=-151.08
## Power ~ Time + Homes + Region + Sequence
## 
##            Df Sum of Sq    RSS     AIC
## - Time      1    0.0708 3.6790 -151.95
## - Region    1    0.0902 3.6983 -151.65
## <none>                  3.6082 -151.08
## - Homes     1    2.3394 5.9475 -124.09
## - Sequence  1    4.6141 8.2223 -105.31
## 
## Step:  AIC=-151.95
## Power ~ Homes + Region + Sequence
## 
##            Df Sum of Sq     RSS      AIC
## - Region    1    0.0847  3.7637 -152.632
## <none>                   3.6790 -151.953
## - Homes     1    4.3768  8.0558 -108.495
## - Sequence  1    8.9427 12.6217  -82.451
## 
## Step:  AIC=-152.63
## Power ~ Homes + Sequence
## 
##            Df Sum of Sq     RSS      AIC
## <none>                   3.7637 -152.632
## - Homes     1    4.3297  8.0934 -110.225
## - Sequence  1    9.0959 12.8596  -83.368
```

From backward selection,the optimal model is

$$Power = Homes + Sequence$$

with the lowest *AIC of -152.63*

```
modelfoward =  stepAIC(power_lm_model,direction="both",trace = TRUE)
```

```
## Start:  AIC=-151.08
## Power ~ Time + Homes + Region + Sequence
## 
##            Df Sum of Sq    RSS     AIC
## - Time      1    0.0708 3.6790 -151.95
## - Region    1    0.0902 3.6983 -151.65
## <none>                  3.6082 -151.08
## - Homes     1    2.3394 5.9475 -124.09
## - Sequence  1    4.6141 8.2223 -105.31
## 
## Step:  AIC=-151.95
## Power ~ Homes + Region + Sequence
## 
```

```
##            Df Sum of Sq     RSS      AIC
## - Region    1    0.0847  3.7637 -152.632
## <none>                   3.6790 -151.953
## + Time      1    0.0708  3.6082 -151.080
## - Homes     1    4.3768  8.0558 -108.495
## - Sequence  1    8.9427 12.6217  -82.451
##
## Step:  AIC=-152.63
## Power ~ Homes + Sequence
##
##            Df Sum of Sq     RSS      AIC
## <none>                   3.7637 -152.632
## + Region    1    0.0847  3.6790 -151.953
## + Time      1    0.0654  3.6983 -151.648
## - Homes     1    4.3297  8.0934 -110.225
## - Sequence  1    9.0959 12.8596  -83.368
```

```
step(modelfoward)
```

```
## Start:  AIC=-152.63
## Power ~ Homes + Sequence
##
##            Df Sum of Sq     RSS      AIC
## <none>                   3.7637 -152.632
## - Homes     1    4.3297  8.0934 -110.225
## - Sequence  1    9.0959 12.8596  -83.368
```

```
##
## Call:
## lm(formula = Power ~ Homes + Sequence, data = dataset1)
##
## Coefficients:
## (Intercept)        Homes     Sequence
##     -2.6955       1.8676       0.5864
```

From both forward and backward selection,the optimal model is

$$Power = 1.8676Homes + 0.5864Sequence - 2.6955$$

with the lowest *AIC of -152.63*