Assignment 2 - Linear Models

Getrude Gichuhi

2022-04-08

Use the dataset attached to do model selection.

Use R Markdown for your submissions.

Ensure you change the variable Region to factor variable before model fitting.

Use the variable name power as your response variable and select the best model using AIC

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

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.1.3
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.1.2

## Warning: package 'readr' was built under R version 4.1.3
## Warning: package 'forcats' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lattice)
```

Loading the DataSet

```
df <- read_excel("Dataset2.xlsx")
print(head(df))
```

```
## # A tibble: 6 x 6
##   Power  Time Homes Region Sequence Rates
##   <dbl> <dbl> <dbl>  <dbl>    <dbl> <dbl>
## 1  18.5   2.5   4.8      1     20.7   6.8
## 2  18.9   2.6   5.2      2     21.0   9.2
## 3  19.3   2.6   5        1     21.6   7  
## 4  19.7   2.6   5.1      2     21.9   9.1
## 5  19.7   2.6   5.1      1     21.9   7.1
## 6  20.1   2.7   5.2      2     22.5   9.2
```

```
summary(df)
```

```
##      Power            Time           Homes            Region     
##  Min.   :18.50   Min.   :2.500   Min.   : 4.800   Min.   :1.000  
##  1st Qu.:22.60   1st Qu.:2.900   1st Qu.: 5.800   1st Qu.:1.000  
##  Median :26.70   Median :3.200   Median : 6.700   Median :2.000  
##  Mean   :29.21   Mean   :3.405   Mean   : 7.226   Mean   :1.586  
##  3rd Qu.:35.17   3rd Qu.:3.900   3rd Qu.: 8.575   3rd Qu.:2.000  
##  Max.   :47.30   Max.   :4.900   Max.   :10.800   Max.   :2.000  
##     Sequence         Rates      
##  Min.   :20.72   Min.   : 6.80  
##  1st Qu.:24.92   1st Qu.: 8.75  
##  Median :28.93   Median : 9.95  
##  Mean   :31.39   Mean   :10.40  
##  3rd Qu.:37.44   3rd Qu.:11.78  
##  Max.   :50.12   Max.   :14.80  
```

Change the Region Variable to the factor variable

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

```
summary (df)
```

```
##      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
```

Build a Model

```
lm_power = lm(Power~ .,data=df)
summary(lm_power)
```

```
##
## Call:
## lm(formula = Power ~ ., data = df)
##
## 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
```

Rates shows no estimates or statistics therefore it’s wise to remove it.

```
df1 = subset(df,select = -c(Rates))

lm_power =  lm(Power~ ., data=df1)
summary(lm_power)
```

```
##
## Call:
## lm(formula = Power ~ ., data = df1)
##
## 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
```

Model selection of AIC 1. Forward Selection 2. Backward Selection 3. Mixed Selection

```
step(lm_power, direction = "forward")
```

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

```
##
## Call:
## lm(formula = Power ~ Time + Homes + Region + Sequence, data = df1)
##
## Coefficients:
## (Intercept)         Time        Homes      Region2     Sequence
##    -3.46083      0.98145      1.70436      0.08236      0.54034
```

The forward stepwise selection shows a model of AIC = -151.08 which includes Power ~ Time + Homes + Region + Sequence. Power as the response Variable.

```
step(lm_power, direction = "backward")
```

```
## 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 = df1)
##
## Coefficients:
## (Intercept)        Homes     Sequence
##     -2.6955       1.8676       0.5864
```

The Backward Stepwise selection shows a model of AIC = -152.63 which includes Power ~ Homes + Sequence.

```
step(lm_power, direction = "both")
```

```
## 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
```

```
##
## Call:
## lm(formula = Power ~ Homes + Sequence, data = df1)
##
## Coefficients:
## (Intercept)        Homes     Sequence
##     -2.6955       1.8676       0.5864
```

The mixed selcetion shows an AIC of -152.63.

The best model to use is the forward selection which has an AIC of -151.08