Data Analysis Tools and Practice(Using R)

2018.05.09

统计II

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

课堂测试时间

课堂测试09

先用电脑完成
*40*分钟　然后誊抄纸上

- 1、某公司想要了解消费者购买牙膏时更追求什么样的目标,于是通过商场拦访对30个人进行访谈，用7级里克特量表询问他们对以下陈述的认同程度(即1表示非常不同意，7表示非常同意,V1:购买预防蛀牙的牙膏是重要的;V2:我喜欢使牙齿亮泽的牙膏；v3:牙膏应当保护牙龈；V4:我喜欢使口气清新的牙膏；V5:预防坏牙不是牙膏提供的一项重要功效；V6:购买牙膏时最重要的考虑是富有魅力的牙齿:
  - ✳ 将调查样本存储于文本文档 yagao.txt。请使用R函数factanal对数据进行因子分析，根据载荷系数矩阵，写出因子和原变量之间的线性关系式。
- 2、某地区农业生态经济系统的各区域单元相关指标数据在文本文件agriculture.txt中，使用R中的主成分分析的函数princomp选取更少的指标来描述该地区的农业生态经济系统。写出主成分和原变量之间的线性关系式。

课堂测试09

先用电脑完成
*40*分钟　然后誊抄纸上

- 3、下表是一个一个村庄儿童年龄和平均身高的统计数据
  - ✳ （1）画出平均身高height和年龄age关系的散点图
  - ✳ （2）建立回归模型并提取结果输出，在(1)中的图中表示生成的模型

平均身高与年龄			
年龄（月）	平均身高（厘米）	年龄（月）	平均身高（厘米）
18	76.1	24	79.9
19	77	25	81.1
20	78.1	26	81.2
21	78.2	27	81.8
22	78.8	28	82.8
23	79.7	29	83.5

- 4、 revenue.txt中记录了财政收入(y)和第一产业GDP X_1、第二产业GDP X_2、第三产业GDPx_3、人口数X_4、社会消费品零售总额X_5 、受灾面积X_6 、等情况的统计数据。要求:写出多元线性回归模型。

上次课程内容回顾

- 基本统计
- 因子和主成分分析

★ cor(), factanal(), princomp(), screeplot(), biplot(), predict(),…

- 回归分析

★ lm(), fitted(), residuals(), scatterplot()…

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

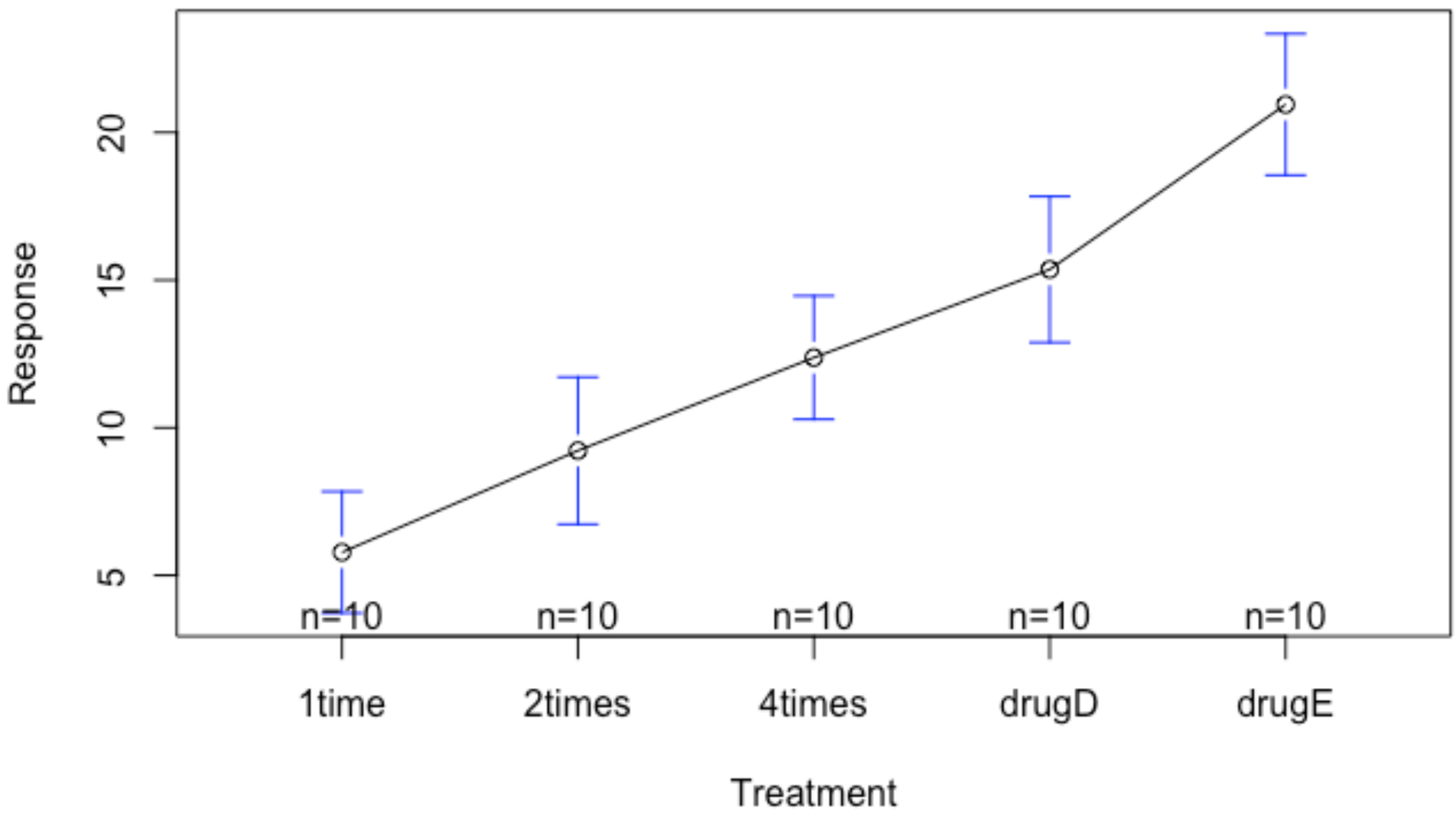
表 7.3: 单因素方差分析表

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

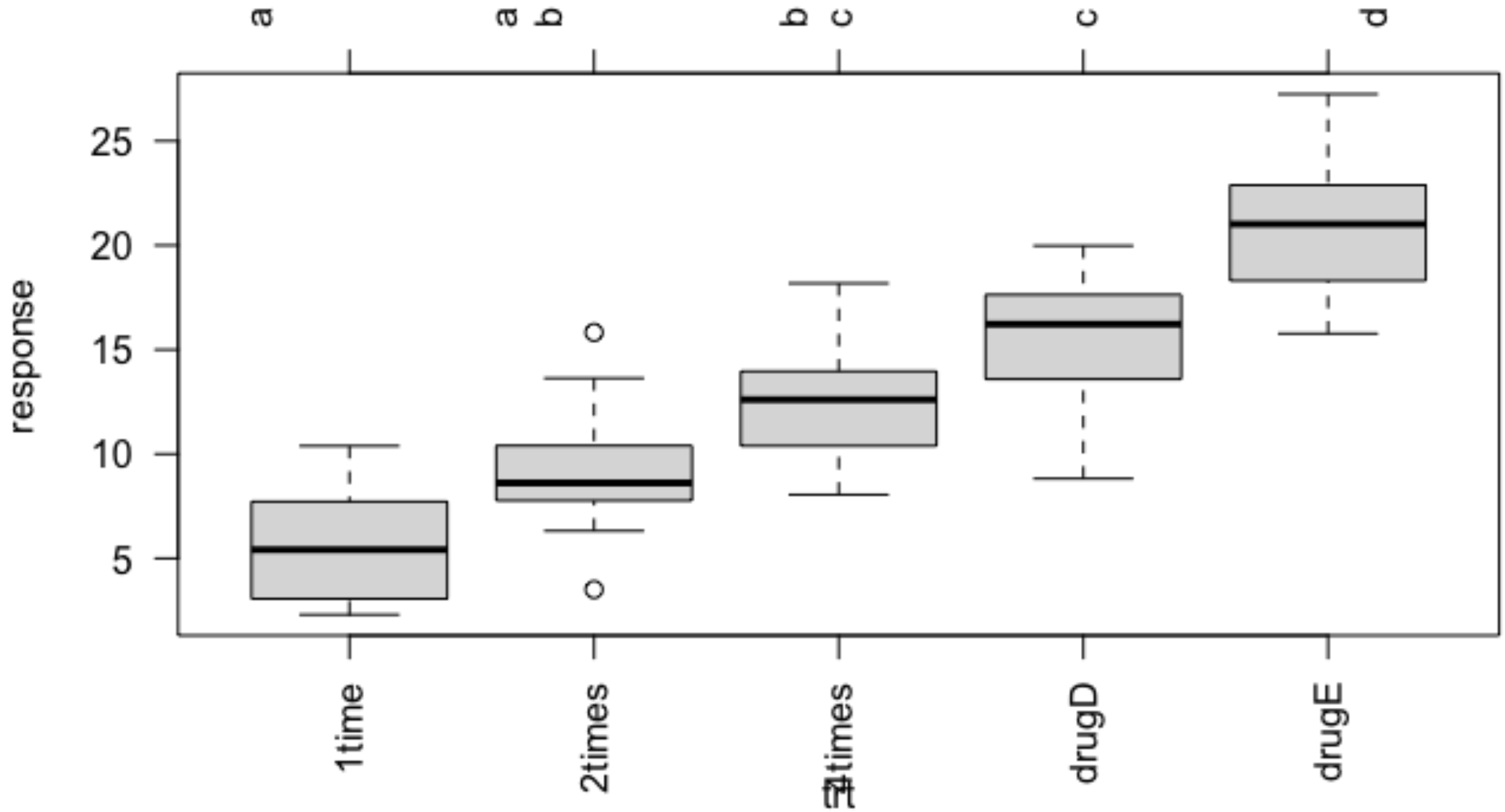
```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

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)
```

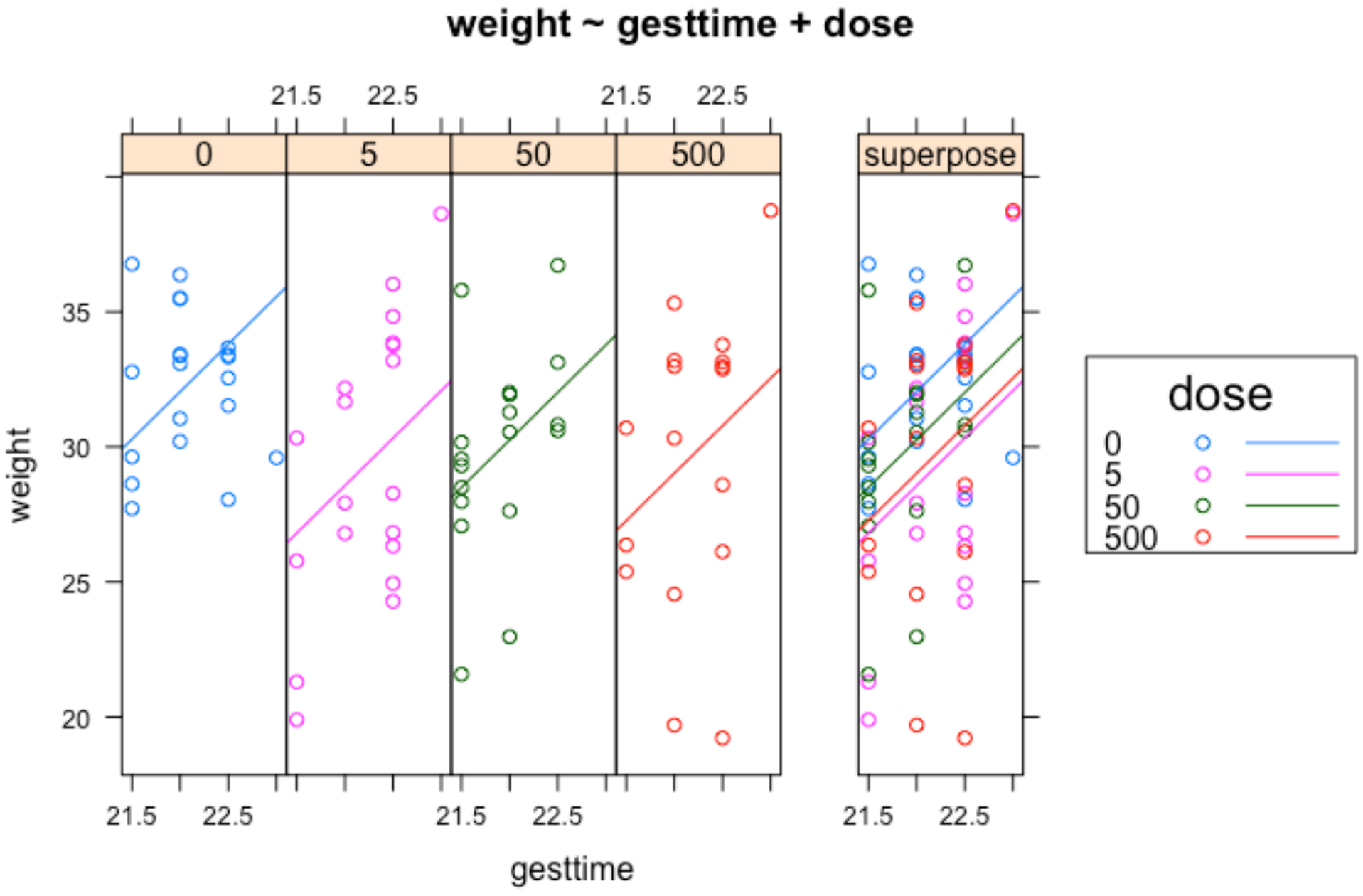
单因素协方差分析

```
> 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

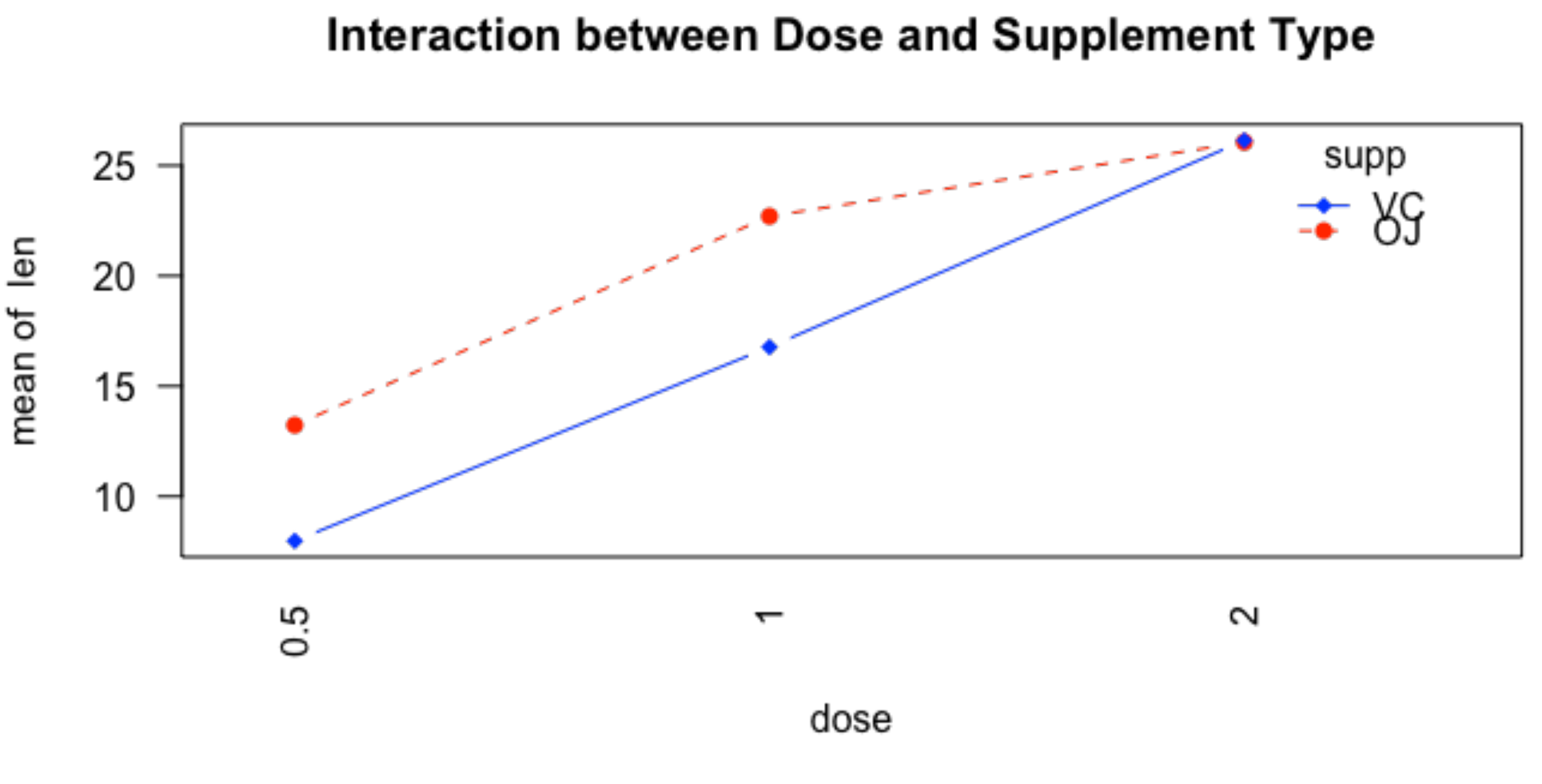
表 7.12: 双因素方差分析表

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

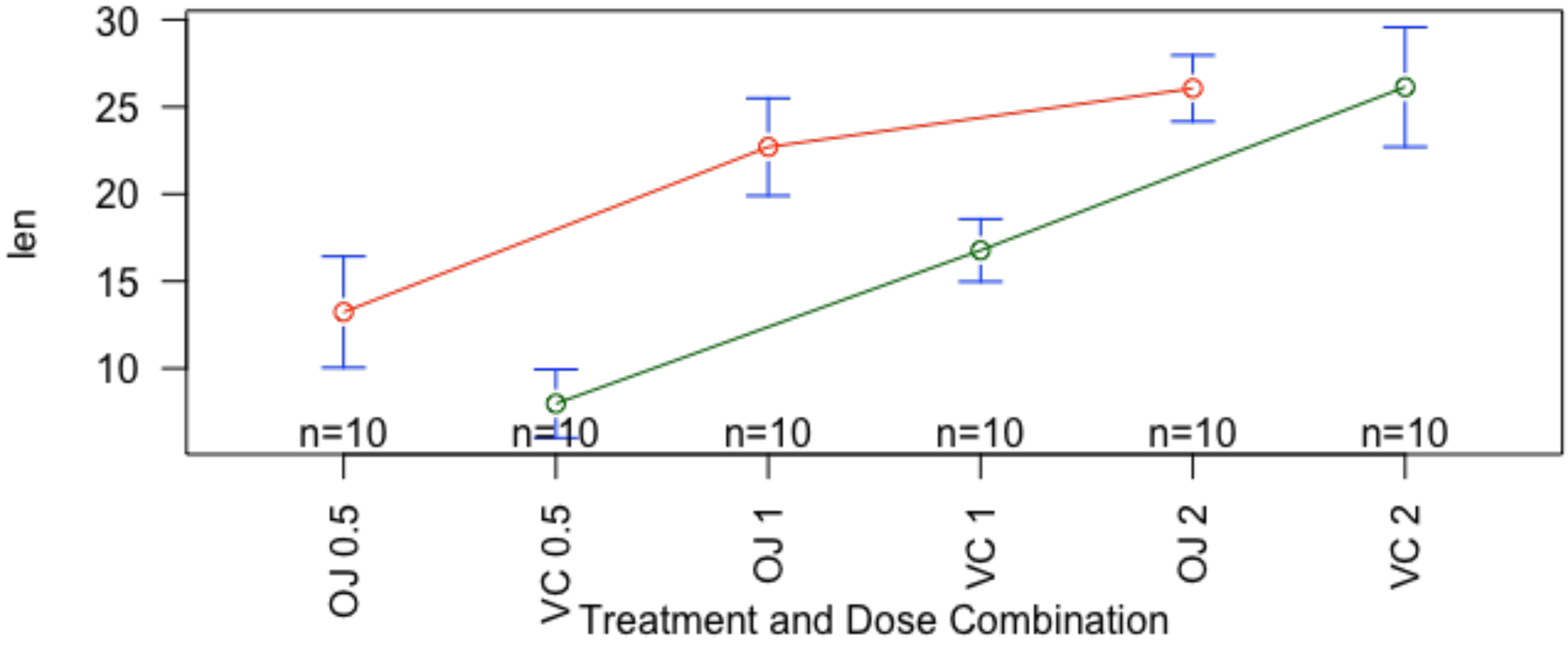
```
interaction.plot(dose, supp, len,
                 type = "b", col = c("red", "blue"), pch = c(16, 18),
                 main = "Interaction between Dose and Supplement Type")
```



双因素方差分析例子

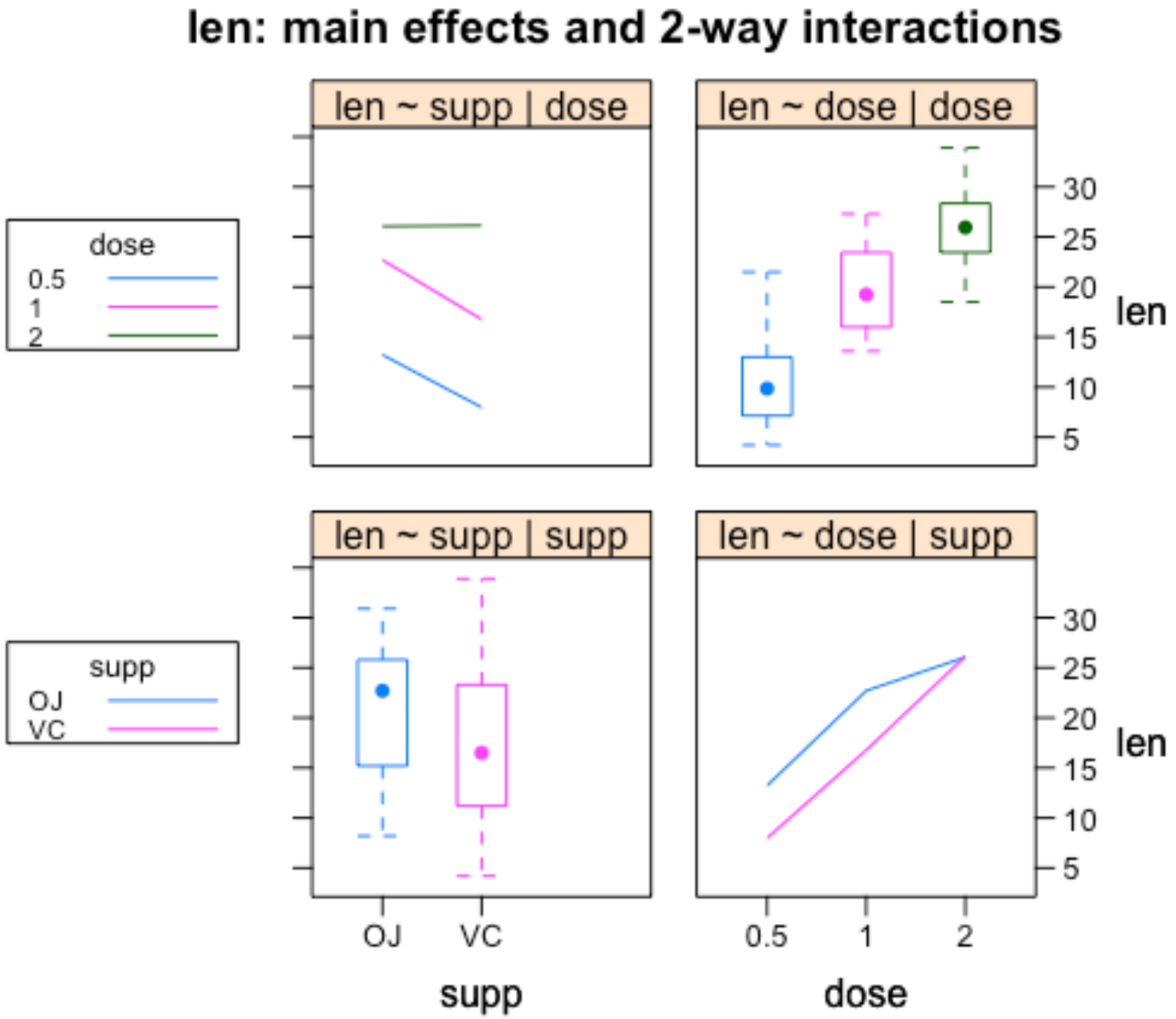
```
library(gplots)
plotmeans(len ~ interaction(supp, dose, sep = " "),
          connect = list(c(1, 3, 5), c(2, 4, 6)), col = c("red", "darkgreen"),
          main = "Interaction Plot with 95% CIs",
          xlab = "Treatment and Dose Combination")
```

交互效应



双因素方差分析例子

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



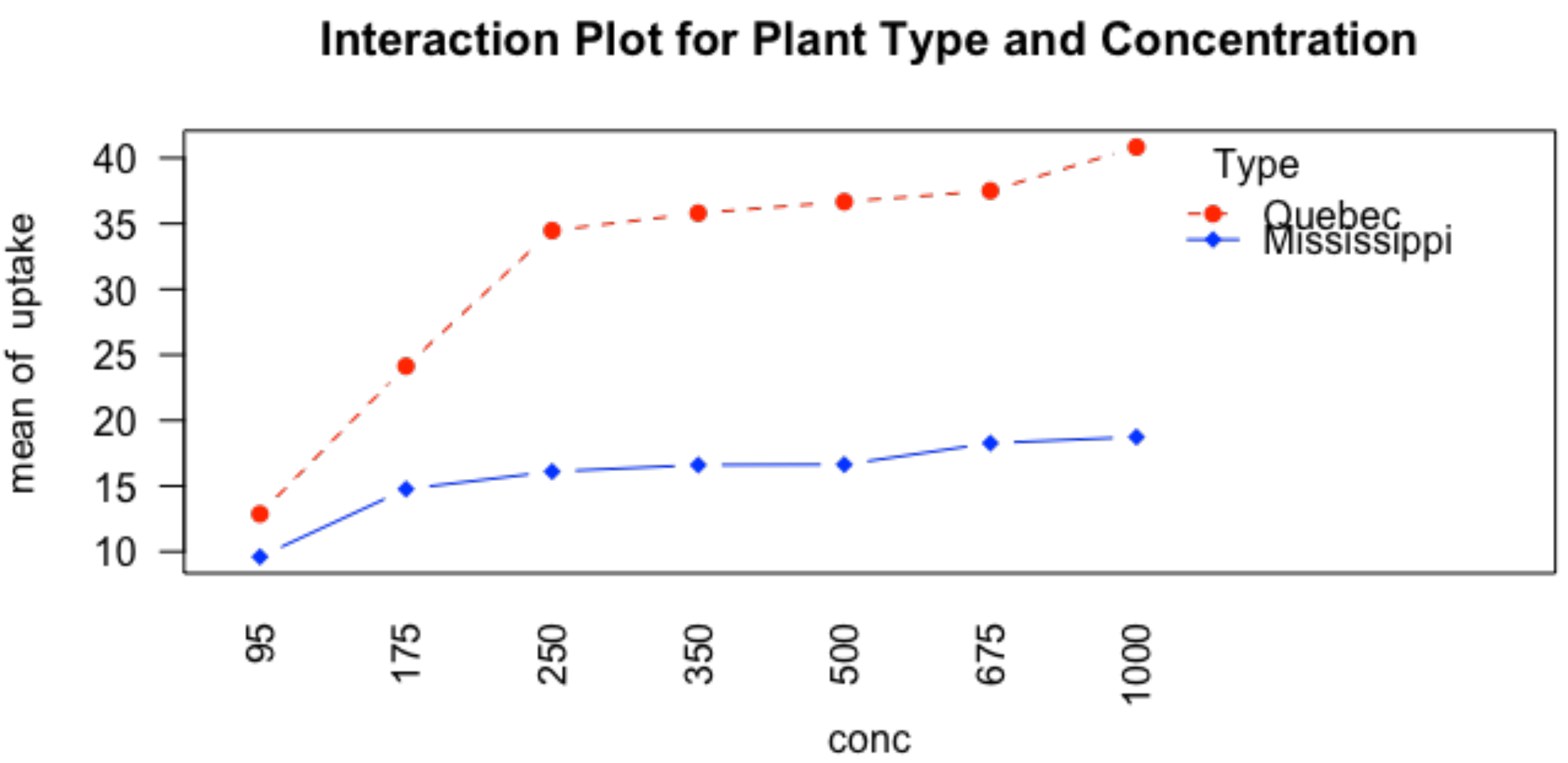
重复测量方差分析

```
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
```

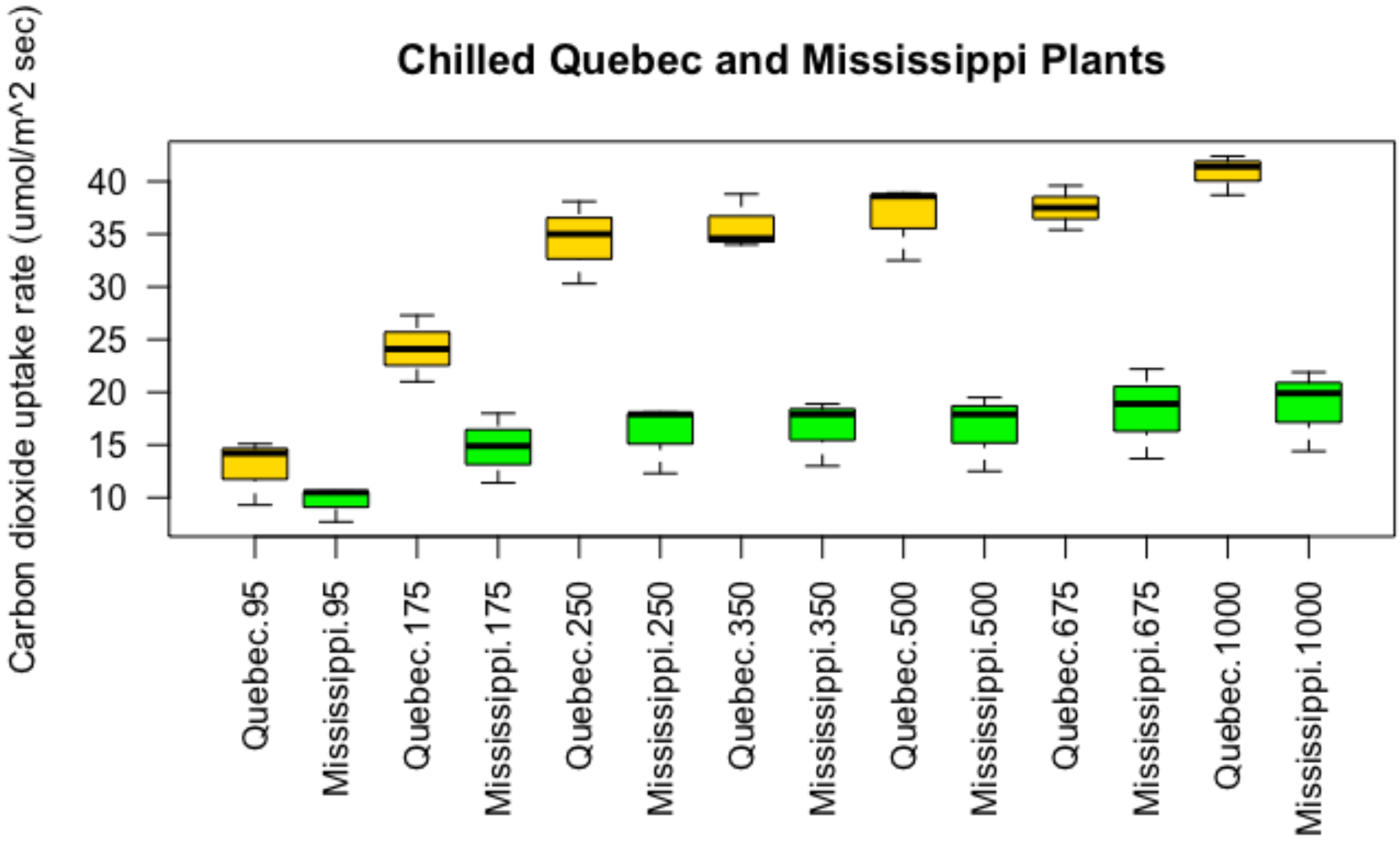
重复测量方差分析

```
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"))
```



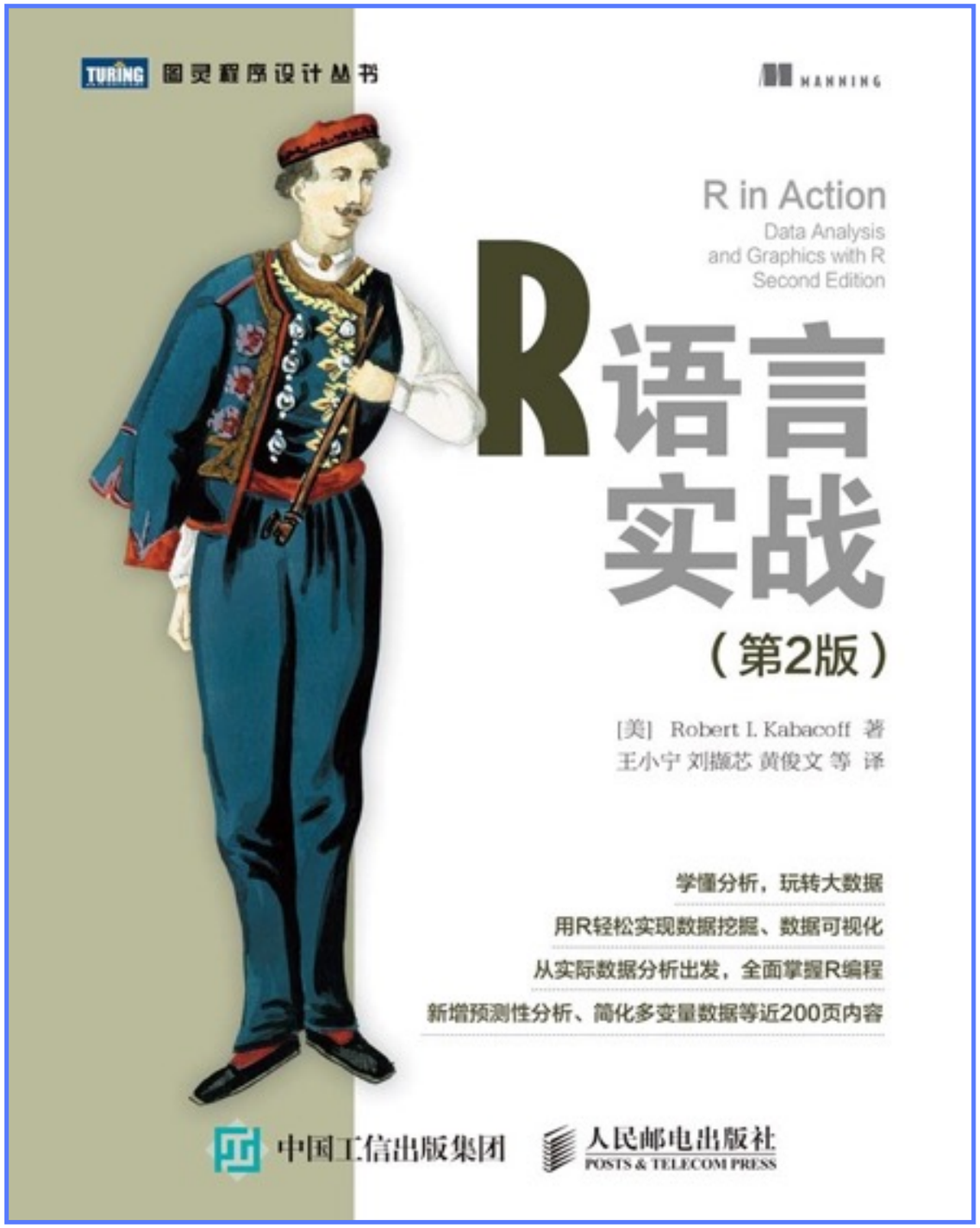
重复测量方差分析例子

```
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)")
```



练习

练习 - 0040


TURING 图灵程序设计丛书
MANNING
R in Action
Data Analysis
and Graphics with R
Second Edition
R语言实战
（第2版）
[美] Robert I. Kabacoff 著
王小宁 刘撷芯 黄俊文 等 译
学懂分析，玩转大数据
用R轻松实现数据挖掘、数据可视化
从实际数据分析出发，全面掌握R编程
新增预测性分析、简化多变量数据等近200页内容
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

第9章

练习 - 0041

例 7.1 利用四种不同配方的材料 A_1 、 A_2 、 A_3 、 A_4 生产出来的元件，测得其使用寿命如表 7.1 所示. 问：四种不同配方下元件的使用寿命有无显著的差异？

表 7.1: 元件寿命数据

材料	使用寿命							
A_1	1600	1610	1650	1680	1700	1700	1780	
A_2	1500	1640	1400	1700	1750			
A_3	1640	1550	1600	1620	1640	1600	1740	1800
A_4	1510	1520	1530	1570	1640	1600		

练习 - 0042

例 7.9 在一个农业试验中，考虑四种不同的种子品种 A_1, A_2, A_3, A_4 和三种不同的施肥方法 B_1, B_2, B_3 得到产量数据如表 7.10 所示 (单位： kg). 试分析种子与施肥对产量有无显著影响?

表 7.10: 农业试验数据

	B_1	B_2	B_3
A_1	325	292	316
A_2	317	310	318
A_3	310	320	318
A_4	330	370	365

练习 - 0043

例 7.11 研究树种与地理位置对松树生长的影响，对四个地区的三种同龄松树的直径进行测量得到数据如下表 7.15 所示 (单位： cm). A_1, A_2, A_3 表示三个不同树种， B_1, B_2, B_3, B_4 表示四个不同地区. 对每一种水平组合，进行了 5 次测量，对此试验结果进行方差分析.

表 7.15: 三种同龄松树的直径测量数据

	B_1	B_2	B_3	B_4
A_1	23 25 21 14 15	20 17 11 26 21	16 19 13 16 24	20 21 18 27 24
A_2	28 30 19 17 22	26 24 21 25 26	19 18 19 20 25	26 26 28 29 23
A_3	18 15 23 18 10	21 25 12 12 22	19 23 22 14 13	22 13 12 22 19

练习 - 0044

7.1 三个工厂生产同一种零件. 现从各厂产品中分别抽取 *4* 件产品作检测, 其检测强度如表 *7.25* 所示.

表 7.25: 产品检测数据

工厂	零 件 强 度			
甲	115	116	98	83
乙	103	107	118	116
丙	73	89	85	97

(1) 对数据作方差分析, 判断三个厂生产的产品的零件强度是否有显著差异;

(2) 求每个工厂生产产品零件强度的均值, 作出相应的区间估计 ($\alpha = 0.05$);

(3) 对数据作多重检验。

练习 - 0045

7.5 为研究人们在催眠状态下对各种情绪的反应国是否有差异，选取了 8 个受试者. 在催眠状态下，要求每人按任意次序做出恐惧、愉快、忧虑和平静 4 种反应. 表 7.29 给出了各受试者在处于这 4 种情绪状态下皮肤的电位变化值. 试在 $\alpha = 0.05$ 下，检验受试者在催眠状态下对这 4 种情绪的反应力是否有显著差异.

表 7.29: 4 种情绪状态下皮肤的电位变化值 (单位： mV)

情绪状态	受 试 者							
	1	2	3	4	5	6	7	8
恐惧	23.1	57.6	10.5	23.6	11.9	54.6	21.0	20.3
愉快	22.7	53.2	9.7	19.6	13.8	47.1	13.6	23.6
忧虑	22.5	53.7	10.8	21.1	13.7	39.2	13.7	16.3
平静	22.6	53.1	8.3	21.6	13.3	37.0	14.8	14.8

大作业

大作业 - 0004

- 完成课后大作业0004
- 提交rmd文档

- 完成练习0041-0045
- 提交rmd文档

确定分组和包

谢谢！

孙惠平

sunhp@ss.pku.edu.cn