

Multiple Regression

Mohit Garg

Loading data form the csv file

```
data <- read.csv('/Users/mohit/Development/My Scripts/modelsEvaluation.csv',stringsAsFactors= T)
summary(data)
```

```
##      NUM
##  Min.   : 1.00
## 1st Qu.: 5.75
##  Median:10.50
##   Mean  :10.50
## 3rd Qu.:15.25
##   Max.  :20.00
##
##                                     PROJ
## f14a_cash_doctor                  : 1
## f14a_gotrla                      : 1
## F14a_mobile_application_for_mobile_controlled_lighting: 1
## F14a_REFERsy                     : 1
## F14a_sharethetraining_com        : 1
## F14a_soccer_data_web_crawler     : 1
## (Other)                          :14
##
##      Effort      Effort_ALY  Effort_Norm      Norm_Factor
##  Min.   : 140.5   Min.   :0    Min.   : 206.6   Min.   :0.6700
## 1st Qu.: 294.0   1st Qu.:0    1st Qu.: 561.3   1st Qu.:0.8975
##  Median : 730.5   Median :0    Median : 831.8   Median :1.0950
##   Mean  :1064.3   Mean   :0    Mean   :1576.7   Mean   :1.1965
## 3rd Qu.:1414.6   3rd Qu.:0    3rd Qu.:2504.1   3rd Qu.:1.4275
##   Max.  :3680.0   Max.   :0    Max.   :5850.4   Max.   :1.8900
##
##      KSLOC      UEUCW_ALY      UEXUCW_ALY      UDUCW_ALY
##  Min.   : 0.552   Min.   : 75.0   Min.   : 42.00   Min.   : 29.00
## 1st Qu.: 2.777   1st Qu.:120.0   1st Qu.: 96.75   1st Qu.: 87.75
##  Median : 4.920   Median :157.5   Median :174.50   Median :148.50
##   Mean  : 5.507   Mean   :225.0   Mean   :193.10   Mean   :191.70
## 3rd Qu.: 7.422   3rd Qu.:292.5   3rd Qu.:250.50   3rd Qu.:236.50
##   Max.  :21.344   Max.   :705.0   Max.   :701.00   Max.   :777.00
##
##      UAW      TCF      TCF_ALY      EF      EF_ALY
##  Min.   : 3.00   Min.   :0.7950   Min.   :0    Min.   :0.8750   Min.   :0
## 1st Qu.: 6.00   1st Qu.:0.8738   1st Qu.:0    1st Qu.:0.9463   1st Qu.:0
##  Median : 9.00   Median :0.9200   Median :0    Median :1.0250   Median :0
##   Mean  : 8.55   Mean   :0.9280   Mean   :0    Mean   :1.0265   Mean   :0
## 3rd Qu.: 9.00   3rd Qu.:0.9363   3rd Qu.:0    3rd Qu.:1.0288   3rd Qu.:0
##   Max.  :14.00   Max.   :1.1750   Max.   :0    Max.   :1.3250   Max.   :0
##
##      EUCP_ALY      EXUCP_ALY      DUCP_ALY      Effort_Norm_UCP
##  Min.   : 77.92   Min.   : 61.57   Min.   : 49.26   Min.   : 194.1
## 1st Qu.:117.55   1st Qu.: 94.24   1st Qu.: 94.71   1st Qu.:528.4
##  Median :151.75   Median :161.15   Median :135.31   Median :681.4
```

```

## Mean : 243.80 Mean : 200.49 Mean : 201.15 Mean :1165.7
## 3rd Qu.: 279.56 3rd Qu.: 215.71 3rd Qu.: 206.74 3rd Qu.:1745.5
## Max. :1067.00 Max. :1061.06 Max. :1173.84 Max. :3265.0
##
## Path_Num UseCase_Num Diagram_Num INT
## Min. : 19.00 Min. : 5.0 Min. : 5.0 0 :4
## 1st Qu.: 33.50 1st Qu.: 8.0 1st Qu.:10.0 13 :2
## Median : 53.00 Median :10.5 Median :12.0 30 :2
## Mean : 62.55 Mean :15.0 Mean :16.7 15 :1
## 3rd Qu.: 76.00 3rd Qu.:19.5 3rd Qu.:21.0 16 :1
## Max. :246.00 Max. :47.0 Max. :47.0 18 :1
## (Other):9
## INT_ALY DM DM_ALY CTRL
## Min. : 2.00 0 :4 Min. : 8.00 0 : 4
## 1st Qu.: 13.75 10 :2 1st Qu.:12.50 51 : 2
## Median : 25.50 13 :2 Median :17.50 18 : 1
## Mean : 29.50 18 :2 Mean :23.15 26 : 1
## 3rd Qu.: 36.25 21 :2 3rd Qu.:30.75 28 : 1
## Max. :119.00 9 :2 Max. :57.00 30 : 1
## (Other):6 (Other):10
## CTRL_ALY EXTIVK EXTIVK_ALY EXTCLL
## Min. : 17.00 0 :11 Min. :0.00 0 :13
## 1st Qu.: 28.50 1 : 2 1st Qu.:0.00 1 : 3
## Median : 49.50 3 : 1 Median :1.00 2 : 2
## Mean : 52.85 5 : 4 Mean :1.50 4 : 1
## 3rd Qu.: 71.25 6 : 1 3rd Qu.:2.25 undefined: 1
## Max. :168.00 undefined: 1 Max. :6.00
##
## EXTCLL_ALY NT NT_ALY NWT_ALY
## Min. : 0.00 undefined:5 Min. : 17.00 Min. : 118.0
## 1st Qu.: 0.00 32 :2 1st Qu.: 28.50 1st Qu.: 232.5
## Median : 0.00 51 :2 Median : 49.50 Median : 329.0
## Mean : 1.35 26 :1 Mean : 52.85 Mean : 520.5
## 3rd Qu.: 2.00 29 :1 3rd Qu.: 71.25 3rd Qu.: 570.2
## Max. :11.00 30 :1 Max. :168.00 Max. :2332.0
## (Other) :8
## NWT_DE_ALY DET RET ILF EIF
## Min. : 116.0 Min. :0 Min. :0.00 Min. :0 Min. :0
## 1st Qu.: 233.5 1st Qu.:0 1st Qu.:0.00 1st Qu.:0 1st Qu.:0
## Median : 328.0 Median :0 Median :2.00 Median :0 Median :0
## Mean : 536.8 Mean :0 Mean :1.65 Mean :0 Mean :0
## 3rd Qu.: 581.5 3rd Qu.:0 3rd Qu.:2.00 3rd Qu.:0 3rd Qu.:0
## Max. :2435.0 Max. :0 Max. :4.00 Max. :0 Max. :0
##
## Type Simple_UC Average_UC Complex_UC
## Mobile App :5 Min. : 0.00 Min. : 0.00 Min. : 0.00
## Mobile Game :1 1st Qu.: 2.00 1st Qu.: 2.75 1st Qu.: 0.00
## Mobile&Web App:4 Median : 5.00 Median : 3.00 Median : 1.00
## web App :2 Mean : 8.80 Mean : 4.15 Mean : 2.05
## Web App :8 3rd Qu.:13.75 3rd Qu.: 6.00 3rd Qu.: 3.25
## Max. :28.00 Max. :10.00 Max. :10.00
##
## Normalized_UC_Effort
## Min. : 8.016

```

```
## 1st Qu.: 16.502
## Median : 38.315
## Mean : 55.634
## 3rd Qu.: 86.024
## Max. :186.051
##
```

Preprocessing the data

Replacing all the NaN with the mean value.

```
data$NT = ifelse(is.na(data$NT), ave(data$NT, FUN = function(x) mean(x, na.rm = TRUE)), data$NT)
data$INT_ALY = ifelse(is.na(data$INT_ALY), ave(data$INT_ALY, FUN = function(x) mean(x, na.rm = TRUE)), data$INT_ALY)
data$INT = ifelse(is.na(data$INT), ave(data$INT, FUN = function(x) mean(x, na.rm = TRUE)), data$INT)
data$DM = ifelse(is.na(data$DM), ave(data$DM, FUN = function(x) mean(x, na.rm = TRUE)), data$DM)
data$CTRL = ifelse(is.na(data$CTRL), ave(data$CTRL, FUN = function(x) mean(x, na.rm = TRUE)), data$CTRL)
data$EXTCLL = ifelse(is.na(data$EXTCLL), ave(data$EXTCLL, FUN = function(x) mean(x, na.rm = TRUE)), data$EXTCLL)
data$EXTIVK = ifelse(is.na(data$EXTIVK), ave(data$EXTIVK, FUN = function(x) mean(x, na.rm = TRUE)), data$EXTIVK)
```

Preparing the independent variables

1. Removing all the variables with zero value for all the observations.
2. Factorizing the type variable
3. Calculating the correlation between all the independent and dependent variables.
4. Choosing all the variables with highest correlation values.

```
x <- data[, 7:45];
x$Type = factor(x$Type, levels = c('Web App', 'Mobile App', 'Mobile&Web App'), labels = c(1, 2, 3))
x$ILF <- NULL
x$EIF <- NULL
x$DET <- NULL
x$EF_ALY <- NULL
x$TCF_ALY <- NULL
x$Type[5] = 1
x$Type[7] = 1
x$Type[20] = 3
x$Type = as.numeric(x$Type)

y = data$Effort
summary(x)
```

##	KSLOC	UEUCW_ALY	UEXUCW_ALY	UDUCW_ALY
##	Min. : 0.552	Min. : 75.0	Min. : 42.00	Min. : 29.00
##	1st Qu.: 2.777	1st Qu.: 120.0	1st Qu.: 96.75	1st Qu.: 87.75
##	Median : 4.920	Median : 157.5	Median : 174.50	Median : 148.50
##	Mean : 5.507	Mean : 225.0	Mean : 193.10	Mean : 191.70
##	3rd Qu.: 7.422	3rd Qu.: 292.5	3rd Qu.: 250.50	3rd Qu.: 236.50
##	Max. : 21.344	Max. : 705.0	Max. : 701.00	Max. : 777.00
##	UAW	TCF	EF	EUCP_ALY
##	Min. : 3.00	Min. : 0.7950	Min. : 0.8750	Min. : 77.92
##	1st Qu.: 6.00	1st Qu.: 0.8738	1st Qu.: 0.9463	1st Qu.: 117.55
##	Median : 9.00	Median : 0.9200	Median : 1.0250	Median : 151.75
##	Mean : 8.55	Mean : 0.9280	Mean : 1.0265	Mean : 243.80

```

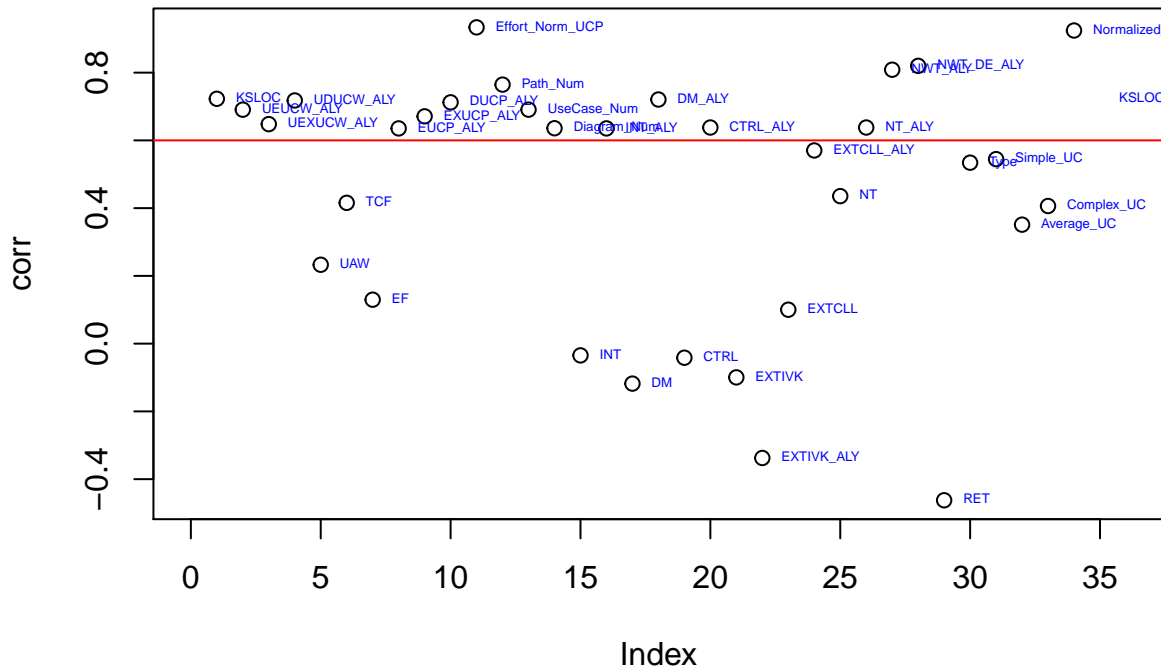
## 3rd Qu.: 9.00    3rd Qu.:0.9363    3rd Qu.:1.0288    3rd Qu.: 279.56
## Max.    :14.00    Max.    :1.1750    Max.    :1.3250    Max.    :1067.00
##      EXUCP_ALY      DUCP_ALY      Effort_Norm_UCP      Path_Num
## Min.    : 61.57    Min.    : 49.26    Min.    : 194.1    Min.    : 19.00
## 1st Qu.: 94.24    1st Qu.: 94.71    1st Qu.: 528.4    1st Qu.: 33.50
## Median : 161.15    Median : 135.31    Median : 681.4    Median : 53.00
## Mean    : 200.49    Mean    : 201.15    Mean    :1165.7    Mean    : 62.55
## 3rd Qu.: 215.71    3rd Qu.: 206.74    3rd Qu.:1745.5    3rd Qu.: 76.00
## Max.    :1061.06    Max.    :1173.84    Max.    :3265.0    Max.    :246.00
##      UseCase_Num      Diagram_Num      INT      INT_ALY
## Min.    : 5.0    Min.    : 5.0    Min.    : 1.00    Min.    : 2.00
## 1st Qu.: 8.0    1st Qu.:10.0    1st Qu.: 2.00    1st Qu.: 13.75
## Median :10.5    Median :12.0    Median : 6.50    Median : 25.50
## Mean    :15.0    Mean    :16.7    Mean    : 6.70    Mean    : 29.50
## 3rd Qu.:19.5    3rd Qu.:21.0    3rd Qu.:10.25    3rd Qu.: 36.25
## Max.    :47.0    Max.    :47.0    Max.    :15.00    Max.    :119.00
##      DM      DM_ALY      CTRL      CTRL_ALY
## Min.    : 1.00    Min.    : 8.00    Min.    : 1.00    Min.    : 17.00
## 1st Qu.: 2.00    1st Qu.:12.50    1st Qu.: 2.75    1st Qu.: 28.50
## Median : 5.00    Median :17.50    Median : 7.50    Median : 49.50
## Mean    : 5.40    Mean    :23.15    Mean    : 7.40    Mean    : 52.85
## 3rd Qu.: 8.25    3rd Qu.:30.75    3rd Qu.:11.25    3rd Qu.: 71.25
## Max.    :12.00    Max.    :57.00    Max.    :16.00    Max.    :168.00
##      EXTIVK      EXTIVK_ALY      EXTCLL      EXTCLL_ALY
## Min.    :1.00    Min.    :0.00    Min.    :1.0    Min.    : 0.00
## 1st Qu.:1.00    1st Qu.:0.00    1st Qu.:1.0    1st Qu.: 0.00
## Median :1.00    Median :1.00    Median :1.0    Median : 0.00
## Mean    :2.25    Mean    :1.50    Mean    :1.7    Mean    : 1.35
## 3rd Qu.:4.00    3rd Qu.:2.25    3rd Qu.:2.0    3rd Qu.: 2.00
## Max.    :6.00    Max.    :6.00    Max.    :5.0    Max.    :11.00
##      NT      NT_ALY      NWT_ALY      NWT_DE_ALY
## Min.    : 1.00    Min.    : 17.00    Min.    : 118.0    Min.    : 116.0
## 1st Qu.: 5.00    1st Qu.: 28.50    1st Qu.: 232.5    1st Qu.: 233.5
## Median : 8.50    Median : 49.50    Median : 329.0    Median : 328.0
## Mean    : 8.60    Mean    : 52.85    Mean    : 520.5    Mean    : 536.8
## 3rd Qu.:13.25    3rd Qu.: 71.25    3rd Qu.: 570.2    3rd Qu.: 581.5
## Max.    :14.00    Max.    :168.00    Max.    :2332.0    Max.    :2435.0
##      RET      Type      Simple_UC      Average_UC
## Min.    :0.00    Min.    :1.00    Min.    : 0.00    Min.    : 0.00
## 1st Qu.:0.00    1st Qu.:1.00    1st Qu.: 2.00    1st Qu.: 2.75
## Median :2.00    Median :1.50    Median : 5.00    Median : 3.00
## Mean    :1.65    Mean    :1.75    Mean    : 8.80    Mean    : 4.15
## 3rd Qu.:2.00    3rd Qu.:2.25    3rd Qu.:13.75    3rd Qu.: 6.00
## Max.    :4.00    Max.    :3.00    Max.    :28.00    Max.    :10.00
##      Complex_UC      Normalized_UC_Effort
## Min.    : 0.00    Min.    : 8.016
## 1st Qu.: 0.00    1st Qu.: 16.502
## Median : 1.00    Median : 38.315
## Mean    : 2.05    Mean    : 55.634
## 3rd Qu.: 3.25    3rd Qu.: 86.024
## Max.    :10.00    Max.    :186.051

```

Correlation

Calculating the correlation and choosing the independent variables with correlation higher than 0.6 with the dependent variable (Effort).

```
corr <- cor(x,y)
plot(corr,xlim=c(0, 36))
text(1:35,corr,row.names(corr),cex=0.4, pos=4, col="blue")
abline(h=0.6,col="red")
```



Looking at the graph, following are the most correlated independent variables:

1. KSLOC
2. UEUCW_ALY
3. UEXUCW_ALY
4. UDUCW_ALY
5. Effort_Norm_UCP
6. Path_Num
7. DUCP_ALY
8. EXUCP_ALY
9. EUCP_ALY
10. UseCase_Num
11. Diagram_Num
12. INT_ALY
13. DM_ALY
14. CTRL_ALY
15. NT_ALY
16. NWT_DE_ALY
17. NWT_ALY

Model Fitting

Using all the above variables except UseCase_NUM and Diagram_Num for fitting the model.

```

independentVar <- data.frame(x$KSLOC,x$UEUCW_ALY,x$UEXUCW_ALY,x$UDUCW_ALY,x$Effort_Norm_UCP,x$Path_Num,
names(independentVar)<- c("KSLOC","UEUCW_ALY","UEXUCW_ALY","UDUCW_ALY","Effort_Norm_UCP","Path_Num","DU

#library(caret)
#set.seed(30)
#model <- train(y~.,data=independentVar,method="lm",trControl = trainControl(method = "cv", number=2,ve

fit <- lm(y~.,data=independentVar)
summary(fit)

##
## Call:
## lm(formula = y ~ ., data = independentVar)
##
## Residuals:
##      1      2      3      4      5      6      7
## 167.0279 -194.0992 -171.8621   8.9569  -0.2924 -118.3572 138.1005
##      8      9     10     11     12     13     14
##  68.4321 185.5159  89.8604 -108.4177 -146.6986  -5.9901   0.8631
##     15     16     17     18     19     20
##   2.8460  27.3099  28.1691  27.0683   6.7885  -5.2211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    183.3461    312.2033   0.587  0.5825
## KSLOC           33.5568     38.4788   0.872  0.4231
## UEUCW_ALY       -1.9482      3.2776  -0.594  0.5781
## UEXUCW_ALY       48.1700     35.3833   1.361  0.2315
## UDOCW_ALY      -45.8973     43.8570  -1.047  0.3433
## Effort_Norm_UCP   0.3988      0.1180   3.380  0.0197 *
## Path_Num        -3.0924     43.3269  -0.071  0.9459
## DUCP_ALY         43.0490     58.1155   0.741  0.4921
## EXUCP_ALY       -49.4032     65.3425  -0.756  0.4837
## EUCP_ALY         2.5930      3.4164   0.759  0.4821
## INT_ALY          84.4573    128.2986   0.658  0.5395
## DM_ALY           77.0080     39.9469   1.928  0.1118
## CTRL_ALY        -68.2261     58.8718  -1.159  0.2988
## NWT_DE_ALY        9.6134      5.2364   1.836  0.1258
## NWT_ALY          -9.3280      8.1313  -1.147  0.3032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 205.5 on 5 degrees of freedom
## Multiple R-squared:  0.988, Adjusted R-squared:  0.9544
## F-statistic: 29.43 on 14 and 5 DF, p-value: 0.0007453

```