Chapter 4 exercises

Nick Lauerman

Question 1

part a

```
#teams <- read.csv("f:/Baseball/data/lahman/teams.csv") # HOME
teams <- read.csv("e:/Baseball/data/lahman/teams.csv") # WORK

teams60 <- subset(teams,
                  yearID >= 1961 & yearID <= 1970,
                  select = c("teamID",
                             "yearID",
                             "lgID",
                             "G",
                             "W",
                             "L",
                             "R",
                             "RA"))
teams60$RD <- with(teams60, R - RA)
teams60$Wpct <- with(teams60, W / (W + L))

teams70 <- subset(teams,
                  yearID >= 1971 & yearID <= 1980,
                  select = c("teamID",
                             "yearID",
                             "lgID",
                             "G",
                             "W",
                             "L",
                             "R",
                             "RA"))
teams70$RD <- with(teams70, R - RA)
teams70$Wpct <- with(teams70, W / (W + L))

teams80 <- subset(teams,
                  yearID >= 1981 & yearID <= 1990,
                  select = c("teamID",
                             "yearID",
                             "lgID",
                             "G",
                             "W",
                             "L",
                             "R",
                             "RA"))
teams80$RD <- with(teams80, R - RA)
teams80$Wpct <- with(teams80, W / (W + L))

teams90 <- subset(teams,
```

```
                  yearID >= 1991 & yearID <= 2000,
                  select = c("teamID",
                             "yearID",
                             "lgID",
                             "G",
                             "W",
                             "L",
                             "R",
                             "RA"))
teams90$RD <- with(teams90, R - RA)
teams90$Wpct <- with(teams90, W / (W + L))

teams00 <- subset(teams,
                  yearID >= 2001 & yearID <= 2010,
                  select = c("teamID",
                             "yearID",
                             "lgID",
                             "G",
                             "W",
                             "L",
                             "R",
                             "RA"))
teams00$RD <- with(teams00, R - RA)
teams00$Wpct <- with(teams00, W / (W + L))


lin60 <- lm(Wpct ~ RD, data = teams60)
lin60
```

```
##
## Call:
## lm(formula = Wpct ~ RD, data = teams60)
##
## Coefficients:
## (Intercept)           RD
##    0.499933     0.000704
```

```
lin70 <- lm(Wpct ~ RD, data = teams70)
lin70
```

```
##
## Call:
## lm(formula = Wpct ~ RD, data = teams70)
##
## Coefficients:
## (Intercept)           RD
##   0.4999884    0.0006375
```

```
lin80 <- lm(Wpct ~ RD, data = teams80)
lin80
```

```
##
```

```
## Call:
## lm(formula = Wpct ~ RD, data = teams80)
##
## Coefficients:
## (Intercept)           RD
##   0.4999448    0.0007014
```

```
lin90 <- lm(Wpct ~ RD, data = teams90)
lin90
```

```
##
## Call:
## lm(formula = Wpct ~ RD, data = teams90)
##
## Coefficients:
## (Intercept)           RD
##   0.4999994    0.0006276
```

```
lin00 <- lm(Wpct ~ RD, data = teams00)
lin00
```

```
##
## Call:
## lm(formula = Wpct ~ RD, data = teams00)
##
## Coefficients:
## (Intercept)           RD
##   0.4999909    0.0006216
```

b

```
wins60 <- as.numeric(coef(lin60)[1] + 10 * coef(lin60)[2])
wins60
```

```
## [1] 0.5069728
```

```
wins70 <- as.numeric(coef(lin70)[1] + 10 * coef(lin70)[2])
wins70
```

```
## [1] 0.5063638
```

```
wins80 <- as.numeric(coef(lin80)[1] + 10 * coef(lin80)[2])
wins80
```

```
## [1] 0.5069586
```

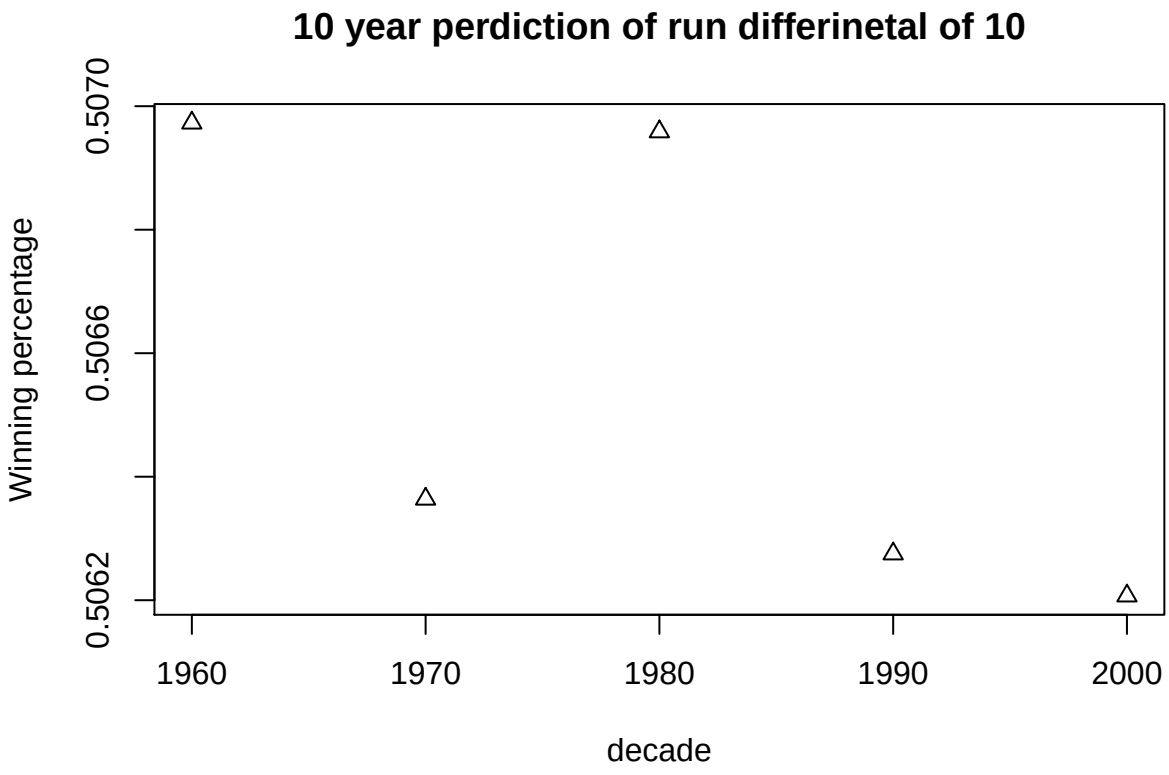
```
wins90 <- as.numeric(coef(lin90)[1] + 10 * coef(lin90)[2])
wins90
```

```
## [1] 0.5062751
```

```r
wins00 <- as.numeric(coef(lin00)[1] + 10 * coef(lin00)[2])
wins00
```

```
## [1] 0.5062071
```

```r
plot(c(seq(from = 1960,
           to = 2000,
           by = 10)),
     c(wins60,
       wins70,
       wins80,
       wins90,
       wins00),
     pch=2,
     xlab = "decade",
     ylab = "Winning percentage",
     main = "10 year perdiction of run differinetal of 10")
```



Question 2

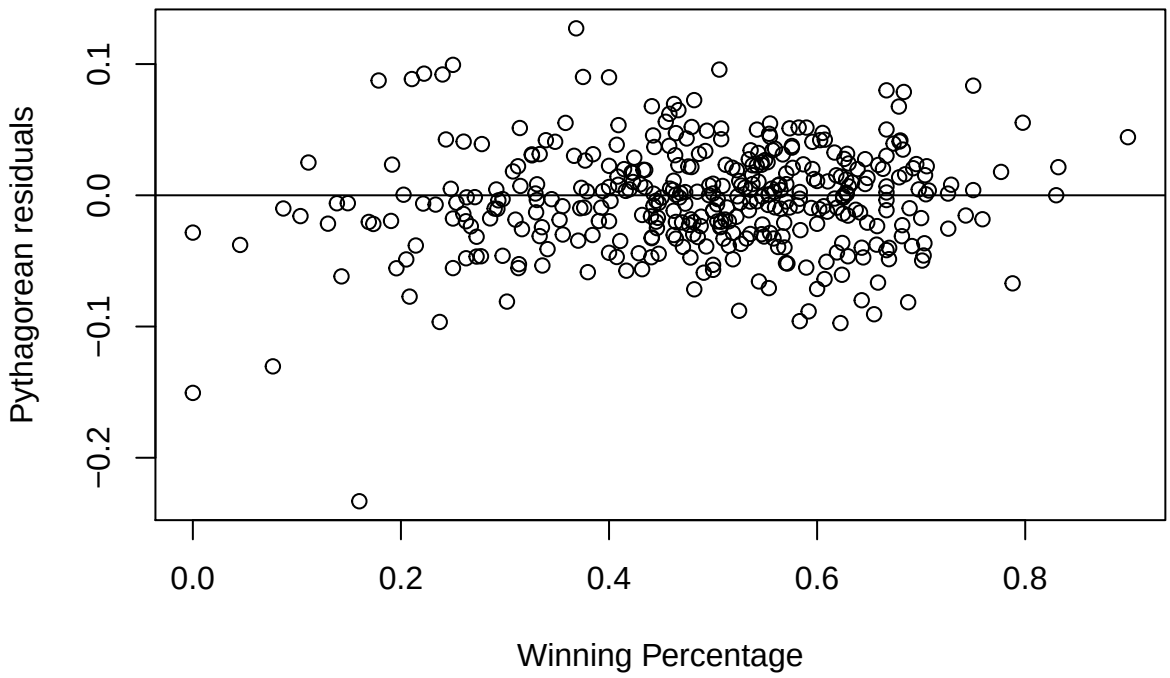
a

```
teams19 <- subset(teams,
                  yearID <= 1900,
                  select = c("teamID",
                             "yearID",
                             "lgID",
                             "G",
                             "W",
                             "L",
                             "R",
                             "RA"))
teams19$RD <- with(teams19, R - RA)
teams19$Wpct <- with(teams19, W / (W + L))

teams19$ptyWpct <- with(teams19, R^2 / (R^2 + RA^2))
teams19$ptyResiduals <- with(teams19, Wpct - ptyWpct)
```

b

```
with(teams19, plot(Wpct, ptyResiduals,
                   ylab = "Pythagorean residuals",
                   xlab = "Winning Percentage"))
abline(h=0)
```



Question 3

a

```
#source("e:/Baseball/scripts/baseball.R")
#parse.retrosheet.php(1990)
#for(season in 1991:2000){
#     parse.retrosheet.php(season)
#}
#
# above not needed. Download GL and unzipped and moved to data/retro

manager <- read.table("e:/Baseball/data/retro/GL1990.txt", sep=",")
for(season in 1991:2000){
     file <- paste0("e:/Baseball/data/retro/GL",season,".txt")
     temp <- read.table(file, sep =",")
     manager <- rbind(manager, temp)
     rm(temp)
     rm(file)
}

glheaders <- read.csv("e:/baseball/data/Book/game_log_header.csv")
names(manager) <- names(glheaders)
```

```
manager <- subset(manager, select = c("VisitingTeam",
                                      "HomeTeam",
                                      "VisitorRunsScored",
                                      "HomeRunsScore",
                                      "VisitorManagerID",
                                      "VisitorManagerName",
                                      "HomeManagerID",
                                      "HomeManagerName"))
make.manager <- function(manager){
     # create a list of teams
     teams <- as.character(levels(manager$HomeTeam))
     for (i in 1:nrow(manager)){
          manager$game_number[i] <- i
     }

     #make a list of home games
     home.teams <- subset(manager, HomeTeam == teams[1])

     for(i in 2:length(teams)){
          temp <- subset(manager, HomeTeam == teams[i])
          home.teams <- rbind(home.teams,temp)
     }
     # Assign verables
     home.teams$Home <- TRUE
     home.teams$manager  <- home.teams$HomeManagerName
     home.teams$ManagerID <- home.teams$HomeManagerID
     home.teams$R  <- home.teams$HomeRunsScore
     home.teams$RA  <- home.teams$VisitorRunsScored
     home.teams$team  <- home.teams$HomeTeam
     for(i in 1:nrow(home.teams)){
          if(home.teams$R[i] > home.teams$RA[i]){
               home.teams$W[i] <- 1
               home.teams$L[i] <- 0
          } else {
               home.teams$W[i] <- 0
               home.teams$L[i] <- 1
          }
     }


     #make a list of away games
     visit.teams <- subset(manager, VisitingTeam == teams[1])
     for(i in 2:length(teams)){
          temp <- subset(manager, VisitingTeam == teams[i])
          visit.teams <- rbind(visit.teams,temp)
     }
     # Assign verables
     visit.teams$Home <- FALSE
     visit.teams$manager  <- visit.teams$VisitorManagerName
     visit.teams$ManagerID <- visit.teams$VisitorManagerID
     visit.teams$RA  <- visit.teams$HomeRunsScore
     visit.teams$R  <- visit.teams$VisitorRunsScored
     visit.teams$team  <- visit.teams$VisitingTeam
```

```
    for(i in 1:nrow(visit.teams)){
        if(visit.teams$R[i] > visit.teams$RA[i]){
            visit.teams$W[i] <- 1
            visit.teams$L[i] <- 0
        } else {
            visit.teams$W[i] <- 0
            visit.teams$L[i] <- 1
        }
    }
    #combine to a single data frame
    manager <- rbind(home.teams,visit.teams)
    manager <- subset(manager,
                      select = c("team",
                                 "game_number",
                                 "Home",
                                 "manager",
                                 "ManagerID",
                                 "R",
                                 "RA",
                                 "W",
                                 "L"))
}

manager.small <- make.manager(manager)
library(reshape2)
manager.melt <- melt(manager.small,
                     id.vars = "manager",
                     measure.vars = c("R",
                                      "RA",
                                      "W",
                                      "L"))

manager.cast <- dcast(manager.melt, manager ~ variable, sum)

# Compute winning percentage (Wpct), Pyth Winning Percentage (pytWpct),
#    and Pyth Residuals (pythResiduals)
manager.cast$games <- with(manager.cast, W + L)
manager.cast$Wpct  <- with(manager.cast, W/(W + L))

manager.cast$pytWpct <- with(manager.cast,
                             R ^ 2 / (R ^ 2  +  RA ^ 2))
manager.cast$pytResiduals <- manager.cast$Wpct - manager.cast$pytWpct

# order the data frame on the residuals
manager.cast <- manager.cast[order(manager.cast$pytResiduals, decreasing = TRUE), ]
```

b

Managers that over performed

```
head(manager.cast, n=10)
```

```
##             manager   R  RA  W  L games      Wpct   pytWpct pytResiduals
```

```
## 56      Bobby Knoop   5  12  1  1     2 0.5000000 0.1479290   0.35207101
## 90     Cookie Rojas   7   4  1  0     1 1.0000000 0.7538462   0.24615385
## 91   Bruce Benedict  10   5  2  0     2 1.0000000 0.8000000   0.20000000
## 72        Mel Queen  25  19  4  1     5 0.8000000 0.6338742   0.16612576
## 33     Bob Schaefer   3   1  1  0     1 1.0000000 0.9000000   0.10000000
## 23 Red Schoendienst  85  88 13 11    24 0.5416667 0.4826642   0.05900249
## 87       Joe Nossek  25  35  3  5     8 0.3750000 0.3378378   0.03716216
## 25       Russ Nixon 267 365 25 40    65 0.3846154 0.3485776   0.03603778
## 75    Glenn Hoffman 351 351 47 41    88 0.5340909 0.5000000   0.03409091
## 62   Mike Jorgensen 353 424 42 54    96 0.4375000 0.4093796   0.02812043
```

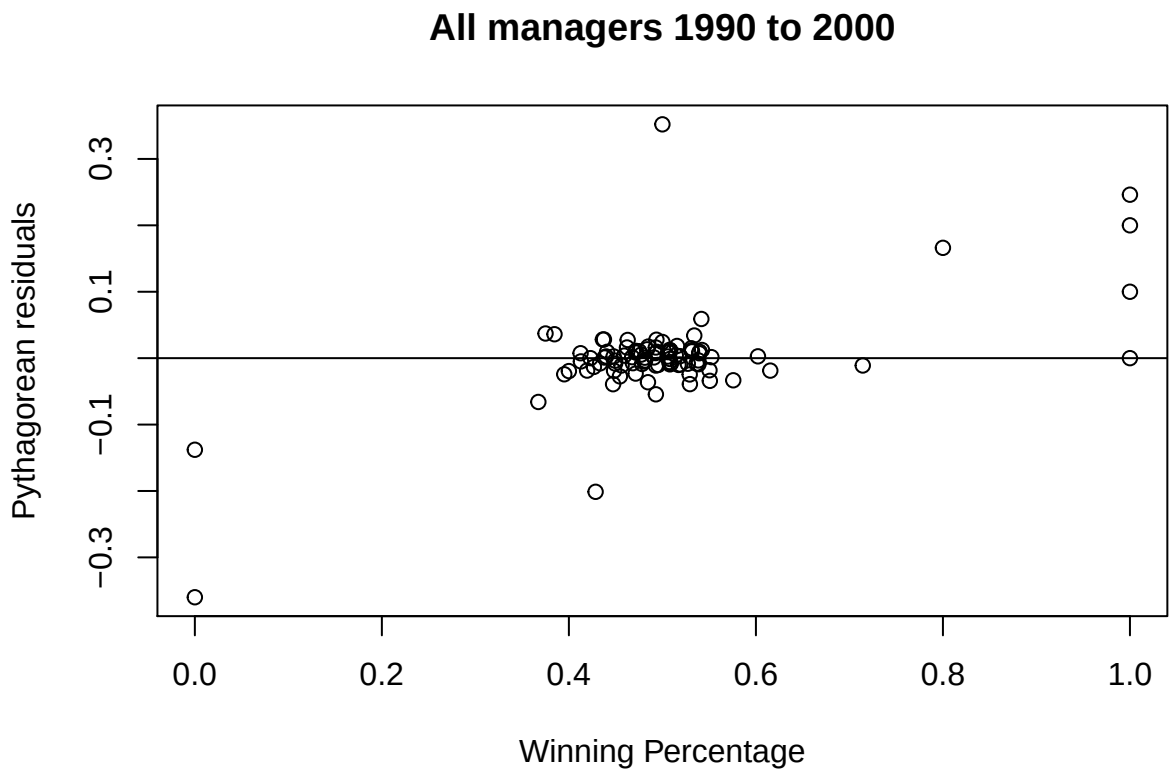
Manager that under performed

```
tail(manager.cast, n=10)
```

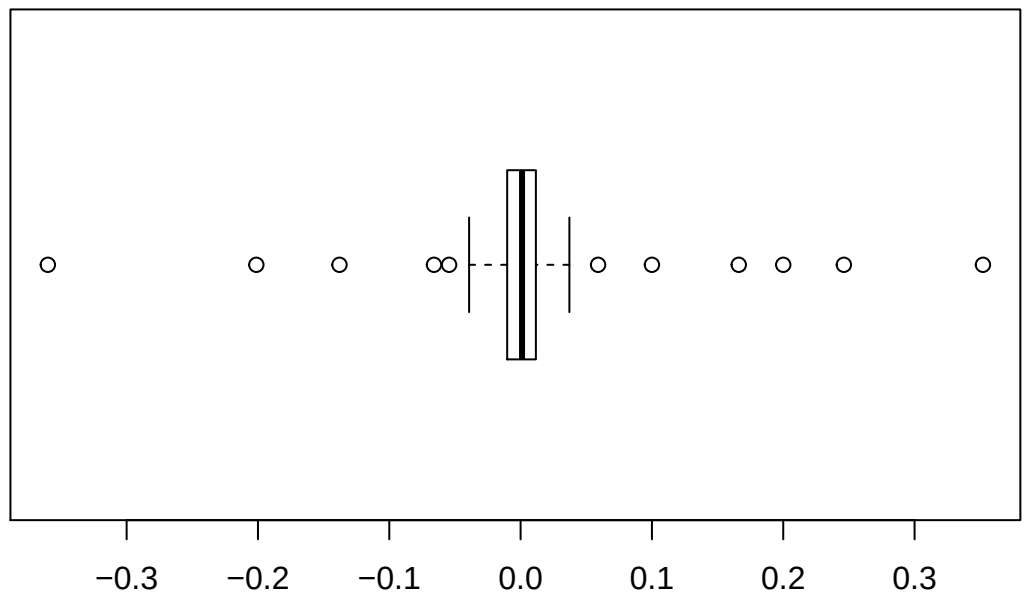
```
##           manager    R   RA   W   L games      Wpct   pytWpct pytResiduals
## 34    Gene Tenace  171  137  19  14    33 0.5757576 0.6090606  -0.03330304
## 71  Larry Dierker 3281 2764 342 279   621 0.5507246 0.5849042  -0.03417955
## 79     Ray Miller 1668 1600 157 167   324 0.4845679 0.5207988  -0.03623093
## 67     Joe Maddon  264  230  27  24    51 0.5294118 0.5685014  -0.03908965
## 50   Dallas Green 2297 2360 229 283   512 0.4472656 0.4864745  -0.03920883
## 63     Phil Regan  704  640  71  73   144 0.4930556 0.5475113  -0.05445576
## 5      Bucky Dent  188  215  18  31    49 0.3673469 0.4333019  -0.06595493
## 93 Fredi Gonzalez    8   20   0   2     2 0.0000000 0.1379310  -0.13793103
## 39   Mike Cubbage   30   23   3   4     7 0.4285714 0.6298111  -0.20123963
## 89  Joe Altobelli    6    8   0   1     1 0.0000000 0.3600000  -0.36000000
```

and some plots

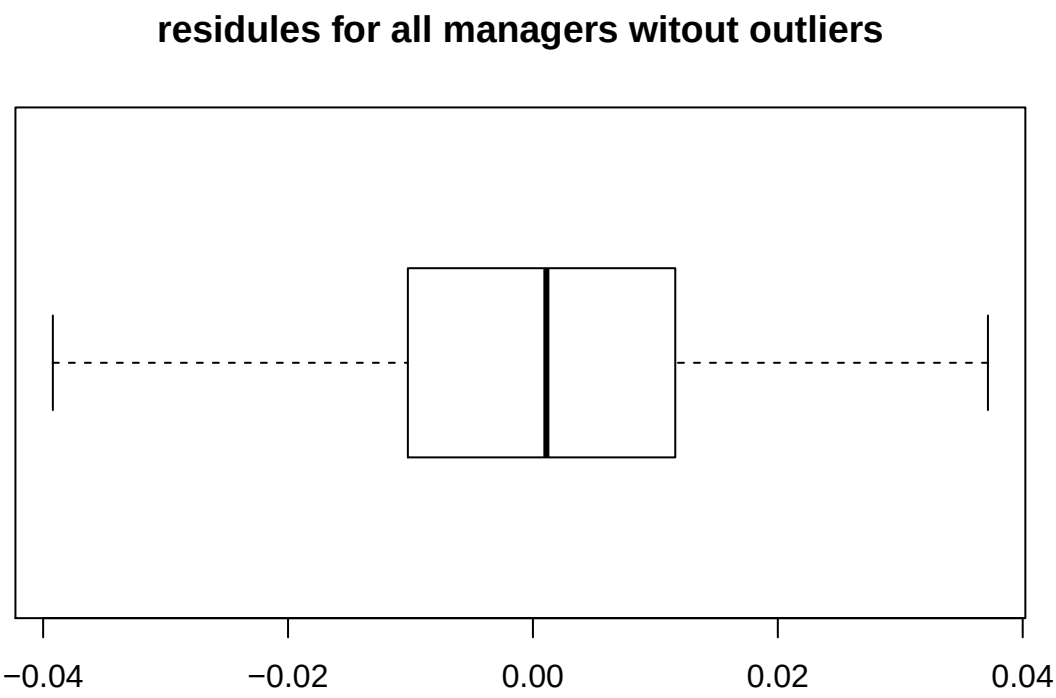
```
with(manager.cast, plot(Wpct, pytResiduals,
                        ylab = "Pythagorean residuals",
                        xlab = "Winning Percentage",
                        main = "All managers 1990 to 2000"))
abline(h=0)
```




```
boxplot(manager.cast$pytResiduals,
        horizontal = TRUE,
        main = "residules for all managers")
```




```
boxplot(manager.cast$pytResiduals,
        horizontal = TRUE,
        main = "residules for all managers witout outliers",
        outline = FALSE)
```



Extra

Remove managers that played ofer 10 games.

```
manager.reduced <- subset(manager.cast, games > 10)
```

Managers that over performed

```
head(manager.reduced, n=10)
```

```
##             manager    R   RA   W   L games      Wpct   pytWpct
## 23 Red Schoendienst   85   88  13  11    24 0.5416667 0.4826642
## 25       Russ Nixon  267  365  25  40    65 0.3846154 0.3485776
## 75    Glenn Hoffman  351  351  47  41    88 0.5340909 0.5000000
## 62   Mike Jorgensen  353  424  42  54    96 0.4375000 0.4093796
## 27    Stump Merrill 1089 1311 120 155   275 0.4363636 0.4082847
## 9        Don Zimmer 1031 1104 116 119   235 0.4936170 0.4658479
## 22       Nick Leyva  700  797  81  94   175 0.4628571 0.4354747
## 76     Jerry Manuel 2591 2605 247 232   479 0.5156576 0.4973056
## 57 Marcel Lachemann 1675 1788 160 170   330 0.4848485 0.4674040
## 16      Jim Leyland 7085 7421 766 788  1554 0.4929215 0.4768496
##    pytResiduals
## 23   0.05900249
## 25   0.03603778
```

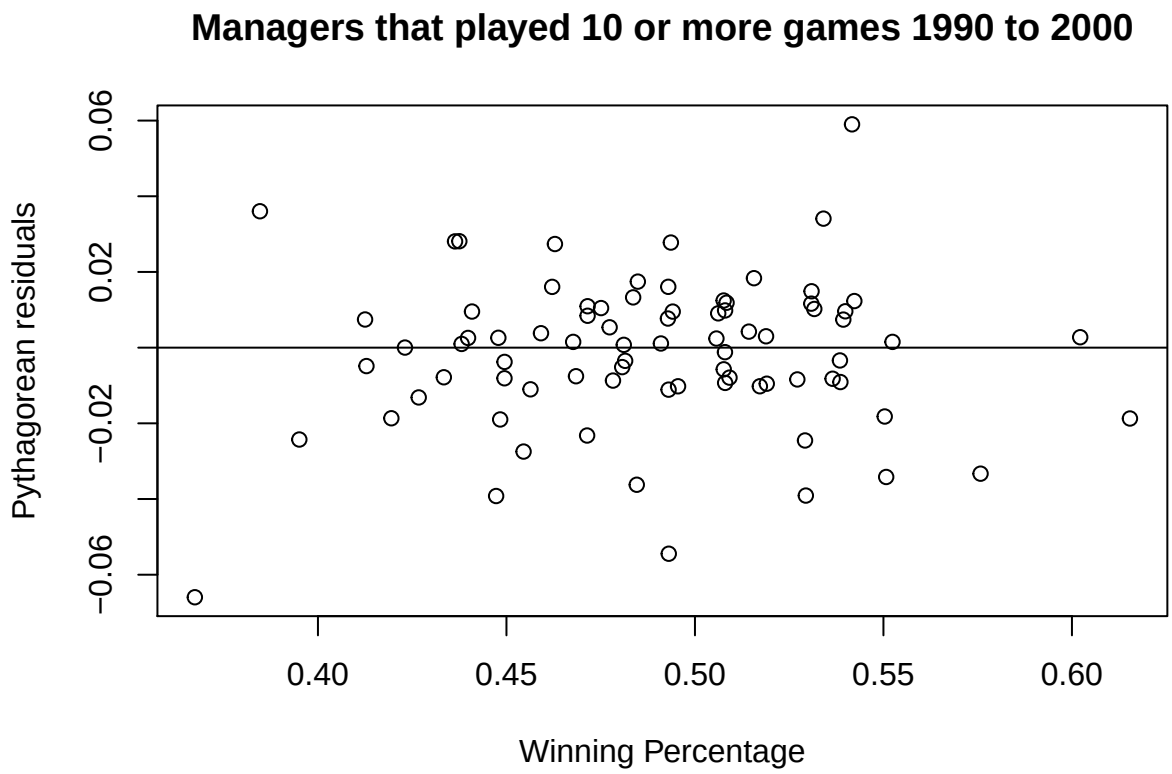
```
## 75   0.03409091
## 62   0.02812043
## 27   0.02807890
## 9    0.02776913
## 22   0.02738249
## 76   0.01835198
## 57   0.01744445
## 16   0.01607190
```

Manager that under performed

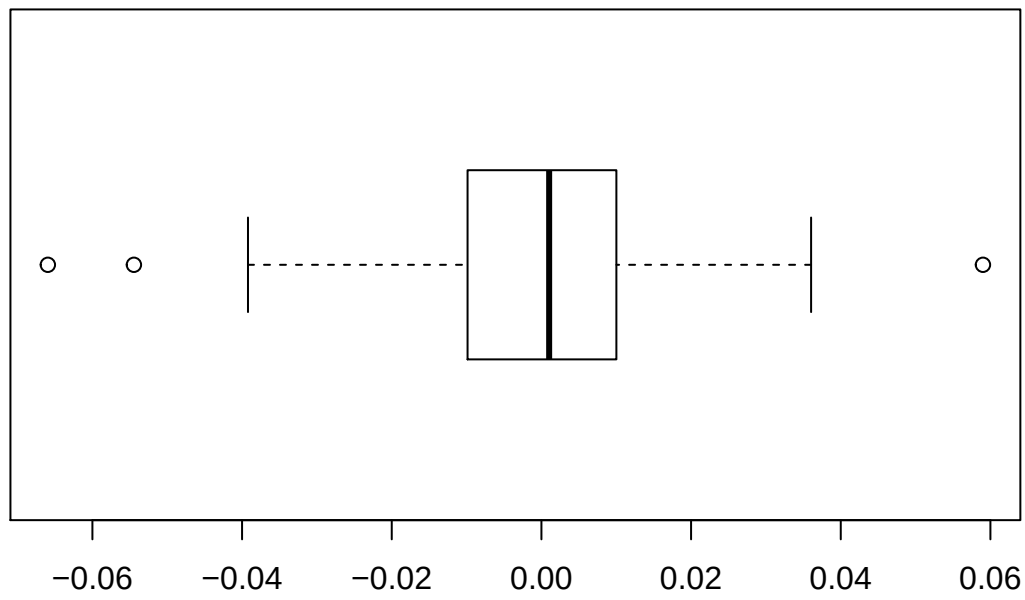
```
tail(manager.reduced, n=10)
```

```
##          manager    R   RA   W   L games      Wpct   pytWpct pytResiduals
## 42  Bill Plummer  679  799  64  98   162 0.3950617 0.4193409  -0.02427917
## 6  Bud Harrelson 1193 1071 145 129   274 0.5291971 0.5537309  -0.02453382
## 55    Tony Perez  191  198  20  24    44 0.4545455 0.4820110  -0.02746551
## 34   Gene Tenace  171  137  19  14    33 0.5757576 0.6090606  -0.03330304
## 71 Larry Dierker 3281 2764 342 279   621 0.5507246 0.5849042  -0.03417955
## 79    Ray Miller 1668 1600 157 167   324 0.4845679 0.5207988  -0.03623093
## 67    Joe Maddon  264  230  27  24    51 0.5294118 0.5685014  -0.03908965
## 50  Dallas Green 2297 2360 229 283   512 0.4472656 0.4864745  -0.03920883
## 63    Phil Regan  704  640  71  73   144 0.4930556 0.5475113  -0.05445576
## 5     Bucky Dent  188  215  18  31    49 0.3673469 0.4333019  -0.06595493
```

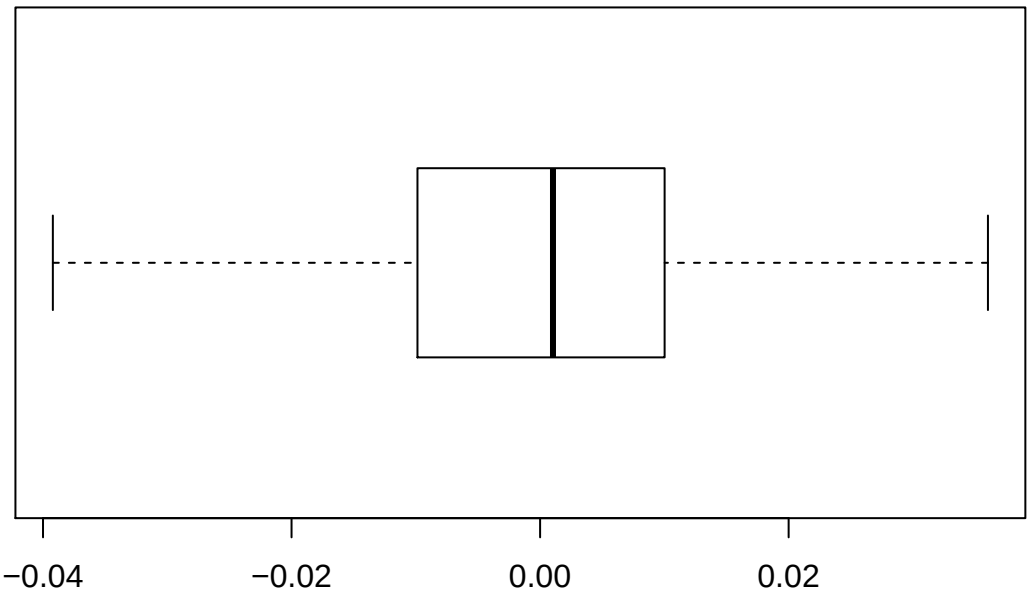
```
with(manager.reduced, plot(Wpct, pytResiduals,
                        ylab = "Pythagorean residuals",
                        xlab = "Winning Percentage",
                        main = "Managers that played 10 or more games 1990 to 2000"))
abline(h=0)
```




```
boxplot(manager.reduced$pytResiduals,
        horizontal = TRUE,
        main = "residules for all managers")
```




```
boxplot(manager.reduced$pytResiduals,
        horizontal = TRUE,
        main = "residules for all managers witout outliers",
        outline = FALSE)
```


residules for all managers witout outliers
−0.04
−0.02
0.00
0.02