hw6

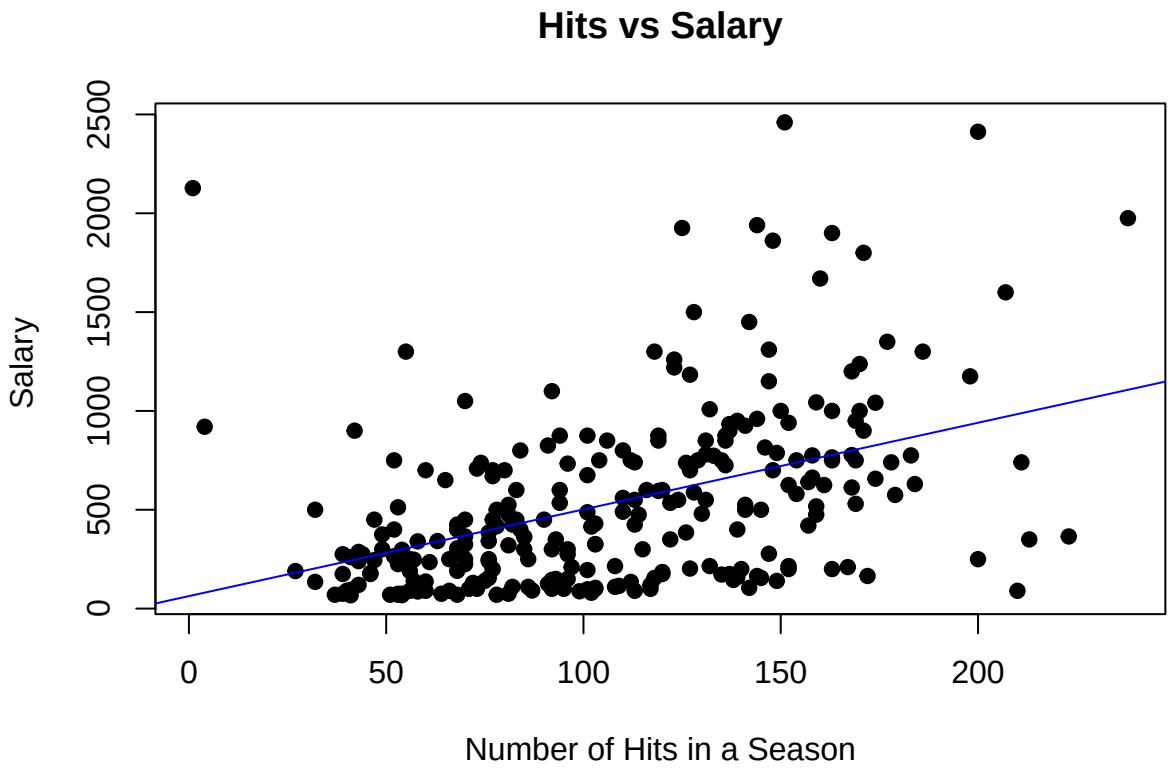
Question #1

```
baseball = read.csv("baseball_5.csv")
head(baseball)
```

```
##   Salary Hits Walks PutOuts CHits
## 1  475.0   81    39     632   835
## 2  480.0  130    76     880   457
## 3  500.0  141    37     200  1575
## 4   91.5   87    30     805   101
## 5  750.0  169    35     282  1133
## 6   70.0   37    21      76    42
```

```
plot(baseball$Hits, baseball$Salary, main="Hits vs Salary",
   xlab="Number of Hits in a Season", ylab="Salary ", pch=19)

abline(lm(baseball$Salary~baseball$Hits, data=baseball), col="blue")
```



```
summary(lm(Salary~Hits, data=baseball))
```

```
##
## Call:
## lm(formula = Salary ~ Hits, data = baseball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -893.99 -245.63  -59.08  181.12 2059.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.0488    64.9822   0.970    0.333
## Hits          4.3854     0.5561   7.886 8.53e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 406.2 on 261 degrees of freedom
## Multiple R-squared:  0.1924, Adjusted R-squared:  0.1893
## F-statistic: 62.19 on 1 and 261 DF,  p-value: 8.531e-14
```

Regression Coefficients: Intercept: 63.04, and Hits: 4.39 Standard Errors: Intercept and hits respectively: 64.9822, 0.056

Residual Sum of Squares

```
deviance(lm(Salary~Hits, data=baseball))
```

```
## [1] 43058621
```

R^2

```
summary((lm(baseball$Salary~baseball$Hits)))$r.squared
```

```
## [1] 0.1924355
```

Question 2

the estimated regression coefficients and standard errors

```
multi_fit = lm(Salary~Hits+Walks+PutOuts+CHits, data=baseball)
summary(multi_fit)$coefficients
```

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -109.8348083 56.44049413 -1.946028 5.273704e-02
## Hits           1.8460077  0.58106103  3.176960 1.669445e-03
## Walks          3.4611108  1.21166094  2.856501 4.632200e-03
## PutOuts        0.2709063  0.07861078  3.446172 6.636175e-04
## CHits          0.3124567  0.03349647  9.328047 5.108227e-18
```

Residual Sum of Squares

```
deviance(multi_fit)
```

```
## [1] 29223384
```

R^2

```
summary(multi_fit)$r.squared
```

```
## [1] 0.4519154
```

The marginal effects of each coefficient: the null hypothesis is that all of the regression coefficients is zero. The alternative hypothesis is that they are not all zero.

```
p_values <- summary(multi_fit)$coefficients[,4]
df <- data.frame(p_values)
p <- c()
for (i in 1:nrow(df)) {
  if (df[i, ] < 0.05) {
    p <- append(p, "reject")
  } else {
    p <- append(p, "fail to reject")
  }
}

df$result <- p
df
```

```
##                 p_values         result
## (Intercept) 5.273704e-02 fail to reject
## Hits        1.669445e-03         reject
## Walks       4.632200e-03         reject
## PutOuts     6.636175e-04         reject
## CHits       5.108227e-18         reject
```

The result is based on alpha = 0.05.

Question 3

```
anova(lm(Salary~Hits, data=baseball), multi_fit)
```

```
## Analysis of Variance Table
##
## Model 1: Salary ~ Hits
## Model 2: Salary ~ Hits + Walks + PutOuts + CHits
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    261 43058621
## 2    258 29223384  3  13835237 40.715 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multivariate model performs better than univariate model. The RSS for model2 is 29223384 which is smaller than 43058621 and the R^2 is 0.45 > 0.19. Based on the test result of anova, we reject null hypothesis and the multivariate model is better.