Linear Regression

Cassidy White

4/6/2022

```
getwd()
```

```
## [1] "C:/Users/Katherine/Documents/872-Data Analytics/KinserOwensWhite_ENV872_EDA_FinalProject/Code"
```

```
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  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

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

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

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

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

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

```
## Warning: package 'stringr' 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(agricolae)
```

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

```
library(corrplot)
```

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

```
## corrplot 0.92 loaded
```

```
library(splitstackshape)
```

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

```
library(matrixStats)
```

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

```
##
## Attaching package: 'matrixStats'
```

```
## The following object is masked from 'package:dplyr':
##
##     count
```

```
#install.packages("gt")
library(gt)
```

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

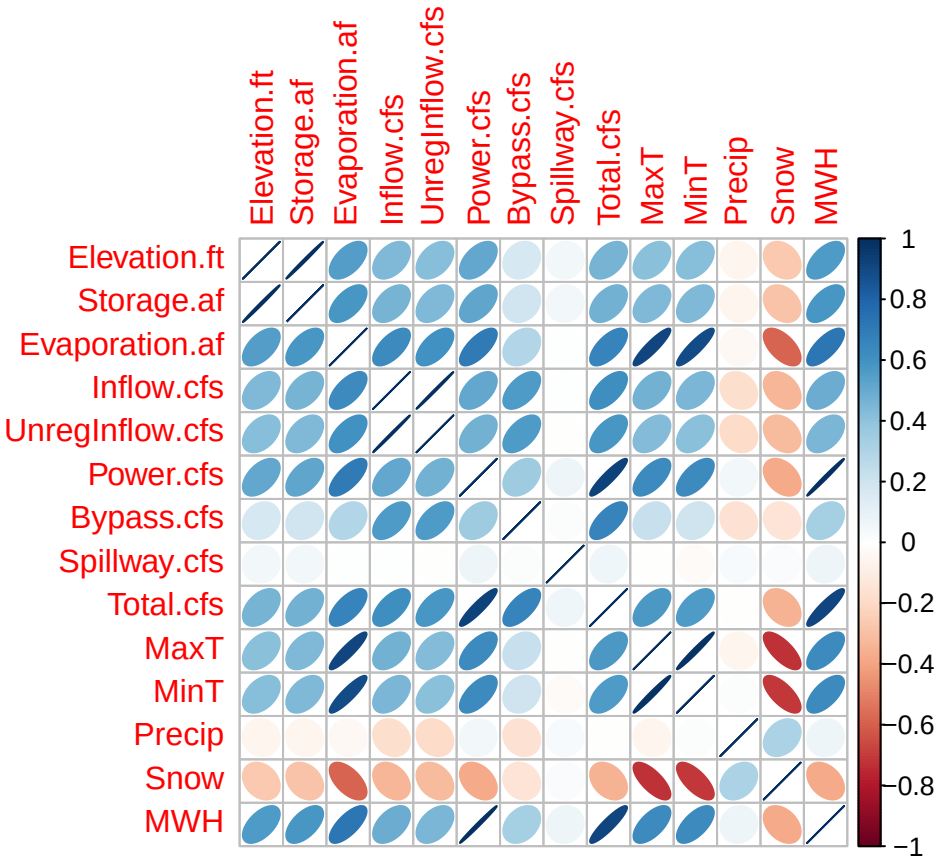
```
library(dplyr)
```

```
getwd()
```

```
## [1] "C:/Users/Katherine/Documents/872-Data Analytics/KinserOwensWhite_ENV872_EDA_FinalProject/Code"
```

```
all.data<-read.csv("../Data/Processed/AllData.csv")
all.data$Date<-as.Date(all.data$Date, format = "%Y-%d-%m")
all.data<-all.data %>%
  na.omit()
all.data.nodate<-select(all.data, -c(X, Date))
```

```
all.data.corr<-cor(all.data.nodate)
corrplot(all.data.corr, method = "ellipse")
```



Elevation (ft) is highly correlated with storage (af). Inflow (cfs) is highly correlated with unregulated inflow (cfs). Total flow (cfs) is somewhat correlated to power flow (cfs). MaxT and MinT are somewhat correlated

with evaporation. MWH is highly correlated with Power.cfs and Total.cfs. MinT and MaxT are highly correlated with each other.

```
#Run a regression with all variables to take a first look
regression.all<-lm(data = all.data, MWH ~ Elevation.ft + Storage.af + Evaporation.af + Inflow.cfs + Unr

summary(regression.all)
```

```
##
## Call:
## lm(formula = MWH ~ Elevation.ft + Storage.af + Evaporation.af +
##     Inflow.cfs + UnregInflow.cfs + Bypass.cfs + Spillway.cfs +
##     MaxT + MinT + Precip + Snow + Total.cfs + Power.cfs, data = all.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2043.45  -522.37   -38.51   505.64  1799.81
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.904e+06  2.862e+05   6.654 3.67e-10 ***
## Elevation.ft    -2.576e+02  3.864e+01  -6.667 3.42e-10 ***
## Storage.af       4.343e-02  5.452e-03   7.965 2.16e-13 ***
## Evaporation.af   1.362e+00  4.628e-01   2.942  0.00371 **
## Inflow.cfs      -1.150e+00  9.911e-01  -1.161  0.24735
## UnregInflow.cfs  4.275e-01  8.867e-01   0.482  0.63035
## Bypass.cfs      -4.946e-01  8.766e-01  -0.564  0.57332
## Spillway.cfs    -2.872e-07  3.545e-07  -0.810  0.41898
## MaxT            -1.641e+01  1.804e+01  -0.910  0.36427
## MinT            -6.669e-01  2.006e+01  -0.033  0.97351
## Precip           1.607e+02  1.104e+02   1.456  0.14725
## Snow             2.215e+00  1.550e+01   0.143  0.88653
## Total.cfs        7.466e-01  6.848e-01   1.090  0.27709
## Power.cfs        1.699e+01  7.298e-01  23.283  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 743.4 on 172 degrees of freedom
## Multiple R-squared:  0.9957, Adjusted R-squared:  0.9953
## F-statistic:  3045 on 13 and 172 DF,  p-value: < 2.2e-16
```

```
#Run a stepwise AIC test to find the model of best fit from all the above variables
step(regression.all)
```

```
## Start:  AIC=2472.8
## MWH ~ Elevation.ft + Storage.af + Evaporation.af + Inflow.cfs +
##     UnregInflow.cfs + Bypass.cfs + Spillway.cfs + MaxT + MinT +
##     Precip + Snow + Total.cfs + Power.cfs
##
##                   Df Sum of Sq       RSS    AIC
## - MinT             1       611  95043712 2470.8
## - Snow             1     11285  95054386 2470.8
## - UnregInflow.cfs  1    128429  95171530 2471.1
## - Bypass.cfs       1    175927  95219028 2471.2
## - Spillway.cfs     1    362662  95405763 2471.5
## - MaxT             1    457259  95500360 2471.7
```

```
## - Total.cfs        1    656909  95700010 2472.1
## - Inflow.cfs       1    744522  95787623 2472.2
## <none>                          95043101 2472.8
## - Precip           1   1171244  96214345 2473.1
## - Evaporation.af   1   4782614  99825715 2479.9
## - Elevation.ft     1  24558350 119601450 2513.6
## - Storage.af       1  35057306 130100407 2529.2
## - Power.cfs        1 299548840 394591941 2735.6
##
## Step:  AIC=2470.8
## MWH ~ Elevation.ft + Storage.af + Evaporation.af + Inflow.cfs +
##     UnregInflow.cfs + Bypass.cfs + Spillway.cfs + MaxT + Precip +
##     Snow + Total.cfs + Power.cfs
##
##                   Df Sum of Sq       RSS    AIC
## - Snow             1     11901  95055613 2468.8
## - UnregInflow.cfs  1    137608  95181320 2469.1
## - Bypass.cfs       1    175584  95219296 2469.2
## - Spillway.cfs     1    366085  95409797 2469.5
## - Total.cfs        1    657652  95701364 2470.1
## - Inflow.cfs       1    785385  95829097 2470.3
## <none>                          95043712 2470.8
## - Precip           1   1346592  96390303 2471.4
## - MaxT             1   2032271  97075983 2472.7
## - Evaporation.af   1   4820082  99863794 2478.0
## - Elevation.ft     1  24571801 119615513 2511.6
## - Storage.af       1  35062815 130106527 2527.2
## - Power.cfs        1 299555938 394599650 2733.6
##
## Step:  AIC=2468.83
## MWH ~ Elevation.ft + Storage.af + Evaporation.af + Inflow.cfs +
##     UnregInflow.cfs + Bypass.cfs + Spillway.cfs + MaxT + Precip +
##     Total.cfs + Power.cfs
##
##                   Df Sum of Sq       RSS    AIC
## - UnregInflow.cfs  1    144939  95200552 2467.1
## - Bypass.cfs       1    167204  95222817 2467.2
## - Spillway.cfs     1    365021  95420633 2467.5
## - Total.cfs        1    646703  95702315 2468.1
## - Inflow.cfs       1    808204  95863816 2468.4
## <none>                          95055613 2468.8
## - Precip           1   1707740  96763353 2470.1
## - MaxT             1   3319637  98375250 2473.2
## - Evaporation.af   1   5312647 100368260 2476.9
## - Elevation.ft     1  24563188 119618800 2509.6
## - Storage.af       1  35051072 130106685 2525.2
## - Power.cfs        1 303010036 398065648 2733.2
##
## Step:  AIC=2467.11
## MWH ~ Elevation.ft + Storage.af + Evaporation.af + Inflow.cfs +
##     Bypass.cfs + Spillway.cfs + MaxT + Precip + Total.cfs + Power.cfs
##
##                   Df Sum of Sq       RSS    AIC
## - Bypass.cfs       1    214906  95415458 2465.5
```

```
## - Spillway.cfs    1    376914  95577466 2465.8
## - Total.cfs       1    730928  95931480 2466.5
## <none>                         95200552 2467.1
## - Precip          1   1663313  96863864 2468.3
## - MaxT            1   3236259  98436811 2471.3
## - Evaporation.af  1   5233580 100434132 2475.1
## - Elevation.ft    1  24438343 119638895 2507.6
## - Storage.af      1  34947837 130148389 2523.3
## - Inflow.cfs      1  51484674 146685225 2545.5
## - Power.cfs       1 313038877 408239428 2735.9
##
## Step:  AIC=2465.53
## MWH ~ Elevation.ft + Storage.af + Evaporation.af + Inflow.cfs +
##     Spillway.cfs + MaxT + Precip + Total.cfs + Power.cfs
##
##                   Df  Sum of Sq        RSS    AIC
## - Spillway.cfs     1     379336   95794794 2464.3
## <none>                           95415458 2465.5
## - Total.cfs        1    1727725   97143183 2466.9
## - Precip           1    1769934   97185392 2466.9
## - MaxT             1    3545521   98960979 2470.3
## - Evaporation.af   1    5639155  101054613 2474.2
## - Elevation.ft     1   24430610  119846068 2505.9
## - Storage.af       1   35188038  130603496 2521.9
## - Inflow.cfs       1   55130057  150545515 2548.3
## - Power.cfs        1 2052378882 2147794340 3042.7
##
## Step:  AIC=2464.27
## MWH ~ Elevation.ft + Storage.af + Evaporation.af + Inflow.cfs +
##     MaxT + Precip + Total.cfs + Power.cfs
##
##                   Df  Sum of Sq        RSS    AIC
## <none>                           95794794 2464.3
## - Total.cfs        1    1732353   97527147 2465.6
## - Precip           1    1739369   97534163 2465.6
## - MaxT             1    3533687   99328481 2469.0
## - Evaporation.af   1    5772302  101567096 2473.2
## - Elevation.ft     1   24097758  119892552 2504.0
## - Storage.af       1   34820401  130615194 2519.9
## - Inflow.cfs       1   54982206  150777000 2546.6
## - Power.cfs        1 2054500628 2150295422 3040.9

##
## Call:
## lm(formula = MWH ~ Elevation.ft + Storage.af + Evaporation.af +
##     Inflow.cfs + MaxT + Precip + Total.cfs + Power.cfs, data = all.data)
##
## Coefficients:
##    (Intercept)    Elevation.ft      Storage.af  Evaporation.af      Inflow.cfs
##      1.837e+06      -2.485e+02       4.212e-02       1.370e+00      -6.834e-01
##           MaxT          Precip       Total.cfs       Power.cfs
##     -1.802e+01       1.659e+02       3.784e-01       1.732e+01
```

```
#Choose the model of best fit from the AIC test and run below
regression.final <- lm(data = all.data, MWH ~ Elevation.ft + Storage.af + Evaporation.af +
```

```
    Inflow.cfs + MaxT + Precip + Total.cfs + Power.cfs)

summary(regression.final)
```

```
##
## Call:
## lm(formula = MWH ~ Elevation.ft + Storage.af + Evaporation.af +
##     Inflow.cfs + MaxT + Precip + Total.cfs + Power.cfs, data = all.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2129.80  -498.26   -14.98   499.99  1799.74
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.837e+06  2.759e+05   6.660 3.34e-10 ***
## Elevation.ft   -2.485e+02  3.724e+01  -6.673 3.11e-10 ***
## Storage.af      4.212e-02  5.251e-03   8.021 1.38e-13 ***
## Evaporation.af  1.370e+00  4.195e-01   3.266  0.00131 **
## Inflow.cfs     -6.834e-01  6.780e-02 -10.079  < 2e-16 ***
## MaxT           -1.802e+01  7.051e+00  -2.555  0.01145 *
## Precip          1.659e+02  9.255e+01   1.793  0.07473 .
## Total.cfs       3.784e-01  2.115e-01   1.789  0.07531 .
## Power.cfs       1.732e+01  2.812e-01  61.612  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 735.7 on 177 degrees of freedom
## Multiple R-squared:  0.9956, Adjusted R-squared:  0.9954
## F-statistic:  5052 on 8 and 177 DF,  p-value: < 2.2e-16
```

```
#write.csv(as.data.frame(summary(regression.final)$coef), file = "./Output/RegressionTable.csv")
```

```
all.data.nospill <-
  select(all.data.nodate, -c(Spillway.cfs)) %>%
  rename("Max Temperature" = MaxT,
         "Min Temperature" = MinT,
         "Precipitation" = Precip,
         "Snowfall" = Snow,
         "Electricity Generation.MWh" = MWH)

summary.table<-all.data.nospill %>%
  summary()
summary.table<-as.data.frame(summary.table) %>% cSplit("Freq", sep = ":", type.convert = FALSE)
summary.table<-summary.table %>%
  select(Var2, Freq_1, Freq_2)
summary.table<-pivot_wider(summary.table, names_from = Freq_1, values_from = Freq_2)
summary.table<-select(summary.table, c(Var2, Mean, Min., Max.))

sd<-colSds(as.matrix(all.data.nospill[sapply(all.data.nospill, is.numeric)]))
sd<-as.data.frame(sd)
summary.table<-cbind(summary.table, sd)

names(summary.table)[names(summary.table)=='Var2']<-'Variable'
```

```
#write.csv(summary.table, file = "./Output/SummaryTable.csv")
```

```
summary.table$Mean <- as.numeric(summary.table$Mean)
summary.table$Max. <- as.numeric(summary.table$Max.)
summary.table$Min. <- as.numeric(summary.table$Min.)
summary.table$sd <- as.numeric(summary.table$sd)

gt(summary.table) %>%
    tab_header(title = "Blue Mesa Reservoir Summary Statistics") %>%
    fmt_number(
      columns = c(Mean, Min., Max., sd), decimals = 2)
```

Blue Mesa Reservoir Summary Statistics

Variable	Mean	Min.	Max.	sd
Elevation.ft	7,486.00	7,438.00	7,519.00	18.51
Storage.af	557,251.00	247,684.00	826,302.00	134,507.54
Evaporation.af	648.60	114.00	1,544.00	443.68
Inflow.cfs	1,139.70	258.00	7,456.00	1,278.05
UnregInflow.cfs	1,137.00	195.00	7,915.00	1,366.55
Power.cfs	1,078.20	186.00	3,118.00	588.99
Bypass.cfs	58.88	0.00	2,379.00	252.86
Total.cfs	1,146.30	186.00	5,939.00	761.36
Max Temperature	55.13	13.40	87.20	20.75
Min Temperature	23.63	−11.80	50.80	16.76
Precipitation	0.73	0.00	4.39	0.61
Snowfall	3.91	0.00	32.90	5.86
Electricity Generation.MWh	18,856.00	2,724.00	54,068.00	10,896.96