Data Analysis Tools and Practice(Using R)

2021.06.01

R语言实战

北京大学 软件与微电子学院
School of Software and Microelectronics, Peking University

Huiping Sun(孙惠平)

sunhp@ss.pku.edu.cn

课堂测试12

先用电脑完成
40分钟 然后誊抄纸上

- US baby names数据集：
- 美国1880年到2008年Top1000的男婴和女婴的名字
- 258000条记录
- year
- name
- sex
- percent
- 见： bnames.csv

year	name	percent	sex
1880	John	0.081541	boy
1880	William	0.080511	boy
1880	James	0.050057	boy
1880	Charles	0.045167	boy
1880	George	0.043292	boy
1880	Frank	0.02738	boy
1880	Joseph	0.022229	boy
1880	Thomas	0.021401	boy

```
> head(bnames,15)
   year     name  percent sex
1  1880     John 0.081541 boy
2  1880  William 0.080511 boy
3  1880    James 0.050057 boy
4  1880  Charles 0.045167 boy
5  1880   George 0.043292 boy
6  1880    Frank 0.027380 boy
7  1880   Joseph 0.022229 boy
8  1880   Thomas 0.021401 boy
9  1880    Henry 0.020641 boy
10 1880   Robert 0.020404 boy
11 1880   Edward 0.019965 boy
12 1880    Harry 0.018175 boy
13 1880   Walter 0.014822 boy
14 1880   Arthur 0.013504 boy
15 1880     Fred 0.013251 boy
```

RiA
Other

课堂测试12

40分钟 先用电脑完成 然后誊抄纸上

- 该数据集中每年有多少记录
- 数据集中男孩和女孩各自排名
- 男孩名和女孩名的Top100

dplyr

- Top100中男孩名和女孩名的所占比例
- 画图显示每一年男孩名和女孩名在Top100的比例
- 哪些名字仅仅在一年中使用，哪些名字每一年都使用
- 显示每个名字的平均百分比
- 那个名字被使用的时间最长

方差分析

方差分析定义

- 方差分析（analysis of variance，ANOVA）是分析各个自变量对因变量影响的一种方法。
- 这里的自变量就是定性变量的因子及可能出现的称为协变量（covariate）的定量变量。
- 分析结果是由一个方差分析表表示的
- 原理为：把因变量的值随着自变量的不同取值而得到的变化进行分解，使得每一个自变量都有一份贡献，最后剩下无法用已知的原因解释的则看成随机误差的贡献。
- 然后用各自变量的贡献和随机误差的贡献进行比较（F检验），以判断该自变量的不同水平是否对因变量的变化有显著贡献。输出就是F-值和检验的一些p-值。

一个例子

表9-1　单因素组间方差分析

治疗方案	
CBT	EMDR
s1	s6
s2	s7
s3	s8
s4	s9
s5	s10

教材RiA
199页

表9-2　单因素组内方差分析

患　　者	时　　间	
	5周	6个月
s1		
s2		
s3		
s4		
s5		
s6		
s7		
s8		
S9		
s10		

一个例子

表9-3　含组间和组内因子的双因素方差分析

		患　　者	时　　间	
			5周	6个月
疗法	CBT	s1 s2 s3 s4 s5		
	EMDR	s6 s7 s8 s9 s10		

协方差分析

多元方差分析

aov()

- **aov**(formula, data = dataframe)

教材RiA 201页

表9-4　R表达式中的特殊符号

符　号	用　法
~	分隔符号，左边为响应变量，右边为解释变量。例如，用A、B和C预测y，代码为y~ A + B + C
+	分隔解释变量
:	表示变量的交互项。例如，用A、B和A与B的交互项来预测y，代码为y~ A + B + A:B
*	表示所有可能交互项。代码y~ A * B * C可展开为y ~ A + B + C + A:B + A:C + B:C + A:B:C
^	表示交互项达到某个次数。代码y ~ (A + B + C)^2可展开为y ~ A + B + C + A:B + A:C + B:C
.	表示包含除因变量外的所有变量。例如，若一个数据框包含变量y、A、B和C，代码y ~ .可展开为y ~ A + B + C

表9-5　常见研究设计的表达式

设　计	表　达　式
单因素ANOVA	y ~ A
含单个协变量的单因素ANCOVA	y ~ x + A
双因素ANOVA	y ~ A * B
含两个协变量的双因素ANCOVA	y ~ x1 + x2 + A*B
随机化区组	y ~ B + A（B是区组因子）
单因素组内ANOVA	y ~ A + Error(Subject/A)
含单个组内因子（W）和单个组间因子（B）的重复测量ANOVA	y ~ B * W + Error(Subject/W)

单因素方差分析

```
> library(multcomp)
> attach(cholesterol)
```

```
> table(trt)
trt
 1time 2times 4times  drugD  drugE
    10     10     10     10     10
> aggregate(response, by = list(trt), FUN = mean)
  Group.1        x
1   1time  5.78197
2  2times  9.22497
3  4times 12.37478
4   drugD 15.36117
5   drugE 20.94752
> aggregate(response, by = list(trt), FUN = sd)
  Group.1        x
1   1time 2.878113
2  2times 3.483054
3  4times 2.923119
4   drugD 3.454636
5   drugE 3.345003
```

```
> cholesterol
      trt response
1   1time   3.8612
2   1time  10.3868
3   1time   5.9059
4   1time   3.0609
5   1time   7.7204
6   1time   2.7139
7   1time   4.9243
8   1time   2.3039
9   1time   7.5301
10  1time   9.4123
11 2times  10.3993
12 2times   8.6027
13 2times  13.6320
14 2times   3.5054
15 2times   7.7703
16 2times   8.6266
17 2times   9.2274
18 2times   6.3159
19 2times  15.8258
20 2times   8.3443
21 4times  13.9621
```

单因素方差分析表

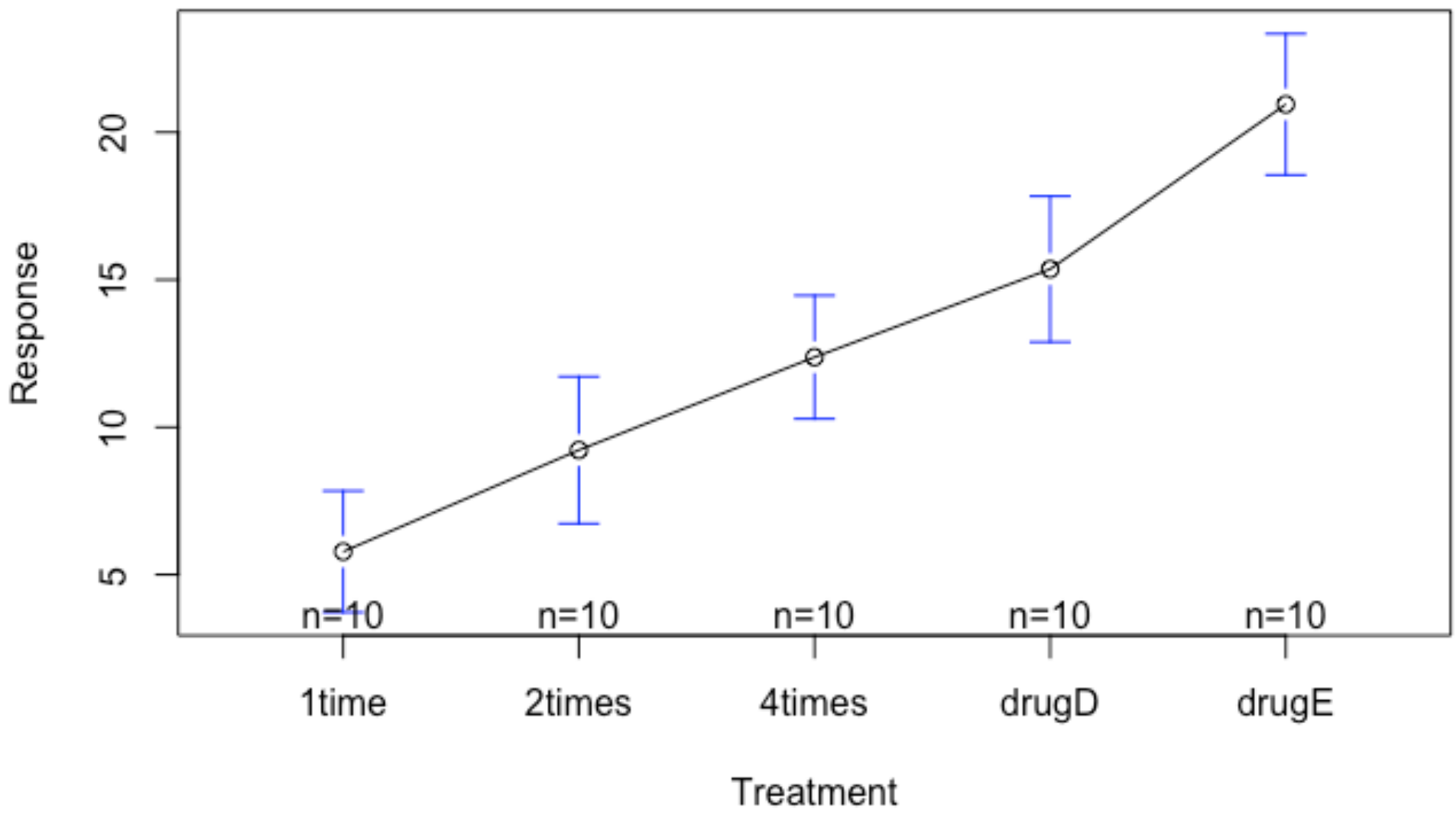
表 7.3: 单因素方差分析表

方差来源	自由度	平方和	均方	F 比	p 值
因素 A	$r-1$	S_A	$MS_A = \frac{S_A}{r-1}$	$F = \frac{MS_A}{MS_E}$	p
误 差	$n-r$	S_E	$MS_E = \frac{S_E}{n-r}$		
总 和	$n-1$	S_T			

```
> fit <- aov(response ~ trt)
> summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
trt          4 1351.4   337.8   32.43 9.82e-13 ***
Residuals   45  468.8    10.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

单因素方差分析例子

```
> library(gplots)
> plotmeans(response ~ trt, xlab = "Treatment", ylab = "Response",
+     main = "Mean Plot\nwith 95% CI")
```

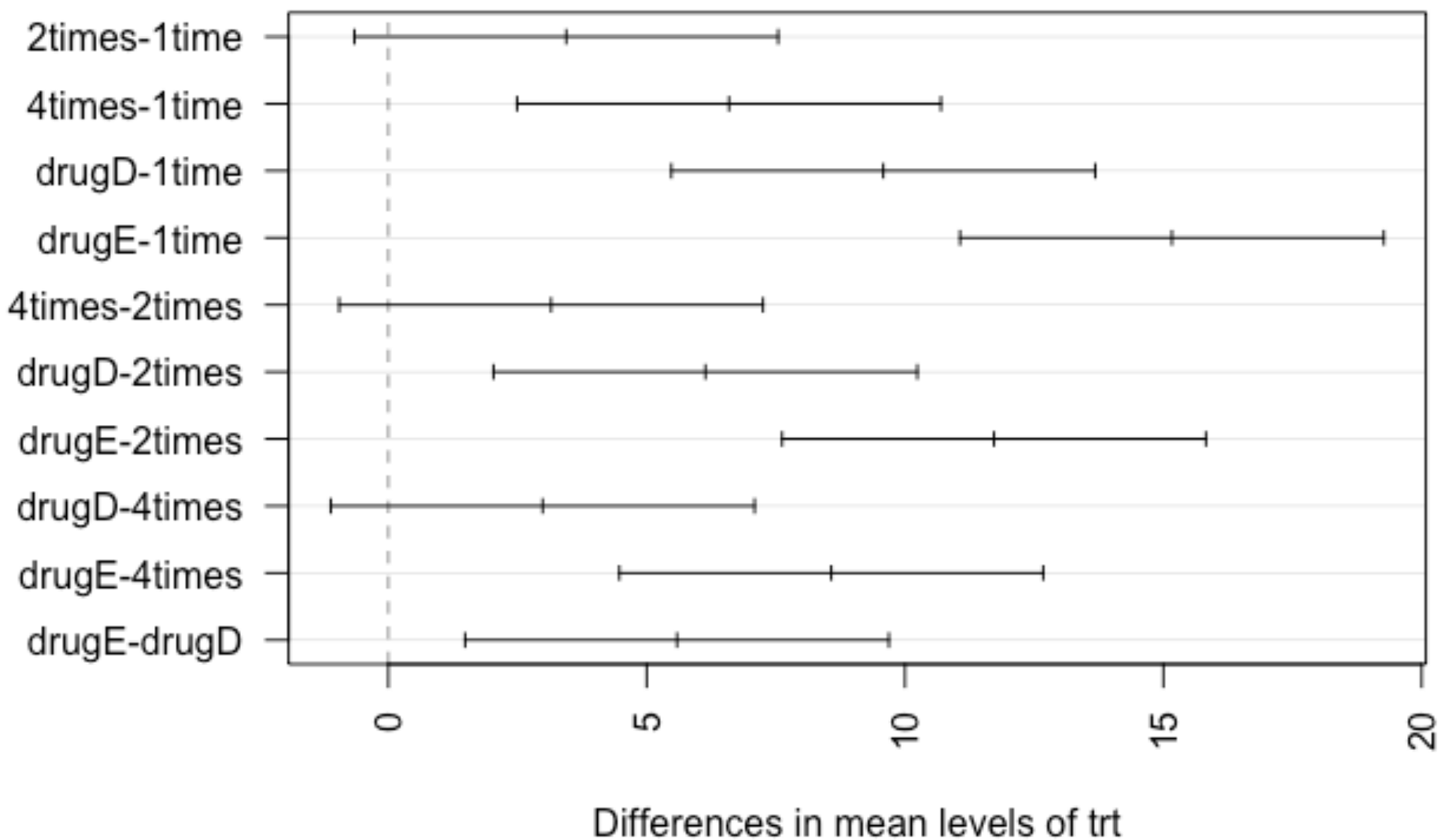


多重比较

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = response ~ trt)

$trt
                    diff        lwr       upr     p adj
2times-1time     3.44300 -0.6582817  7.544282 0.1380949
4times-1time     6.59281  2.4915283 10.694092 0.0003542
drugD-1time      9.57920  5.4779183 13.680482 0.0000003
drugE-1time     15.16555 11.0642683 19.266832 0.0000000
4times-2times    3.14981 -0.9514717  7.251092 0.2050382
drugD-2times     6.13620  2.0349183 10.237482 0.0009611
drugE-2times    11.72255  7.6212683 15.823832 0.0000000
drugD-4times     2.98639 -1.1148917  7.087672 0.2512446
drugE-4times     8.57274  4.4714583 12.674022 0.0000037
drugE-drugD      5.58635  1.4850683  9.687632 0.0030633
```

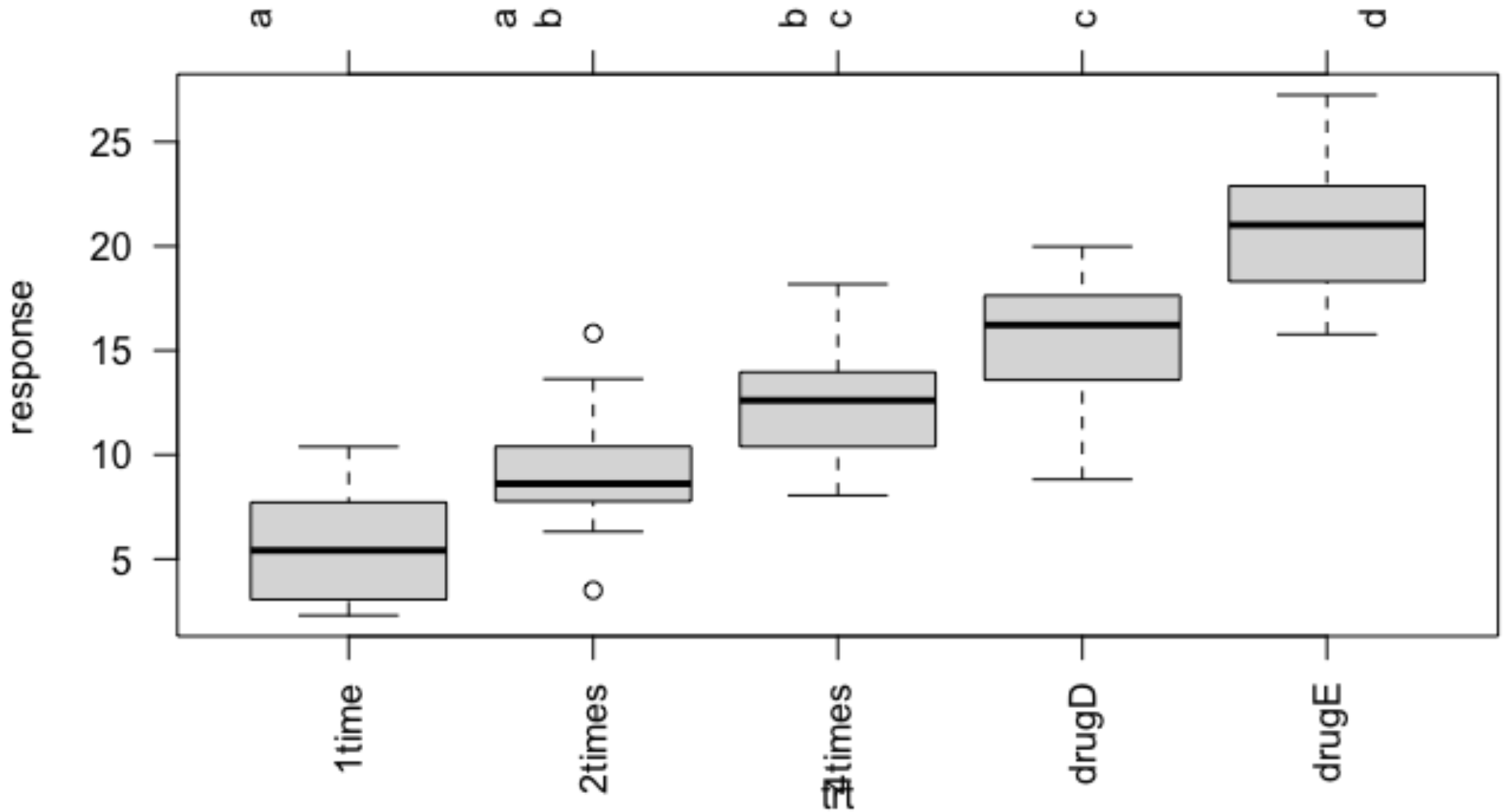
多重比较

```
> par(las = 2)
> par(mar = c(5, 8, 4, 2))
> plot(TukeyHSD(fit))
> par(opar)
```



多重比较

```
library(multcomp)
par(mar = c(5, 4, 6, 2))
tuk <- glht(fit, linfct = mcp(trt = "Tukey"))
plot(cld(tuk, level = 0.05), col = "lightgrey")
par(opar)
```



正态假设检验

```
library(car)
qqPlot(lm(response ~ trt, data = cholesterol), simulate = TRUE,
       main = "QQ Plot", labels = FALSE)
```

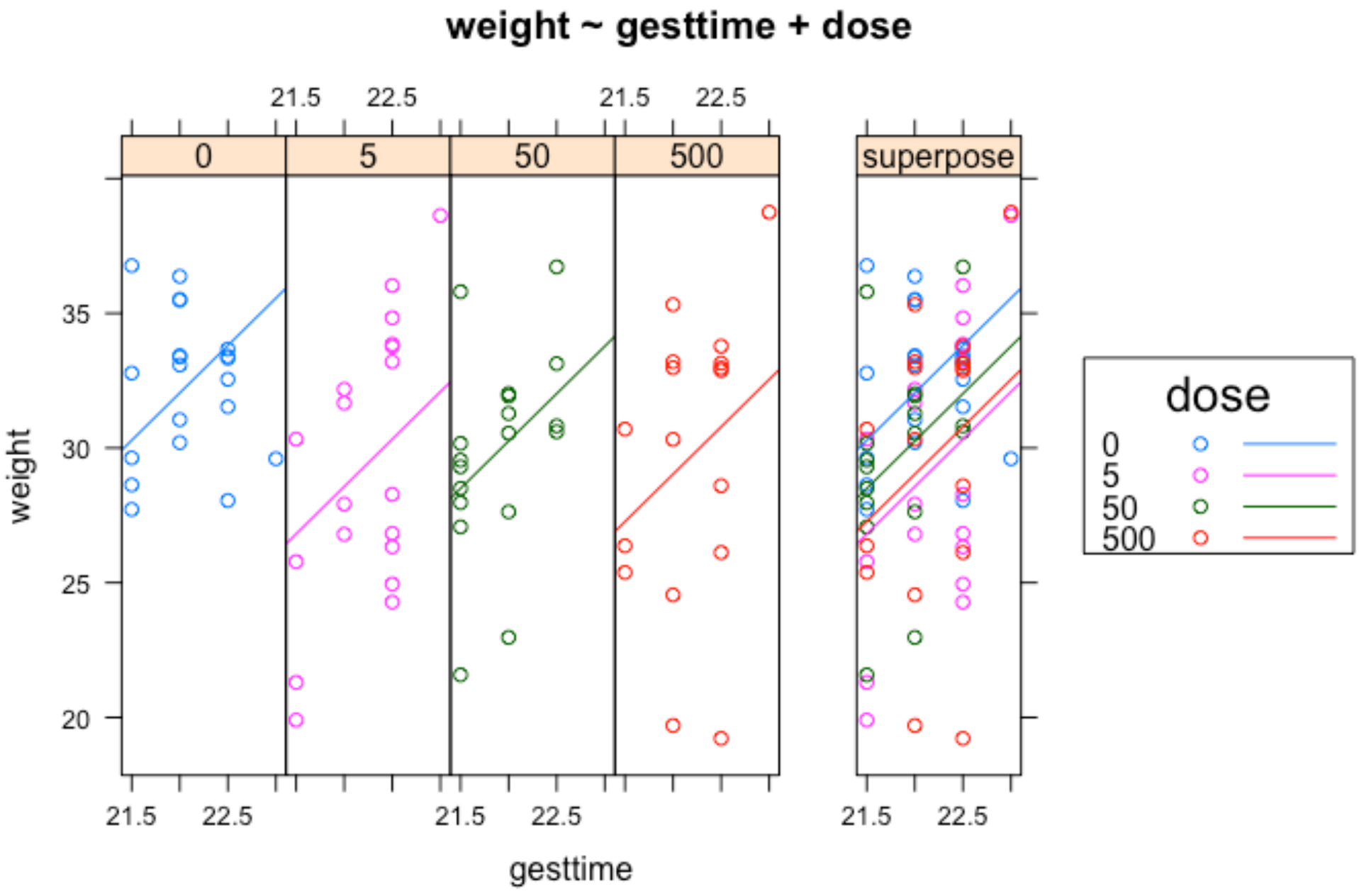
单因素协方差分析

```
> data(litter, package = "multcomp")
> attach(litter)
> table(dose)
dose
  0   5  50 500
 20  19  18  17
> aggregate(weight, by = list(dose), FUN = mean)
  Group.1        x
1       0 32.30850
2       5 29.30842
3      50 29.86611
4     500 29.64647
> fit <- aov(weight ~ gesttime + dose)
> summary(fit)
            Df Sum Sq Mean Sq F value  Pr(>F)
gesttime     1  134.3  134.30   8.049 0.00597 **
dose         3  137.1   45.71   2.739 0.04988 *
Residuals   69 1151.3   16.69
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> litter
   dose weight gesttime number
1     0  28.05     22.5     15
2     0  33.33     22.5     14
3     0  36.37     22.0     14
4     0  35.52     22.0     13
5     0  36.77     21.5     15
6     0  29.60     23.0      5
7     0  27.72     21.5     16
8     0  33.67     22.5     15
9     0  32.55     22.5     14
10    0  32.78     21.5     15
11    0  31.05     22.0     12
12    0  33.40     22.5     15
13    0  30.20     22.0     16
14    0  28.63     21.5      7
15    0  33.38     22.0     15
16    0  33.43     22.0     13
17    0  29.63     21.5     14
18    0  33.08     22.0     15
19    0  31.53     22.5     16
20    0  35.48     22.0      9
```

单因素协方差分析

```
library(HH)
ancova(weight ~ gesttime + dose, data = litter)
```



双因素方差分析

```
> table(supp, dose)
    dose
supp 0.5  1  2
  OJ  10 10 10
  VC  10 10 10
> aggregate(len, by = list(supp, dose), FUN = mean)
  Group.1 Group.2     x
1      OJ     0.5 13.23
2      VC     0.5  7.98
3      OJ     1.0 22.70
4      VC     1.0 16.77
5      OJ     2.0 26.06
6      VC     2.0 26.14
> aggregate(len, by = list(supp, dose), FUN = sd)
  Group.1 Group.2        x
1      OJ     0.5 4.459709
2      VC     0.5 2.746634
3      OJ     1.0 3.910953
4      VC     1.0 2.515309
5      OJ     2.0 2.655058
6      VC     2.0 4.797731
```

双因素方差分析表

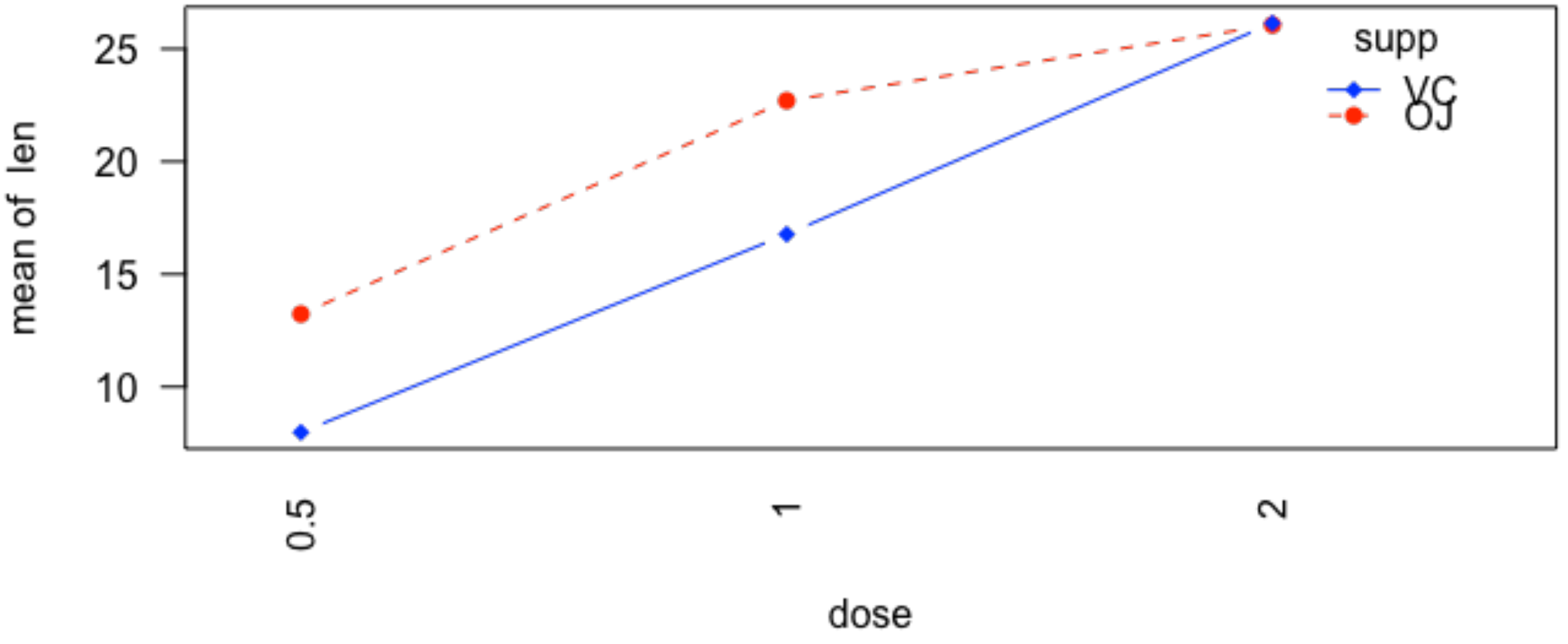
表 7.12: 双因素方差分析表

方差来源	自由度	平方和	均方	F 比	p 值
因素 A	$r-1$	S_A	$MS_A = \frac{S_A}{r-1}$	$F_A = \frac{MS_A}{MS_E}$	p_A
因素 B	$s-1$	S_B	$MS_B = \frac{S_B}{s-1}$	$F_B = \frac{MS_B}{MS_E}$	p_B
误 差	$(r-1)(s-1)$	S_E	$MS_E = \frac{S_E}{(r-1)(s-1)}$		
总 和	$rs-1$	S_T			

```
> fit <- aov(len ~ supp * dose)
> summary(fit)
             Df Sum Sq Mean Sq F value   Pr(>F)
supp          1  205.4   205.4  12.317 0.000894 ***
dose          1 2224.3  2224.3 133.415  < 2e-16 ***
supp:dose     1   88.9    88.9   5.333 0.024631 *
Residuals    56  933.6    16.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

双因素方差分析例子

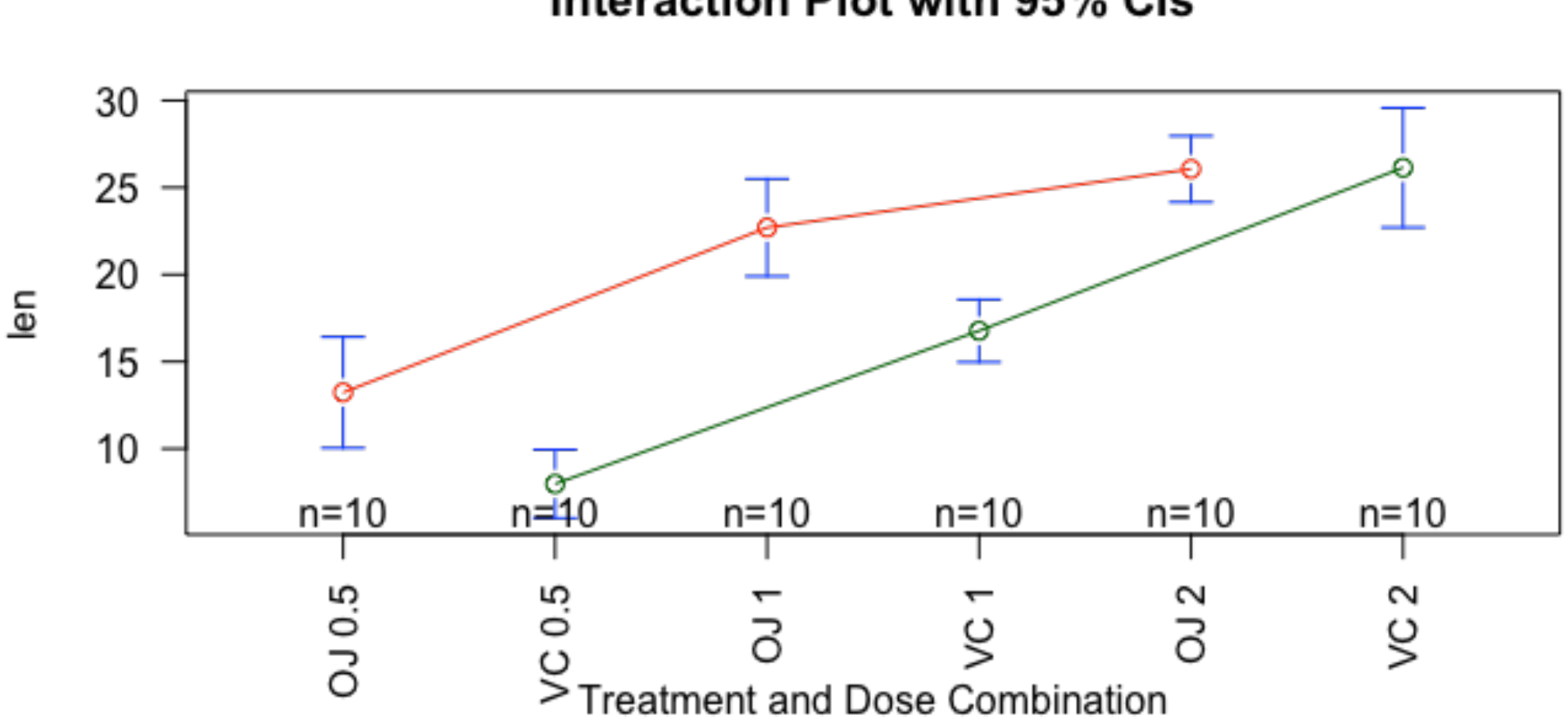
```
interaction.plot(dose, supp, len,
                 type = "b", col = c("red", "blue"), pch = c(16, 18),
                 main = "Interaction between Dose and Supplement Type")
```



双因素方差分析例子

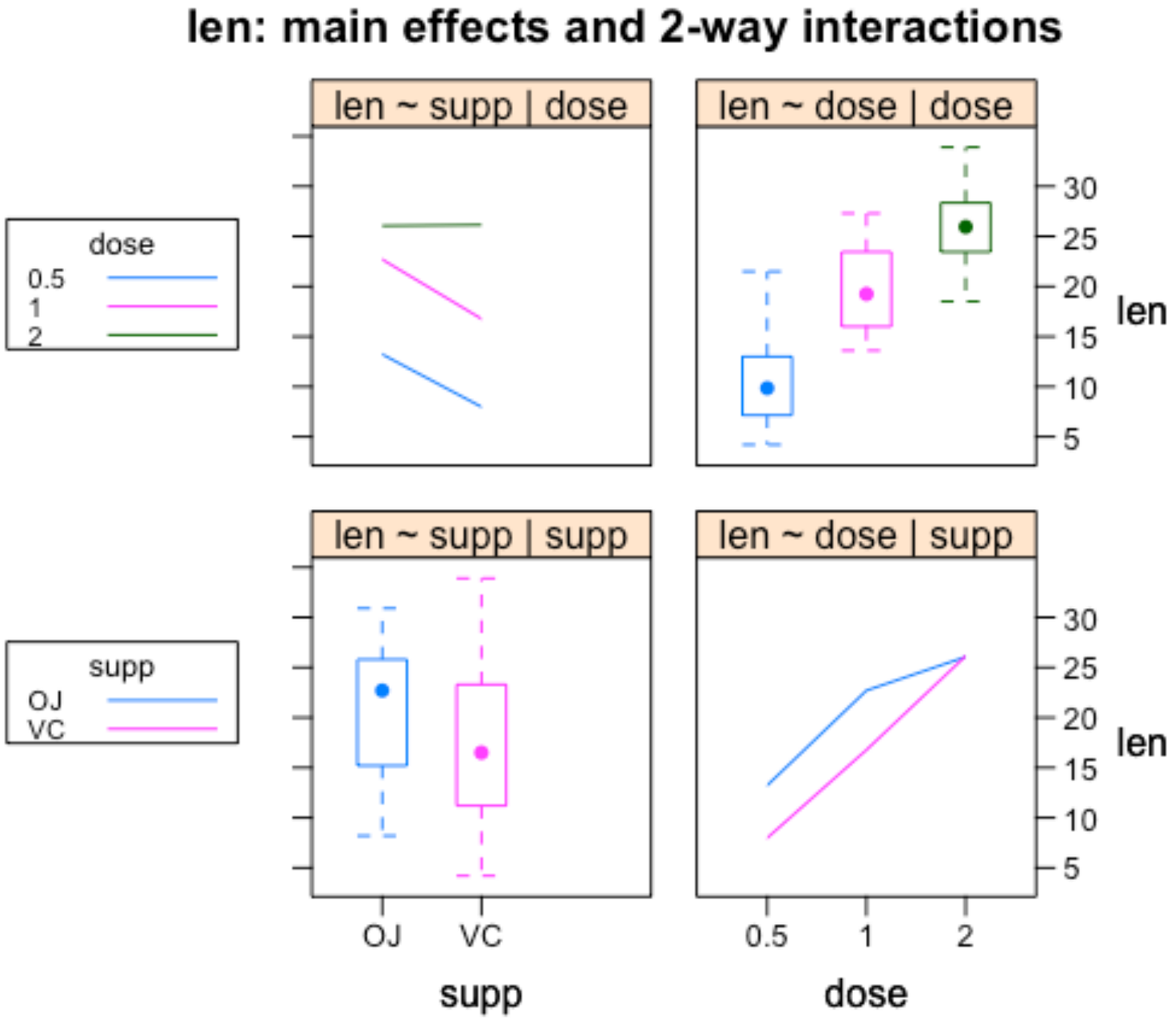
```
library(gplots)
plotmeans(len ~ interaction(supp, dose, sep = " "),
          connect = list(c(1, 3, 5), c(2, 4, 6)), col = c("red", "darkgreen"),
          main = "Interaction Plot with 95% CIs",
          xlab = "Treatment and Dose Combination")
```

交互效应



双因素方差分析例子

```
library(HH)
interaction2wt(len ~ supp * dose)
```



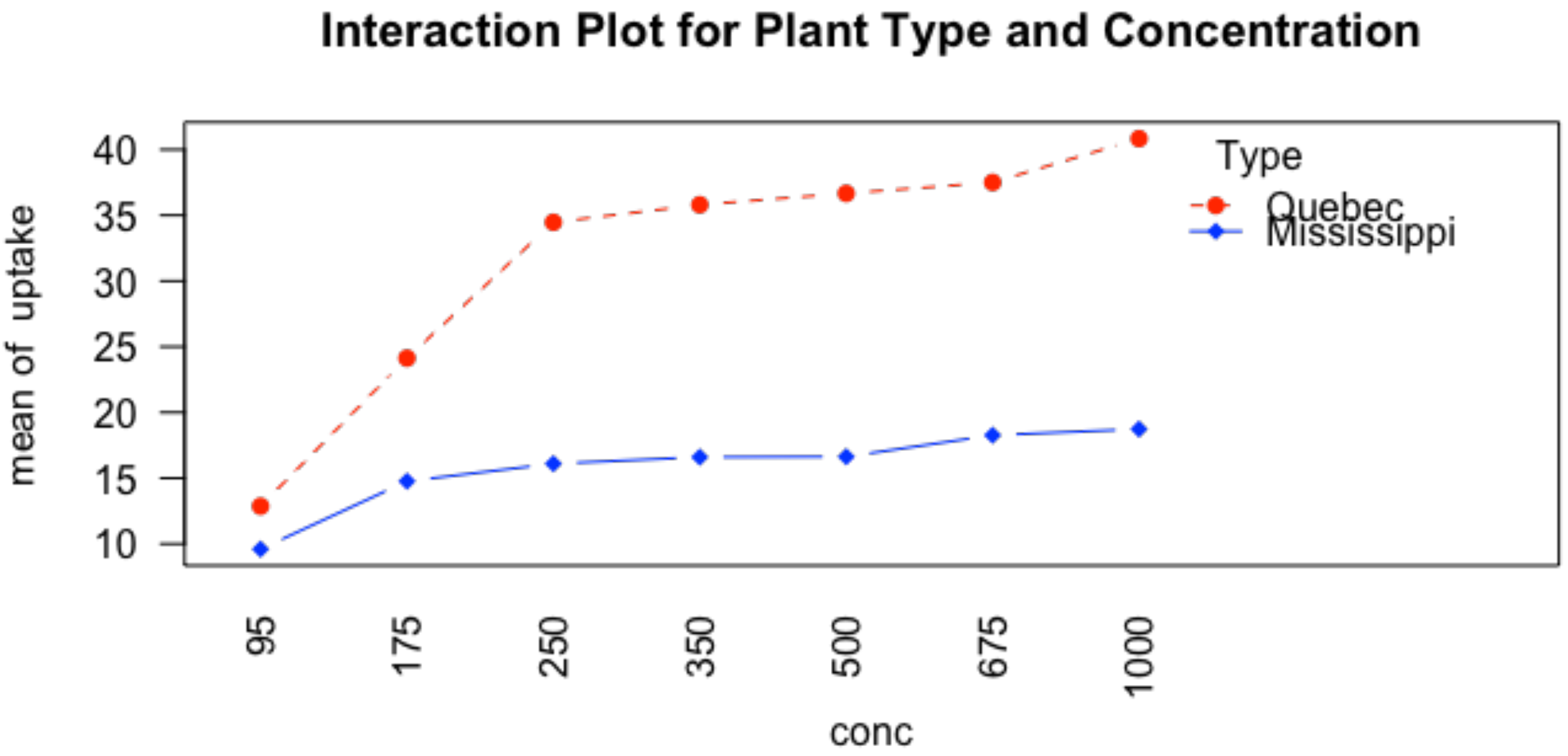
重复测量方差分析

```
w1b1 <- subset(CO2, Treatment == "chilled")
fit <- aov(uptake ~ (conc * Type) + Error(Plant/(conc)),
    w1b1)
summary(fit)
```

```
> summary(fit)

Error: Plant
          Df Sum Sq Mean Sq F value  Pr(>F)
Type       1 2667.2  2667.2   60.41 0.00148 **
Residuals  4  176.6    44.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Plant:conc
          Df Sum Sq Mean Sq F value   Pr(>F)
conc       1  888.6   888.6  215.46 0.000125 ***
conc:Type  1  239.2   239.2   58.01 0.001595 **
Residuals  4   16.5     4.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 30    869   28.97
```

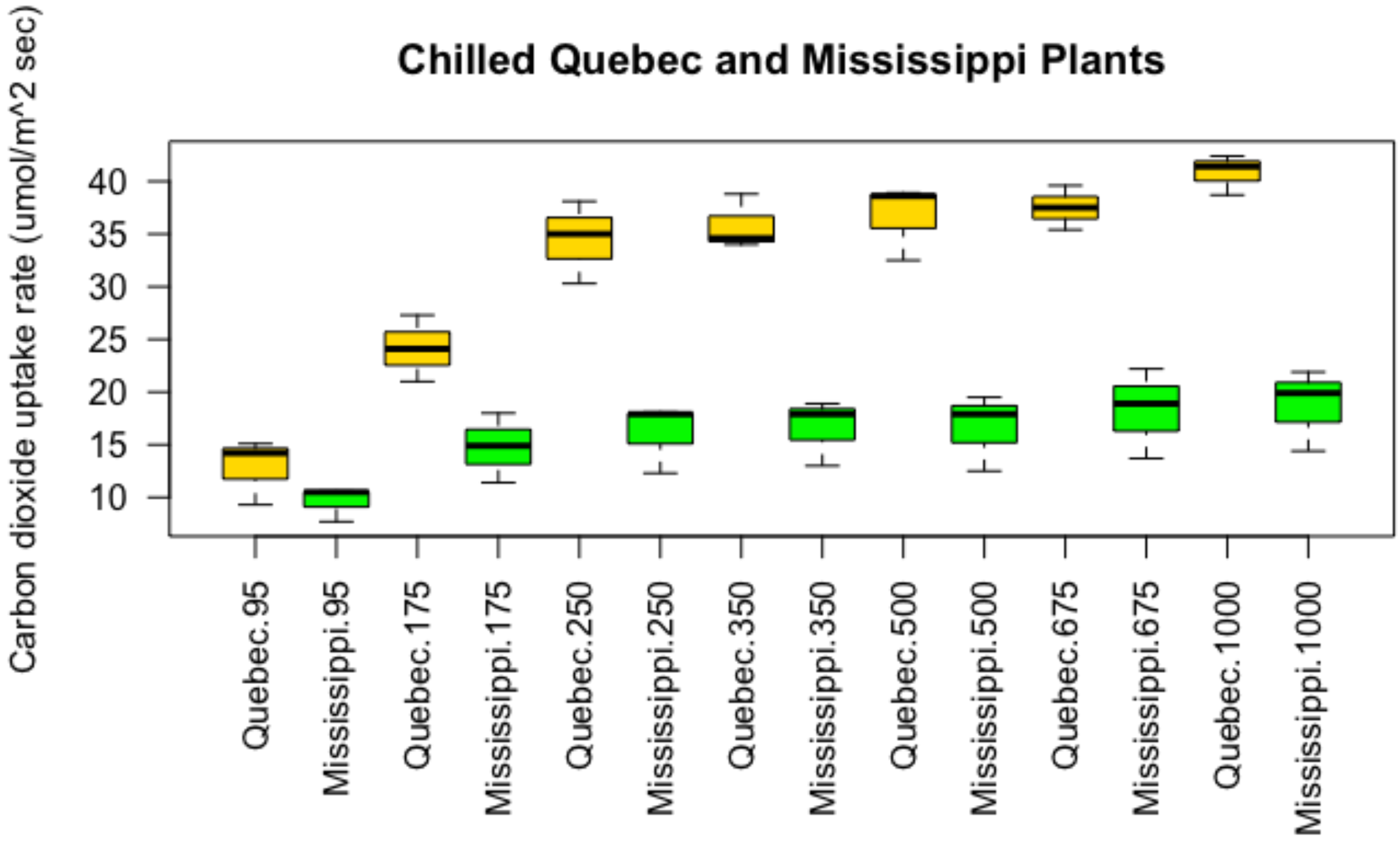
重复测量方差分析

```
par(las = 2)
par(mar = c(10, 4, 4, 2))
with(w1b1, interaction.plot(conc, Type, uptake, type = "b",
       col = c("red", "blue"), pch = c(16, 18),
       main = "Interaction Plot for Plant Type and Concentration"))
```



重复测量方差分析例子

```
boxplot(uptake ~ Type * conc, data = w1b1,
        col = (c("gold", "green")),
        main = "Chilled Quebec and Mississippi Plants",
        ylab = "Carbon dioxide uptake rate (umol/m^2 sec)")
```



缺失值处理

缺失值处理方法

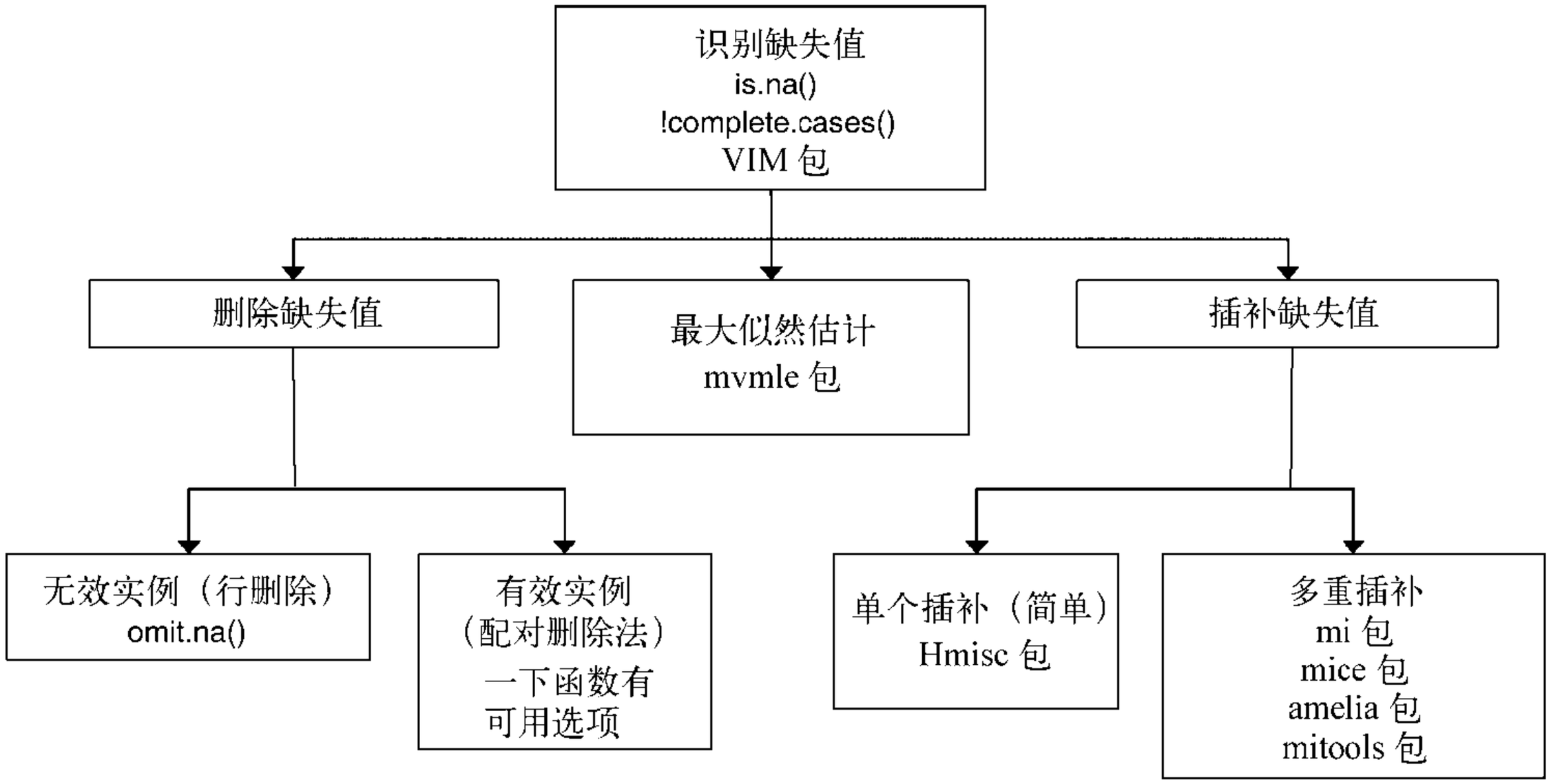


图18-1　处理不完整数据的方法，以及R中相关的包和函数

缺失值处理方法

表18-1　is.na()、is.nan()和is.infinite()函数的返回值示例

x	is.na(x)	is.nan(x)	is.infinite(x)
x <- NA	TRUE	FALSE	FALSE
x <- 0 / 0	TRUE	TRUE	FALSE
x <- 1 / 0	FALSE	FALSE	TRUE

mice包中的md.pattern()函数可生成一个以矩阵或数据框形式展示缺失值模式的表格。将函数应用到sleep数据集，可得到：

```
> library(mice)
> data(sleep, package="VIM")
> md.pattern(sleep)
   BodyWgt BrainWgt Pred Exp Danger Sleep Span Gest Dream NonD
42       1        1    1   1      1     1    1    1     1    0
 2       1        1    1   1      1     1    0    1     1    1
 3       1        1    1   1      1     1    1    0     1    1
 9       1        1    1   1      1     1    1    1     0    2
 2       1        1    1   1      1     0    1    1     1    2
 1       1        1    1   1      1     1    0    0     1    2
 2       1        1    1   1      1     0    1    1     0    3
 1       1        1    1   1      1     1    0    1     0    3
         0        0    0   0      0     4    4    4    12   14 38
```

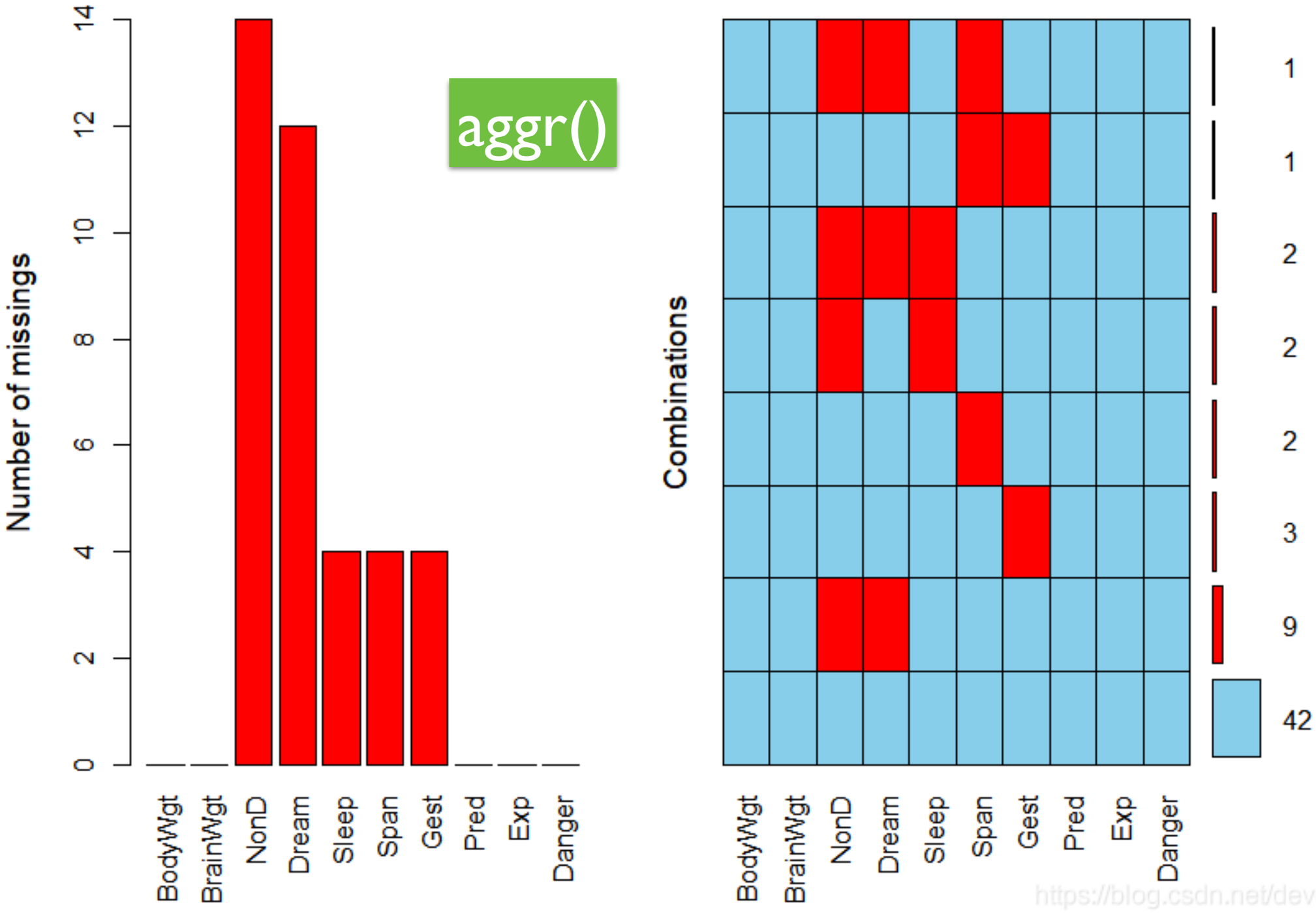
缺失值统计

mice包中的md.pattern()函数可生成一个以矩阵或数据框形式展示缺失值模式的表格。将函数应用到sleep数据集，可得到：

```
> library(mice)
> data(sleep, package="VIM")
> md.pattern(sleep)
   BodyWgt BrainWgt Pred Exp Danger Sleep Span Gest Dream NonD
42       1        1    1   1      1     1    1    1     1    0
 2       1        1    1   1      1     1    0    1     1    1
 3       1        1    1   1      1     1    1    0     1    1
 9       1        1    1   1      1     1    1    1     0    2
 2       1        1    1   1      1     0    1    1     0    2
 1       1        1    1   1      1     1    0    0     1    2
 2       1        1    1   1      1     0    1    1     0    3
 1       1        1    1   1      1     1    0    1     0    3
         0        0    0   0      0     4    4    4    12   14   38
```

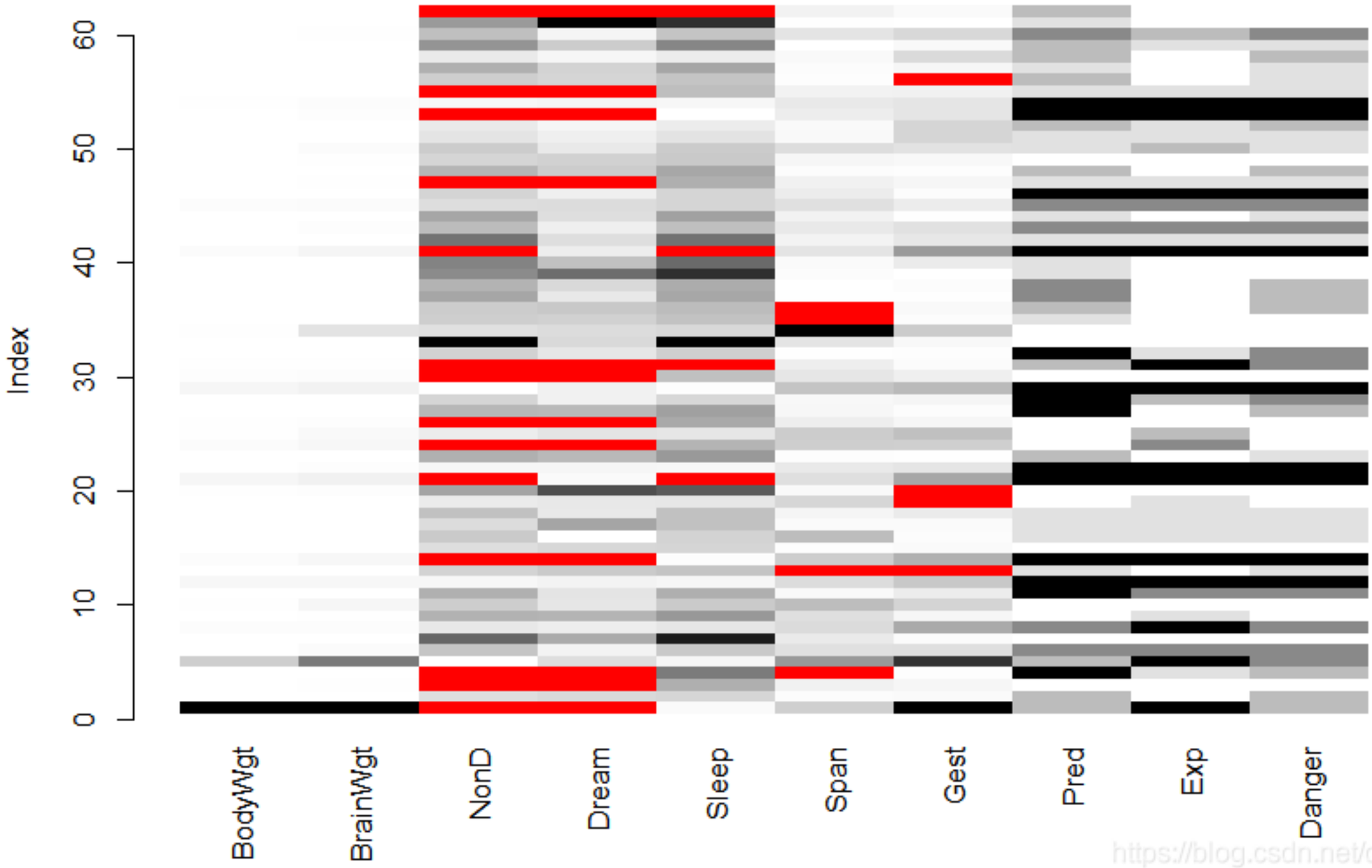
缺失值图形显示

aggr()

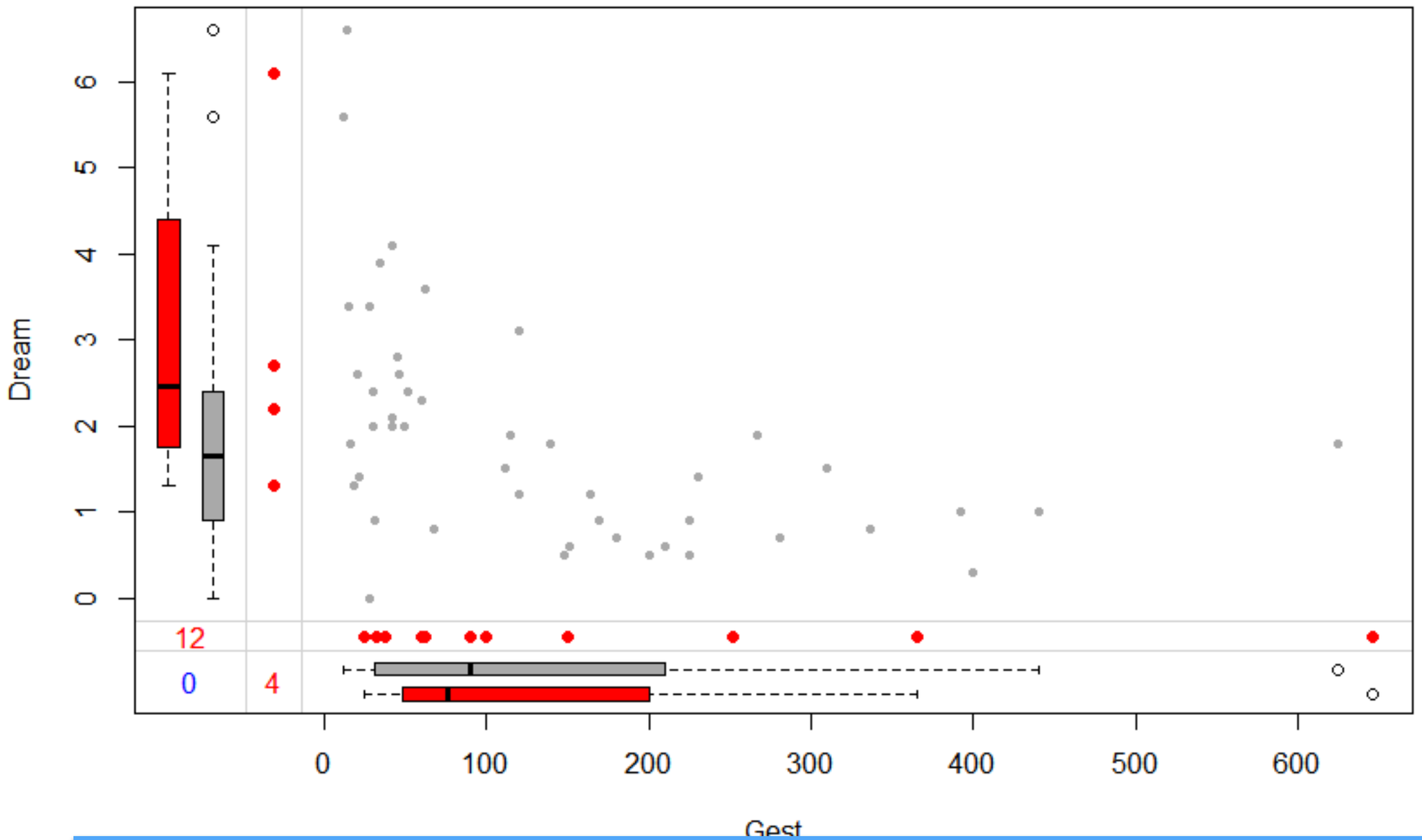


缺失值图形显示

matrixplot()



缺失值图形显示




```
marginplot(sleep[c("Gest", "Dream")], pch=c(20),
col=c("darkgray", "red", "blue"))
```

处理确实数据

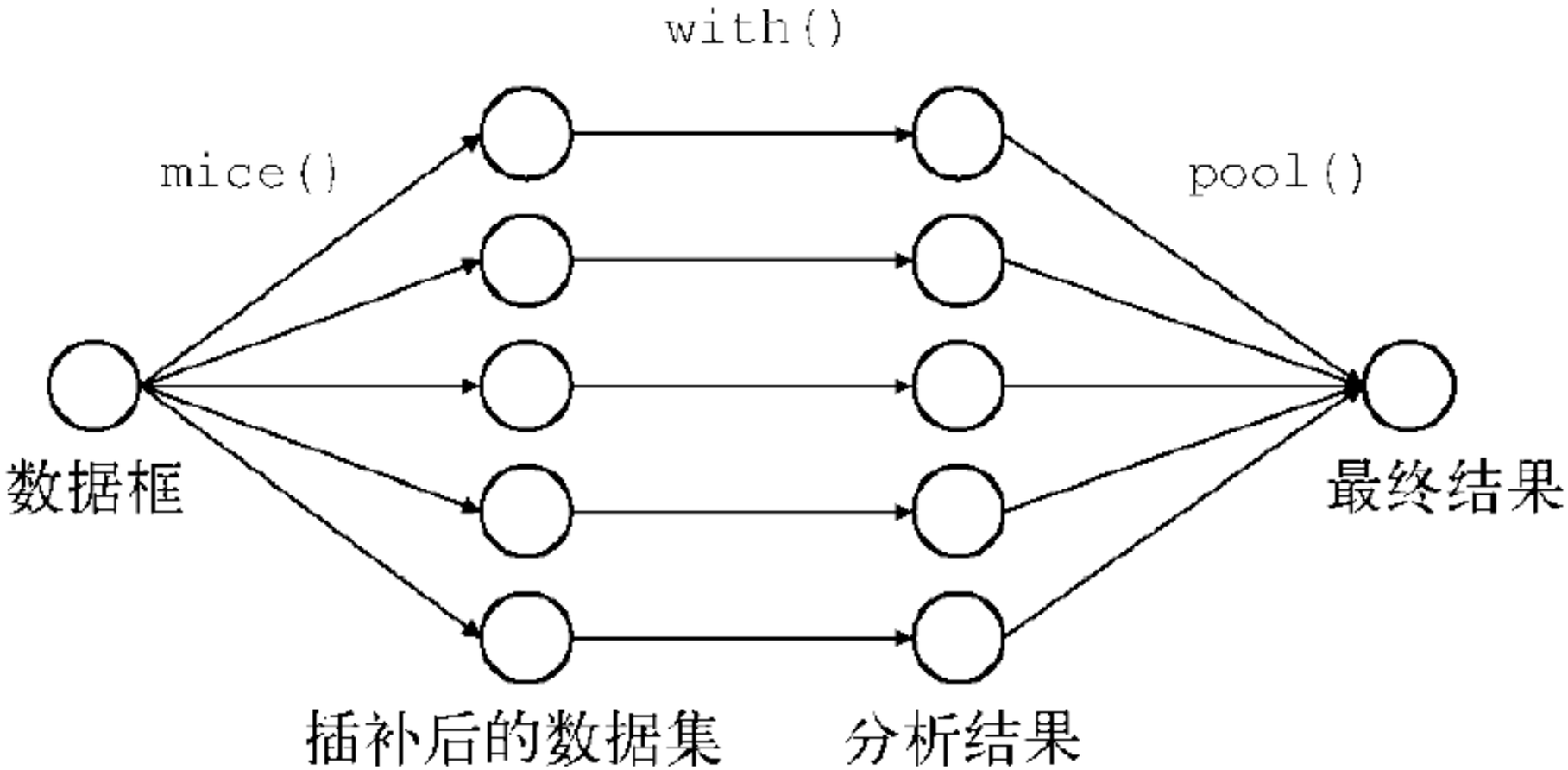


图18-5　通过mice包应用多重插补的步骤

完整行删除　多重插补　成对删除　简单插补

RCurl

Curl和RCurl

- curl：利用URL语法在命令行方式工作的开源文件传输工具
- libCurl：实现Curl的开源库，主要实现：获取页面、认证、上传、下载、信息搜索等

https://curl.haxx.se/

- RCurl：提供了R到libcurl的接口，从而实现了HTTP的一些功能。例如：服务器下载文件、保持连接、上传文件、采用二进制格式读取、句柄重定向、密码认证等

http://anson.ucdavis.edu/~duncan/

getURL()

```
> url.exists(url="www.baidu.com")
[1] TRUE
```

```
> d <- debugGatherer()
> tmp <- getURL(url="www.baidu.com", debugfunction = d$update, verbose = TRUE)
> names(d$value())
[1] "text"       "headerIn"   "headerOut"  "dataIn"     "dataOut"    "sslDataIn"
[7] "sslDataOut"
```

```
> cat(d$value()[1])
Rebuilt URL to: www.baidu.com/
  Trying 119.75.218.70...
Connected to www.baidu.com (119.75.218.70) port 80 (#0)
Connection #0 to host www.baidu.com left intact
```

getURL()

```
> cat(d$value()[2])
HTTP/1.1 200 OK
Date: Wed, 25 May 2016 02:58:38 GMT
Content-Type: text/html
Content-Length: 14613
Last-Modified: Wed, 03 Sep 2014 02:48:32 GMT
Connection: Keep-Alive
Vary: Accept-Encoding
Set-Cookie: BAIDUID=2CB1A15F480DC65312CC359A6DE5ADBC:FG=1; expires=Thu, 31-Dec-37 23:55:55 GMT
max-age=2147483647; path=/; domain=.baidu.com
Set-Cookie: BIDUPSID=2CB1A15F480DC65312CC359A6DE5ADBC; expires=Thu, 31-Dec-37 23:55:55 GMT; max
-age=2147483647; path=/; domain=.baidu.com
Set-Cookie: PSTM=1464145118; expires=Thu, 31-Dec-37 23:55:55 GMT; max-age=2147483647; path=/;
omain=.baidu.com
P3P: CP=" OTI DSP COR IVA OUR IND COM "
Server: BWS/1.1
X-UA-Compatible: IE=Edge,chrome=1
Pragma: no-cache
Cache-control: no-cache
Accept-Ranges: bytes
```

getURL()

```
> cat(d$value()[3])
GET / HTTP/1.1
Host: www.baidu.com
Accept: */*
```

```
> d$reset()
> d$value()
      text   headerIn  headerOut     dataIn    dataOut  sslDataIn sslDataOut
        ""         ""          ""         ""         ""         ""         ""
```

getURL()

```
> h <- basicTextGatherer()
> txt <- getURL("http://www.baidu.com", headerfunction = h$update)
> names(h$value())
NULL
> h$value()
[1] "HTTP/1.1 200 OK\r\nDate: Wed, 25 May 2016 03:35:42 GMT\r\nContent-Type: text/html\r\nContent-Length: 14613\r\
nLast-Modified: Wed, 03 Sep 2014 02:48:32 GMT\r\nConnection: Keep-Alive\r\nVary: Accept-Encoding\r\nSet-Cookie: BA
IDUID=E077EAC981524636A14600D379007DBF:FG=1; expires=Thu, 31-Dec-37 23:55:55 GMT; max-age=2147483647; path=/; doma
in=.baidu.com\r\nSet-Cookie: BIDUPSID=E077EAC981524636A14600D379007DBF; expires=Thu, 31-Dec-37 23:55:55 GMT; max-a
ge=2147483647; path=/; domain=.baidu.com\r\nSet-Cookie: PSTM=1464147342; expires=Thu, 31-Dec-37 23:55:55 GMT; max-
age=2147483647; path=/; domain=.baidu.com\r\nP3P: CP=\" OTI DSP COR IVA OUR IND COM \"\r\nServer: BWS/1.1\r\nX-UA-
Compatible: IE=Edge,chrome=1\r\nPragma: no-cache\r\nCache-control: no-cache\r\nAccept-Ranges: bytes\r\n\r\n"
```

```
> cat(h$value())
HTTP/1.1 200 OK
Date: Wed, 25 May 2016 03:35:42 GMT
Content-Type: text/html
Content-Length: 14613
Last-Modified: Wed, 03 Sep 2014 02:48:32 GMT
Connection: Keep-Alive
Vary: Accept-Encoding
Set-Cookie: BAIDUID=E077EAC981524636A14600D379007DBF:FG=1; expires=Thu, 31-Dec-37 23:55:55 GMT; ma
x-age=2147483647; path=/; domain=.baidu.com
Set-Cookie: BIDUPSID=E077EAC981524636A14600D379007DBF; expires=Thu, 31-Dec-37 23:55:55 GMT; max-ag
e=2147483647; path=/; domain=.baidu.com
Set-Cookie: PSTM=1464147342; expires=Thu, 31-Dec-37 23:55:55 GMT; max-age=2147483647; path=/; doma
in=.baidu.com
P3P: CP=" OTI DSP COR IVA OUR IND COM "
Server: BWS/1.1
X-UA-Compatible: IE=Edge,chrome=1
Pragma: no-cache
Cache-control: no-cache
Accept-Ranges: bytes
```

getURL()

```
> cat(h$value())
HTTP/1.1 200 OK
Date: Wed, 25 May 2016 03:35:42 GMT
Content-Type: text/html
Content-Length: 14613
Last-Modified: Wed, 03 Sep 2014 02:48:32 GMT
Connection: Keep-Alive
Vary: Accept-Encoding
Set-Cookie: BAIDUID=E077EAC981524636A14600D379007DBF:FG=1; expires=Thu, 31-Dec-37 23:55:55 GMT; ma
x-age=2147483647; path=/; domain=.baidu.com
Set-Cookie: BIDUPSID=E077EAC981524636A14600D379007DBF; expires=Thu, 31-Dec-37 23:55:55 GMT; max-ag
e=2147483647; path=/; domain=.baidu.com
Set-Cookie: PSTM=1464147342; expires=Thu, 31-Dec-37 23:55:55 GMT; max-age=2147483647; path=/; doma
in=.baidu.com
P3P: CP=" OTI DSP COR IVA OUR IND COM "
Server: BWS/1.1
X-UA-Compatible: IE=Edge,chrome=1
Pragma: no-cache
Cache-control: no-cache
Accept-Ranges: bytes
```

getURL()

```
> h <- basicHeaderGatherer()
> txt <- getURL(url="http://www.baidu.com", headerfunction = h$update)
> names(h$value())
 [1] "Date"            "Content-Type"     "Content-Length" "Last-Modified"  "Connection"
 [6] "Vary"            "Set-Cookie"       "Set-Cookie"     "Set-Cookie"     "P3P"
[11] "Server"          "X-UA-Compatible" "Pragma"          "Cache-control"  "Accept-Ranges"
[16] "status"          "statusMessage"
```

```
> h$value()
                                                                                              Date
                                                                   "Wed, 25 May 2016 03:34:03 GMT"
                                                                                      Content-Type
                                                                                       "text/html"
                                                                                    Content-Length
                                                                                           "14613"
                                                                                     Last-Modified
                                                                   "Wed, 03 Sep 2014 02:48:32 GMT"
                                                                                        Connection
                                                                                      "Keep-Alive"
                                                                                              Vary
                                                                                 "Accept-Encoding"
                                                                                        Set-Cookie
"BAIDUID=D7ACEDAEAA0295848111148DA18F535C:FG=1; expires=Thu, 31-Dec-37 23:55:55 GMT; max-age=2147483647; path=/; domain=.baidu.com"
                                                                                        Set-Cookie
    "BIDUPSID=D7ACEDAEAA0295848111148DA18F535C; expires=Thu, 31-Dec-37 23:55:55 GMT; max-age=2147483647; path=/; domain=.baidu.com"
                                                                                        Set-Cookie
                           "PSTM=1464147243; expires=Thu, 31-Dec-37 23:55:55 GMT; max-age=2147483647; path=/; domain=.baidu.com"
                                                                                               P3P
                                                              "CP=\" OTI DSP COR IVA OUR IND COM \""
                                                                                            Server
                                                                                        "BWS/1.1"
                                                                                   X-UA-Compatible
                                                                               "IE=Edge,chrome=1"
                                                                                            Pragma
                                                                                        "no-cache"
```

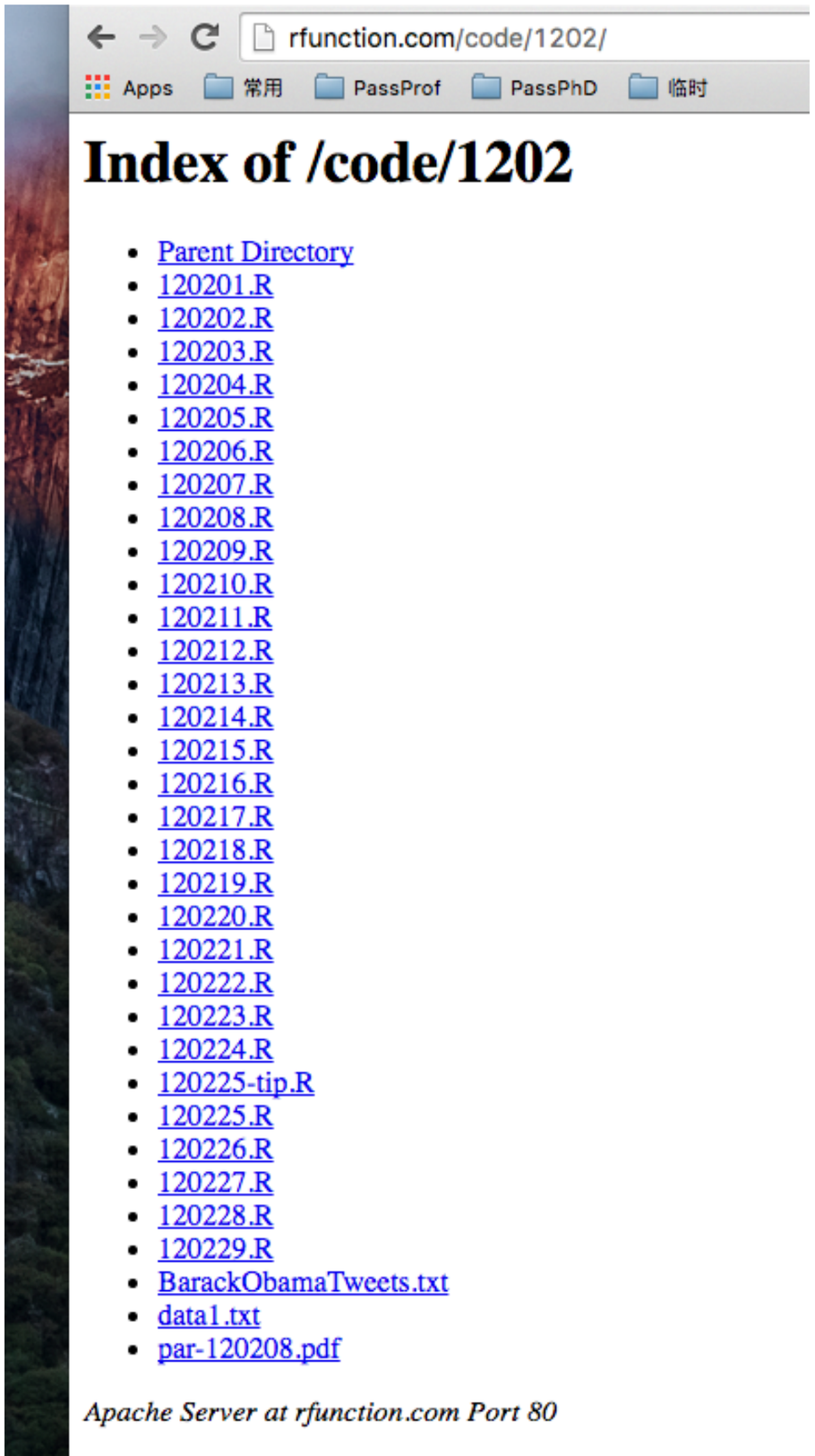
Curl参数

```
> listCurlOptions()
  [1] "address.scope"              "append"                  "autoreferer"
  [4] "buffersize"                 "cainfo"                  "capath"
  [7] "certinfo"                   "closepolicy"             "connect.only"
 [10] "connecttimeout"             "connecttimeout.ms"       "conv.from.network.function"
 [13] "conv.from.utf8.function"    "conv.to.network.function" "cookie"
 [16] "cookiefile"                 "cookiejar"               "cookielist"
 [19] "cookiesession"              "copypostfields"          "crlf"
 [22] "crlfile"                    "customrequest"           "debugdata"
 [25] "debugfunction"              "dirlistonly"             "dns.cache.timeout"
 [28] "dns.use.global.cache"       "egdsocket"               "encoding"
 [31] "errorbuffer"                "failonerror"             "file"
 [34] "filetime"                   "followlocation"          "forbid.reuse"
 [37] "fresh.connect"              "ftp.account"             "ftp.alternative.to.user"
 [40] "ftp.create.missing.dirs"    "ftp.filemethod"          "ftp.response.timeout"
 [43] "ftp.skip.pasv.ip"           "ftp.ssl"                 "ftp.ssl.ccc"
 [46] "ftp.use.eprt"               "ftp.use.epsv"            "ftpappend"
 [49] "ftplistonly"                "ftpport"                 "ftpsslauth"
 [52] "header"                     "headerdata"              "headerfunction"
 [55] "http.content.decoding"      "http.transfer.decoding"  "http.version"
 [58] "http200aliases"             "httpauth"                "httpget"
 [61] "httpheader"                 "httppost"                "httpproxytunnel"
 [64] "ignore.content.length"      "infilesize"              "infilesize.large"
 [67] "interface"                  "ipresolve"               "issuercert"
 [70] "keypasswd"                  "krb4level"               "krblevel"
 [73] "localport"                  "localportrange"          "low.speed.limit"
 [76] "low.speed.time"             "mail.auth"               "mail.from"
 [79] "mail.rcpt"                  "maxconnects"             "maxfilesize"
 [82] "maxfilesize.large"          "maxredirs"               "mute"
 [85] "netrc"                      "netrc.file"              "new.directory.perms"
 [88] "new.file.perms"             "nobody"                  "noprogress"
 [91] "nosignal"                   "password"                "port"
 [94] "post"                       "postfields"              "postfieldsize"
 [97] "postfieldsize.large"        "postquote"               "prequote"
[100] "private"                    "progressdata"            "progressfunction"
[103] "proxy"                      "proxy.transfer.mode"     "proxyauth"
[106] "proxyport"                  "proxytype"               "proxyuserpwd"
[109] "put"                        "quick.exit"              "quote"
[112] "random.file"                "range"                   "readdata"
[115] "readfunction"               "redir.protocols"         "referer"
[118] "resolve"                    "resume.from"             "resume.from.large"
[121] "rtsp.client.cseq"           "rtsp.request"            "rtsp.server.cseq"
[124] "rtsp.session.id"            "rtsp.stream.uri"         "rtsp.transport"
[127] "sasl.ir"                    "service.name"            "share"
[130] "shared"                     "skip.peer.verification"  "skip.proxy.verification"
[133] "socks5.gssapi.nec"          "socks5.gssapi.service"   "ssh.auth.types"
[136] "ssh.host.public.key.md5"    "ssh.knownhosts"          "ssh.private.keyfile"
[139] "ssh.public.keyfile"         "ssl.cipher.list"         "ssl.ctx.data"
[142] "ssl.ctx.function"           "ssl.options"             "ssl.sessionid.cache"
[145] "ssl.verifyhost"             "ssl.verifypeer"          "sslcert"
[148] "sslcerttype"                "sslengine"               "sslengine.default"
[151] "sslkey"                     "sslkeytype"              "sslversion"
[154] "stderr"                     "tcp.keepalive"           "tcp.keepidle"
[157] "tcp.keepintvl"              "tcp.nodelay"             "telnetoptions"
[160] "timeout"                    "timeout.ms"              "timevalue"
[163] "transfertext"               "unrestricted.auth"       "upload"
[166] "url"                        "use.ssl"                 "useragent"
[169] "username"                   "userpwd"                 "verbose"
[172] "writedata"                  "writefunction"           "writeheader"
[175] "writeinfo"
```

其余

- getForm()
- postForm()

批量下载




```
html=getURL("http://rfunction.com/code/1202/")
temp =strsplit(html, "<li><a href=\"")[[1]]
files =strsplit(temp, "\"")
files=lapply(files, function(x){ x[1] })
files= unlist(files)

files = files[-(1:2)]
```

```
base="http://rfunction.com/code/1202/"
for(i in 1:length(files)){
  URL = paste(base, files[i], sep="")


  bin <- getBinaryURL(URL)
  con <- file(paste("1202", files[i], sep="."), open = "wb")
  writeBin(bin, con)
  close(con)
  Sys.sleep(2)
}
```

使用XML

Mus musculus

No.	Ensembl ID	Gene ID	Symbol	Family
1	ENSMUSG00000029313	17355	Aff1	AF-4
2	ENSMUSG00000031189	14266	Aff2	AF-4
3	ENSMUSG00000037138	16764	Aff3	AF-4
4	ENSMUSG00000049470	93736	Aff4	AF-4
5	ENSMUSG00000046532	11835	Ar	Androgen re
6	ENSMUSG00000021359	21418	Tcfap2a	AP-2
7	ENSMUSG00000025927	21419	Tcfap2b	AP-2
8	ENSMUSG00000028640	21420	Tcfap2c	AP-2
9	ENSMUSG00000042477	332937	Tcfap2e	AP-2
10	ENSMUSG00000042596	226896	Tcfap2d	AP-2
11	ENSMUSG00000004661	56380	Arid3b	ARID

```
> head(tables)
  No.         Ensembl ID Gene ID  Symbol              Family
1   1 ENSMUSG00000029313   17355    Aff1               AF-4
2   2 ENSMUSG00000031189   14266    Aff2               AF-4
3   3 ENSMUSG00000037138   16764    Aff3               AF-4
4   4 ENSMUSG00000049470   93736    Aff4               AF-4
5   5 ENSMUSG00000046532   11835      Ar Androgen receptor
6   6 ENSMUSG00000021359   21418 Tcfap2a               AP-2
```

```
url="http://www.bioguo.org/AnimalTFDB/BrowseAllTF.php?spe=Mus_musculus"
wp <- getURL(url)
doc  <- htmlParse(wp, asText= TRUE)
tables  <- readHTMLTable(doc,which=5)
```

其余

- 使用正则表达式来匹配文本

  https://en.wikipedia.org/wiki/Regular_expression

- 其余包
  - Rvest
  - httr

HTTP: https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol

练习 - 0034

- **天眼查**每个企业均有详细信息和企业评分
- 写一个爬虫，下载1000个企业信息
- 根据企业信息项，分析预测天眼查现有企业评分的计算方法，用下载的数据进行回归分析，再爬一些企业作为测试数据，分析方法的有效性
- 自己设计一个企业评分算法，说明设计原理，并用数据检验

企业最好针对一个行业

法定代表人	季 季昊 任职 8 家企业，分布如下 北京（共8家） 北京国际大数据交易… 等	经营状态	存续	天眼评分	评分 68
		成立日期	2020-04-29		
		注册资本	625万人民币		
		实缴资本	-	工商注册号	110108028700442
统一社会信用代码	91110108MA01R2FT36	纳税人识别号	91110108MA01R2FT36	组织机构代码	MA01R2FT3
营业期限	2020-04-29 至 2070-04-28	纳税人资质	-	核准日期	2021-01-06
公司类型	其他有限责任公司	行业	软件和信息技术服务业	人员规模	-
保人数	-	登记机关	北京市海淀区市场监督管理局		
名	-	英文名称	-		
地址	北京市海淀区丹棱街1号院1号楼26层2601室 附近公司				

下周上课前提交方式和以前一样

谢谢！

孙惠平

sunhp@ss.pku.edu.cn