Chapter 4

Nick Lauerman

Section 4.2

```
teams <- read.csv("E:/Baseball/data/lahman/teams.csv")
tail(teams)
```

```
##      yearID lgID teamID franchID divID Rank   G Ghome  W  L DivWin WCWin
## 2680   2011   NL    FLO      FLA     E    5 162    NA 72 90      N     N
## 2681   2011   NL    ARI      ARI     W    1 162    NA 94 68      Y     N
## 2682   2011   NL    SFN      SFG     W    2 162    NA 86 76      N     N
## 2683   2011   NL    LAN      LAD     W    3 161    NA 82 79      N     N
## 2684   2011   NL    COL      COL     W    4 162    NA 73 89      N     N
## 2685   2011   NL    SDN      SDP     W    5 162    NA 71 91      N     N
##      LgWin WSWin   R   AB    H X2B X3B  HR  BB   SO  SB CS HBP SF  RA  ER
## 2680     N     N 625 5508 1358 274  30 149 542 1244  95 41  51 42 702 640
## 2681     N     N 731 5421 1357 293  37 172 531 1249 133 55  61 33 662 609
## 2682     N     N 570 5486 1327 282  24 121 448 1122  85 51  52 43 578 522
## 2683     N     N 644 5436 1395 237  28 117 498 1087 126 40  45 43 612 563
## 2684     N     N 735 5544 1429 274  40 163 555 1201 118 42  57 44 774 713
## 2685     N     N 593 5417 1284 247  42  91 501 1320 170 44  48 47 611 551
##       ERA CG SHO SV IPouts   HA HRA BBA  SOA   E  DP    FP
## 2680 3.95  7  11 40   4379 1403 149 500 1218  93 126 0.985
## 2681 3.80  5  12 58   4330 1414 159 442 1058  90 130 0.985
## 2682 3.21  3  12 52   4404 1260  96 559 1316 104 127 0.983
## 2683 3.56  7  17 40   4296 1287 132 507 1265  85 121 0.986
## 2684 4.44  5   7 41   4343 1471 176 522 1118  98 156 0.984
## 2685 3.43  0  10 44   4348 1324 125 521 1139  94 138 0.985
##                      name             park attendance BPF PPF teamIDBR
## 2680      Florida Marlins Sun Life Stadium    1477462  99 100      FLA
## 2681 Arizona Diamondbacks      Chase Field    2105432 107 106      ARI
## 2682 San Francisco Giants        AT&T Park    3387303  89  89      SFG
## 2683  Los Angeles Dodgers   Dodger Stadium    2935139  98  98      LAD
## 2684     Colorado Rockies      Coors Field    2909777 116 116      COL
## 2685     San Diego Padres       Petco Park    2143018  92  92      SDP
##      teamIDlahman45 teamIDretro
## 2680            FLO         FLO
## 2681            ARI         ARI
## 2682            SFN         SFN
## 2683            LAN         LAN
## 2684            COL         COL
## 2685            SDN         SDN
```

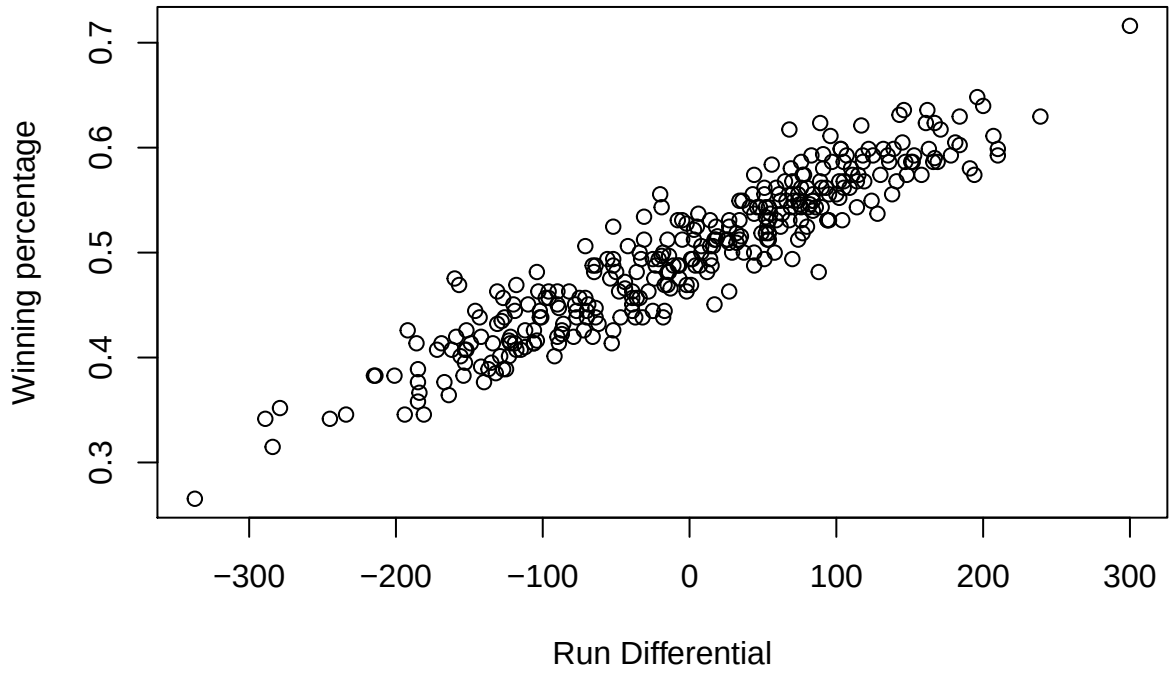
```
myteams <- subset(teams,
                  yearID > 2000,
                  select = c("teamID",
                             "yearID",
                             "lgID",
```

```
                              "G",
                              "W",
                              "L",
                              "R",
                              "RA"))
tail(myteams)
```

```
##      teamID yearID lgID   G  W  L   R  RA
## 2680    FLO   2011   NL 162 72 90 625 702
## 2681    ARI   2011   NL 162 94 68 731 662
## 2682    SFN   2011   NL 162 86 76 570 578
## 2683    LAN   2011   NL 161 82 79 644 612
## 2684    COL   2011   NL 162 73 89 735 774
## 2685    SDN   2011   NL 162 71 91 593 611
```

```
myteams$RD <- with(myteams, R - RA)
myteams$Wpct <- with(myteams, W / (W + L))

plot(myteams$RD, myteams$Wpct,
     xlab = "Run Differential",
     ylab = "Winning percentage")
```

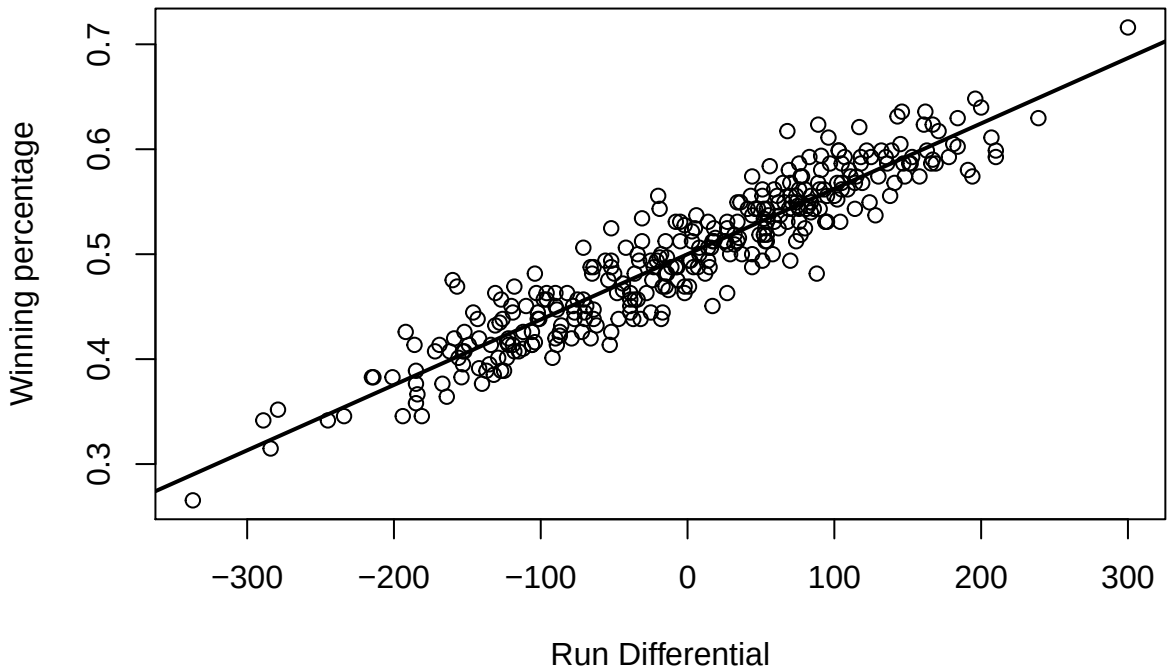


Section 4.3

```
linfit <- lm(Wpct ~ RD, data = myteams)
linfit
```

```
##
## Call:
## lm(formula = Wpct ~ RD, data = myteams)
##
## Coefficients:
## (Intercept)           RD
##    0.499992     0.000623
```

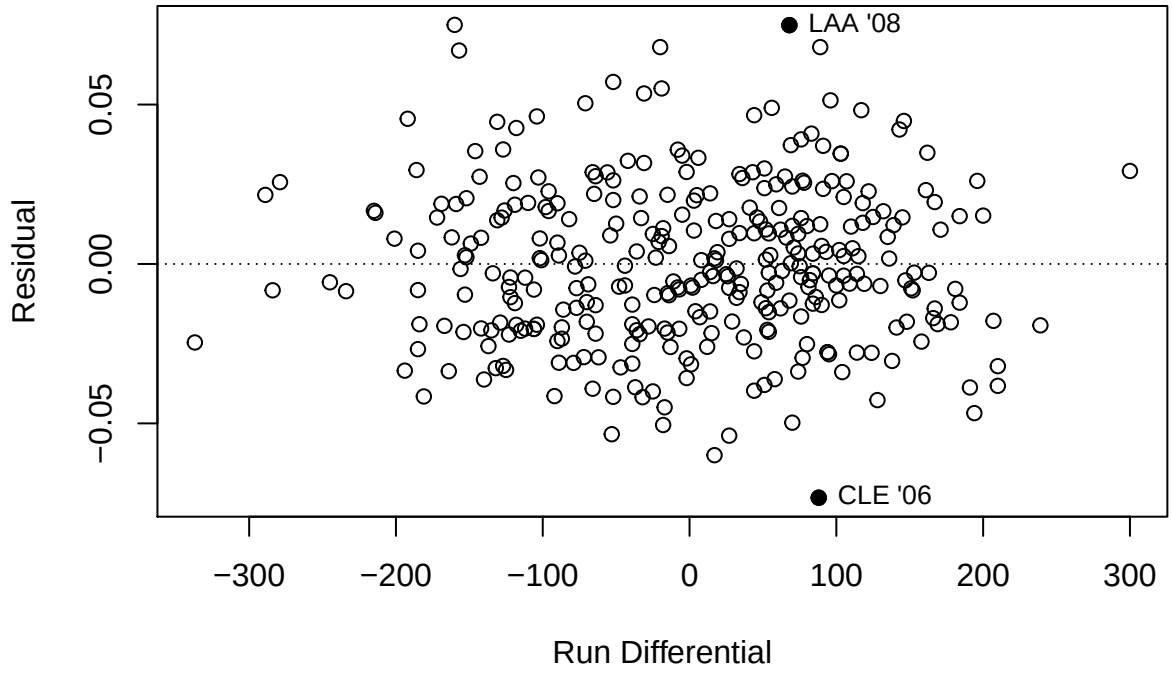
```
plot(myteams$RD, myteams$Wpct,
     xlab = "Run Differential",
     ylab = "Winning percentage")
# add a linear line to the scatter plot
abline(a = coef(linfit)[1],
       b = coef(linfit)[2],
       lwd = 2)
```




```
#add linear model predictions and residuals to data fram
myteams$linWpct <- predict(linfit)
myteams$linResiduals <- residuals(linfit)
```

```r
#plot the residuals
plot(myteams$RD, myteams$linResiduals,
     xlab = "Run Differential",
     ylab = "Residual")
abline(h=0, lty = 3)

#indicate some outlayers
points(c(68,88),c(0.0749, -0.0733), pch = 19)
text(68, 0.0749, "LAA '08", pos = 4, cex = 0.8)
text(88, -0.0733, "CLE '06", pos = 4, cex = 0.8)
```




```r
#compute the mean and Root Mean Square Error
mean(myteams$linResiduals)
```

```
## [1] -2.952603e-19
```

```r
linRMSE <- sqrt(mean(myteams$linResiduals ^ 2))
linRMSE
```

```
## [1] 0.02507176
```

```r
#error intervials
nrow(subset(myteams,
```

```
            abs(linResiduals) < linRMSE)) /
     nrow(myteams)
```

```
## [1] 0.6757576
```

```
nrow(subset(myteams,
            abs(linResiduals) < 2 * linRMSE)) /
     nrow(myteams)
```

```
## [1] 0.9545455
```

Section 4.4

```
myteams$pytWpct <- with(myteams,
                        R ^ 2 / (R ^ 2  +  RA ^ 2))
myteams$pytResiduals <- myteams$Wpct - myteams$pytWpct

#calculate RMSE
sqrt(mean(myteams$pytResiduals ^ 2))
```

```
## [1] 0.02545247
```

Section 4.5

```
myteams$logWratio <- log(myteams$W / myteams$L)
myteams$logRratio <- log(myteams$R / myteams$RA)
pytFit <- lm(logWratio ~ 0 + logRratio, data = myteams)
pytFit
```

```
##
## Call:
## lm(formula = logWratio ~ 0 + logRratio, data = myteams)
##
## Coefficients:
## logRratio
##     1.903
```

Section 4.6

Need to review and take care of how to get data from Retorsheets.ORG outside of the function already developed in chapter 3

```
gl2011 <- read.table("e:/baseball/data/Book/gl2011.txt",
                     sep = ",")
glheaders <- read.csv("e:/baseball/data/Book/game_log_header.csv")
names(gl2011) <- names(glheaders)

BOS2011 <- subset(gl2011,
                  HomeTeam == "BOS" | VisitingTeam == "BOS",
                  select = c("VisitingTeam",
                             "HomeTeam",
                             "VisitorRunsScored",
                             "HomeRunsScore"))

head(BOS2011)
```

```
##    VisitingTeam HomeTeam VisitorRunsScored HomeRunsScore
## 16          BOS      TEX                 5            9
## 31          BOS      TEX                 5           12
## 45          BOS      TEX                 1            5
## 61          BOS      CLE                 1            3
## 76          BOS      CLE                 4            8
## 88          BOS      CLE                 0            1
```

```
BOS2011$ScoreDiff <- with(BOS2011,
                          ifelse(HomeTeam == "BOS",
                                 HomeRunsScore - VisitorRunsScored,
                                 VisitorRunsScored - HomeRunsScore))

BOS2011$W <- BOS2011$ScoreDiff >0

aggregate(abs(BOS2011$ScoreDiff),
        list(W = BOS2011$W),
        summary)
```

```
##       W x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.
## 1 FALSE  1.000     1.000    3.000  3.458     4.000 11.000
## 2  TRUE  1.000     2.000    4.000  4.300     6.000 14.000
```

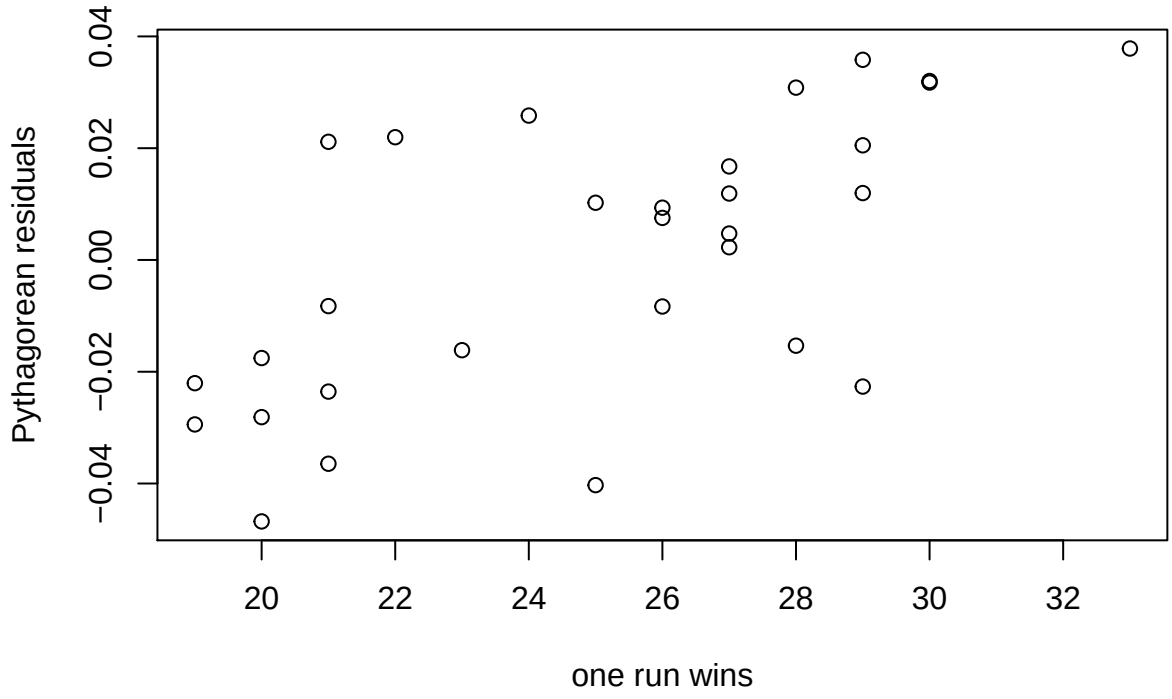
```
results <- gl2011[,c("VisitingTeam",
                     "HomeTeam",
                     "VisitorRunsScored",
                     "HomeRunsScore")]

results$winner <- ifelse(results$HomeRunsScore > results$VisitorRunsScored,
                         as.character(results$HomeTeam),
                         as.character(results$VisitingTeam))

results$diff <- abs(results$VisitorRunsScored - results$HomeRunsScore)

onerungames <- subset(results, diff == 1)
onerunwins <- as.data.frame(table(onerungames$winner))
names(onerunwins) <- c("teamID", "onerunW")
```

```
teams2011 <- subset(myteams, yearID == 2011)
teams2011[teams2011$teamID == "LAA", "teamID"] <- "ANA"
teams2011 <- merge(teams2011, onerunwins)
plot(teams2011$onerunW, teams2011$pytResiduals,
     xlab = "one run wins",
     ylab = "Pythagorean residuals")
```




```
pit <- read.csv("e:/Baseball/data/lahman/Pitching.csv")
top_closers <- subset(pit,
                      GF > 50 & ERA < 2.5,
                      select = c("playerID",
                                 "yearID",
                                 "teamID"))

teams_top_closers <- merge(myteams, top_closers)
summary(teams_top_closers$pytResiduals)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.048690 -0.011660  0.003359  0.005189  0.022990  0.071400
```

Section 4.7