Assignment2_GR2_TEAM10

Group 2 - TEAM10

15/07/2019

Data Cleaning

The categorical variables such as age is coded as a number and hence when we read it using R, they are coded as numeric variables. As a first step we need to convert all the relevant categorical variables as factors in R.

```
library("readxl")
ipl.data <- read_xlsx("IMB381IPL2013.xlsx",sheet = 3)
ipl.data$AGE<-factor(ipl.data$AGE)
levels(ipl.data$AGE)<-c(" -less than 25 yrs", " -between 25-35 yrs", " -more Than 35 yrs")
ipl.data$COUNTRY<-factor(ipl.data$COUNTRY)
ipl.data$TEAM<-factor(ipl.data$TEAM)
ipl.data$`PLAYING ROLE`<-factor(ipl.data$`PLAYING ROLE`)
ipl.data$`CAPTAINCY EXP`<-factor(ipl.data$`CAPTAINCY EXP`)
levels(ipl.data$`CAPTAINCY EXP`)<-c(" -no", " -yes")
```

Question 1

We notice a statistical difference among the coefficents for the different age groups. However we do not see a trend between the age and selling price. Younger players are valued more, but the older players (age > 35yrs) are also valued more than the middle aged players(25-35 yrs).

```
price.age.model <- lm(`SOLD PRICE` ~ AGE -1 , data=ipl.data)
summary(price.age.model)
```

```
##
## Call:
## lm(formula = `SOLD PRICE` ~ AGE - 1, data = ipl.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -696250 -307035  -82714  190465 1315465
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## AGE -less than 25 yrs     720250     100684   7.154 5.96e-11 ***
## AGE -between 25-35 yrs    484535      43428  11.157  < 2e-16 ***
## AGE -more Than 35 yrs     520179      76110   6.835 3.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 402700 on 127 degrees of freedom
## Multiple R-squared:  0.6365, Adjusted R-squared:  0.6279
## F-statistic: 74.12 on 3 and 127 DF,  p-value: < 2.2e-16
```

Question 2

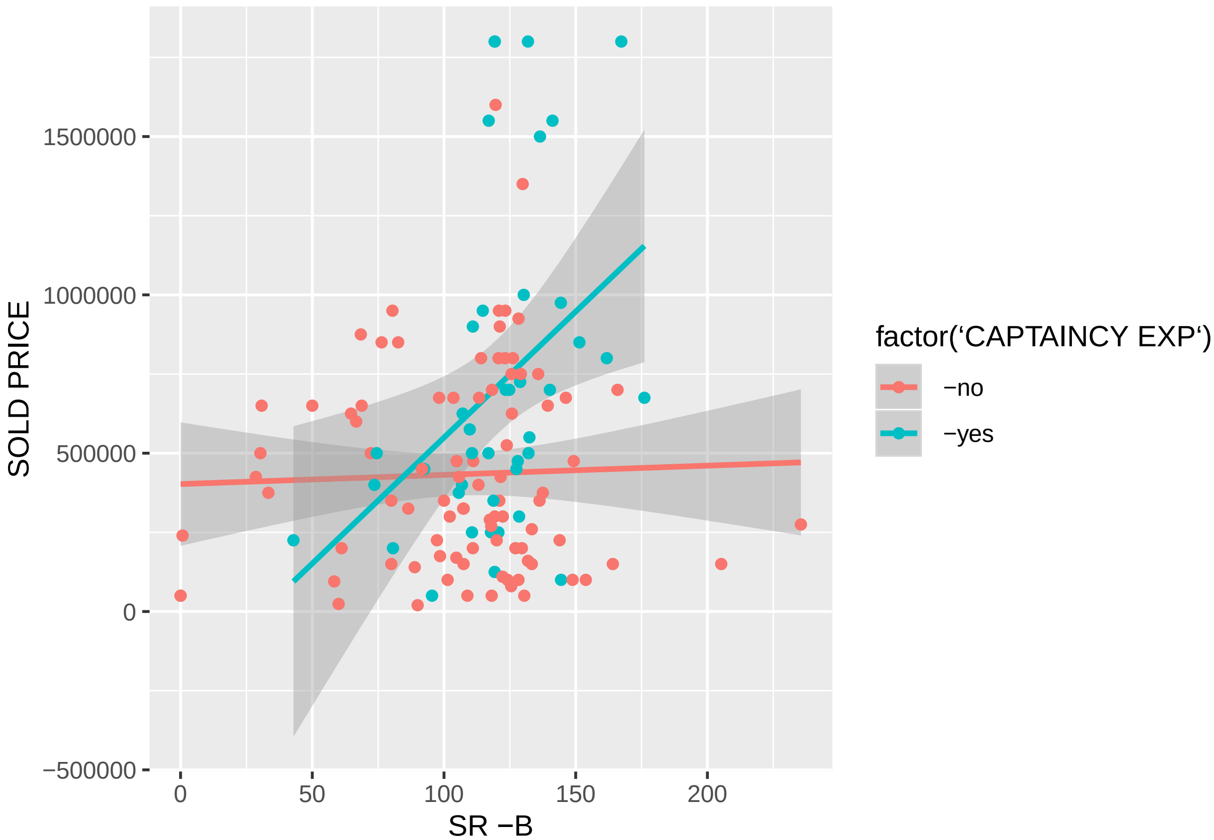
Here plot the relation between the 'Sold Price' and 'Strike Rate', 'Captaincy Experience' with interactions only for the batsman role players.

From the graph we notice that 'Strike Rate' and 'Captaincy Experience' are not statistically significant individually but their interaction however is significant.

```r
##Question 2

data.slice<-subset(ipl.data,`PLAYING ROLE`=="Batsman")
model2<-lm(`SOLD PRICE`~`SR -B` + `CAPTAINCY EXP` + `SR -B`:`CAPTAINCY EXP`, data = data.slice)
summary(model2)
```

```
##
## Call:
## lm(formula = `SOLD PRICE` ~ `SR -B` + `CAPTAINCY EXP` + `SR -B`:`CAPTAINCY EXP`,
##     data = data.slice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -964447 -324110 -107077  214087 1034944
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1207910    1117976   1.080   0.2873
## `SR -B`                         -5178       8924  -0.580   0.5654
## `CAPTAINCY EXP` -yes         -1858111    1214877  -1.529   0.1351
## `SR -B`:`CAPTAINCY EXP` -yes    17049       9787   1.742   0.0903 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 479300 on 35 degrees of freedom
## Multiple R-squared:  0.2242, Adjusted R-squared:  0.1577
## F-statistic: 3.372 on 3 and 35 DF,  p-value: 0.02916
```

```r
library(ggplot2)
plot <- ggplot(data=ipl.data, aes(x=`SR -B`, y=`SOLD PRICE`, colour=factor(`CAPTAINCY EXP`)))
plot + stat_smooth(method=lm, fullrange=FALSE) + geom_point()
```



Question 3

From our model summary we find that batting average is statistically significant on regressing with the player selling price.

```
model3<-lm(`SOLD PRICE`~`AVE` , data = ipl.data)
summary(model3)
```

```
##
## Call:
## lm(formula = `SOLD PRICE` ~ AVE, data = ipl.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -616140 -272858  -71338  213141 1168281
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   249062      64680   3.851 0.000185 ***
## AVE            14539       2975   4.886 3.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 374900 on 128 degrees of freedom
## Multiple R-squared:  0.1572, Adjusted R-squared:  0.1506
## F-statistic: 23.88 on 1 and 128 DF,  p-value: 3.012e-06
```

Quesetion 4

The adjusted R- squared has increased from 15% to 20%. Hence we can conclude that the including 'SIXERS' has improved the model. How ever since 'SIXERS' and batting average are correlated, we find that the significance of the 'AVE' variable has decreased.

```
model4<-lm(`SOLD PRICE`~`AVE`+`SIXERS` , data = ipl.data)
summary(model4)
```

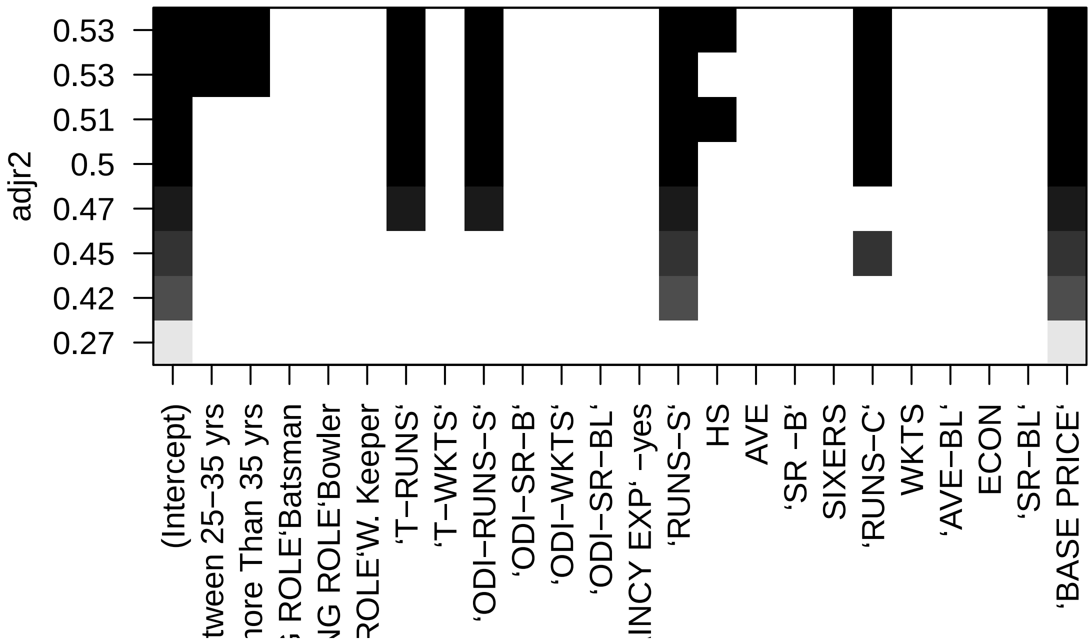
```
##
## Call:
## lm(formula = `SOLD PRICE` ~ AVE + SIXERS, data = ipl.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -547885 -264124  -83989  238794 1131999
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   311016      65828   4.725 6.01e-06 ***
## AVE             5740       4066   1.412  0.16048
## SIXERS          5808       1893   3.068  0.00264 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 363200 on 127 degrees of freedom
## Multiple R-squared:  0.2154, Adjusted R-squared:  0.203
## F-statistic: 17.43 on 2 and 127 DF,  p-value: 2.049e-07
```

Question 5

Variable Selection

In order to select the appropriate variable we use all subsets regression where we test the models by taking all the variables and dropping the variables one by one (backward step). We then use the leaps package to plot only one best model in each scenario.

```
library(leaps)
leaps<-regsubsets(`SOLD PRICE`~ `AGE`+  `PLAYING ROLE`+ `T-RUNS`+ `T-WKTS`+ `ODI-RUNS-S`+ `ODI-SR-B`+ `(
plot(leaps, scale="adjr2")
```



Ideal Model

From the above graph we find that the best model consists of the following variables : AGE, T-RUNS, ODI-RUNS, RUNS-S, HS, RUNS-C, BASE PRICE. Our ideal model explains 53% of the variance between the predicted and observed sellign prices.

```r
#Ideal Model

ideal<-lm(`SOLD PRICE`~ `AGE`+  `T-RUNS`+ `ODI-RUNS-S`+ `RUNS-S`+`HS`+ `RUNS-C`+ `BASE PRICE`, data = i
summary(ideal)
```

```
##
## Call:
## lm(formula = `SOLD PRICE` ~ AGE + `T-RUNS` + `ODI-RUNS-S` + `RUNS-S` +
##     HS + `RUNS-C` + `BASE PRICE`, data = ipl.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -606449 -165536  -78426  168615  951020
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.087e+05  8.672e+04   3.560 0.000532 ***
## AGE -between 25-35 yrs -2.133e+05  7.893e+04  -2.702 0.007876 **
## AGE -more Than 35 yrs  -2.271e+05  1.008e+05  -2.253 0.026087 *
## `T-RUNS`              -5.874e+01  1.759e+01  -3.340 0.001116 **
## `ODI-RUNS-S`           5.170e+01  1.650e+01   3.133 0.002172 **
## `RUNS-S`               3.401e+02  7.781e+01   4.371 2.63e-05 ***
## HS                    -2.161e+03  1.272e+03  -1.698 0.092034 .
## `RUNS-C`               1.064e+02  4.670e+01   2.278 0.024493 *
## `BASE PRICE`           1.442e+00  1.803e-01   7.993 8.81e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 277900 on 121 degrees of freedom
## Multiple R-squared:  0.5621, Adjusted R-squared:  0.5332
## F-statistic: 19.42 on 8 and 121 DF,  p-value: < 2.2e-16
```

Mallow's Cp

We find use the formula and calculate it manually as follows.

```
#Full Model
full.model<-lm(`SOLD PRICE`~ `AGE`+  `PLAYING ROLE`+ `T-RUNS`+ `T-WKTS`+ `ODI-RUNS-S`+ `ODI-SR-B`+ `ODI-
#Manually calculating the Mallow's Cp
SSE<-sum(ideal$residuals**2) # for the ideal model
MSE<-7.7746e+10  # Taken from the full.model annova table
n<-nrow(ipl.data)
p<-9; p#no. of parameters in our ideal model
```

```
## [1] 9
```

```
Cp<-(SSE/MSE)-(n-(2*p)); Cp
```

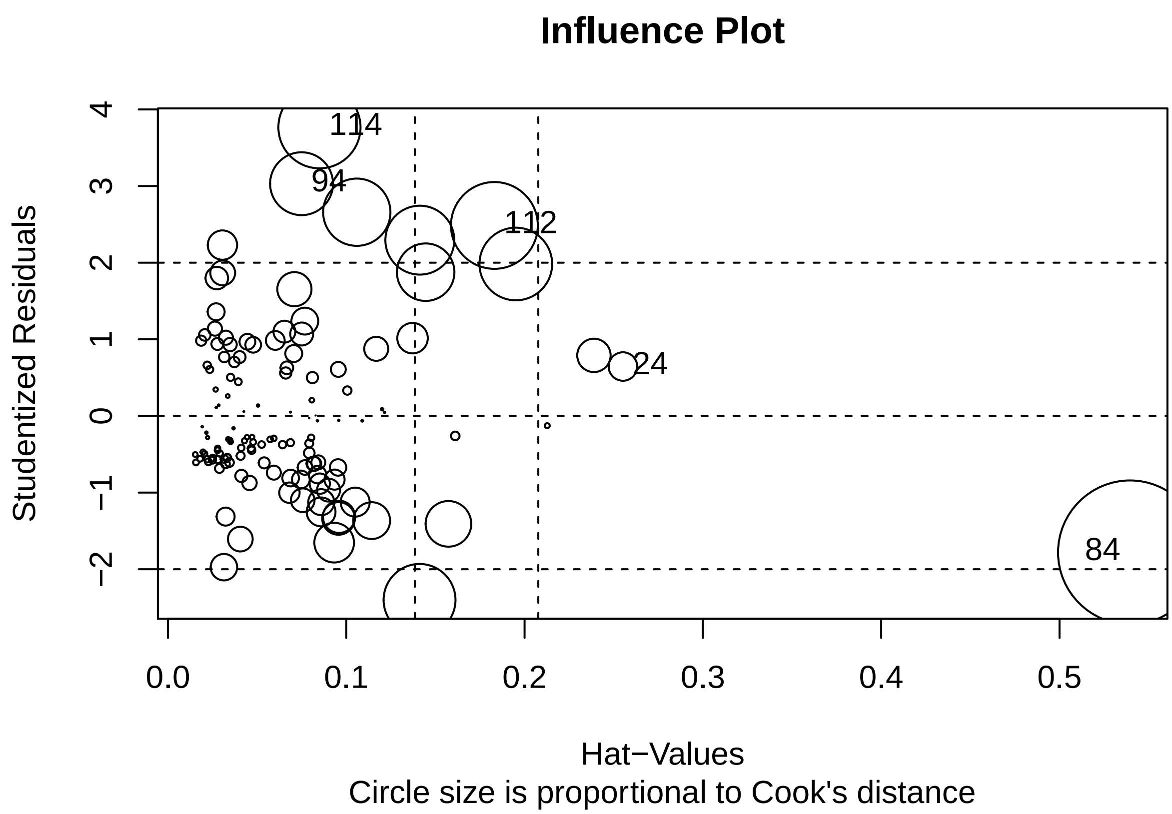
```
## [1] 8.235006
```

Outliers

We use the car package and influence plot to find the ourliers.

```
#To Find outliers in our model
library(carData)
library(car)
influencePlot(ideal, id.method = "identify", main="Influence Plot", sub="Circle size is proportional to
```

```
## Warning in plot.window(...): "id.method" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "id.method" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" is
## not a graphical parameter

## Warning in box(...): "id.method" is not a graphical parameter

## Warning in title(...): "id.method" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.method" is not a
## graphical parameter
```




```
##        StudRes        Hat      CookD
## 24   0.6452135 0.25516897 0.01592335
## 84  -1.7806304 0.53955206 0.40554219
## 94   3.0305640 0.07493119 0.07742275
## 112  2.4860062 0.18307216 0.14756881
## 114  3.7668335 0.08497108 0.13201248
```

```
#From the graph we identify the outliers
outlier.pts<-ipl.data[c(84,24,94,112,114),];outlier.pts
```

```
## # A tibble: 5 x 26
##   Sl.NO. `PLAYER NAME` AGE   COUNTRY TEAM  `PLAYING ROLE` `T-RUNS` `T-WKTS`
##    <dbl> <chr>         <fct> <fct>   <fct> <fct>             <dbl>    <dbl>
## 1     84 Pietersen, KP " -b~ ENG     RCB+  Batsman            6654        5
## 2     24 Flintoff, A   " -b~ ENG     CSK   Allrounder         3845      226
## 3     94 Sehwag, V     " -b~ IND     DD    Batsman            8178       40
## 4    112 Tendulkar, SR " -m~ IND     MI    Batsman           15470       45
## 5    114 Tiwary, SS    " -l~ IND     MI+   Batsman               0        0
## # ... with 18 more variables: `ODI-RUNS-S` <dbl>, `ODI-SR-B` <dbl>,
## #   `ODI-WKTS` <dbl>, `ODI-SR-BL` <dbl>, `CAPTAINCY EXP` <fct>,
## #   `RUNS-S` <dbl>, HS <dbl>, AVE <dbl>, `SR -B` <dbl>, SIXERS <dbl>,
## #   `RUNS-C` <dbl>, WKTS <dbl>, `AVE-BL` <dbl>, ECON <dbl>, `SR-BL` <dbl>,
## #   `AUCTION YEAR` <dbl>, `BASE PRICE` <dbl>, `SOLD PRICE` <dbl>
```

Question 6

Part (a)

Since we got around 53% R-square using full model - we believe that the data is sufficient to explain the variation in price of IPL players.

Part (b)

No. of matches played can be a good stat that could improve the model. Stats from frist class cricket. Stats of international T20 matches played by players.

Part (c)

```r
par(mfrow=c(2,2))
plot(ideal)
```