# Exercise_4

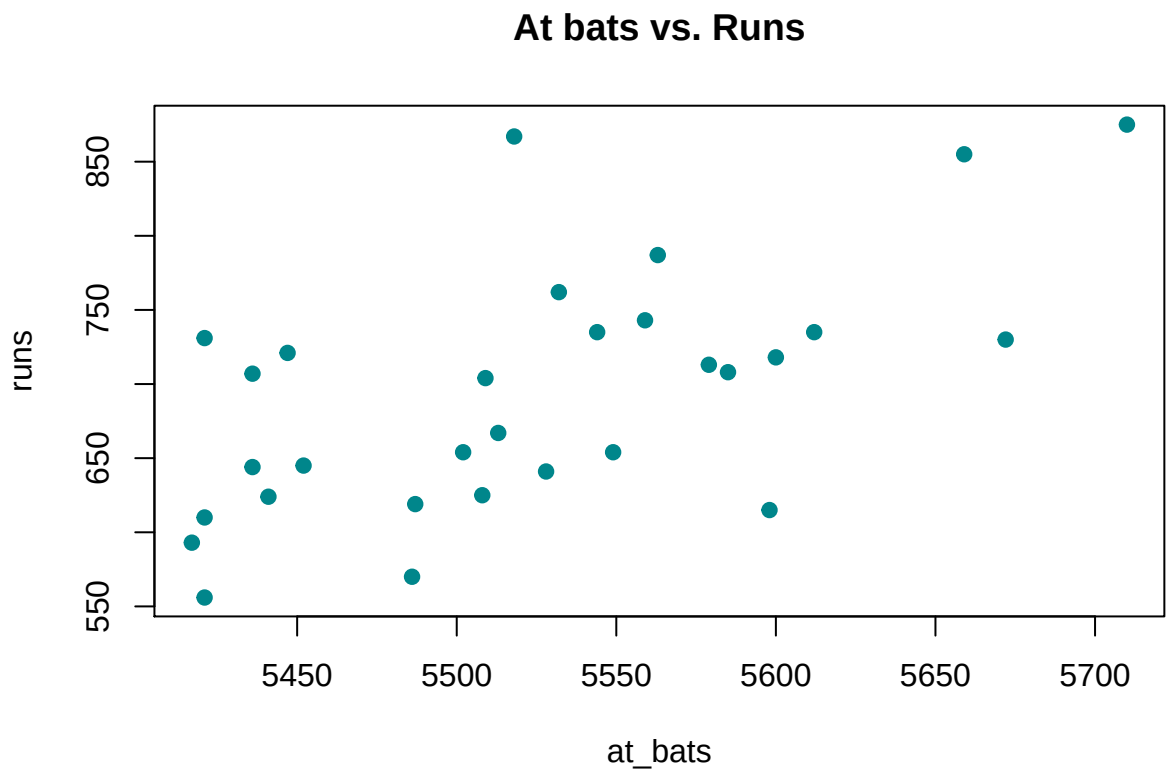
*Mansoor Baba Shaik*

*November 23, 2016*

```
rm(list=ls())
setwd("~/IS665/Exercise_4")
baseball = read.csv(file="baseball.csv", header=TRUE, sep=",", quote="\"", dec=".")
```

## 1

```
plot(runs ~ at_bats, data=baseball, xlab="at_bats", ylab="runs", pch=19, col="turquoise4",
     main="At bats vs. Runs")
```




## 2

```
cor(baseball$runs, baseball$at_bats)
```

```
## [1] 0.610627
```

The relationship between runs and at_bats is positive and moderately strong.

## 3

```
m1 <- lm(runs ~ at_bats, data=baseball)
summary(m1)
```

```
##
## Call:
## lm(formula = runs ~ at_bats, data = baseball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -125.58  -47.05  -16.59   54.40  176.87
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2789.2429   853.6957  -3.267 0.002871 **
## at_bats         0.6305     0.1545   4.080 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.47 on 28 degrees of freedom
## Multiple R-squared:  0.3729, Adjusted R-squared:  0.3505
## F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003388
```

```
m1$coefficients
```

```
## (Intercept)     at_bats
## -2789.24289     0.63055
```

The linear function ($\hat{y} = c + mx$) that describes the relationship between runs and at_bats in the given baseball dataset is $\hat{runs}$ = -2789.24289 + 0.63055 * at_bats.

## 4

```
m2 <- lm(runs ~ homeruns, data=baseball)
m2$coefficients
```

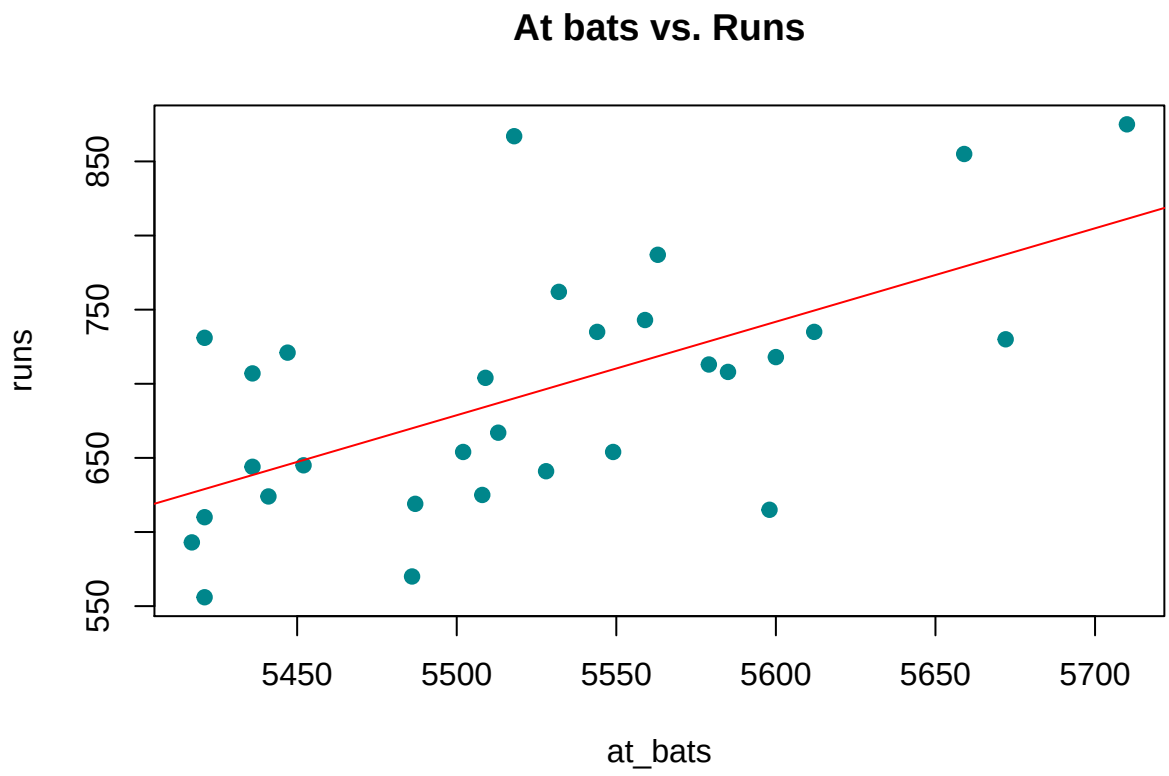
```
## (Intercept)    homeruns
##  415.238885    1.834542
```

The equation of the regression line is $\hat{runs}$ = 415.238885 + 1.834542 * homeruns.

Slope Interpretations in the context of the relationship between success of a team and its home runs:-
*For each additional homeruns scored, we would expect the runs to be increased by 1.834542.* Therefore, when played on home ground the home team is likely to score more runs which may result in the success of the team.

## 5

```
plot(runs ~ at_bats, data=baseball, xlab="at_bats", ylab="runs", pch=19, col="turquoise4",
     main="At bats vs. Runs")
abline(m1, col="red")
```




If the prediction is done for a value of x that is outside of the range of the original datatset on the basis of the relationship between the explanatory variable and the response variable by applying a model estimate, then the process is called "extrapolation". Sometimes the intercept might be an extrapolation. Problems with extrapolation is that it subjects to greater uncertainty and a higher risk of producing meaningless results.

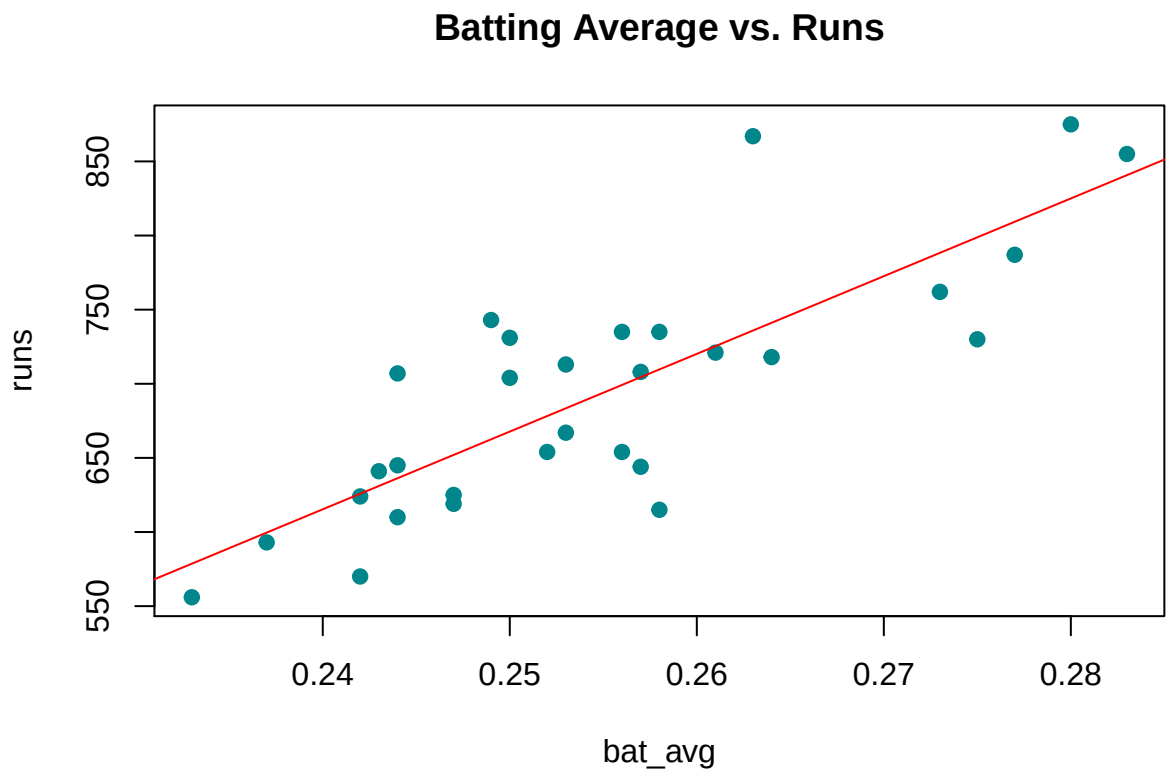
```
newdata = data.frame("at_bats"=c(5578))
predictedValue = as.numeric(predict(m1, newdata))
observedClosest_at_bats = min(baseball$at_bats[which(baseball$at_bats>=newdata$at_bats)])
observedValue = as.numeric(baseball$runs[baseball$at_bats==observedClosest_at_bats])
residual = observedValue - predictedValue
residual
```

```
## [1] -14.96497
```

Therefore, the team manager overestimated the runs by about 14 runs.

## 6

```
m3 = lm(runs~bat_avg, data=baseball)
plot(runs~bat_avg, data=baseball, xlab="bat_avg", ylab="runs", pch=19, col="turquoise4",
     main="Batting Average vs. Runs")
abline(m3, col="red")
```




There seems to be a linear relationship between runs and bat_avg when looked at a glance on the plot. Also, we can see that most of the points on the scatterplot are close to the abline.

## 7

```
R2_runsVSat_bats = summary(m1)$r.squared
R2_runsVSat_bats
```

```
## [1] 0.3728654
```

```
R2_runsVSbat_avg = summary(m3)$r.squared
R2_runsVSbat_avg
```

```
## [1] 0.6560771
```

The R-squared value for the linear model of runs and batting average(here, m3) is higher than the R-squared value for the linear model of runs and at bats(here, m1). Therefore, the linear model of runs and batting average(here, m3) is best model for prediction, and hence we can say that the variable bat_avg predicts runs better than at_bats.