Linear Regression

Nikhil Muthukrishnan

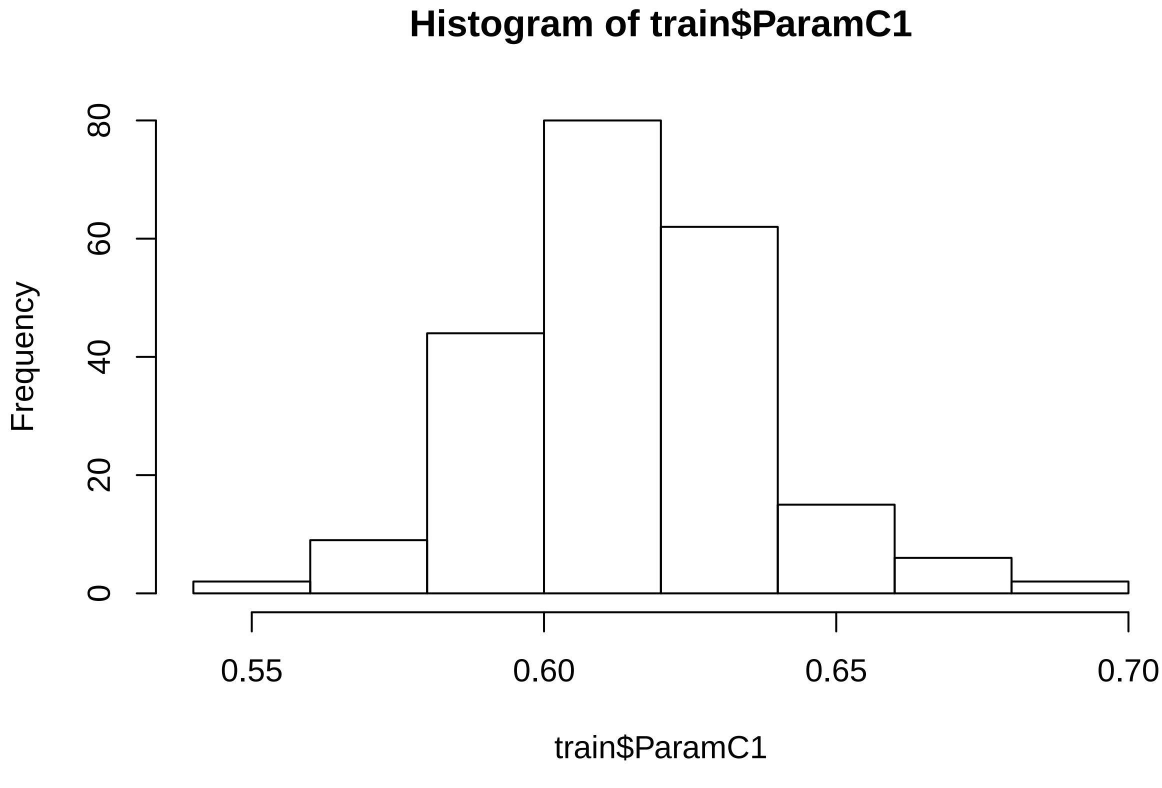
28 February 2019

Objective:Fit General Linear Model on Line2 Material2 of aggregated 2017 process variables to predict ABD value(ParamC1) from PI tags from different phases.

The fully assembeled dataset contains Time,Batch Number,Process Order,PI tags,Grab Samples,Weather and batch duration.

```
##                 df.t1 df.BatchNumber df.hours df.ParamC1
## 1 2017-02-07 00:00:00     3022002216       12      0.601
## 2 2017-02-07 00:01:00     3022002216       12      0.601
##  [ reached getOption("max.print") -- omitted 4 rows ]
```

The below data is aggregated by batchnumber and subsetted for only PI tags and ABD value which will be modelled.



The below glm summary shows the Adjusted R squared of .35 and other metrics of the all batches and all PI tags with no transformation.

```
## [1] 0.3543567047
```

```
##          RMSE      Rsquared           MAE
## 0.01640757610 0.46933427782 0.01258959536
```

Using step regression the number of variables were reduced to 22 and adjusted r square incerased to 0.38 indicating that some noise has been filtered.

```
## [1] 0.3894266278
```

```
##          RMSE      Rsquared           MAE
## 0.01669221212 0.45076276566 0.01286337268
```

Apart from removing variables that dont add information to the model further data transformations can also be applied such as log and sqaure which make linear fit more likely.

squared(SC2__FIC20469.pv__Ph__4) squared(SC2__FIC20463.pv__Ph__2) log(SC2__TIC20762.pv__Ph__2) cubed(SC2__TIC20710.pv__Ph__6)

The above transformations reduced the skewness of certain tags which increased adjusted r square to .41 with 200 degress of freedom.

```
## [1] 0.4009275272
```

```
##          RMSE      Rsquared           MAE
## 0.01653425560 0.46110832356 0.01276146986
```