Baseball Analytics

Zhe Wang

Syllabus

- Where to find the data?
- How to generate the basic graphs?
- What's the relation between runs and wins?
- Career Trajectories
- Run Expectancy

What is Sabermetrics?

Sabermetrics is the empirical analysis of baseball, especially baseball statistics that measure in-game activity.

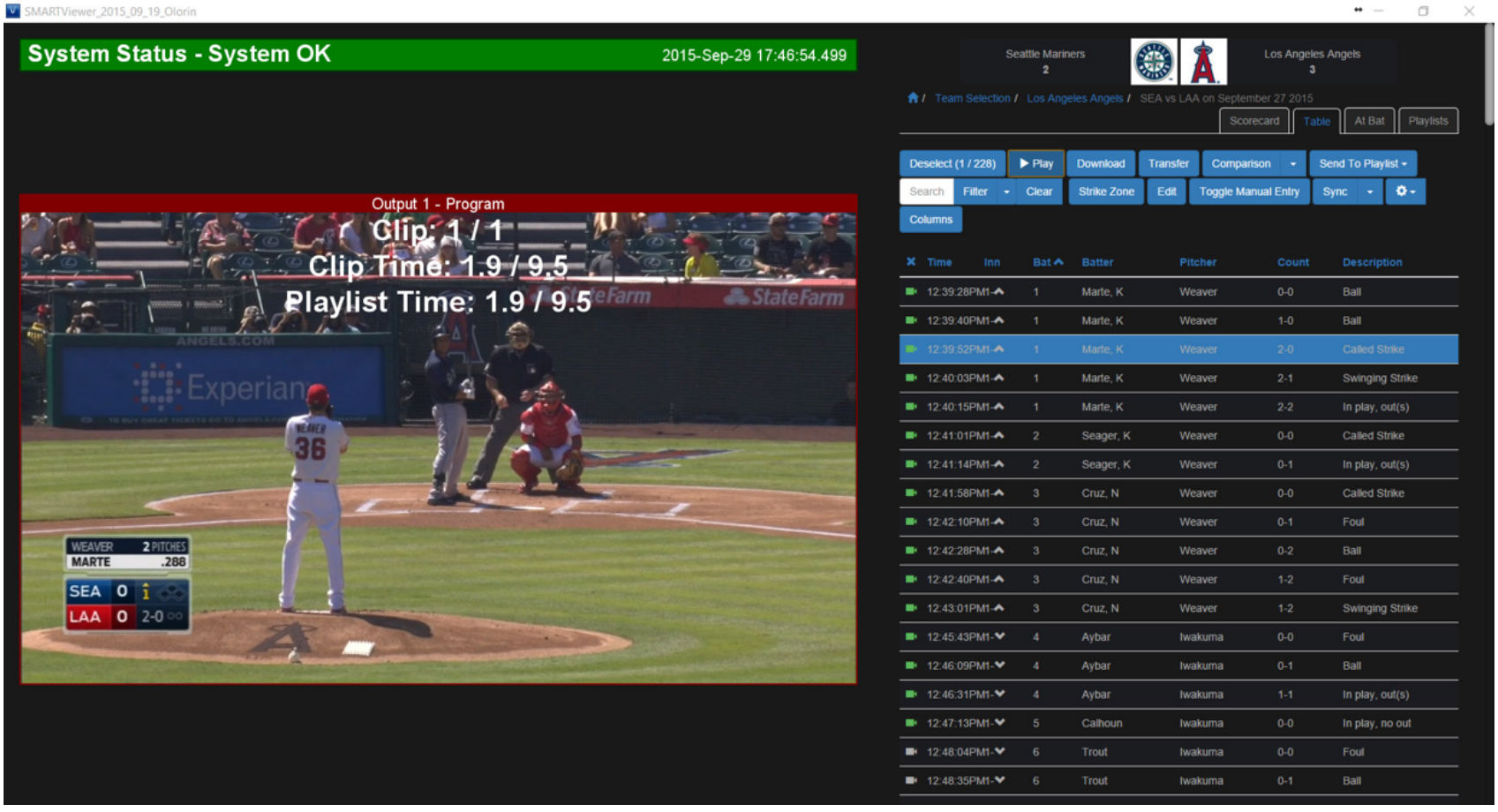


Figure 1: Sabermetrics

What is Sabermetrics?

Sabermetrics is the empirical analysis of baseball, especially baseball statistics that measure in-game activity.

Lahman's Baseball Database

http://www.seanlahman.com/baseball-archive/statistics

Basic Graphs

Need package: “graphics”

Need datafile: “hofbatting.csv”

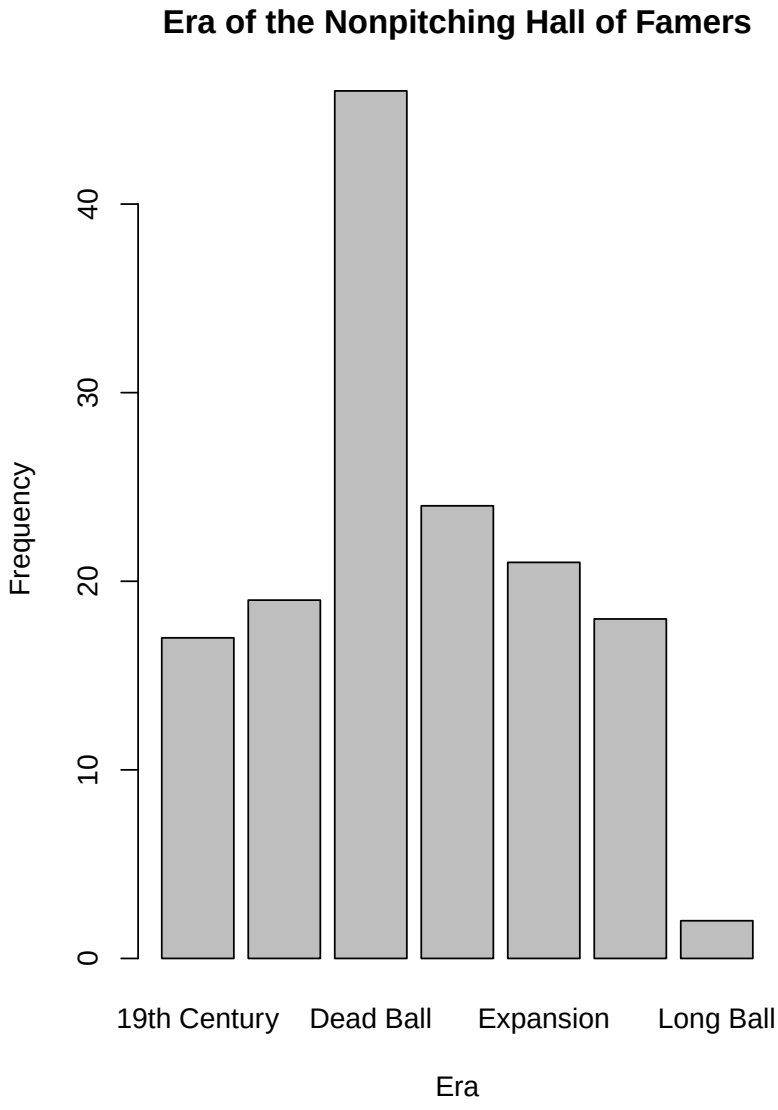
- Generate traditional graphs for factor variable and numeric variable.
- Scatter plots, pie pots, histogram, boxplots, etc.
- Identify particular points from a plot.
- Title, legend, axis lable, etc., for a graph.

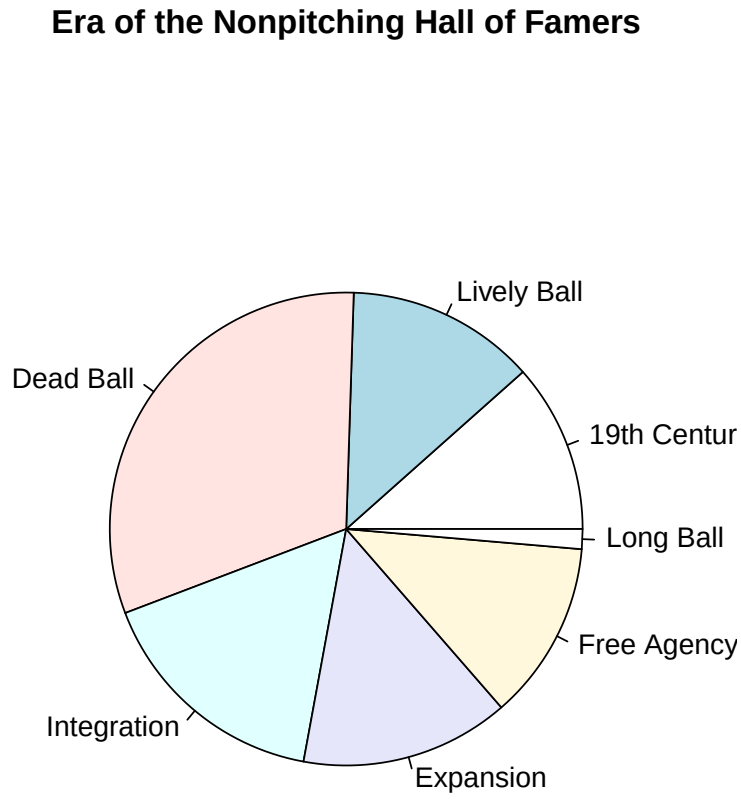
Graphs for Factor Variables

- Create a new factor variable "Era" from numeric variable "MidCareer".
- **Frequency Table**

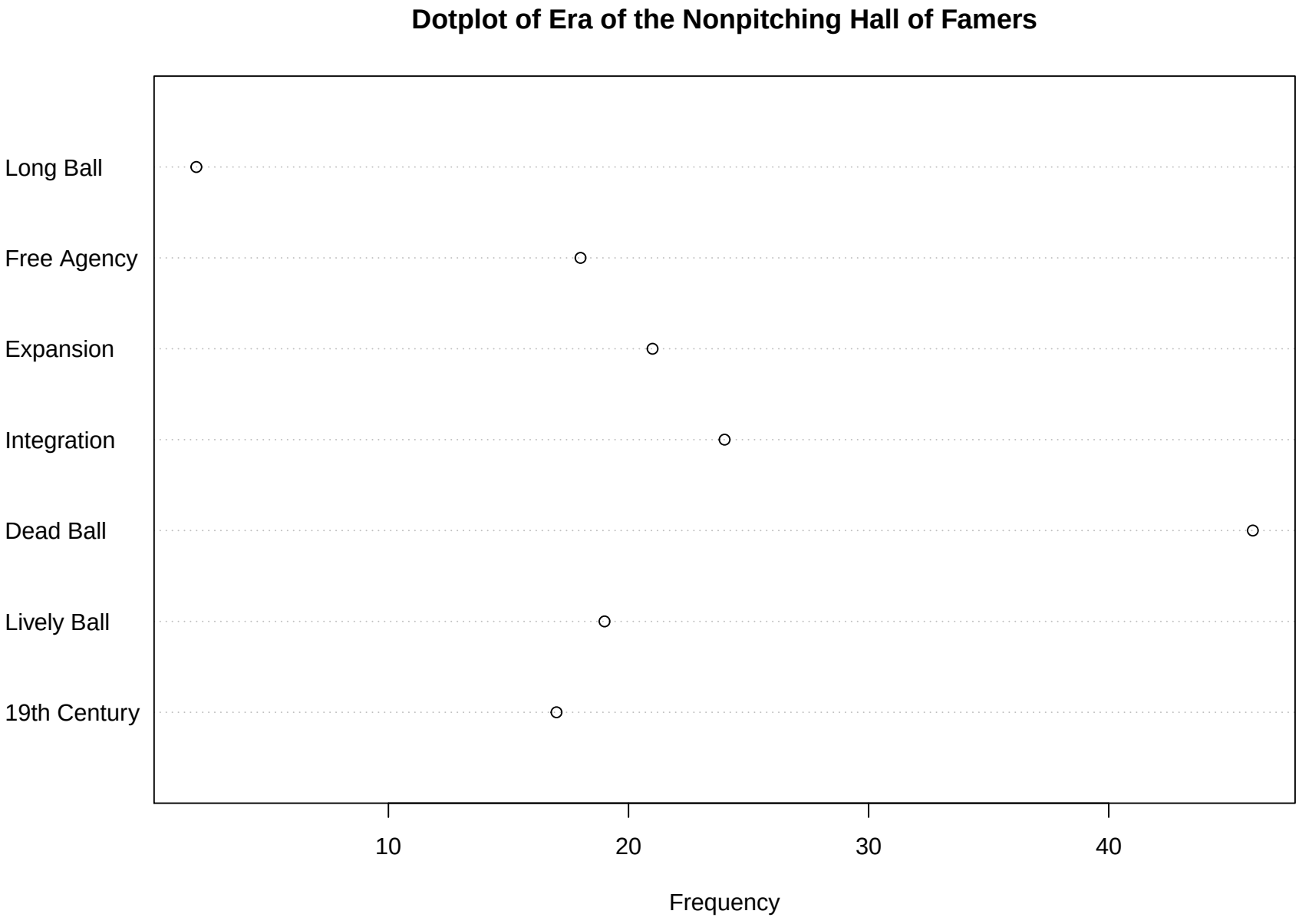
```
##
## 19th Century  Lively Ball    Dead Ball  Integration    F
##           17            19           46           24
##  Free Agency    Long Ball
##           18             2
```

Bar Graph & Pie Graph



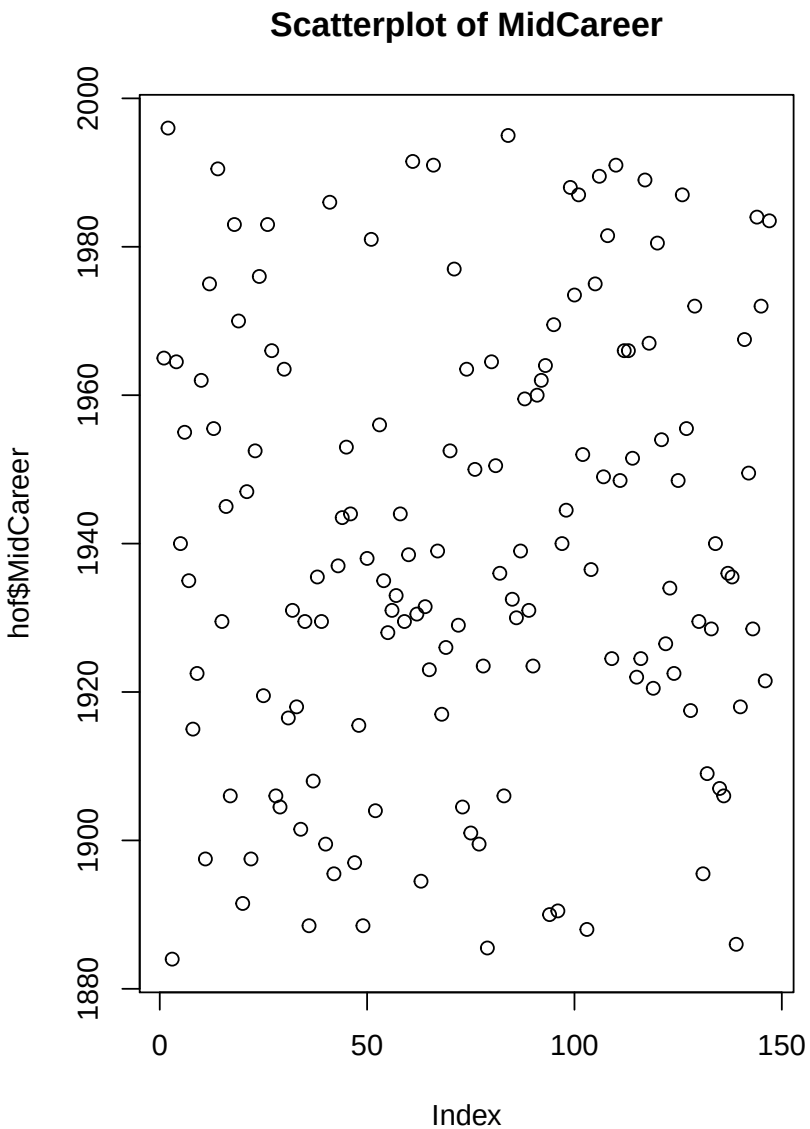


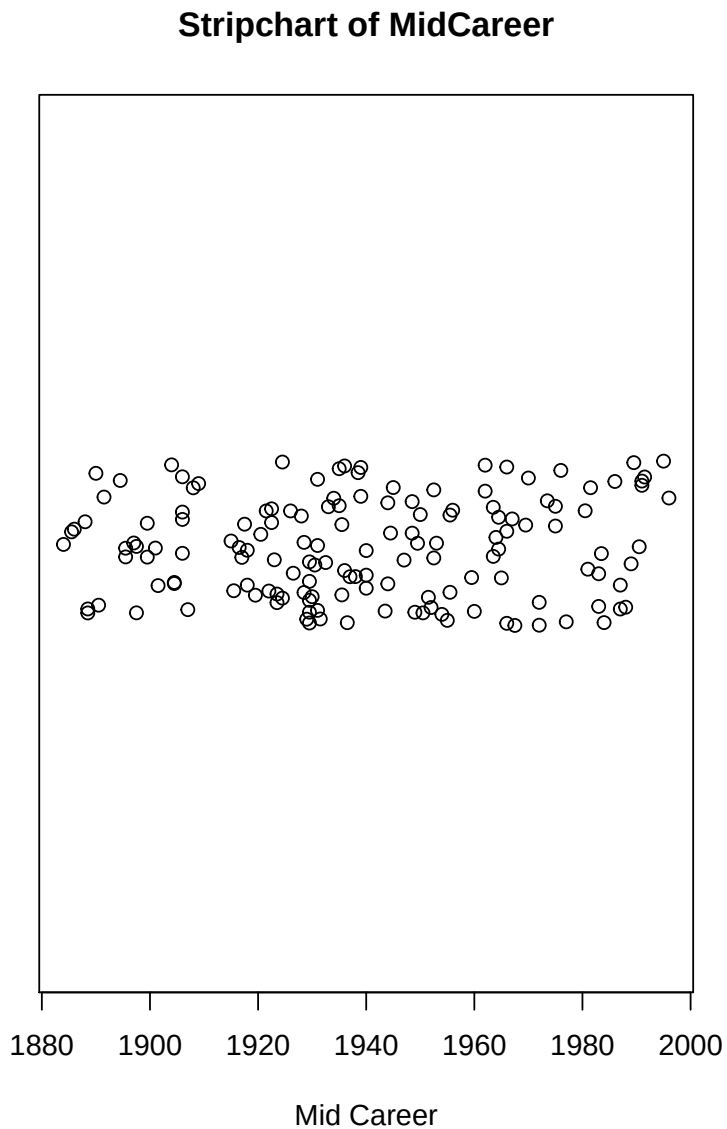
Dotplot


Dotplot of Era of the Nonpitching Hall of Famers
Long Ball
Free Agency
Expansion
Integration
Dead Ball
Lively Ball
19th Century
10
20
30
40
Frequency

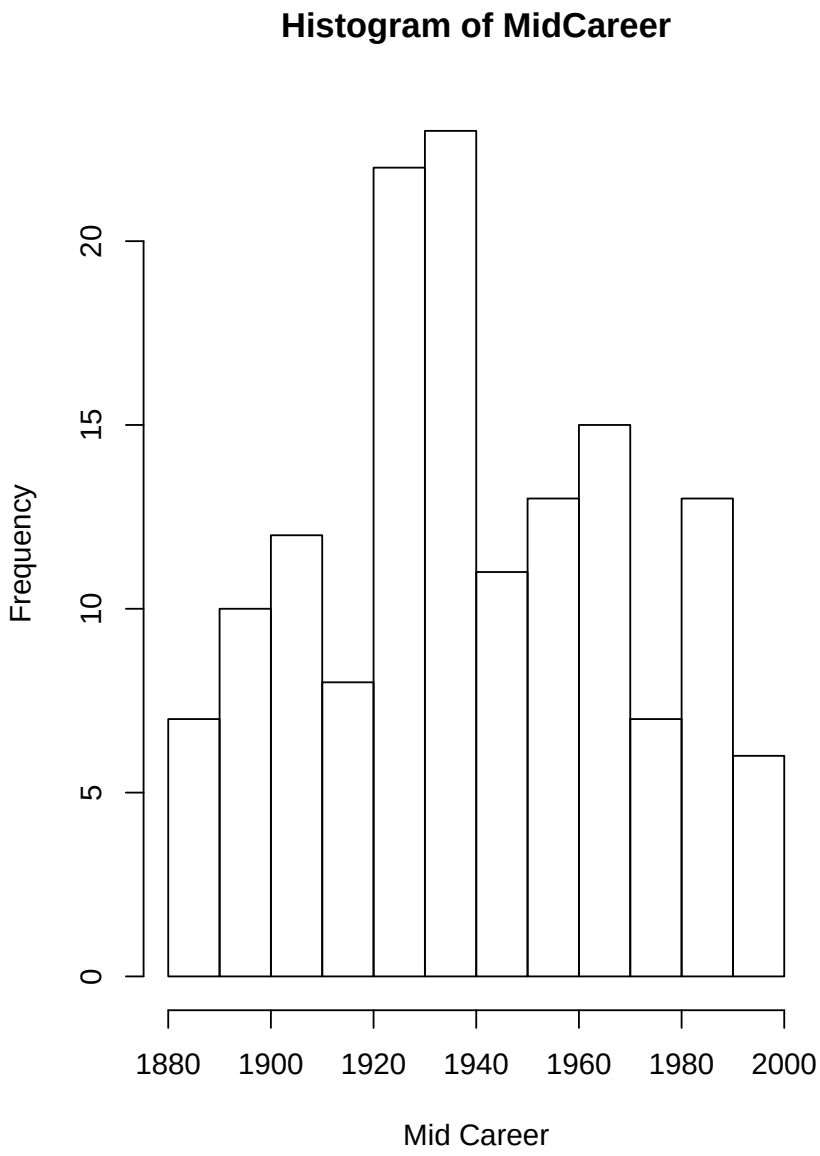
Graphs for Numeric Variables

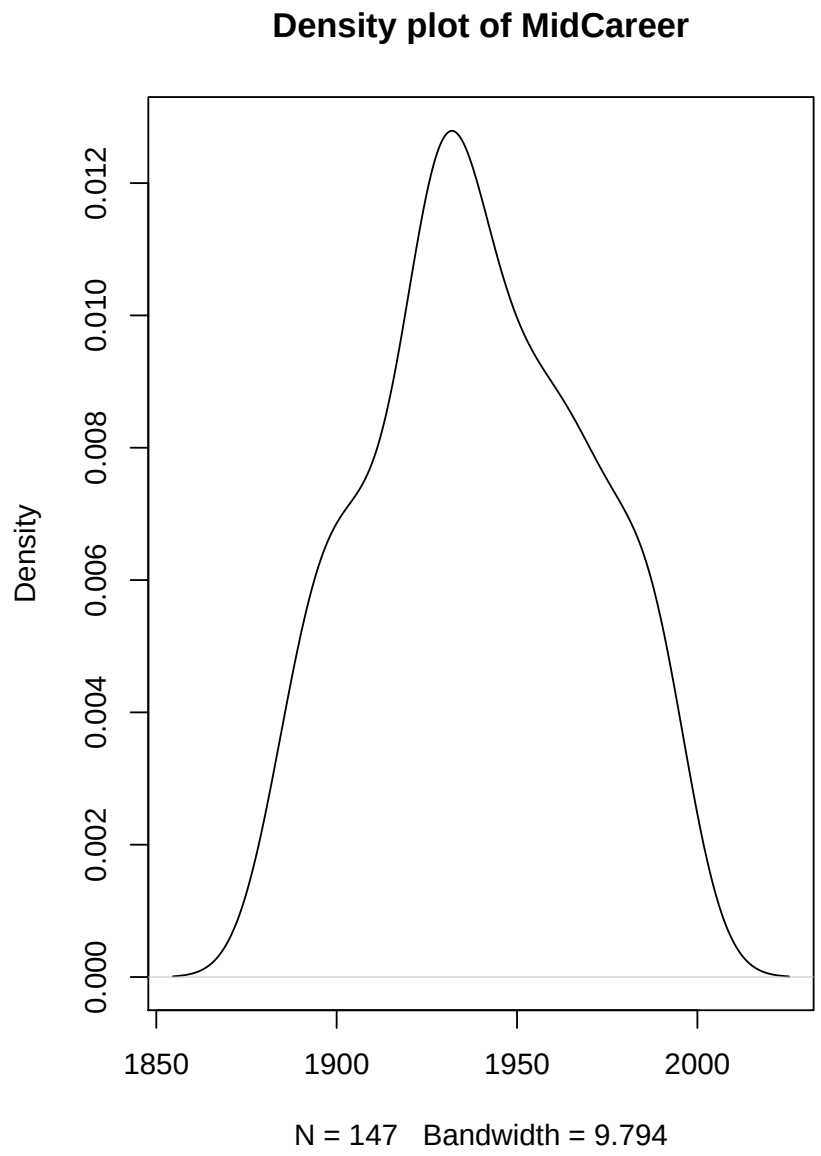
- **Scatterplot & Stripchart**





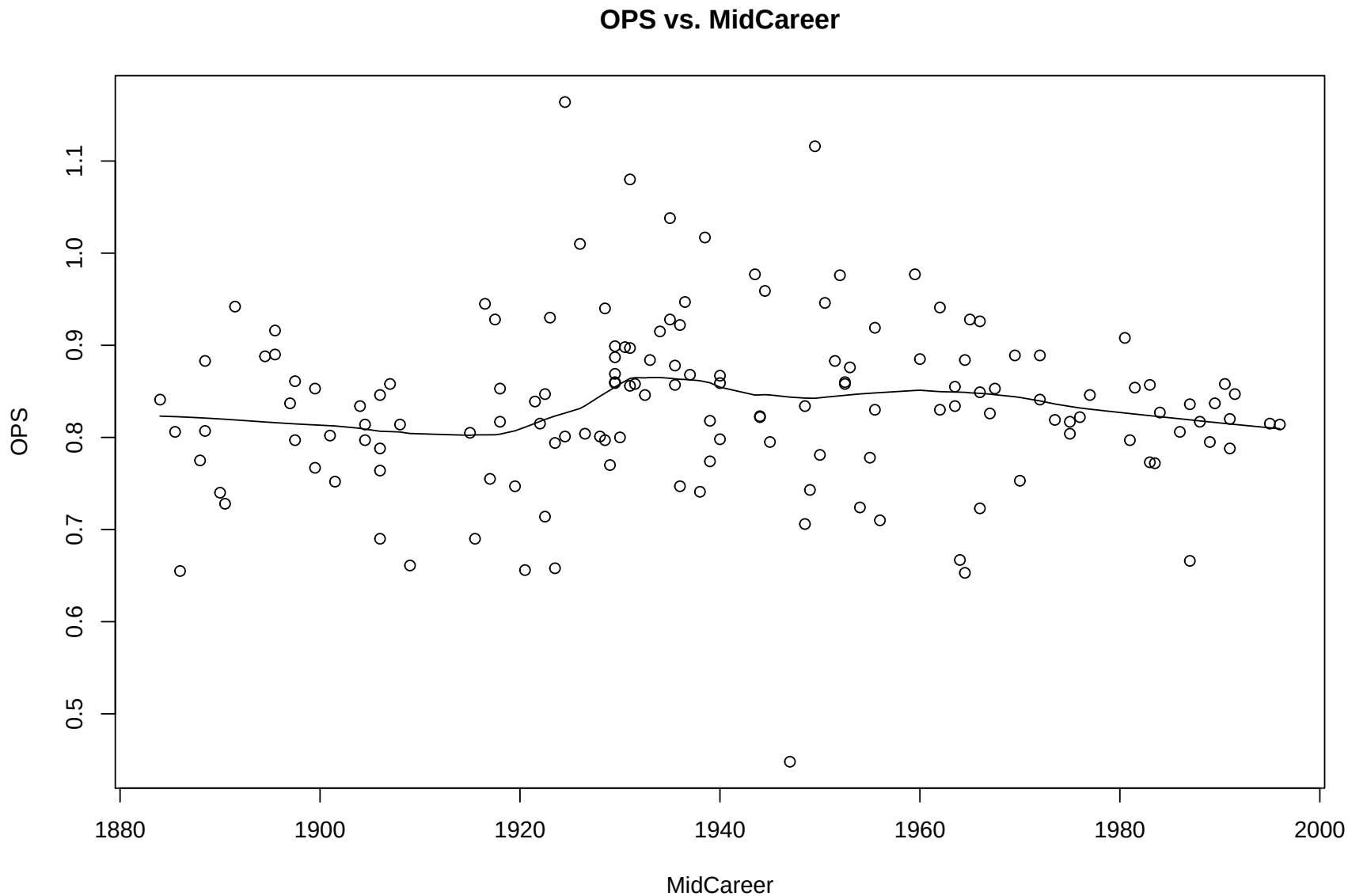
Histogram & Density Plot



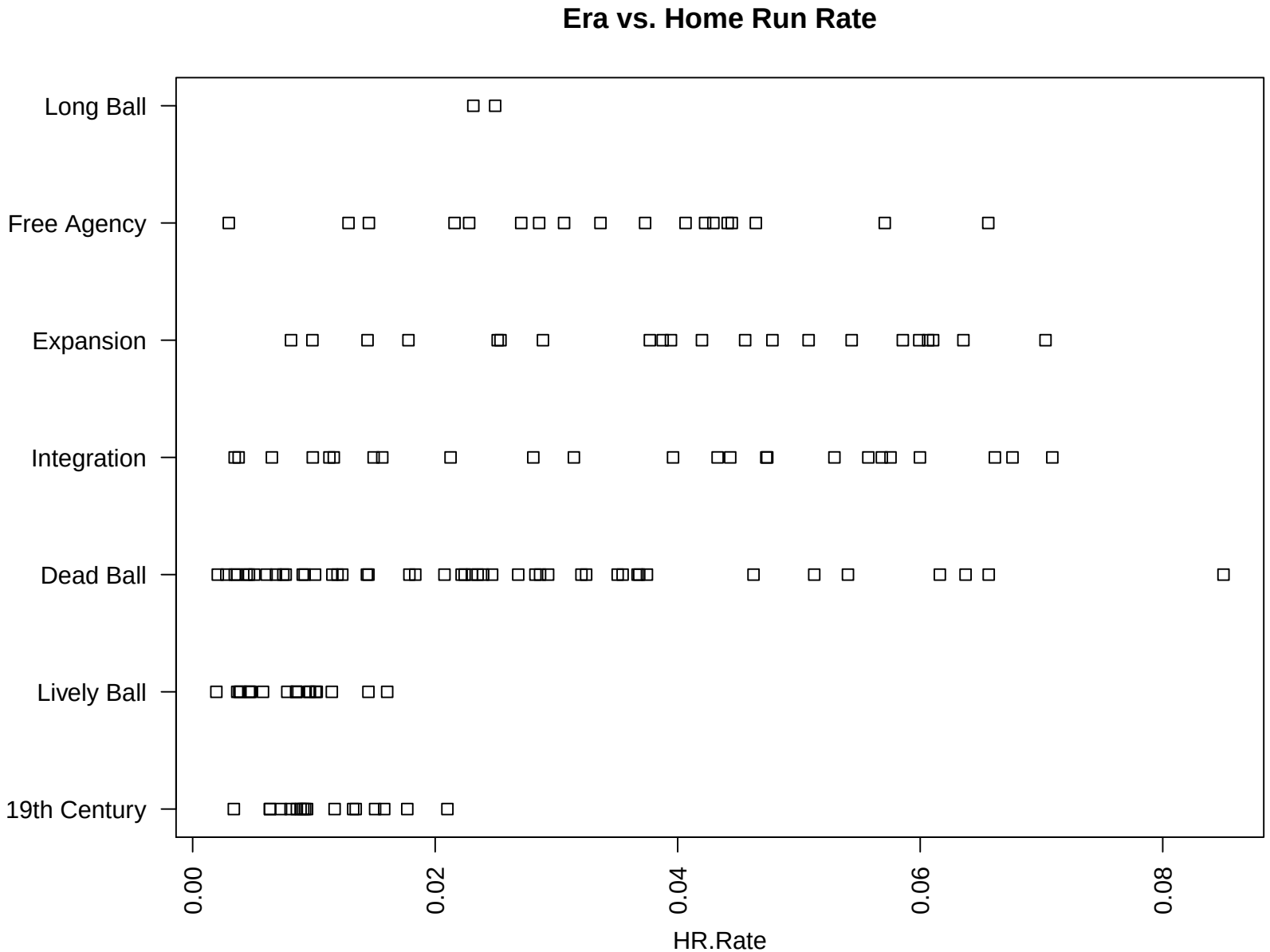


Graphs for Two Variables

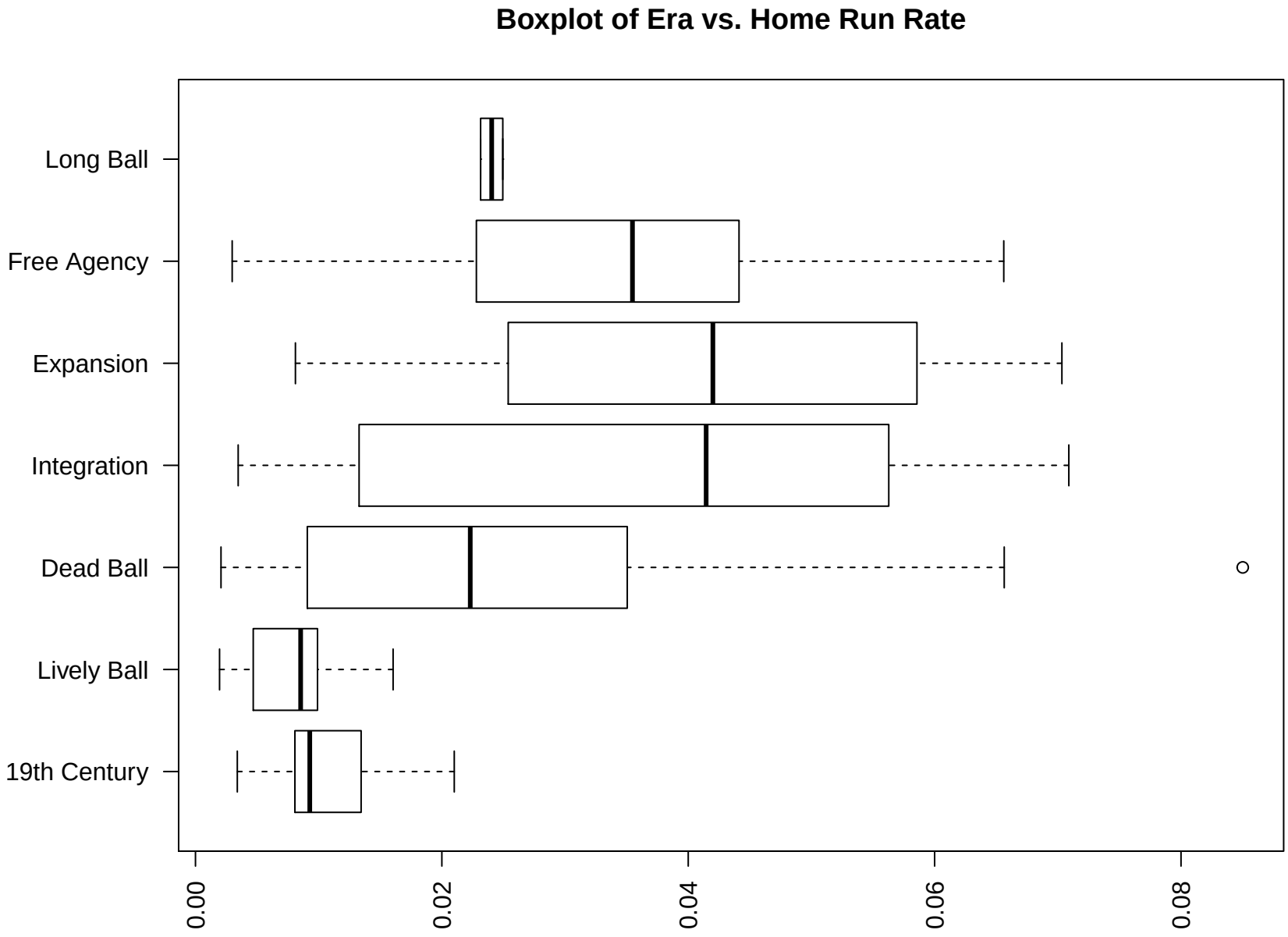
- **Scatterplot & Smoothing Curve**
- How to identify specific points on a graph?



Parallel Stripcharts



Side-by-Side Boxplot



How to save the graphs?

Exercises

Graph the followings all together and save it as a pdf file.

- Scatter plot of Middle Career
- Density plot of Middle Career
- Parallel Stripcharts: Era vs. Home Run Rate
- Side-by-side boxplot: Era vs. Home Run Rate
- Use the records in the year 2018

The Relation between Runs and Wins

Need package: “stats”

Need datafile: “teams.csv”

- Introduction to linear regression.
- The Pythagorean formula for winning percentage.
- Predictions.

The *teams.csv* file from Lahman's database contains seasonal stats for major league teams.

Select the subset and calculate the new variables of interest.

```
##      teamID yearID lgID   G  W  L   R  RA  RD      Wpct
## 2710    PHI   2012   NL 162 81 81 684 680   4 0.5000000
## 2711    PIT   2012   NL 162 79 83 651 674 -23 0.4876543
## 2712    SDN   2012   NL 162 76 86 651 710 -59 0.4691358
## 2713    SFN   2012   NL 162 94 68 718 649  69 0.5802469
## 2714    SLN   2012   NL 162 88 74 765 648 117 0.5432099
## 2715    WAS   2012   NL 162 98 64 731 594 137 0.6049383
```

Linear Regression

A *Simple Linear Model*(SLM) has the following formula:

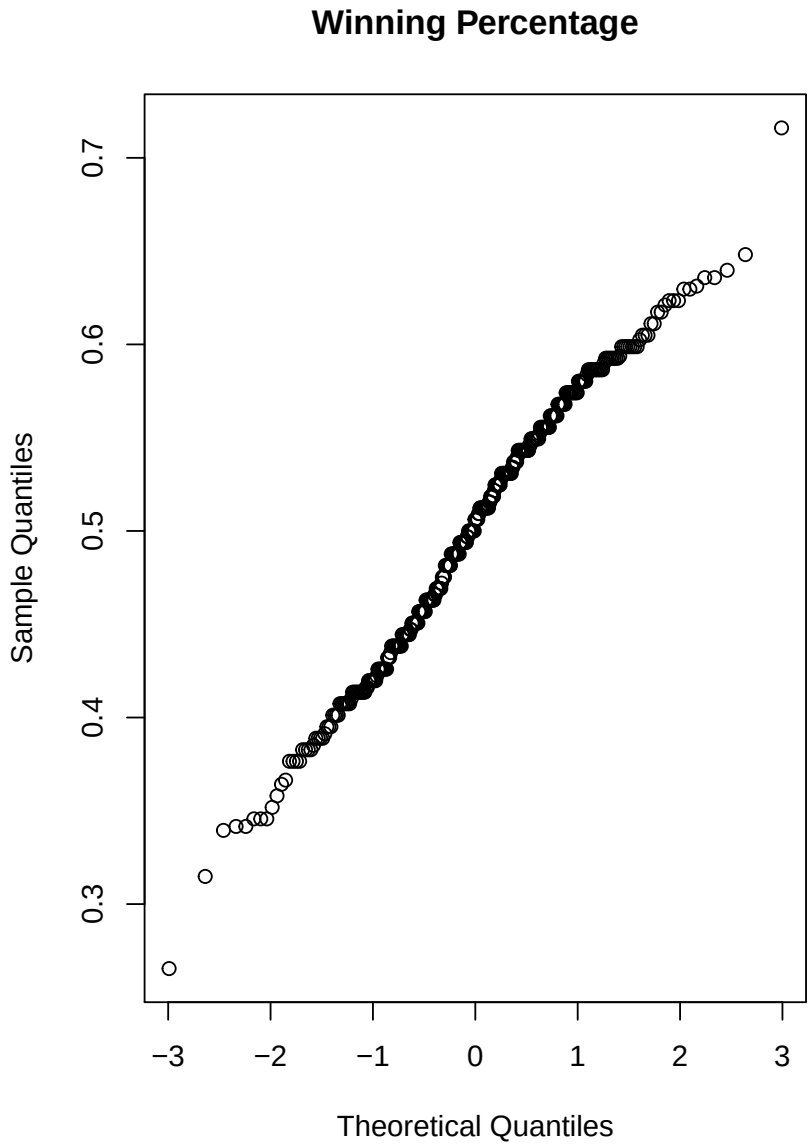
$$Response = a + b \times Predictor + \epsilon$$

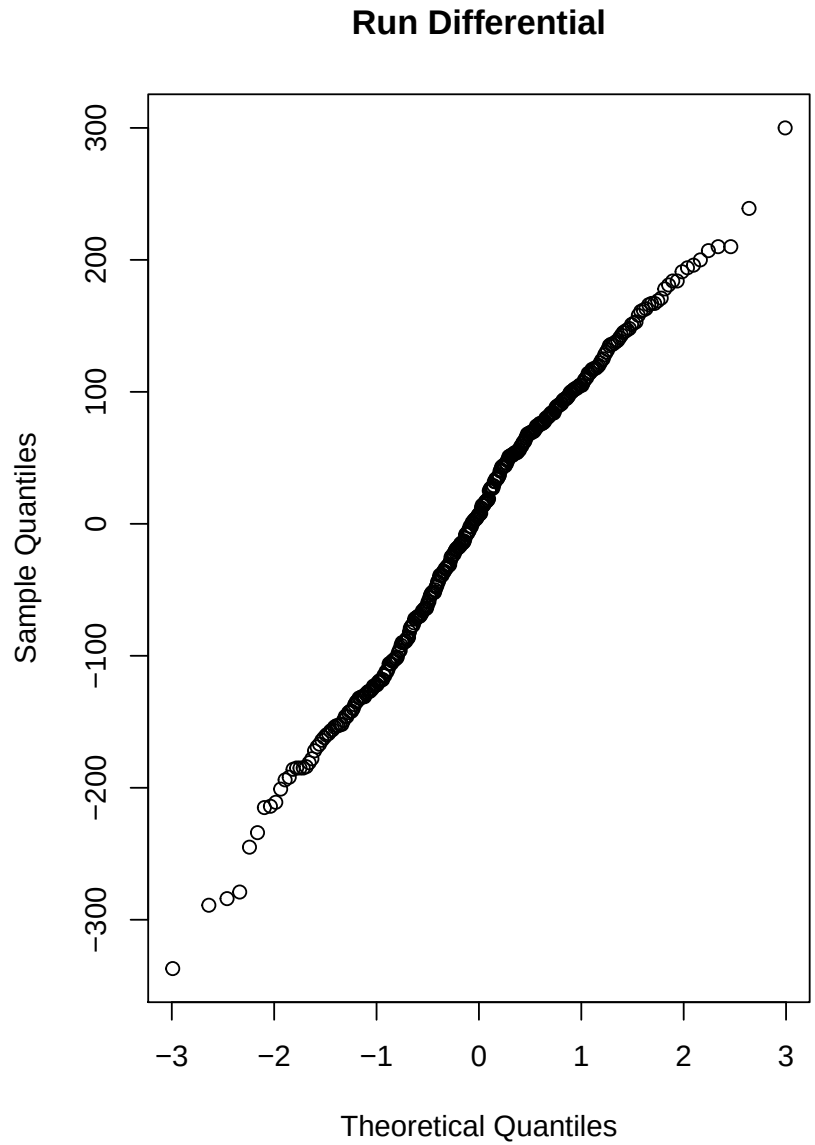
where a and b are unknown constants and ϵ is the error term which captures all other factors influencing the dependent variable (Response).

- Linear relationship
- Normality
- No multicollinearity
- No auto-correlation
- Homoscedasticity

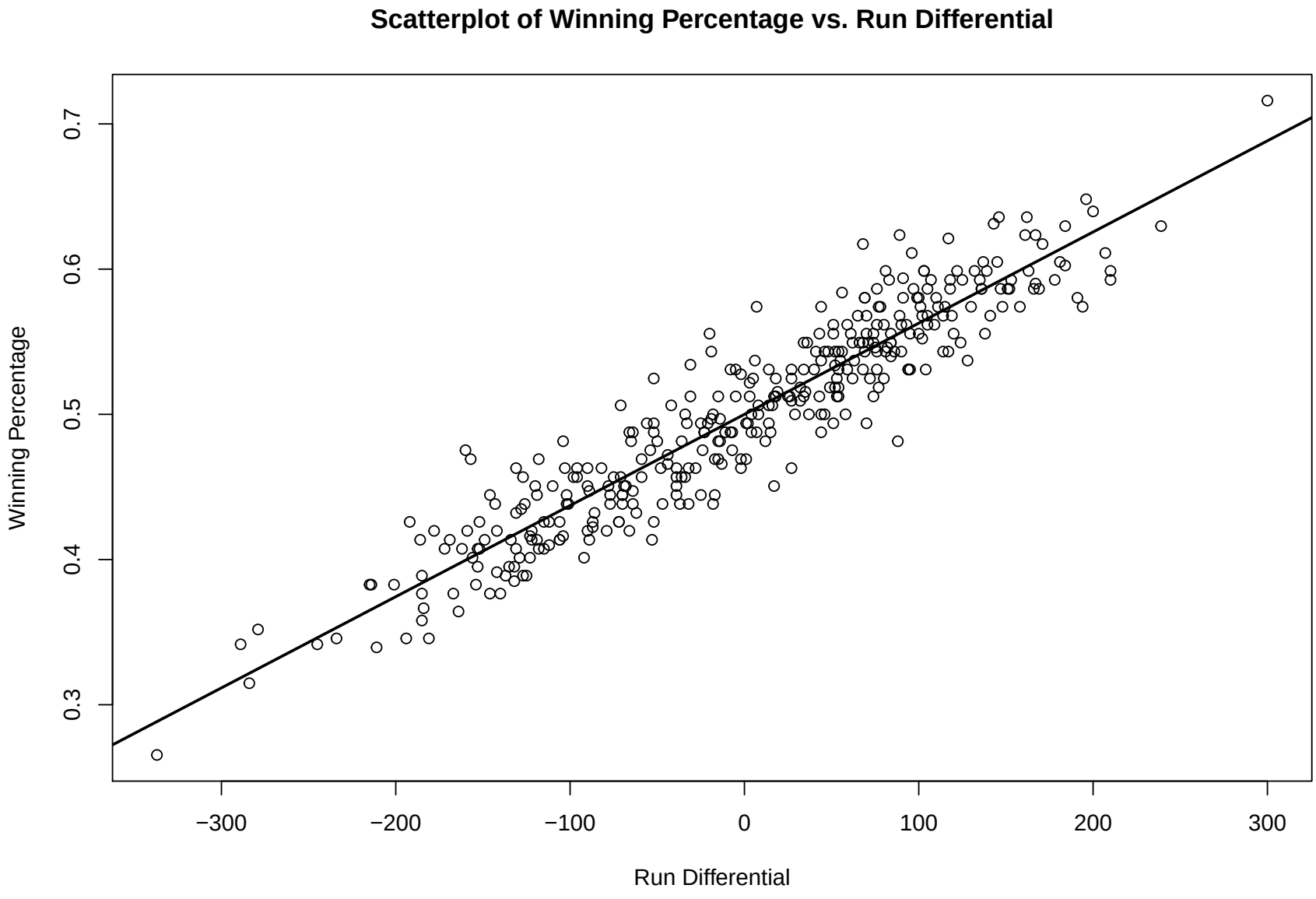
To predict a team's *winning percentage* using *runs scored* with SLM, check the normality by generating QQ plots

$$Wpct = a + b \times RD + \epsilon$$

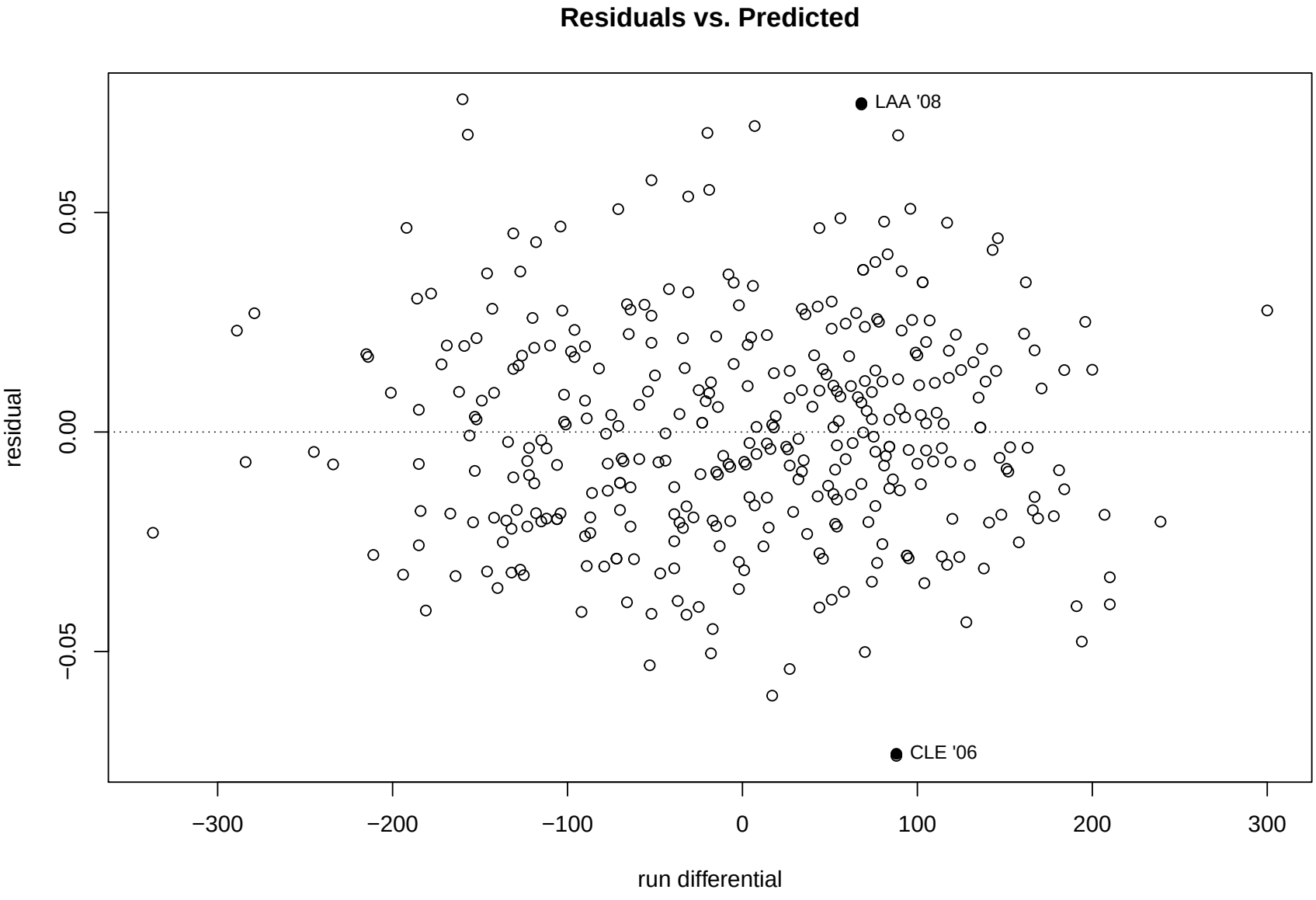




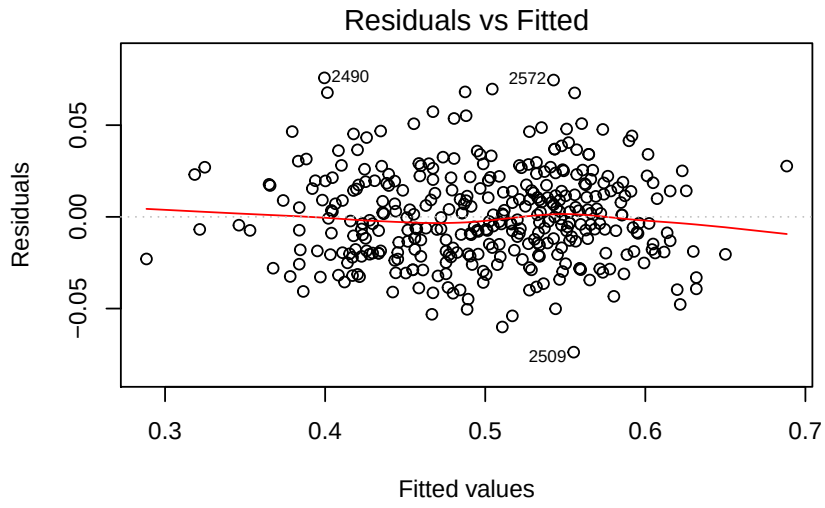
Fit the regression line by “lm” function in “stats” package.

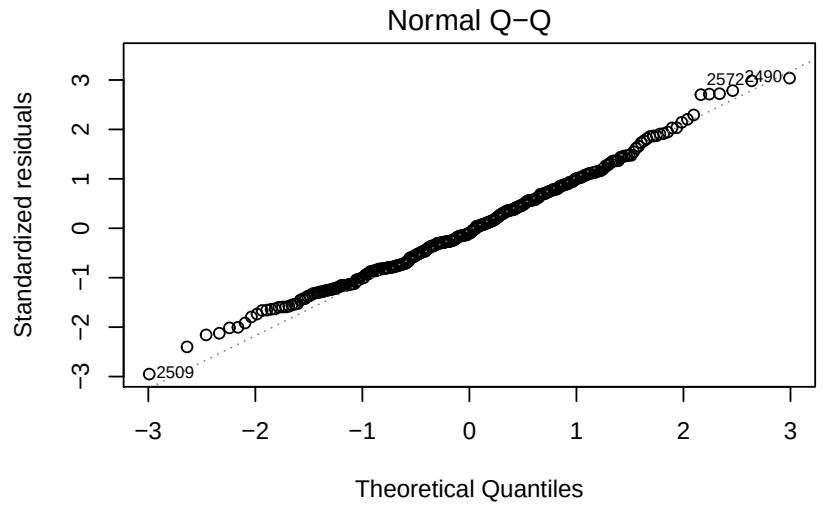


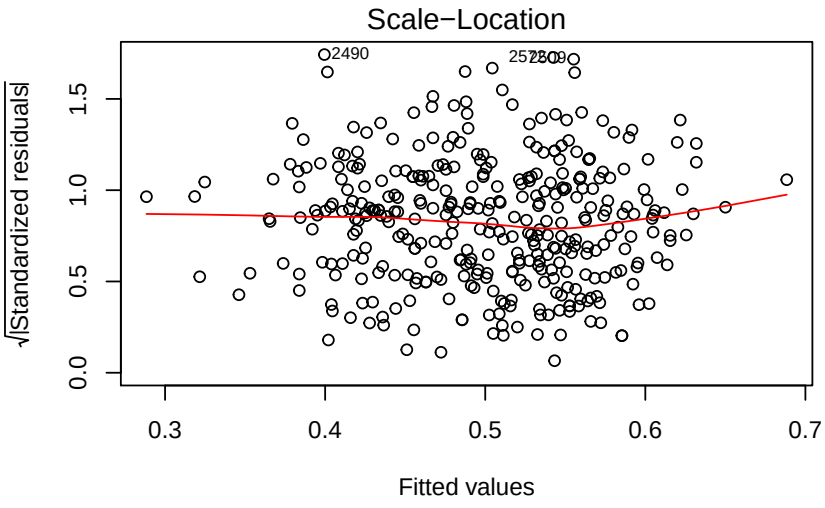
Residual Plot

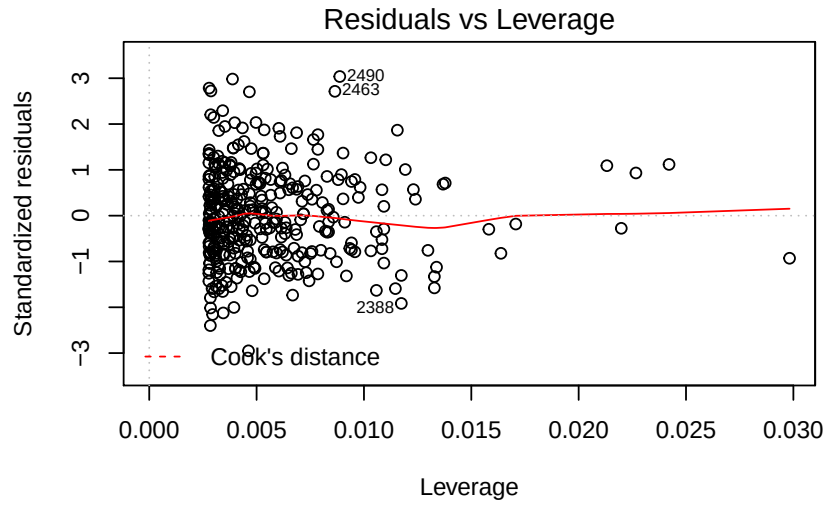


Diagnostic Plot





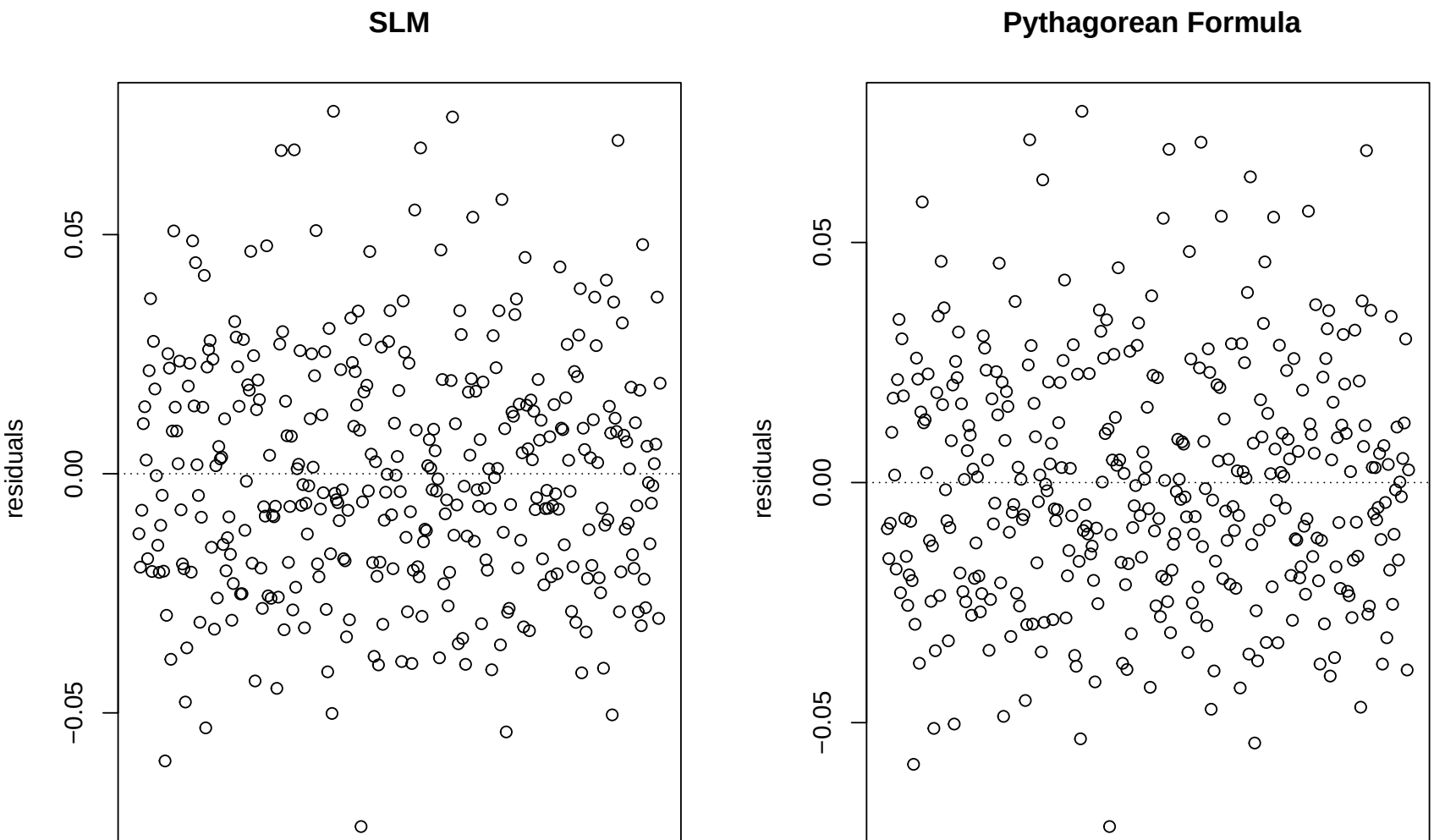




Pythagorean Formula

Bill James empirically derived the non-linear formula to estimate winning percentage, called the Pythagorean expectation

$$Wpct = \frac{R^k}{R^k + RA^k}$$



Career Trajectories

Need packages: “car”, “plyr”, “ggplot2”

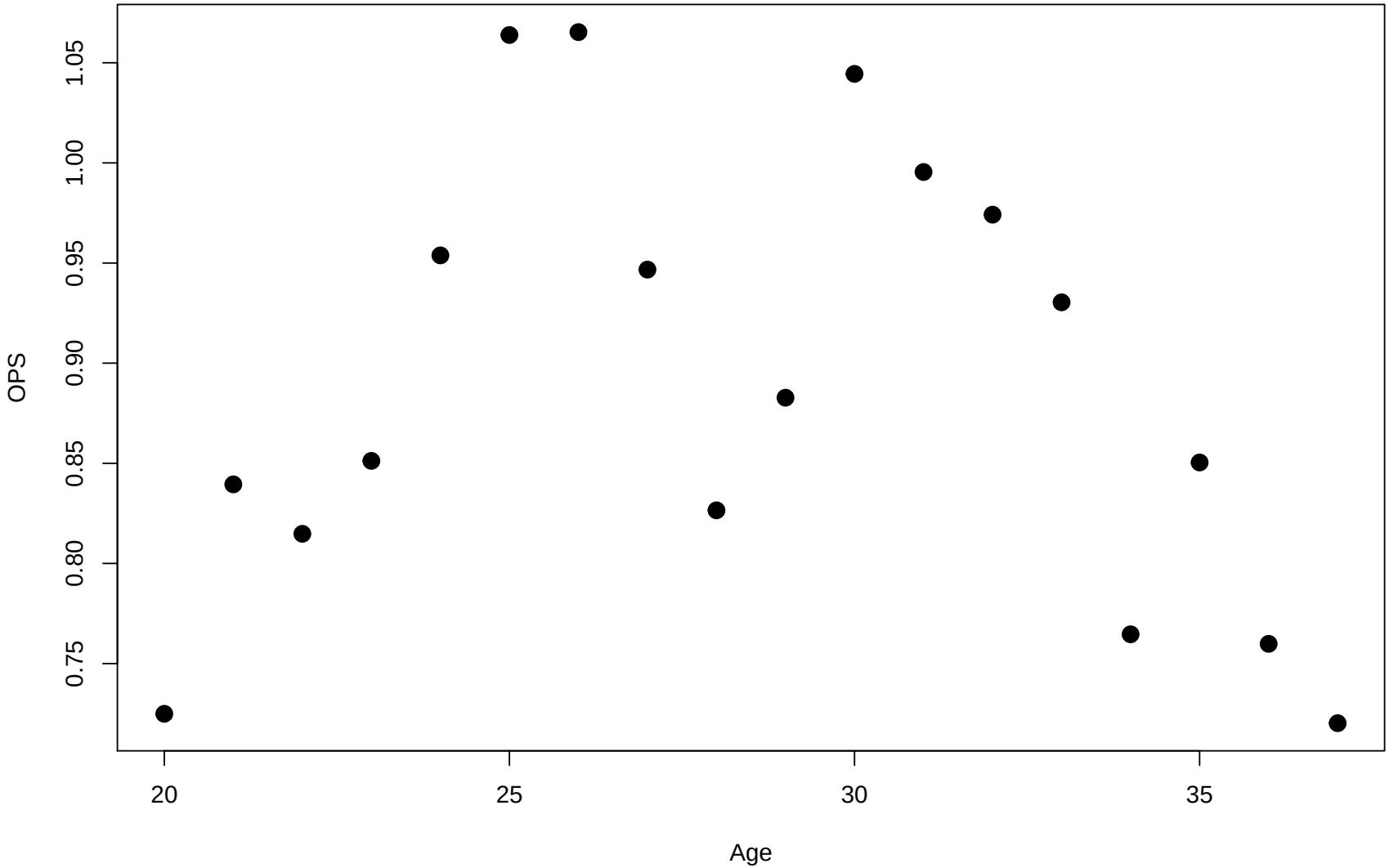
Need datafiles: “Batting.csv”, “Master.csv”, “Fielding.csv”

- Get the general information for players from a dataset.
- Fit and plot trajectories.
- Compare different players by computing similarity scores.
- Find the peak age for players.

Mickey Mantel's Batting Trajectory

It is believed that most players peak in their late 20s. While Mickey Mantle made an immediate impact on the New York Yankees at age 19. But injuries took a toll on Mantel's performance and his hitting declined until his retirement at age 36.

```
## Loading required package: carData
```



Smooth Curve

A convenient choice of smooth curve is a quadratic function of the form (why subtract 30?):

$$A + B(Age - 30) + C(Age - 30)^2$$

Answer the following questions:

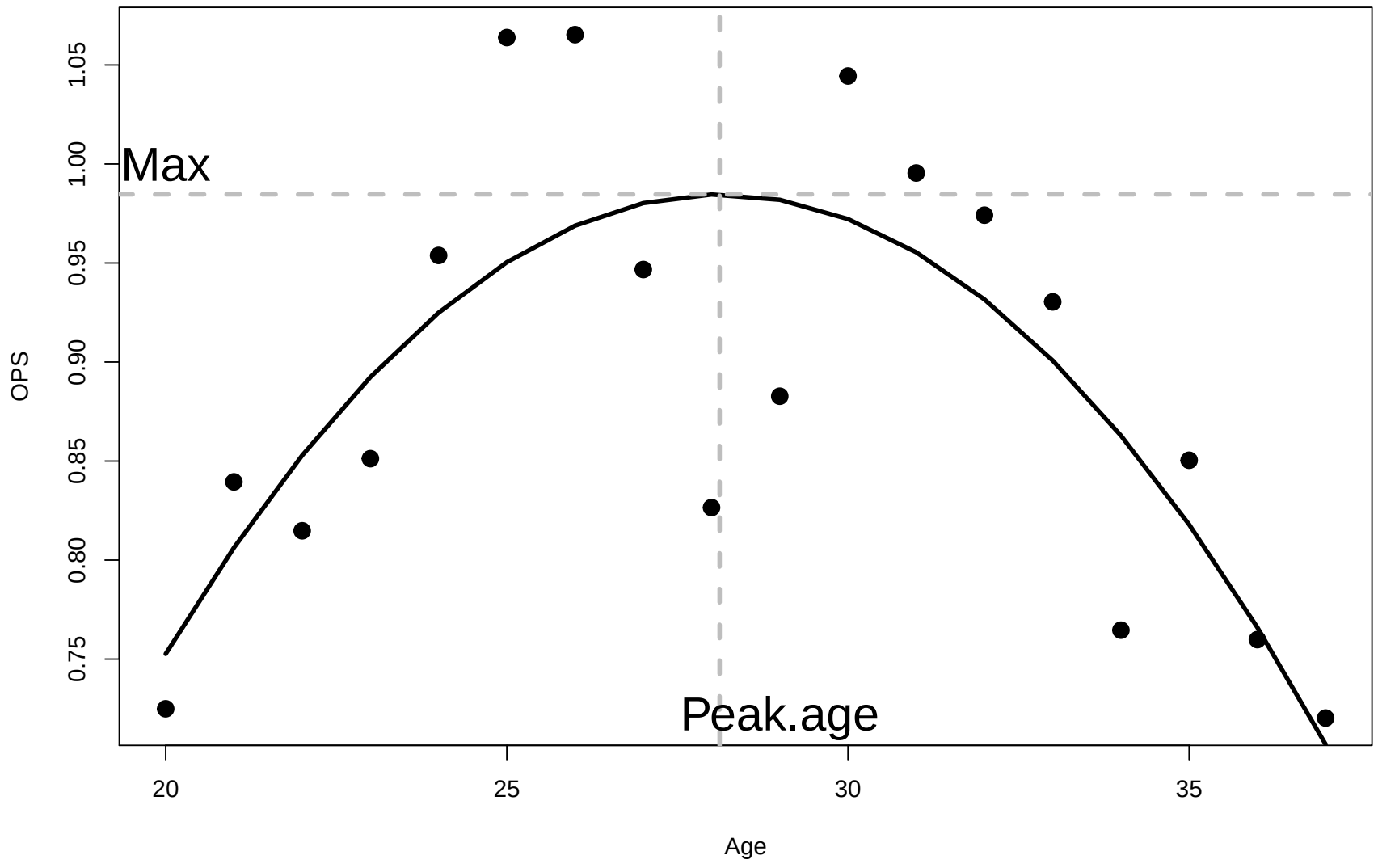
- What's the pratical meaning of the constant A?
- What's the pratical meaning of the coefficient C?
- When does the function reach its largest value?
- What is the maximum value of the curve?

Fit the Smooth Curve

```
##      (Intercept)      I(Age - 30) I((Age - 30)^2)
##      0.972202241    -0.013248087    -0.003520738
```

```
## I(Age - 30) (Intercept)
##   28.118564    0.984665
```

Fit the Smooth Curve



Fit the Smooth Curve

```
summary(F2$fit)
```

```
##
## Call:
## lm(formula = OPS ~ I(Age - 30) + I((Age - 30)^2), data =
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.15809 -0.03694  0.02108  0.03832  0.11344
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.9722022  0.0258845  37.559 2.98e-16 *
## I(Age - 30)      -0.0132481  0.0040638  -3.260 0.005274 *
## I((Age - 30)^2) -0.0035207  0.0007379  -4.772 0.000247 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

Compare Trajectories

To compare the trajectories of different players, it's reasonable to compute and compare the career statistics. Toward this goal, one needs to compute the career games played, at-bats, runs, hit, etc., for each player in the database.

How to compute the career slugging percentage?

$$\text{SP} = \frac{\textit{Total Bases}}{\textit{At Bats}} = \frac{\textit{Hits}+\textit{Doubles}+2\times\textit{Triples}+3\times\textit{HR}}{\textit{AB}}$$

Bill James introduced the concept of **Similarity Scores** to facilitate the comparison of players on the basis of career statistics.

Similarity Score

To compare two hitters, one starts at 1000 points and subtracts points based on the differences in different statistical categories. 1 points is subtracted for each of the following differences:

20 games played, 75 at-bats, 10 runs scored, 15 hits, 5 doubles, 4 triples, 2 home runs, 10 runs batted in, 25 walks, 150 strikeouts, 20 stolen bases, 0.001 in batting average, 0.002 in slugging percentage

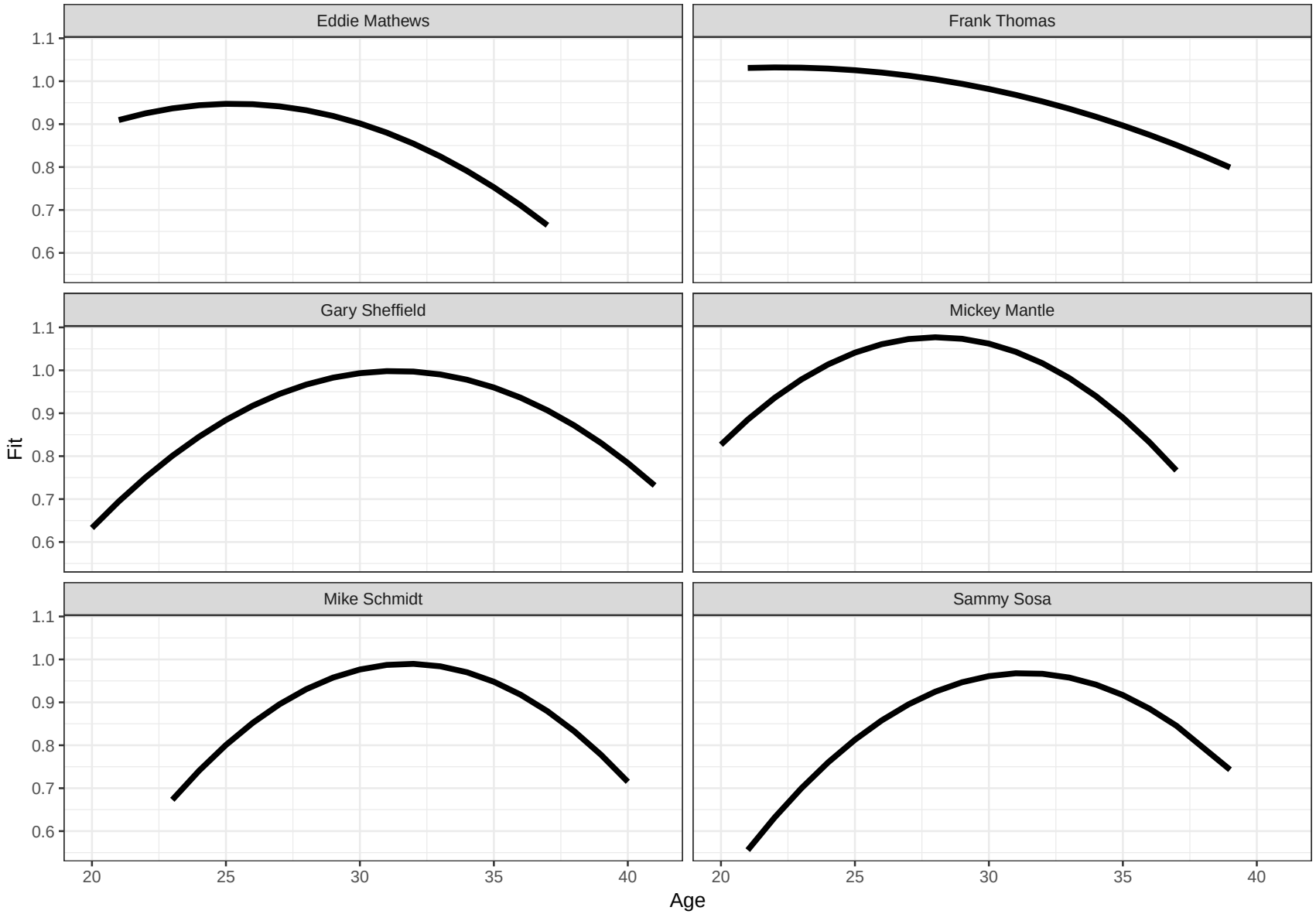
In addition, one adds the difference between the fielding position values of the two players.

Similarity Score

```
##       playerID  C.G C.AB  C.R  C.H C.2B C.3B C.HR C.RBI
## 1293 mantlmi01 2401 8102 1677 2415  344   72  536  1509
## 1319 matheed01 2391 8537 1509 2315  354   72  512  1453
## 1828 schmimi01 2404 8352 1506 2234  408   59  548  1595
## 1867 sheffga01 2576 9217 1636 2689  467   27  509  1676
## 2038 thomafr04 2322 8199 1494 2468  495   12  521  1704
## 1924  sosasa01 2354 8813 1475 2408  379   45  609  1667
##          C.AVG     C.SLG POS Value.POS   SS
## 1293 0.2980745 0.5567761  OF        48 1000
## 1319 0.2711725 0.5094295  3B        84  853
## 1828 0.2674808 0.5272989  3B        84  848
## 1867 0.2917435 0.5139416  OF        48  847
## 2038 0.3010123 0.5549457  DH         0  844
## 1924 0.2732327 0.5337569  OF        48  831
```

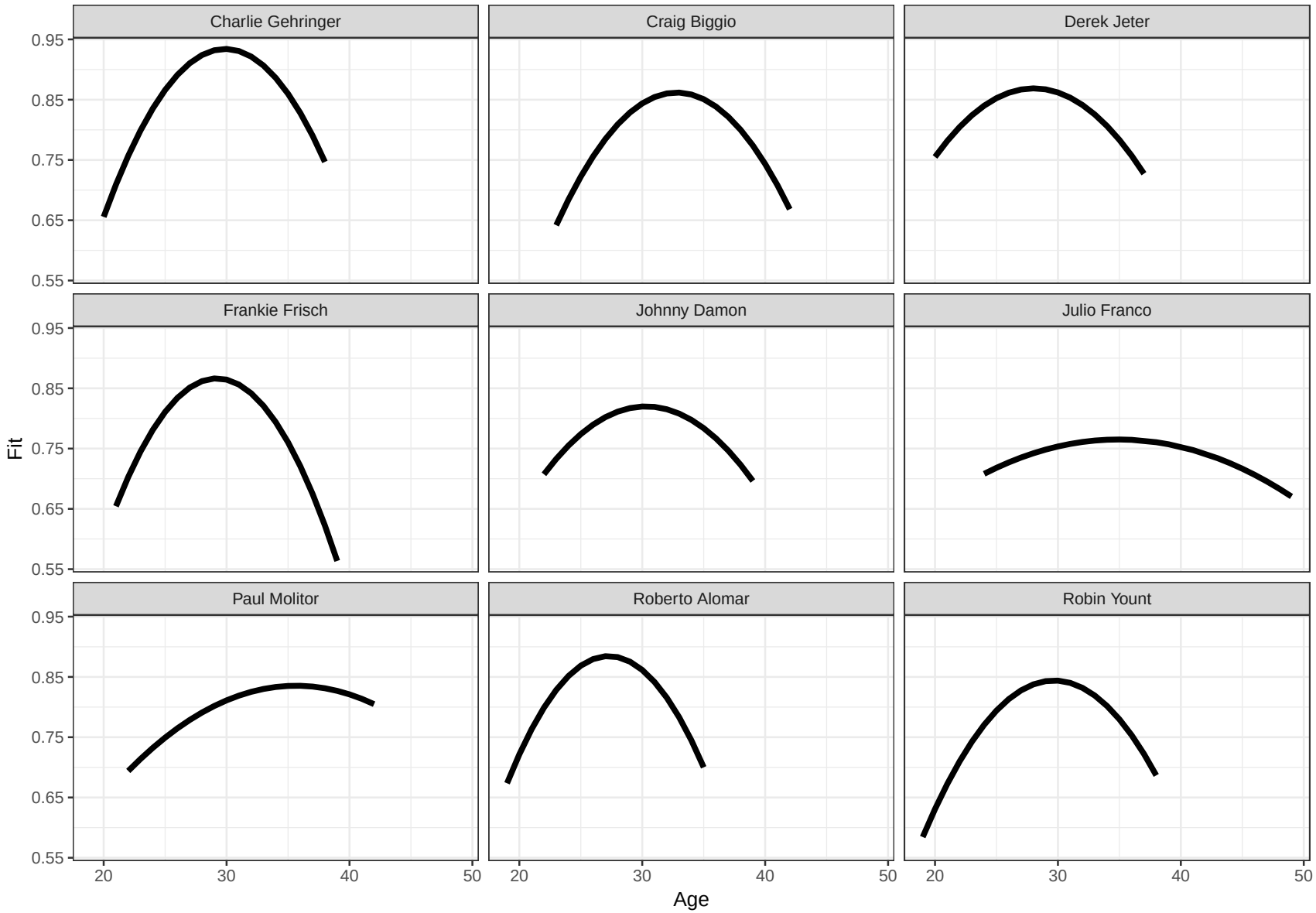
Fit and Plot Trajectories

- Mickey Mantle



Fit and Plot Trajectories

- Derek Jeter



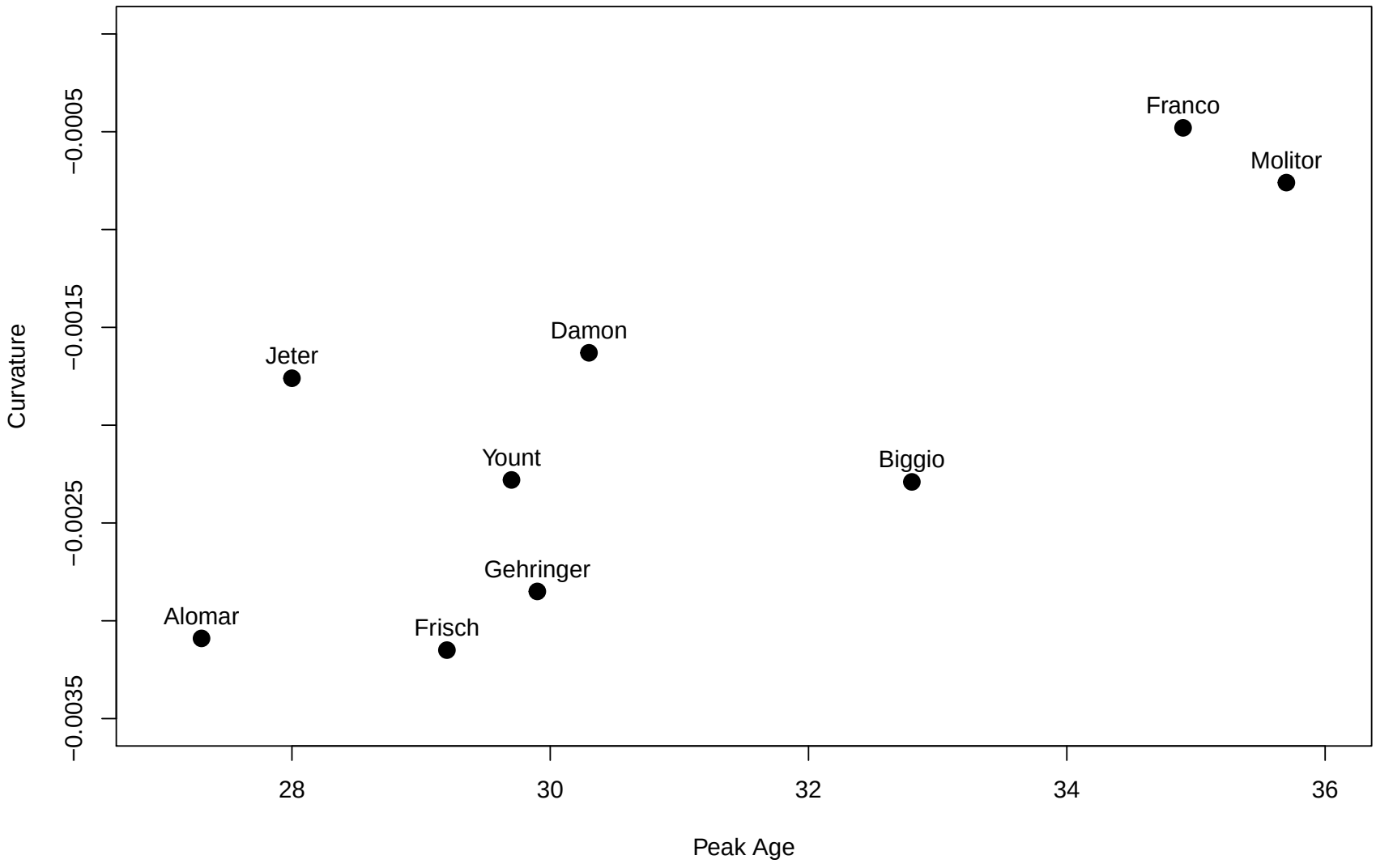
Summary

- There are players such as Eddie Mathews, Frank Thomas, Mickey Mantle, and Roberto Alomar who appeared to peak early in their careers.
- In contrast, other players such as Mike Schmidt, Craig Biggio, and Julio Franco who peaked in their 30s.
- The players also show differneces in the shape of the trajectory. Johnny Damon and Julio Franco had relatively constant trajectories, and Frankie Frisch and Roberto Alomar had trajectories with high curvature.

Summary

```
##    playerID Age.max   Max    Curve
## 1 alomaro01    27.3 0.885 -0.00309
## 2 biggicr01    32.8 0.862 -0.00229
## 3 damonjo01    30.3 0.820 -0.00163
## 4 francju01    34.9 0.765 -0.00048
## 5 friscfr01    29.2 0.866 -0.00315
## 6 gehrich01    29.9 0.934 -0.00285
## 7 jeterde01    28.0 0.869 -0.00176
## 8 molitpa01    35.7 0.835 -0.00076
## 9 yountro01    29.7 0.844 -0.00228
```

Summary


Curvature
−0.0005
−0.0015
−0.0025
−0.0035
28
30
32
34
36
Peak Age
Franco
Molitor
Damon
Jeter
Yount
Biggio
Gehringer
Alomar
Frisch

Run Expectancy

Need package: “plyr”

Need datafiles: “all2011.csv”, “fields.csv”, “roster2011.csv”

- Find the Runs Expectancy Matrixs
- Case Study (See example code)

The Runs Expectancy Matrix

Each base can be occupied by a runner or empty.

The number of outs can be 0, 1, or 2.

For each combination, one is interested in computing the average number of runs scored in the remainder of the inning.

Arrange the average runs as a table classified by runners and outs, this display is called the **Runs Expectancy Matrix**.

```
##      0 outs 1 out 2 outs 0 outs 1 out 2 outs
## 000    0.47  0.25    0.10    0.51  0.27    0.10
## 001    1.45  0.94    0.32    1.40  0.94    0.36
## 010    1.06  0.65    0.31    1.14  0.68    0.32
## 011    1.93  1.34    0.54    1.96  1.36    0.63
## 100    0.84  0.50    0.22    0.90  0.54    0.23
## 101    1.75  1.15    0.49    1.84  1.18    0.52
## 110    1.41  0.87    0.42    1.51  0.94    0.45
## 111    2.17  1.47    0.76    2.33  1.51    0.78
```

It is remarkable that these run expectancy values have not changed over the recent history of baseball. This indicates that there have been little changes in the average runscoring tendencies of this team between 2002 and 2011.