Math 189 HW6 Solutions

Q1.

First read in the dataset:

```
> baseball = read.csv("baseball_5.csv")
> head(baseball,n = 10)
      Salary Hits Walks PutOuts CHits
1   475.000   81    39     632   835
2   480.000  130    76     880   457
3   500.000  141    37     200  1575
4    91.500   87    30     805   101
5   750.000  169    35     282  1133
6    70.000   37    21      76    42
7   100.000   73     7     121   108
8    75.000   81     8     143    86
9  1100.000   92    65       0  1332
10  517.143  159    59     238  1300
> attach(baseball)
```

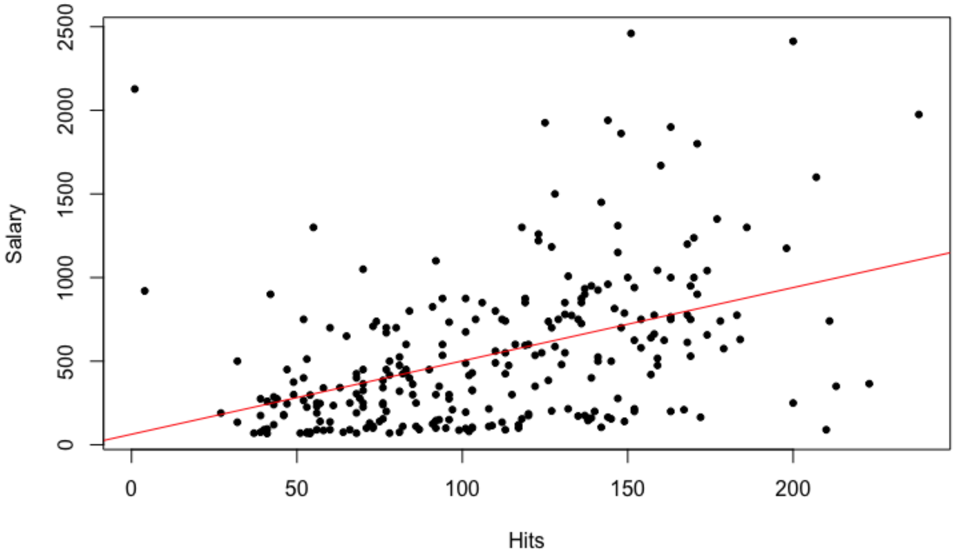
Then fit the linear model:

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

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
```

The Residual Sum of Squares (RSS) and R^2 can be drawn out using the following codes:

```
> RSS.univ = deviance(lm.univ)
> RSS.univ
[1] 43058621
> R2.univ = summary(lm.univ)$r.squared
> R2.univ
[1] 0.1924355
```

Finally, make a plot with the fitted line:



We can see that the line is not a good fit here. Also, the R^2 is too low.

Q2.

We fit a multivariate linear model with all the other variables:

```
> lm.multi = lm(Salary ~ ., data = baseball)
> summary(lm.multi)

Call:
lm(formula = Salary ~ ., data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max 
-811.49 -169.57  -40.38  108.18 2211.38 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) -109.83481   56.44049  -1.946 0.052737 . 
Hits           1.84601    0.58106   3.177 0.001669 **
```

```
Walks          3.46111    1.21166   2.857 0.004632 **
PutOuts        0.27091    0.07861   3.446 0.000664 ***
CHits          0.31246    0.03350   9.328  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 336.6 on 258 degrees of freedom
Multiple R-squared:  0.4519,  Adjusted R-squared:  0.4434
F-statistic: 53.18 on 4 and 258 DF,  p-value: < 2.2e-16
```

The RSS and R^2:

```
> RSS.multi = deviance(lm.multi)
> RSS.multi
[1] 29223384
> R2.multi = summary(lm.multi)$r.squared
> R2.multi
[1] 0.4519154
```

Based on the result of summary function, all variables are statistically significant at significance level 0.05.

Q3.

We can directly apply `anova` function:

```
> anova(lm.univ, lm.multi)
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
```

By the result above, we reject the null and conclude that the multivariable linear model is more adequate.