MATH 189 HW5

Zijian Su
Zelong Zhou
Xiangyi Lin

Last Updated: February 17, 2023

Concrete contributions

All problems were done by Zijian Su, Zelong Zhou, Xiangyi Lin. All contributing equally to this assignment. Everyone put in enough effort.

Overview

The baseball dataset collected the statistics of 263 players in Major League Baseball in the season 1986-1987. This dataset (baseball_5.csv) contains 5 variables selected from the original baseball dataset. The variable names and descriptions are listed in the table below.

Variable	Description
Salary	1987 annual salary on opening day in thousands of dollars
Hits	Number of hits in 1986
Walks	Number of walks in 1986
PutOuts	Number of put outs in 1986
CHits	Number of hits during his career

Packages

```r
#install.packages("rmarkdown")
#install.packages("scatterplot3d")
```

Question 1

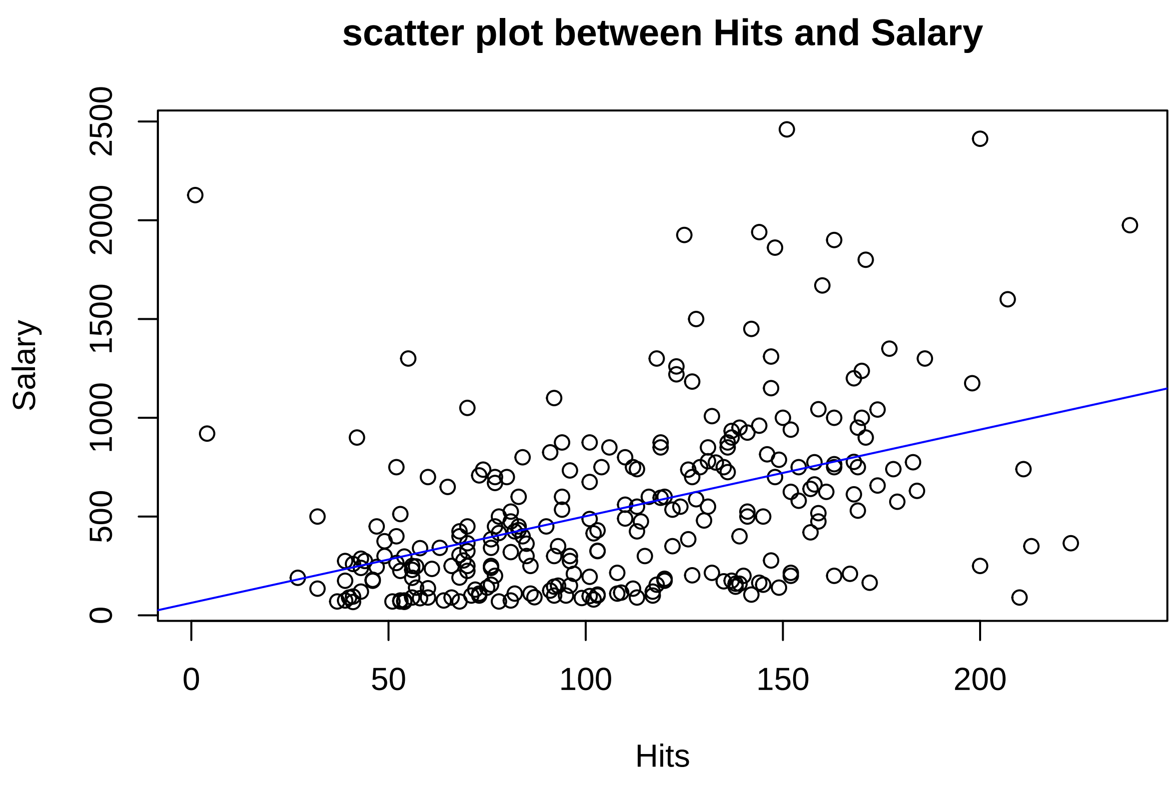
Draw a scatter plot between Hits and Salary. Consider a simple linear regression using Hits as predictor. Estimate the regression coefficients and their standard errors. Add a line to the scatter plot according to the predicted linear curve. Do you think this line fits the data well? Calculate Residual Sum of Squares (RSS) and ***R*^2**

Answer:

```
baseball <- read.csv("baseball_5.csv")

plot(baseball$Hits, baseball$Salary, xlab = 'Hits', ylab = 'Salary', main = 'scatter plot between Hits a

model1 <- lm(Salary ~ Hits, data = baseball)
abline(model1, col = 'blue')
```




```
model1
```

```
##
## Call:
## lm(formula = Salary ~ Hits, data = baseball)
##
## Coefficients:
```

```
## (Intercept)         Hits
##      63.049        4.385
```

```r
summary(model1)
```

```
##
## 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 regression coefficients B0 is 63.049, B1 is 4.385. The standard errors for B0 is 64.98, for B1 is 0.5561.

```r
RSS_1 <- sum(model1$residuals^2)
TSS_1 <- sum((baseball$Salary - mean(baseball$Salary))^2)
R2_1 <- 1 - RSS_1/TSS_1
cat("RSS: ", RSS_1, "\n")
```

```
## RSS:  43058621
```

```r
cat("R^2: ", R2_1)
```

```
## R^2:  0.1924355
```

Residual Sum of Squares (RSS) is about 43058621 and the ***R^2*** is about 0.1924355.

Through the value of R^2, we can know whether the regression line fits the data. When R^2 is closer to 1, it fits better. The closer to 0, the less the match. We get an R^2 of only about 0.19, which indicates that only about 19% of the points are on this line. Therefore, we don't think this line fits the data very well.

Question 2

Consider a multivariate linear model using Hits, Walks, PutOuts and CHits as predictors. Report the estimated regression coefficients and their standard errors. Calculate Residual Sum of Squares (RSS) and $\boldsymbol{R}$**^2**. Test the marginal effects of each coefficient.

Answer:

```
model2 <- lm(Salary ~ Hits+Walks+PutOuts+CHits, data = baseball)
model2
```

```
##
## Call:
## lm(formula = Salary ~ Hits + Walks + PutOuts + CHits, data = baseball)
##
## Coefficients:
## (Intercept)         Hits        Walks      PutOuts        CHits
##   -109.8348       1.8460       3.4611       0.2709       0.3125
```

The regression coefficients B0 is -109.8348, B1 is 1.8460, B2 is 3.4611, B3 is 0.2709, B4 is 0.3125 . The standard errors for B0 is 56.44049, for B1 is 00.58106, for B2 is 1.21166., for B3 is 0.07861, for B4 is 0.03350.

```
summary(model2)
```

```
##
## Call:
## lm(formula = Salary ~ Hits + Walks + PutOuts + CHits, 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
```

```
RSS_2 <- sum(model2$residuals^2)
TSS_2 <- sum((baseball$Salary - mean(baseball$Salary))^2)
R2_2 <- 1 - RSS_2/TSS_2
cat("RSS: ", RSS_2, "\n")
```

```
## RSS:  29223384
```

```r
cat("R^2: ", R2_2)
```

```
## R^2:  0.4519154
```

Residual Sum of Squares (RSS) is about 29223384 and the ***R^2*** is about 0.4519154.

Test the marginal effects of each coefficient:

Consider null and alternative hypotheses: H_0 Bj = 0 versus H_1 Bj != 0, for some j = 1, . . . ,p. (a= 0.05),

```r
rse <- summary(model2)$sigma
b1 <- summary(model2)$coefficients[2]
b2 <- summary(model2)$coefficients[3]
b3 <- summary(model2)$coefficients[4]
b4 <- summary(model2)$coefficients[5]
b <- c(summary(model2)$coefficients[2:5])

## get the matrix
v1 <- model.matrix(Salary ~ Hits + Walks + PutOuts + CHits, data = baseball)
v1_t <- t(v1)
v1_t_v1_in <- solve((v1_t %*% v1))
d1 <- diag(v1_t_v1_in)


## find p1
t1 <- b1 / (sqrt(d1[2])* rse)
#t1
p1 <- 2*pt(t1, 258, lower.tail = FALSE)
#p1

## find p2
t2 <- b2 / (sqrt(d1[3])* rse)
#t2
p2 <- 2*pt(t2, 258, lower.tail = FALSE)
#p2

## find p3
t3 <- b3 / (sqrt(d1[4])* rse)
#t3
p3 <- 2*pt(t3, 258, lower.tail = FALSE)
#p3

#find p4
t4 <- b4 / (sqrt(d1[5])* rse)
#t4
p4 <- 2*pt(t4, 258, lower.tail = FALSE)
#p4
```

```r
cat("t-statistic for all coefficient:\n")
```

```
## t-statistic for all coefficient:
```

```r
t1
```

```
##    Hits
## 3.17696
```

```r
t2
```

```
##    Walks
## 2.856501
```

```r
t3
```

```
##  PutOuts
## 3.446172
```

```r
t4
```

```
##    CHits
## 9.328047
```

```r
cat("p-value for all coefficient:\n")
```

```
## p-value for all coefficient:
```

```r
p1
```

```
##        Hits
## 0.001669445
```

```r
p2
```

```
##     Walks
## 0.0046322
```

```r
p3
```

```
##      PutOuts
## 0.0006636175
```

```r
p4
```

```
##        CHits
## 5.108227e-18
```

We do some test above.

Hit: P -value < 0.05, I reject H_0 and support H_1.
Walks: P -value < 0.05, I reject H_0 and support H_1.
PutOuts: P -value < 0.05, I reject H_0 and support H_1.
CHits: P -value < 0.05, I reject H_0 and support H_1.

Therefore, we can believe Bj != 0, for some j = 1, . . . ,p. (a= 0.05)

Question 3

Compare the model fitted in 2 and 1 by their RSS and ***R*^2**. Test the model adequacy by letting the simple linear model as the null model and the multivariate linear model in 2 as the alternative model. What can you conclude?

Answer:

lets say the null hypothesis(H0) is the model from Q1, and the alternative hypothesis(H1) is the model from Q2.

```
test <- anova(model1, model2)
test
```

```
## 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
```

```
summary(anova(model1, model2))
```

```
##      Res.Df           RSS                   Df      Sum of Sq
##  Min.   :258.0   Min.   :29223384   Min.   :3   Min.   :13835237
##  1st Qu.:258.8   1st Qu.:32682193   1st Qu.:3   1st Qu.:13835237
##  Median :259.5   Median :36141003   Median :3   Median :13835237
##  Mean   :259.5   Mean   :36141003   Mean   :3   Mean   :13835237
##  3rd Qu.:260.2   3rd Qu.:39599812   3rd Qu.:3   3rd Qu.:13835237
##  Max.   :261.0   Max.   :43058621   Max.   :3   Max.   :13835237
##                                     NA's   :1   NA's   :1
##        F              Pr(>F)
##  Min.   :40.72   Min.   :0
##  1st Qu.:40.72   1st Qu.:0
##  Median :40.72   Median :0
##  Mean   :40.72   Mean   :0
##  3rd Qu.:40.72   3rd Qu.:0
##  Max.   :40.72   Max.   :0
##  NA's   :1       NA's   :1
```

We can know that the F-statistic value is 40.715. And at the significance level a = 0.05, the p-value is 2.2e-16 < 0.0001. So, we reject the null hypothesis(H0).

Therefore, the multivariate linear model is better in this data.