Assignment 4

Shinjon Ghosh

2025-02-07

```r
library (ISLR)
data(Hitters)
names(Hitters)
```

```
##  [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"       "Walks"
##  [7] "Years"     "CAtBat"    "CHits"     "CHmRun"    "CRuns"     "CRBI"
## [13] "CWalks"    "League"    "Division"  "PutOuts"   "Assists"   "Errors"
## [19] "Salary"    "NewLeague"
```

```r
head(Hitters)
```

```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits
CHmRun
## -Andy Allanson      293   66     1   30  29    14     1    293    66
1
## -Alan Ashby         315   81     7   24  38    39    14   3449   835
69
## -Alvin Davis        479  130    18   66  72    76     3   1624   457
63
## -Andre Dawson       496  141    20   65  78    37    11   5628  1575
225
## -Andres Galarraga   321   87    10   39  42    30     2    396   101
12
## -Alfredo Griffin    594  169     4   74  51    35    11   4408  1133
19
##                   CRuns CRBI CWalks League Division PutOuts Assists Errors
## -Andy Allanson       30   29     14      A        E     446      33     20
## -Alan Ashby         321  414    375      N        W     632      43     10
## -Alvin Davis        224  266    263      A        W     880      82     14
## -Andre Dawson       828  838    354      N        E     200      11      3
## -Andres Galarraga    48   46     33      N        E     805      40      4
## -Alfredo Griffin    501  336    194      A        W     282     421     25
##                   Salary NewLeague
## -Andy Allanson        NA         A
## -Alan Ashby        475.0         N
## -Alvin Davis       480.0         A
## -Andre Dawson      500.0         N
## -Andres Galarraga   91.5         N
## -Alfredo Griffin   750.0         A
```

```r
str(Hitters)
```

```
## 'data.frame':    322 obs. of  20 variables:
##  $ AtBat    : int  293 315 479 496 321 594 185 298 323 401 ...
```

```
##  $ Hits     : int  66 81 130 141 87 169 37 73 81 92 ...
##  $ HmRun    : int  1 7 18 20 10 4 1 0 6 17 ...
##  $ Runs     : int  30 24 66 65 39 74 23 24 26 49 ...
##  $ RBI      : int  29 38 72 78 42 51 8 24 32 66 ...
##  $ Walks    : int  14 39 76 37 30 35 21 7 8 65 ...
##  $ Years    : int  1 14 3 11 2 11 2 3 2 13 ...
##  $ CAtBat   : int  293 3449 1624 5628 396 4408 214 509 341 5206 ...
##  $ CHits    : int  66 835 457 1575 101 1133 42 108 86 1332 ...
##  $ CHmRun   : int  1 69 63 225 12 19 1 0 6 253 ...
##  $ CRuns    : int  30 321 224 828 48 501 30 41 32 784 ...
##  $ CRBI     : int  29 414 266 838 46 336 9 37 34 890 ...
##  $ CWalks   : int  14 375 263 354 33 194 24 12 8 866 ...
##  $ League   : Factor w/ 2 levels "A","N": 1 2 1 2 2 1 2 1 2 1 ...
##  $ Division : Factor w/ 2 levels "E","W": 1 2 2 1 1 2 1 2 2 1 ...
##  $ PutOuts  : int  446 632 880 200 805 282 76 121 143 0 ...
##  $ Assists  : int  33 43 82 11 40 421 127 283 290 0 ...
##  $ Errors   : int  20 10 14 3 4 25 7 9 19 0 ...
##  $ Salary   : num  NA 475 480 500 91.5 750 70 100 75 1100 ...
##  $ NewLeague: Factor w/ 2 levels "A","N": 1 2 1 2 2 1 1 1 2 1 ...
```

```
Hitters <- na.omit(Hitters)
```

1.
```
h1 <- lm(Salary ~ AtBat + Hits + HmRun + Runs, data = Hitters)
summary(h1)
```

```
##
## Call:
## lm(formula = Salary ~ AtBat + Hits + HmRun + Runs, data = Hitters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -884.01 -217.15  -62.26  174.28 1990.28
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 160.5700    76.4265   2.101 0.036614 *
## AtBat        -1.5825     0.6479  -2.442 0.015264 *
## Hits          8.1318     2.2528   3.610 0.000368 ***
## HmRun         9.1861     3.6970   2.485 0.013600 *
## Runs          0.5579     2.6039   0.214 0.830517
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 399.6 on 258 degrees of freedom
## Multiple R-squared:  0.2273, Adjusted R-squared:  0.2153
## F-statistic: 18.97 on 4 and 258 DF,  p-value: 1.091e-13
```

2. The p-value of Runs (0.830517) represents the probability of observing an intercept/slope estimate at least as extreme as what we observe here when the null hypothesis is true.

```
SSTotal <- sum((Hitters$Salary - mean(Hitters$Salary))^2)
print(paste("Total Sum of Squares (SSTotal):", SSTotal))
```

```
## 3. "Total Sum of Squares (SSTotal): 53319112.7886453"
```

```
BIC_value <- BIC(h1)
print(paste("Bayesian Information Criterion (BIC):", BIC_value))
```

```
## 4. "Bayesian Information Criterion (BIC): 3925.75825266015"
```

```
new_sample <- data.frame(AtBat = 380, Hits = 96, HmRun = 8, Runs = 48)
prediction <- predict(h1, newdata = new_sample, interval = "prediction",
level = 0.95)
print("95% Prediction Interval:")
```

```
## 5. "95% Prediction Interval:"
```

```
print(prediction)
```

```
##        fit       lwr      upr
## 1 440.1358 -348.8147 1229.086
```

Prediction Interval solved by manually and get[- 349, 1229].