Pairs Plots and Correlation

JuJu

## Grab the Data

We need some data with a number of variables. Using the Lahman package, we can download a year’s worth of team data.

 p\_load(Lahman)
 data(Teams)
 MLB = subset.data.frame(Teams, subset=(yearID==1961 & lgID %in% c("AL","NL")))
 head(MLB)

## yearID lgID teamID franchID divID Rank G Ghome W L DivWin WCWin LgWin
## 1360 1961 AL BAL BAL <NA> 3 163 82 95 67 <NA> <NA> N
## 1361 1961 AL BOS BOS <NA> 6 163 82 76 86 <NA> <NA> N
## 1362 1961 AL CHA CHW <NA> 4 163 81 86 76 <NA> <NA> N
## 1363 1961 NL CHN CHC <NA> 7 156 78 64 90 <NA> <NA> N
## 1364 1961 NL CIN CIN <NA> 1 154 77 93 61 <NA> <NA> Y
## 1365 1961 AL CLE CLE <NA> 5 161 81 78 83 <NA> <NA> N
## WSWin R AB H X2B X3B HR BB SO SB CS HBP SF RA ER ERA CG SHO
## 1360 N 691 5481 1393 227 36 149 581 902 39 30 NA NA 588 526 3.22 54 21
## 1361 N 729 5508 1401 251 37 112 647 847 56 36 NA NA 792 687 4.29 35 6
## 1362 N 765 5556 1475 216 46 138 550 612 100 40 NA NA 726 653 4.06 39 3
## 1363 N 689 5344 1364 238 51 176 539 1027 35 25 NA NA 800 689 4.48 34 6
## 1364 N 710 5243 1414 247 35 158 423 761 70 33 NA NA 653 575 3.78 46 12
## 1365 N 737 5609 1493 257 39 150 492 720 34 11 NA NA 752 665 4.15 35 12
## SV IPouts HA HRA BBA SOA E DP FP name
## 1360 33 4413 1226 109 617 926 126 173 0.980 Baltimore Orioles
## 1361 30 4326 1472 167 679 831 143 140 0.977 Boston Red Sox
## 1362 33 4344 1491 158 498 814 128 138 0.980 Chicago White Sox
## 1363 25 4155 1492 165 465 755 183 175 0.970 Chicago Cubs
## 1364 40 4110 1300 147 500 829 134 124 0.977 Cincinnati Reds
## 1365 23 4329 1426 178 599 801 139 142 0.977 Cleveland Indians
## park attendance BPF PPF teamIDBR teamIDlahman45 teamIDretro
## 1360 Memorial Stadium 951089 96 96 BAL BAL BAL
## 1361 Fenway Park II 850589 102 103 BOS BOS BOS
## 1362 Comiskey Park 1146019 99 97 CHW CHA CHA
## 1363 Wrigley Field 673057 101 104 CHC CHN CHN
## 1364 Crosley Field 1117603 102 101 CIN CIN CIN
## 1365 Cleveland Stadium 725547 97 98 CLE CLE CLE

 names(MLB)

## [1] "yearID" "lgID" "teamID" "franchID"
## [5] "divID" "Rank" "G" "Ghome"
## [9] "W" "L" "DivWin" "WCWin"
## [13] "LgWin" "WSWin" "R" "AB"
## [17] "H" "X2B" "X3B" "HR"
## [21] "BB" "SO" "SB" "CS"
## [25] "HBP" "SF" "RA" "ER"
## [29] "ERA" "CG" "SHO" "SV"
## [33] "IPouts" "HA" "HRA" "BBA"
## [37] "SOA" "E" "DP" "FP"
## [41] "name" "park" "attendance" "BPF"
## [45] "PPF" "teamIDBR" "teamIDlahman45" "teamIDretro"

 MLB$lgID = factor(MLB$lgID)

## Pairs Plots

Pairs plots make it possible to look at the relationship between multiple variables at the same time. Generally we look at relationships between quantitative variables. However, comparisons using qualitative data generate parallel dotplots which are sometimes informative.

 pairs(MLB[,c(2,9,10,15,16,17,18,19,20,21)])



## Correlation

Data where when plotted we see that if one variable increases the other will increase — or equivalently when one decreases the other decreases — are said to be ***positively*** associated. If when one variable increases the other variable decreases we say the variables are ***negatively*** associated.

A measure of the strength and direction of the ***linear*** association between two variables is the **correlation**. Correlation is essentially the almost average product of the z-scores of the two variables. In R it is easy to compute the correlation between two (or more) variables. To find the correlation between hits, ***H***, and runs, ***R***, we can use R to compute the correlation “by hand.”

 zh = (MLB$H - mean(MLB$H))/sqrt(var(MLB$H))
 zh

## [1] 0.006959129 0.132223454 1.290918454 -0.447124047 0.335777981
## [6] 1.572763184 1.384866698 -0.791600939 -0.963839385 -0.541072290
## [11] -0.619362493 -0.431466006 1.071705887 -1.997270062 0.868151360
## [16] -0.212253439 0.680254873 -1.339632358

 zr = (MLB$R - mean(MLB$R))/sqrt(var(MLB$R))
 zr

## [1] -0.4512746 0.1611695 0.7413798 -0.4835086 -0.1450526 0.2901051
## [7] 1.9662681 -0.5802103 0.4029238 0.2578712 -0.1934034 -0.1128187
## [13] 1.7406308 -2.1757885 -0.4029238 0.8703154 -0.2578712 -1.6278121

 cor.rh = sum(zh\*zr)/(length(zh)-1)
 cor.rh

## [1] 0.7119815

Alternatively, we can use the internal function ***cor***.

 # For hits and runs this is
 cor(MLB$H,MLB$R)

## [1] 0.7119815

 # For all of the variables that were ploted (except lgID) we use
 cor(MLB[,c(9,10,15,16,17,18,19,20,21)])

## W L R AB H X2B
## W 1.00000000 -0.969570439 0.80863191 0.320836049 0.7045472 0.016427476
## L -0.96957044 1.000000000 -0.74046762 -0.105680982 -0.6528911 0.007618194
## R 0.80863191 -0.740467616 1.00000000 0.497382881 0.7119815 0.031894294
## AB 0.32083605 -0.105680982 0.49738288 1.000000000 0.5885888 0.283990796
## H 0.70454721 -0.652891137 0.71198146 0.588588835 1.0000000 0.437289268
## X2B 0.01642748 0.007618194 0.03189429 0.283990796 0.4372893 1.000000000
## X3B -0.19259119 0.152371525 -0.19491536 -0.007360244 0.2019594 -0.017286952
## HR 0.60840098 -0.602646208 0.66475810 0.176830482 0.2525095 -0.240654442
## BB 0.09293399 0.062188409 0.35991868 0.439793320 -0.1156499 -0.165058955
## X3B HR BB
## W -0.192591188 0.6084010 0.09293399
## L 0.152371525 -0.6026462 0.06218841
## R -0.194915357 0.6647581 0.35991868
## AB -0.007360244 0.1768305 0.43979332
## H 0.201959439 0.2525095 -0.11564993
## X2B -0.017286952 -0.2406544 -0.16505896
## X3B 1.000000000 -0.4308045 -0.30224067
## HR -0.430804506 1.0000000 0.18589154
## BB -0.302240666 0.1858915 1.00000000

The values computed by hand and using the internal function are the same. Since 0.7112 is positive, we see that as hits go up runs go up. The magnitude of 0.7112 inidicates a moderately strong linear association (while sociology majors might be happy, physics majors would not be impressed).

Wins and losses are strongly negatively associated (r=-0.9696). Since if you are not winning then you are losing, this is not surprising. At bats are positively associated with runs. This makes sense, but it should be noted that the correlation of 0.4974 is fairly weak.