

Introduction to ddply

Cleaning and Summarizing Data

Iowa State University

Outline

- ▶ conditionals & subsets
- ▶ for loops
- ▶ avoiding for loops with `ddply`

Baseball Data

- ▶ The `plyr` package contains the data set `baseball`
- ▶ seasonal batting statistics of all major league players (through 2007)

```
library(plyr)
help(baseball)
head(baseball)
```

	id	year	stint	team	lg	g	ab	r	h	X2b	X3b	hr	rbi	sb	cs	bb	so	ibb	hbp	sh	sf	gidp
4	ansonca01	1871	1	RC1		25	120	29	39	11	3	0	16	6	2	2	1	NA	NA	NA	NA	NA
44	forceda01	1871	1	WS3		32	162	45	45	9	4	0	29	8	0	4	0	NA	NA	NA	NA	NA
68	mathebo01	1871	1	FW1		19	89	15	24	3	1	0	10	2	1	2	0	NA	NA	NA	NA	NA
99	startjo01	1871	1	NY2		33	161	35	58	5	1	1	34	4	2	3	0	NA	NA	NA	NA	NA
102	suttoez01	1871	1	CL1		29	128	35	45	3	7	3	23	3	1	1	0	NA	NA	NA	NA	NA
106	whitede01	1871	1	CL1		29	146	40	47	6	5	1	21	2	2	4	1	NA	NA	NA	NA	NA

Baseball Data

- ▶ We would like to create career summary statistics for each player
- ▶ Plan: subset on a player, and compute statistics

```
ss <- subset(baseball, id=="sosasa01")  
head(ss)
```

	id	year	stint	team	lg	g	ab	r	h	X2b	X3b	hr	rbi	sb	cs	bb	so	ibb	hbp	sh	sf	gidp
66822	sosasa01	1989	1	TEX	AL	25	84	8	20	3	0	1	3	0	2	0	20	0	0	4	0	3
66823	sosasa01	1989	2	CHA	AL	33	99	19	27	5	0	3	10	7	3	11	27	2	2	1	2	3
67907	sosasa01	1990	1	CHA	AL	153	532	72	124	26	10	15	70	32	16	33	150	4	6	2	6	10
69018	sosasa01	1991	1	CHA	AL	116	316	39	64	10	1	10	33	13	6	14	98	2	2	5	1	5
70599	sosasa01	1992	1	CHN	NL	67	262	41	68	7	2	8	25	15	7	19	63	1	4	4	2	4
71757	sosasa01	1993	1	CHN	NL	159	598	92	156	25	5	33	93	36	11	38	135	6	4	0	1	14

```
mean(ss$h/ss$ab)  
## [1] 0.2682
```

Baseball Data

- ▶ We would like to create career summary statistics for each player
- ▶ Plan: subset on a player, and compute statistics

```
ss <- subset(baseball, id=="sosasa01")
head(ss)
```

	id	year	stint	team	lg	g	ab	r	h	X2b	X3b	hr	rbi	sb	cs	bb	so	ibb	hbp	sh	sf	gidp
66822	sosasa01	1989	1	TEX	AL	25	84	8	20	3	0	1	3	0	2	0	20	0	0	4	0	3
66823	sosasa01	1989	2	CHA	AL	33	99	19	27	5	0	3	10	7	3	11	27	2	2	1	2	3
67907	sosasa01	1990	1	CHA	AL	153	532	72	124	26	10	15	70	32	16	33	150	4	6	2	6	10
69018	sosasa01	1991	1	CHA	AL	116	316	39	64	10	1	10	33	13	6	14	98	2	2	5	1	5
70599	sosasa01	1992	1	CHN	NL	67	262	41	68	7	2	8	25	15	7	19	63	1	4	4	2	4
71757	sosasa01	1993	1	CHN	NL	159	598	92	156	25	5	33	93	36	11	38	135	6	4	0	1	14

```
mean(ss$h/ss$ab)
## [1] 0.2682
```

We need an automatic way to calculate this

for loops

- ▶ Idea: repeat the same (set of) statement(s) for each element of an index set
- ▶ Setup:
 - ▶ Introduce counter variable (sometimes named *i*)
 - ▶ Reserve space for results
- ▶ Generic Code:

```
result <- rep(NA, length(indexset))
for(i in indexset){
  ... some statments ...
  result[i] <- ...
}
```

for loops for Baseball

- ▶ Index set: player id
- ▶ Setup:

```
# Index set
players <- unique(baseball$id)
n <- length(players)

# Place to store data
ba <- rep(NA, n)

# Loop
for(i in 1:n){
  career <- subset(baseball, id==players[i])
  ba[i] <- with(career, mean(h/ab, na.rm=T))
}
```

```
# Results
summary(ba)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.000	0.183	0.246	0.223	0.270	0.500	6

for loops for Baseball

- ▶ Index set: player id
- ▶ $i=0$

```
# Index set  
players <- unique(baseball$id)  
n <- length(players)
```

```
# Place to store data  
ba <- rep(NA, n)
```

```
head(ba)  
## [1] NA NA NA NA NA NA NA
```


for loops for Baseball

- ▶ Index set: player id
- ▶ $i=1$

```
# Index set
players <- unique(baseball$id)
n <- length(players)

# Place to store data
ba <- rep(NA, n)

# Loop
for(i in 1:n){
  career <- subset(baseball, id==players[i])
  ba[i] <- with(career, mean(h/ab, na.rm=T))
}
i
## [1] 1

head(ba)
## [1] 0.3371      NA      NA      NA      NA      NA
```

for loops for Baseball

- ▶ Index set: player id
- ▶ $i=2$

```
# Index set
players <- unique(baseball$id)
n <- length(players)

# Place to store data
ba <- rep(NA, n)

# Loop
for(i in 1:2){
  career <- subset(baseball, id==players[i])
  ba[i] <- with(career, mean(h/ab, na.rm=T))
}
i
## [1] 2

head(ba)
## [1] 0.3371 0.2489      NA      NA      NA      NA
```

Your Turn

- ▶ MLB rules for the greatest all-time hitters are that players have to have played at least 1000 games with at least as many at-bats in order to be considered
- ▶ Extend the for loop above to collect the additional information
Introduce and collect data for two new variables: `games` and `atbats`

How did it go? What was difficult?

- ▶ household chores (declaring variables, setting values each time) distract from real work
- ▶ indices are error-prone
- ▶ loops often result in slow code because R can compute quantities using entire vectors in an optimized way

plyr package

- ▶ Routines from the plyr package help us to avoid for loops

- ▶ usage:

```
ddply(.data, .variables, .fun=NULL, ...)
```

- ▶ Split-apply-combine approach:
 1. SPLIT data into subsets on each element of an index set
 2. APPLY the same statements for each subset
 3. COMBINE the results into a new data frame

Example

- ▶ Goal: Compute mean statistics for each player
- ▶ Split the dataset by player ID, compute the mean for each column

```
allstats <- ddpby(baseball, .(id), mean)
```

```
head(allstats)
```

```
##           id V1  
## 1 aaronha01 NA  
## 2 abernte02 NA  
## 3 adairje01 NA  
## 4 adamsba01 NA  
## 5 adamsbo03 NA  
## 6 adcocjo01 NA
```

What went wrong?

Not all data is numeric!

Summarize

- A special function: summarise or summarize

```
summarise(baseball, ab=mean(h/ab, na.rm=T))  
##      ab  
## 1 0.234
```

```
summarize(baseball,  
           ba = mean(h/ab, na.rm=T),  
           games = sum(g, na.rm=T),  
           hr = sum(hr, na.rm=T),  
           ab = sum(ab, na.rm=T))  
##      ba  games    hr    ab  
## 1 0.234 1580070 113577 4891061
```

```
summarize(subset(baseball, id=="sosasa01"),  
           ba = mean(h/ab, na.rm=T),  
           games = sum(g, na.rm=T),  
           hr = sum(hr, na.rm=T),  
           ab = sum(ab, na.rm=T))  
##      ba games    hr    ab  
## 1 0.2682  2354  609 8813
```

ddply + Summarize

A powerful combination to create summary statistics

```
careers <- ddply(baseball, .(id), summarize,  
  ba = mean(h/ab, na.rm=T),  
  games = sum(g, na.rm=T),  
  homeruns = sum(hr, na.rm=T),  
  atbats = sum(ab, na.rm=T))
```

```
head(careers)
```

##	id	ba	games	homeruns	atbats
## 1	aaronha01	0.3011	3298	755	12364
## 2	abernte02	0.1824	681	0	181
## 3	adairje01	0.2363	1165	57	4019
## 4	adamsba01	0.2097	482	3	1019
## 5	adamsbo03	0.2378	1281	37	4019
## 6	adcocjo01	0.2752	1959	336	6606

Your Turn

- ▶ Find some summary statistics for each of the teams (variable team)
 - ▶ How many different (unique) players has the team had?
 - ▶ What was the team's first/last season?
- ▶ Challenge:
Find the number of players on each team over time. Does the number change?

```
careers <- ddply(baseball, .(id), summarize,  
  ba = mean(h/ab, na.rm=T),  
  games = sum(g, na.rm=T),  
  homeruns = sum(hr, na.rm=T),  
  atbats = sum(ab, na.rm=T))  
  
head(careers)
```