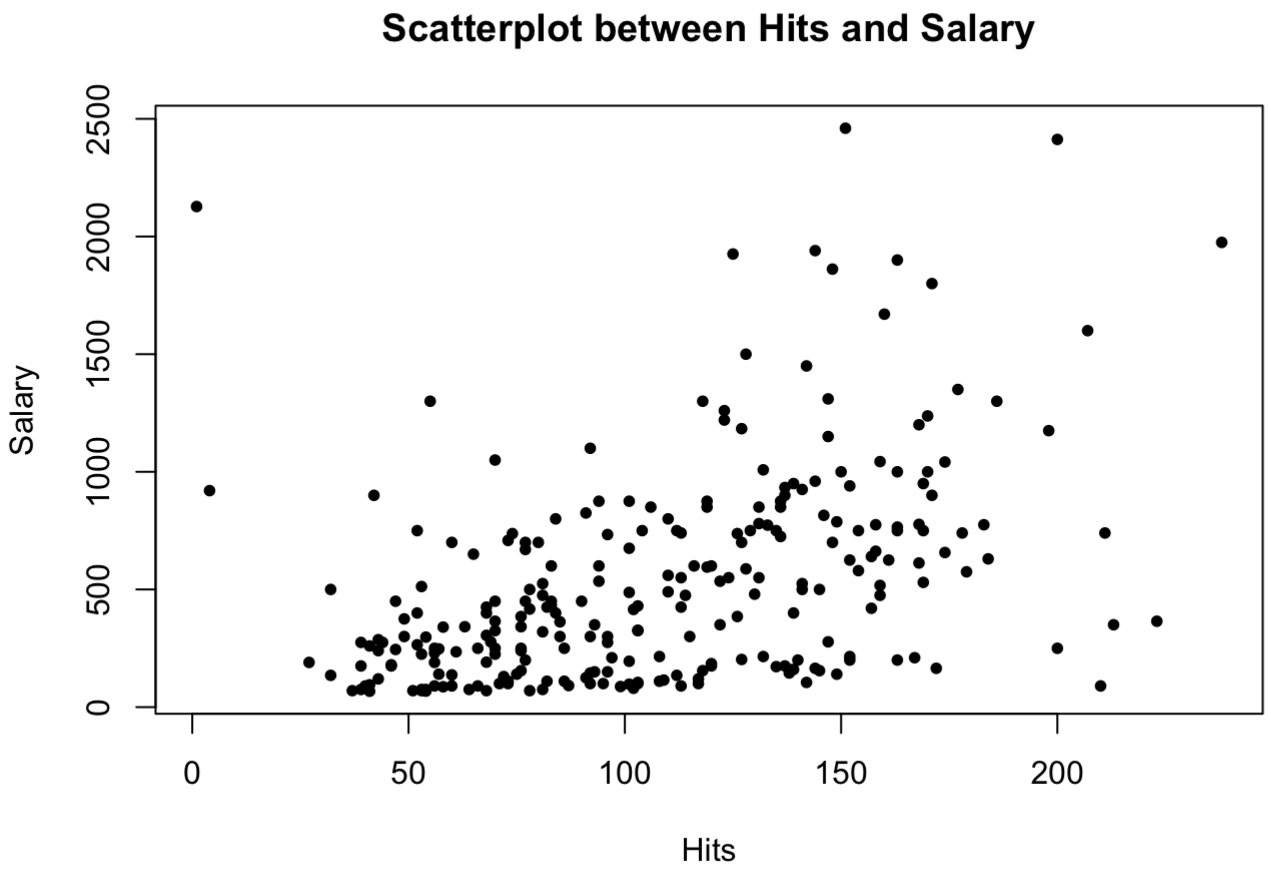
HW7

Runye Hu A13410385
Yijie Fan A13485989
Xinyi He A13561164
Yifan Wu A14060535
Yiwen Cai A13530685

```
data = read.csv("baseball_5.csv")
```

Q1

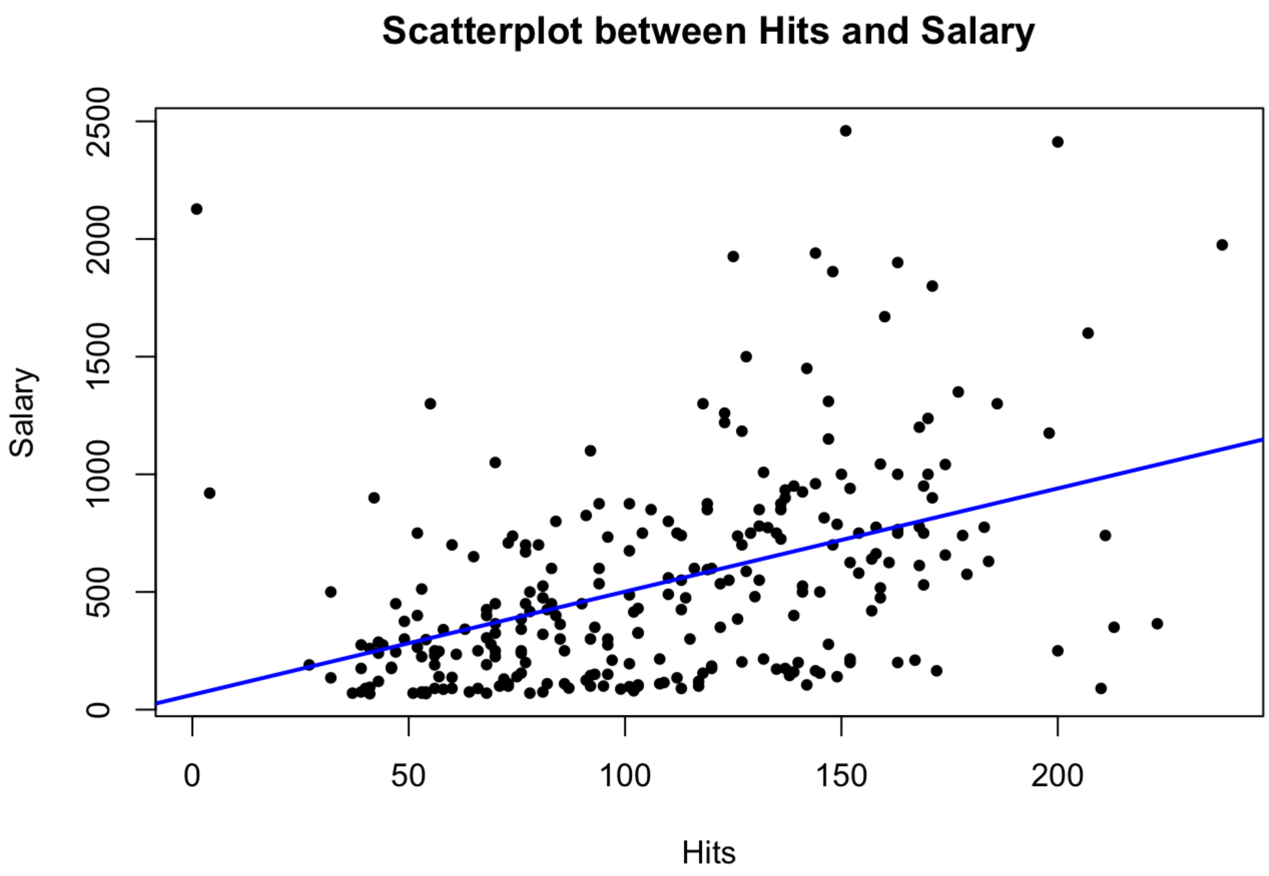
```
plot(data$Hits,data$Salary, main="Scatterplot between Hits and Salary", xlab="Hits", ylab="Salary",pch = 20)
```



```
slm.fit =lm(Salary ~ Hits,data=data)
#regression coefficients,standard errors
summary(slm.fit)$coefficients[,1:2]
```

```
##                Estimate Std. Error
## (Intercept) 63.048819  64.982225
## Hits          4.385439   0.556082
```

```
plot(data$Hits,data$Salary, main="Scatterplot between Hits and Salary", xlab="Hits", yla
b="Salary",pch = 20)
abline(slm.fit, col="blue", lwd=2)
```



```
predicted = predict(slm.fit, data)
RSS = sum((data$Salary - predicted)^2)
RSS
```

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

```
mean = mean(data$Salary)
TSS = sum((data$Salary - mean)^2)
R2 = 1-RSS/TSS
R2
```

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

Answer:

The regression coefficients and standard errors are shown above. RSS is 43058621, R^2 is 0.1924355. This linear regression fits the data in some way, because it follows that the larger the Hits is, the larger Salary is. Also, there are sufficient points having a tedency to spread around the linear regression. However, due to the obviously high variability of the data, many outliers dispersedly distribute in the graph. We can also conclude from the small R^2 that this line does not fit well in some way. Thus, salary may not only depend on Hits, but may depend on other factors which will be discussed in Q2.

Q2

```
slm_mult =lm(Salary ~ Hits + Walks + PutOuts +  CHits,data=data)
summary(slm_mult)$coefficients[,1:2]
```

```
##                  Estimate  Std. Error
## (Intercept) -109.8348083 56.44049413
## Hits           1.8460077  0.58106103
## Walks          3.4611108  1.21166094
## PutOuts        0.2709063  0.07861078
## CHits          0.3124567  0.03349647
```

```
predicted_mult = predict(slm_mult, data)
RSS_mult = sum((data$Salary - predicted_mult)^2)
RSS_mult
```

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

```
TSS_mult = sum((data$Salary - mean)^2 )
R2_mult = 1-RSS_mult/TSS_mult
R2_mult
```

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

```
summary(slm_mult)$coefficients[,3:4]
```

```
##               t value     Pr(>|t|)
## (Intercept) -1.946028 5.273704e-02
## Hits         3.176960 1.669445e-03
## Walks        2.856501 4.632200e-03
## PutOuts      3.446172 6.636175e-04
## CHits        9.328047 5.108227e-18
```

```
summary(slm_mult)$coefficients[,4] < 0.05
```

```
## (Intercept)        Hits       Walks     PutOuts       CHits 
##       FALSE        TRUE        TRUE        TRUE        TRUE
```

Answer:

The regression coefficients and standard errors are shown above. RSS is 29223384. R^2 is 0.4519154.T_statstics for all four coefficients are Hits Walks PutOuts CHits 3.176960 2.856501 3.446172 9.328047, and we reject null hypotheses for all four coefficients(meaning they are not equal to zero).

Q3

```
p = 4
p0 = 1
n = nrow(data)
F_stat = ((RSS-RSS_mult)/(p-p0))/(RSS_mult/(n-p-1))
F_stat
```

```
## [1] 40.71501
```

```
p = pf(F_stat, p-p0, n-p-1, lower.tail = FALSE)
p
```

```
## [1] 1.417223e-21
```

Answer:
In model 2, R^2 of multivariate regression is greater and RSS of multivariate regression is smaller. Thus, the model 2 is better than the model 1. As p_value is 1.417223e-21, which is really small, we reject the null hypothesis that a subset of the covariates have zero regression coefficients. Thus, multivariate regression is better.