多层线性模型/混合线性模型工作坊

——R语言中的实现

张光耀
（中科院心理所 李兴珊课题组）
2019年11月5日

R语言中的实现

- 数据包及函数
- 数据基本操作
- 模型优化
- 主效应与固定效应
- 简单效应分析
- 因子对比方式与模型回归系数
- planned contrasts

个人简介

- 研究方向：
  - 工作记忆与语义加工；眼动的脑机制；
- GitHub：
  - https://github.com/usplos
  - Eye-movement-related
    - Shiny dashboard
  - DPEEM
- 知乎：
  - https://www.zhihu.com/people/Psych.ZhangGuangyao/
  - 「中国R语言社区」

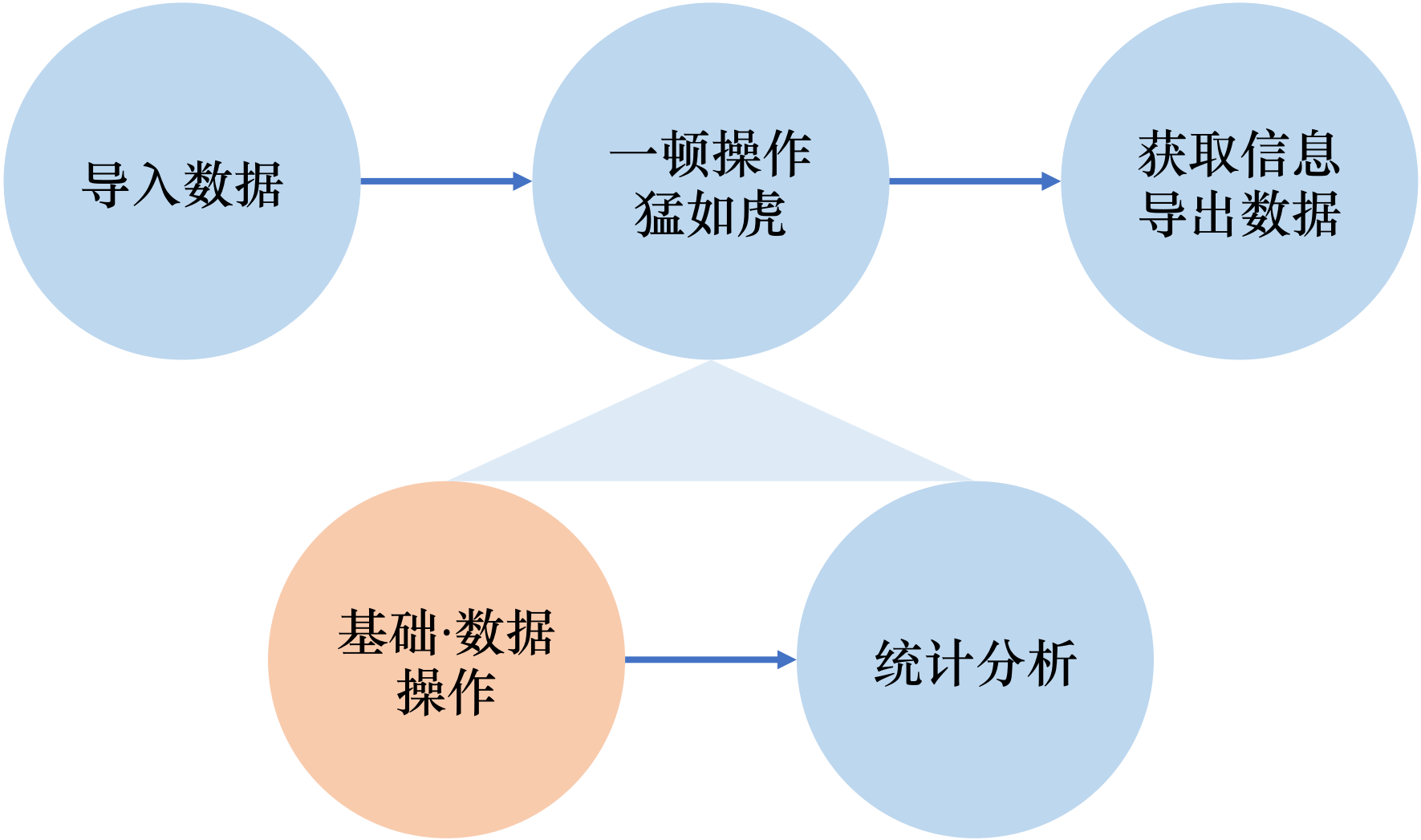
数据包及函数

- lmerTest package　　强烈推荐
- lme4 package → 不能输出回归系数的 p 值

```
lmer(data = Data,
    formula = DV ~ IV + (1 + RandomSlope | Cluster),
    contrasts = ...,
    control = ...,
    ...)
```

```
glmer(data = , formula = ,  contrasts = , family = , ...)
```

Wait！！Let's start from the very basic data operations.


导入数据
一顿操作
猛如虎
获取信息
导出数据
基础·数据
操作
统计分析

数据基本操作——数据结构

	创建	示例
原子向量	c( )	Vector = c(2,3,4)
矩阵	matrix()	Matrix = matrix(data = 1:8, nrow = 2, ncol = 4, byrow = F)
数组	array()	Array = array(data = 1:8, dim = c(2,2,2))
数据框	data.frame()	DataFrame = data.frame(Name = c('Bob','Tom'), Score = c(80,90))
列表	list()	List = list(DataFrame, Vector)

数据基本操作——数据类型

	创建	强制转变	判断
数值型	double()	as.double()	is.double()
整形	integer()	as.integer()	is.integer()
逻辑型	logical()	as.logical()	is.logical()
因子型	factor()	as.factor()	is.factor()
字符型	character()	as.character()	is.character()
矩阵	matrix()	as.matrix()	is.matrix()
数据框	data.frame()	as.data.frame()	is.data.frame()
列表	list()	as.list()	is.list()
		……	

数据基本操作——数据导入与导出

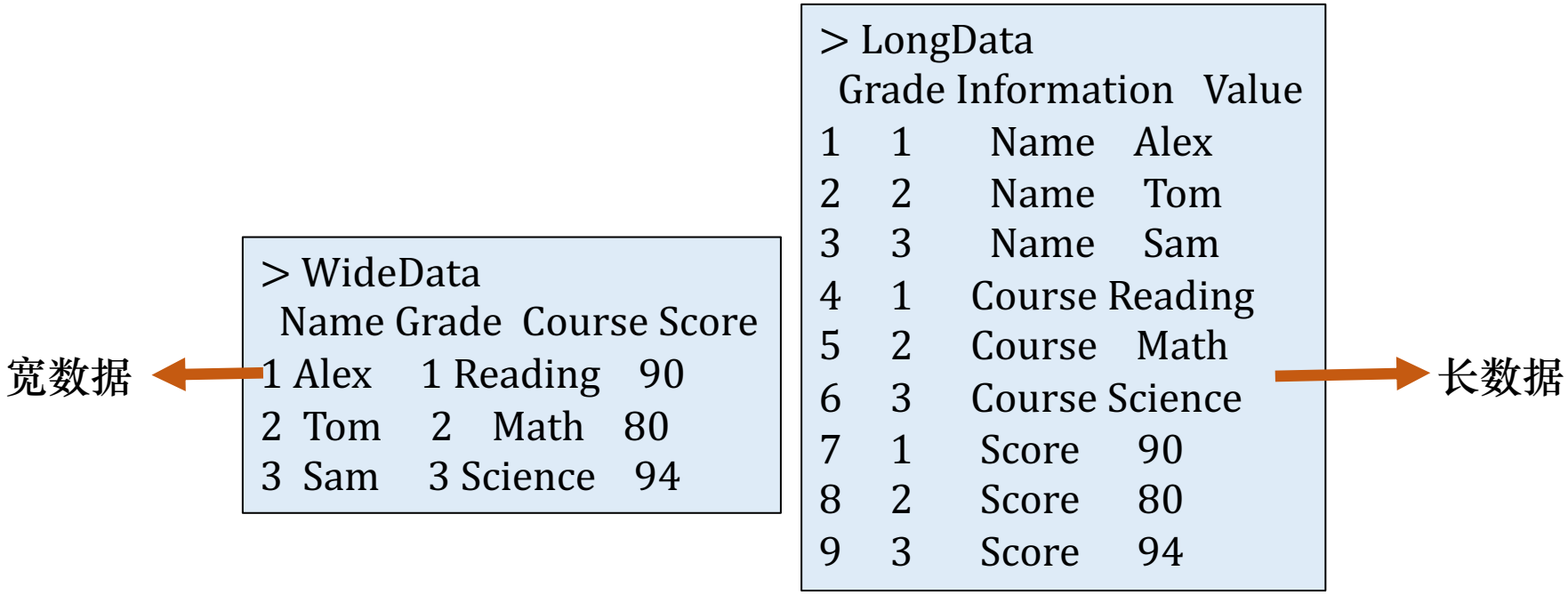
rio package

- 数据导入
  - import('Data.csv')
  - import('Data.xlsx')
  - import('Data..sav')
  - import('Data.dat')
  - import('Data.txt')
- 数据导出
  - export('Data.csv')
  - export('Data.xlsx')
  - export('Data..sav')
  - export('Data.dat')

other packages

- 数据导入
  - readr::read_csv('Data.csv')
  - readr::read_csv('*URL*')
  - readxl::read_excel('Data.xlsx')
  - haven::read_dta('Data.dta')
  - haven::read_sav('Data.sav')
- 数据导出
  - readr::write_csv('Data.csv')
  - haven::write_sav('Data.sav')
  - haven::write_dta('Data.dta')

数据基本操作——长宽数据转换



数据基本操作——长宽数据转换

- 宽变长 tidyr::gather(data = , key = , value = , ...)

tidyr package

```
> gather(data = WideData, key = Information, value = Value, -Grade)
  Grade Information   Value
1     1        Name    Alex
2     2        Name     Tom
3     3        Name     Sam
4     1      Course Reading
5     2      Course    Math
6     3      Course Science
.....
```

数据基本操作——长宽数据转换

- 宽变长 tidyr::gather(data = , key = , value = , ...)
- 长变宽 tidyr::spread(data = , key = , value = , ...)

```
> spread(data = LongData, key = Information, value = Value)
  Grade  Course Name Score
1     1 Reading Alex    90
2     2    Math  Tom    80
3     3 Science  Sam    94
```

- 其他函数：data.table::melt()

数据基本操作——数据框操作1

创建新变量

```
> mutate(WideData, Gender = c('M','M','M'))
  Name    Grade  Course      Score      Gender
1 Alex    1      Reading     90         M
2  Tom    2      Math        80         M
3  Sam    3      Science     94         M
```

dplyr package

筛选变量

```
> select(WideData, c(Name,Score))
  Name Score
1 Alex   90
2  Tom   80
3  Sam   94
```

筛选数据

```
> filter(WideData, Score>=90)
  Name    Grade  Course      Score
1 Alex    1      Reading     90
2  Sam    3      Science     94
```

数据基本操作——数据框操作2

数据排序

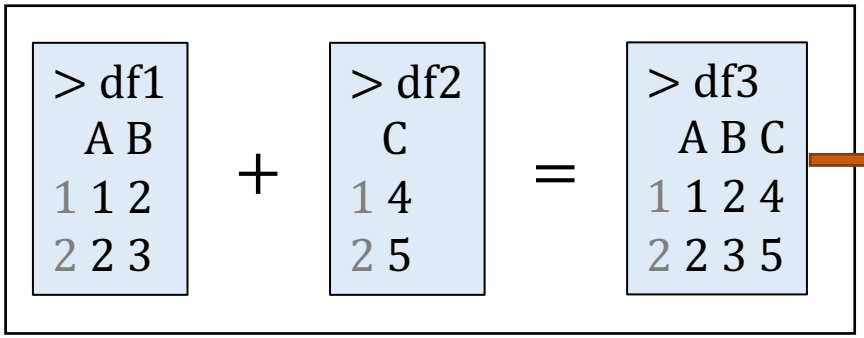
```
> arrange(WideData, Score)
  Name Grade  Course Score
1  Tom     2    Math    80
2 Alex     1 Reading    90
3  Sam     3 Science    94
> arrange(WideData, -Score)
  Name Grade  Course Score
1  Sam     3 Science    94
2 Alex     1 Reading    90
3  Tom     2    Math    80
```

变量重命名

```
> rename(WideData,
                Subject = Course,
                Value = Score)
  Name Grade Subject Value
1 Alex     1 Reading    90
2  Tom     2    Math    80
3  Sam     3 Science    94
```

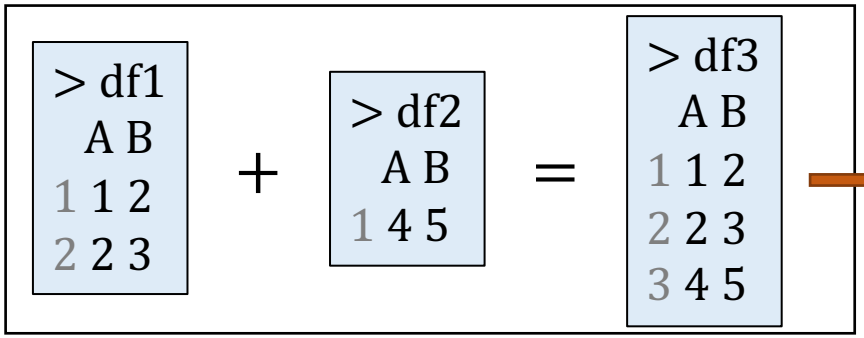
数据基本操作——数据合并.1

dplyr package



按行合并/合并列

dplyr::bin_col( )

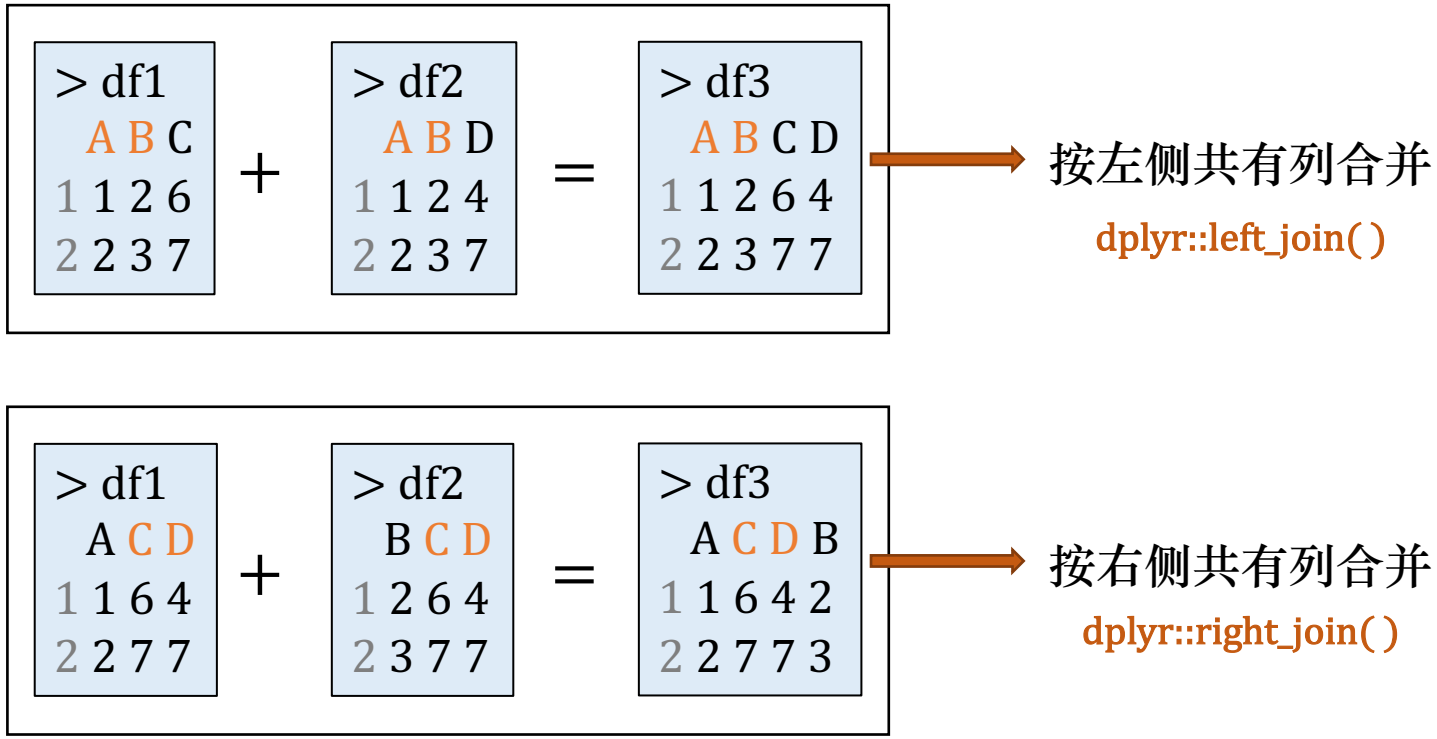


按列合并/合并行

dplyr::bin_row( )

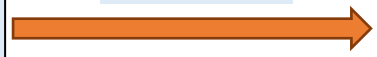
数据基本操作——数据合并.2

dplyr package



数据基本操作——变量中心化

```
> Data
[1] 12  9 14  6 11 12  7  3
> mean(Data)
[1] 9.25
> sd(Data)
[1] 3.69
```




```
> scale(Data)
       [,1]
[1,]  0.7445
[2,] -0.0677
[3,]  1.2860
[4,] -0.8799
[5,]  0.4738
[6,]  0.7445
[7,] -0.6092
[8,] -1.6921
attr(,"scaled:center")
[1] 9.25
attr(,"scaled:scale")
[1] 3.69
```

数据基本操作——分组中心化

```
> df
  Score Class
1    93     A
2    97     A
3    97     A
4    92     A
5    97     A
6   100     A
7    96     B
8    85     B
9   100     B
10   87     B
11   80     B
12  100     B
```

dplyr package

```
% >%
group_by( )
mutate( )
```

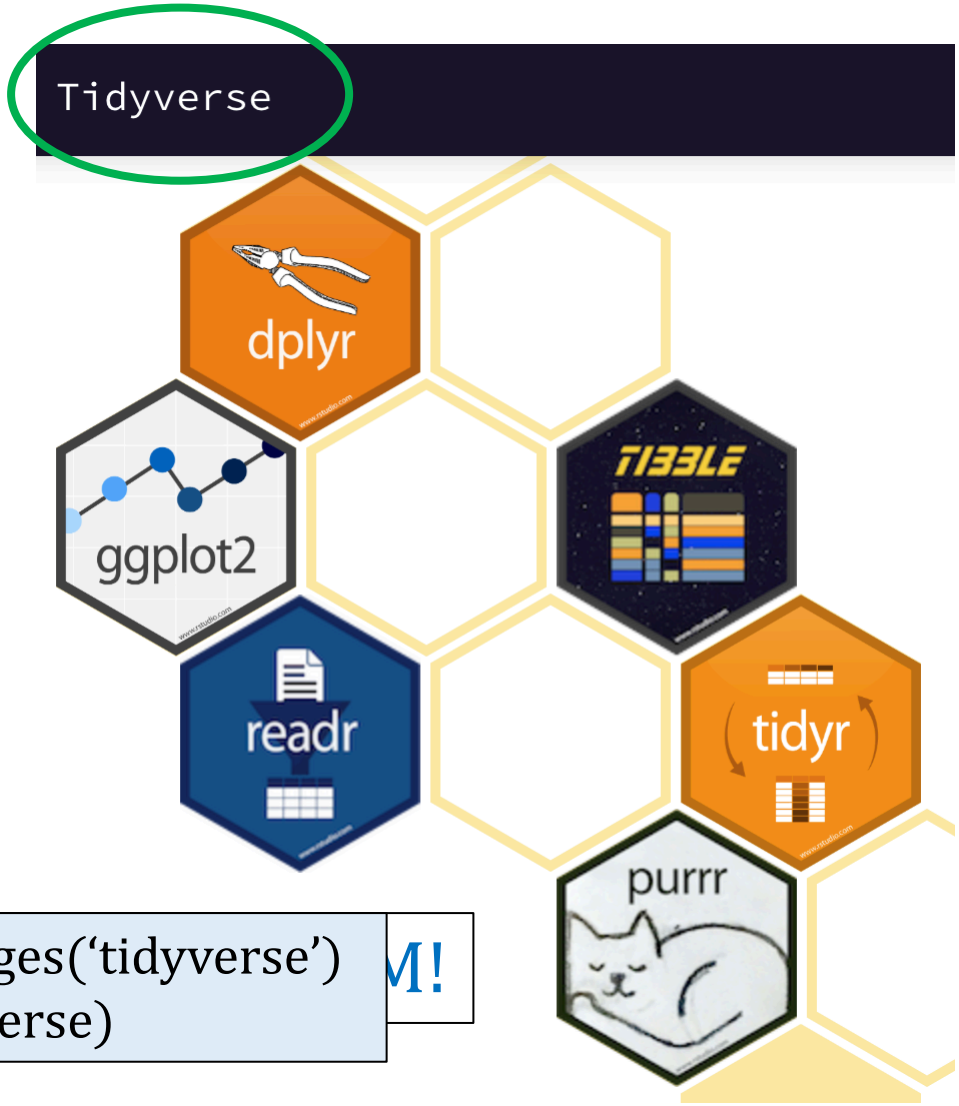
```
> df %>%
      group_by(Class) %>%
      mutate(ScoreNew = scale(Score))
```

Score	Class	ScoreNew
<int>	<fct>	<dbl>
1 93	A	-1.01
2 97	A	0.337
3 97	A	0.337
4 92	A	-1.35
5 97	A	0.337
6 100	A	1.35
7 96	B	0.550
8 85	B	-0.747
9 100	B	1.02
10 87	B	-0.511
11 80	B	-1.34
12 100	B	1.02

数据基本操作——小结

- 数据结构与数据类别
  - 有哪几种？如何创建？如何转换？如何判断？
- 读入和导出数据
  - 从文件夹读入数据
  - 从网页读入数据
  - 将R的工作空间的数据导出到文件夹
- 数据框操作
  - 数据框内：创建变量，筛选变量，筛选数据，数据排序，变量重命名
  - 数据框间：按列合并，按行合并，按左/右侧共有列合并
- 数据中心化
  - 分组&不分组
- 建议利用某些数据包来快速整理数据

数据基本操作——小结



数据整理

数据可视化

数据导入输出

数据框

长宽数据

泛函与映射

```
> install.packages('tidyverse')
> library(tidyverse)
```

从读数据开始

```
> library(tidyverse) # 会用到读取数据的函数
> library(lmerTest) # 加载 lmerTest 包
```

```
> Data = read_csv('https://raw.githubusercontent.com/usplos/Eye-movement-related/master/DataforShiny.csv')
```

```
> head(Data,3)
# A tibble: 3 x 5
   Sub A    B     Item     Y
  <dbl> <chr> <chr> <dbl> <dbl>
1    1 A2    B2        1   254
2    1 A2    B2        1   341
3    1 A1    B1        2   189
```

两因素被试内设计

改变变量类别

```
> Data[c('Sub','A','B','Item')] = lapply(Data[c('Sub','A','B','Item')],factor)
```

等价

```
> Data$Sub = factor(Data$Sub)
> Data$A = factor(Data$A)
> Data$B = factor(Data$B)
> Data$Item = factor(Data$Item)
```

```
> head(Data,3)
# A tibble: 3 x 5
  Sub   A     B     Item     Y
  <fct> <fct> <fct> <fct> <dbl>
1 1     A2    B2    1       254
2 1     A2    B2    1       341
3 1     A1    B1    2       189
```

建模

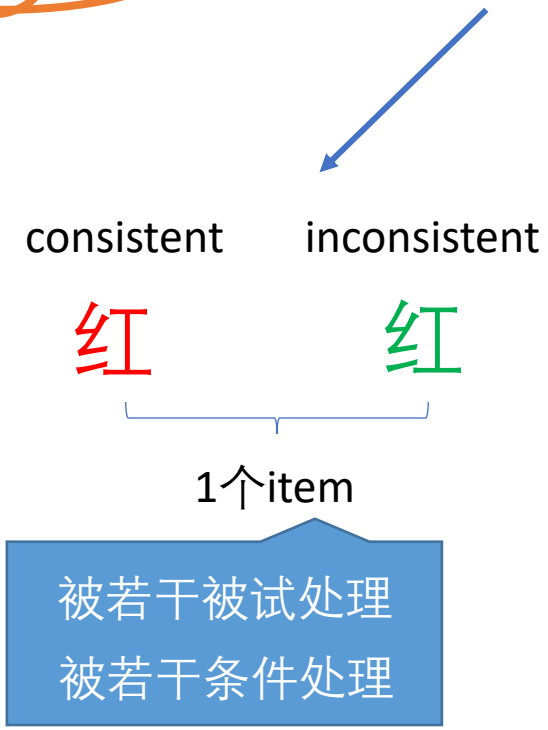
```
> Model = lmer(data = Data, Y ~ A*B + (1 + A*B |Sub) + (1+A*B|Item))
```

固定效应： 一般为要考察的自变量。

随机因子： beyond the first level。e.g. 认知实验中的被试，项目；社会调查中的城市、省份

随机截距： 因变量在随机因子的每个单位上的分布是不同的。 e.g. 被试的平均反应时

随机斜率： 随机因子的每个单位上，某因素（一般为固定效应）与因变量的关系是不同的。 e.g. 噪音大小对学习效率的影响在不同被试上的差异



建模

```
> Model = lmer(data = Data, Y ~ A*B + (1 + A*B |Sub)+(1+A*B|Item))
```

But……

```
> Model = lmer(data = Data, Y ~ A*B + (1 + A*B |Sub)+(1+A*B|Item))
boundary (singular) fit: see ?isSingular
Warning message:
Model failed to converge with 2 negative eigenvalues: -1.5e-01 -2.0e-01
```

全模型（包含尽可能多的随机斜率）往往不能收敛或出现畸形协方差矩阵；

零模型（只包含随机截距）往往不出现以上问题。

Q1. 建模从零模型还是全模型开始?

- 模型应包含尽可能多的随机斜率(Barr, 2013)
- 从零模型开始，逐渐加随机斜率 → p-Harking!
- 从全模型开始！从全模型开始！！从全模型开始！！！（说三遍）再解决问题。

模型优化

模型不能收敛

畸形协方差矩阵

1. **数据不能支持如此多的随机斜率**
2. **某个随机斜率的效应太小（与其他斜率存在共线性/方差过小）**

```
> summary(Model)$varcor
 Groups   Name          Std.Dev.   Corr
 Item     (Intercept)   12.5670
          AA2           25.1545   -0.948
          BB2           34.8443   -0.994  0.908
          AA2:BB2       49.2017    0.996 -0.971 -0.981
 Sub      (Intercept)   26.9739
          AA2            9.4915    0.179
          BB2            8.2699    0.202 -0.705
          AA2:BB2       19.6023    0.315 -0.856  0.634
 Residual                85.3361
```

查看模型信息

固定效应&随机效应：**summary(Model)**

主效应&交互作用：**anova(Model)**

判断模型是否出现畸形协方差

```
> Model = lmer(data = Data, Y ~ A*B + (1 + A*B |Sub)+(1+A*B|Item))
boundary (singular) fit: see ?isSingular
Warning message:
Model failed to converge with 2 negative eigenvalues: -1.5e-01 -2.0e-01
```

→ 建模时输出warning

```
> isSingular(Model, tol = 1e-05)
[1] TRUE
```

→ 利用函数判断
tol设为 1e-04 可能更好

Q2.1 如何删减随机斜率（优化模型）？

模型优化

Q2.1 删减随机斜率（优化模型）的步骤?

1. 如果随机斜率中包含交互作用，优先删除交互作用 (Barr, 2013)

```
> Model = lmer(data = Data, Y ~ A*B + (1 + A*B |Sub)+(1+A*B|Item))
```

等价

A+B+A:B

```
> Model = lmer(data = Data, Y ~ A*B + (1 + A:B |Sub)+(1+A*B|Item))
```

模型优化

Q2.1 删减随机斜率（优化模型）的步骤?

1. **如果随机斜率中包含交互作用，优先删除交互作用 (Barr, 2013)**
2. **当需从两个主效应的随机斜率删除某一个时，应考虑分别删除后的模型**

```
> Model = lmer(data = Data, Y ~ A*B + (1 + A |Sub)+(1+A*B|Item))
```

考虑并比较

```
> Model = lmer(data = Data, Y ~ A*B + (1 + B |Sub)+(1+A*B|Item))
```

模型优化

Q2.1 删减随机斜率（优化模型）的步骤?

1. 如果随机斜率中包含交互作用，优先删除交互作用 (Barr, 2013)
2. 当存在两个主效应的随机斜率时，应同时考虑分别删除后的模型
3. 优先考虑删除差异较小(between units)的随机因子上的斜率

```
> Model = lmer(data = Data, Y ~ A*B + (1 + A*B |Sub)+(1+A*B|Item))
```

被试间的差异较大

项目间的差异较小

优先删除项目上的斜率

模型优化

Q2.1 删减随机斜率（优化模型）的步骤?

1. 如果随机斜率中包含交互作用，优先删除交互作用 (Barr, 2013)
2. 当存在两个主效应的随机斜率时，应同时考虑分别删除后的模型
3. 优先考虑删除差异较小(between units)的随机因子上的斜率
4. 对删减后的新模型，应考察它与全模型是否有显著差异（理论上应没有）

```
> Model = lmer(data = Data, Y ~ A*B + (1 + A*B |Sub)+(1+A*B|Item))
```

```
> Model2 = lmer(data = Data, Y ~ A*B + (1 + B |Sub)+(1|Item))
```

```
> anova(Model2, Model)
refitting model(s) with ML (instead of REML)
```

Df	AIC	BIC	logLik	deviance	Chisq	Chi	Df	Pr(>Chisq)
Model2	9	14218	14264	-7100.2	14200			
Model	25	14245	14372	-7097.4	14195	5.6199	16	0.9917

```
anova(Model1, Model2) # 比较模型差异
```

模型优化

Q2.1 删减随机斜率（优化模型）的步骤?

1. **如果随机斜率中包含交互作用，优先删除交互作用 (Barr, 2013)**
2. **当存在两个主效应的随机斜率时，应同时考虑分别删除后的模型**
3. **优先考虑删除差异较小(between units)的随机因子上的斜率**
4. **对删减后的新模型，应考察它与全模型是否有显著差异（理论上应没有）**

But……

随着随机因子数量和随机斜率数量的增加，可能存在的模型数量急剧增加!

- 两因素+单个随机因子 → 8个
- 两因素+两个随机因子 → 35个
- 三因素+两个随机因子→ 288个

→ **Q2.2 如何更高效地优化模型?**

模型优化

Q2.2 如何更高效地优化模型?

1. **寻求较强的理论支持，限定要考察的随机斜率**
   - e.g. 只从包含某个随机斜率的模型中考察
2. **从方法上优化**

模型优化

Q2.2.1 模型遍历的思路?

1. 自动/手动产生所有可能的模型；
   - 两因素/三因素+两个随机因子
2. 自动运行这些模型，并考察它们是否能收敛、是否出现畸形协方差；
3. 根据输出的结果筛选；
   - 好的模型的标准：能收敛、无畸形协方差、包含尽可能多的斜率。

Q2.2.2 利用主成分分析优化模型的思路？（Bate et al., 2015）

- 如果模型的随机斜率组成是合理的，那么每个随机斜率都能代表某个独特的效应/独特的成分；
- 这个随机斜率应该表现出在某个成分的载荷量非常大，而在其他成分的载荷量非常小；
- 如果某个随机斜率在任何成分的载荷量都不大，说明这个成分是多余的，同时这个随机斜率也可能是不合理。

利用主成分分析（PCA）来优化模型

```
> Model = lmer(data = Data, Y ~ A*B + (1 + A*B |Sub)+(1+A*B|Item))
> ModelPCA = rePCA(Model)
> ModelPCA
```

```
$Item
Standard deviations (1, .., p=4):
[1] 0.77 0.1 0.01 0

Rotation (n x k) = (4 x 4):
     [,1]   [,2]  [,3]   [,4]
[1,] -0.19 -0.085  0.19 -0.9591
[2,]  0.37 -0.733 -0.56 -0.1195
[3,]  0.52  0.664 -0.47 -0.2566
[4,] -0.75  0.123 -0.65  0.0064
```

```
$Sub
Standard deviations (1, .., p=4):
[1] 0.33 0.24 0.07 0

Rotation (n x k) = (4 x 4):
     [,1]  [,2]   [,3]  [,4]
[1,] -0.91 -0.39 -0.035 -0.15
[2,]  0.04 -0.45  0.139  0.88
[3,] -0.11  0.24 -0.923  0.28
[4,] -0.40  0.76  0.356  0.36
```

```
> ModelPCAItem = ModelPCA$Item$rotation
```

利用主成分分析（PCA）来优化模型

```
> ModelPCAItem
     [,1]   [,2]  [,3]    [,4]
[1,] -0.19 -0.085  0.19 -0.9591
[2,]  0.37 -0.733 -0.56 -0.1195
[3,]  0.52  0.664 -0.47 -0.2566
[4,] -0.75  0.123 -0.65  0.0064
```

```
> rownames(ModelPCAItem) = c('Intercept','A','B','A:B')
> colnames(ModelPCAItem) = c('Comp1','Comp2','Comp3','Comp4')
> ModelPCAItem
          Comp1  Comp2  Comp3  Comp4
Intercept -0.19  -0.085  0.19  -0.9591
A          0.37  -0.733 -0.56  -0.1195
B          0.52   0.664 -0.47  -0.2566
A:B       -0.75   0.123 -0.65   0.0064
```

利用主成分分析（PCA）来优化模型

```
> ModelPCAItem[abs(ModelPCAItem) < 0.9] = NA
> ModelPCAItem
            Comp1 Comp2 Comp3 Comp4
Intercept   NA    NA    NA    -0.96
A           NA    NA    NA    NA
B           NA    NA    NA    NA
A:B         NA    NA    NA    NA
```

同理……

```
> ModelPCASub
            Comp1 Comp2 Comp3 Comp4
Intercept   -0.91 NA    NA    NA
A           NA    NA    NA    NA
B           NA    NA    -0.92 NA
A:B         NA    NA    NA    NA
```

```
> ModelNew = lmer(data = Data, Y ~ A*B + (B|Sub) + (1|Item))
```

利用主成分分析（PCA）来优化模型

```
> anova(ModelNew, Model)
refitting model(s) with ML (instead of REML)
```

Df	AIC	BIC	logLik	deviance	Chisq	Chi	Df	Pr(>Chisq)
ModelNew	9	14218	14264	-7100	14200			
Model	25	14245	14372	-7097	14195	5.62	16	0.99

PCA：推荐使用，but注意一些问题：

- 0.9的标准
- 在多水平/多因素交互时是否有变化
- 目前没有固定的优化标准(提供优化的依据，而不是规定优化的流程)

模型优化完成：

- 被试(Sub)上：随机截距+B Slope
- 项目(Item)上：随机截距

Q3 如何查看模型信息和因素的效应？

主效应与固定效应

summary(Model) → 两个水平的比较(t 检验)

Estimate	Std.	Error	df	t value	Pr(>|t|)
(Intercept)	239.7	10.5	13	22.89	5.2e-12
AA2	11.1	8.0	145	1.39	1.7e-01
BB2	11.1	9.4	28	1.17	2.5e-01
AA2:BB2	-9.5	12.2	57	-0.78	4.4e-01

anova(Model) → F检验　（特殊情况：因素只有两个水平）

Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
A	7999	7999	1	25.5	1.10	0.30
B	4591	4591	1	11.1	0.63	0.44
A:B	4904	4904	1	16.1	0.67	0.42

WHY Pr(>|t|) ≠ Pr(>F) → 稍后讨论……

Q4.1 如果主效应显著怎么办/如何进行事后检验?

事后检验

Q4.1如何进行事后检验?

emmeans package

emmeans(model = , pairwise~, adjust =)

```
> emmeans(ModelNew, pairwise~A, adjust='none')
$contrasts
 contrast estimate   SE         df      t.ratio   p.value
 A1 - A2    -6.34      5.11       1088  -1.243  0.2143

Results are averaged over the levels of: B
Degrees-of-freedom method: kenward-roger
```

```
> p.adjust.methods
[1] "holm"      "hochberg"   "hommel"     "bonferroni" "BH"
[6] "BY"        "fdr"        "none"
```

Q4.2 交互作用显著了怎么办?

简单（主）效应

Q4.2 交互作用显著了怎么办?

emmeans package

joint_test(object = , by =)

```
> joint_tests(ModelNew, by = 'A')
A = A1:
 model term df1    df2     F.ratio   p.value
 B           1     24.67   1.362     0.2544

A = A2:
 model term df1    df2     F.ratio   p.value
 B           1     24.05   0.028     0.8692
```

Q4.3 简单主效应显著了怎么办/如何进行简单主效应的事后比较?

简单效应的事后检验

Q4.3 如何进行简单主效应的事后比较?

emmeans package

emmeans(model = , pairwise~ A|B)

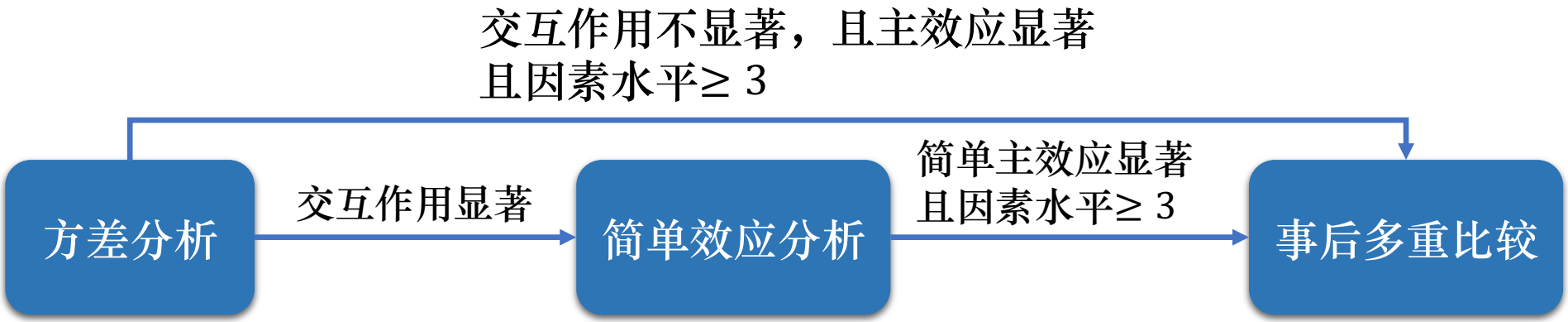
```
> emmeans(ModelNew, pairwise~A|B, adjust='none')
$emmeans
......

$contrasts
B = B1:
 contrast  estimate  SE   df    t.ratio  p.value
 A1 - A2   -11.1      8.0 144   -1.380   0.1700
B = B2:
 contrast  estimate  SE   df    t.ratio  p.value
 A1 - A2    -1.6      7.87 134  -0.203   0.8394
......
```

Q4.4 简单效应、事后检验、事后多重比较的关系?

简单效应 & 事后检验 & 事后多重比较

Q4.4 简单效应、事后检验、事后多重比较的关系?

- **若交互作用显著，则分析简单主效应**
- **若简单主效应显著，且因素水平超过2个，则事后检验需要矫正**
- **若交互作用不显著但主效应显著，且因素水平超过2个，则主效应的事后检验需要矫正**



小结

Q1. 建模从零模型还是全模型开始?

- 从全模型开始

Q2.1 删减随机斜率（优化模型）的步骤?

1. 如果随机斜率中包含交互作用，优先删除交互作用 (Barr, 2013)
2. 当存在两个主效应的随机斜率时，应同时考虑分别删除后的模型
3. 优先考虑删除差异较小(between units)的随机因子上的斜率
4. 对删减后的新模型，应考察它与全模型是否有显著差异（理论上应没有）

Q2.2 如何更高效地优化模型?

1. 寻求较强的理论支持，限定要考察的随机斜率
2. 从方法上优化

小结

Q2.2.1 模型遍历的思路?

- 自编函数运行所有可能的模型，并根据模型信息筛选

Q2.2.2 利用主成分分析优化模型的思路?

- 如果模型的随机斜率组成是合理的，那么每个随机斜率都能代表某个独特的效应/独特的成分；
- 这个随机斜率应该表现出在某个成分的载荷量非常大，而在其他成分的载荷量非常小；
- 如果某个随机斜率在任何成分的载荷量都不大，说明这个成分是多余的，同时这个随机斜率也可能是不合理。

Q3 如何查看模型信息和因素的效应?

- summary() & anova()

Q4 简单效应、事后检验、多重比较矫正的关系?

回到固定效应与主效应……

summary(Model)

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	239.7	10.5	13	22.89	5.2e-12
AA2	11.1	8.0	145	1.39	1.7e-01
BB2	11.1	9.4	28	1.17	2.5e-01
AA2:BB2	-9.5	12.2	57	-0.78	4.4e-01

anova(Model)

Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
A	7999	7999	1	25.5	1.10	0.30
B	4591	4591	1	11.1	0.63	0.44
A:B	4904	4904	1	16.1	0.67	0.42

WHY Pr(>|t|) ≠ Pr(>F) → ~~稍后讲述~~……现在解决！

固定效应中的回归系数与比较方式的关系

- 自变量(X)为连续变量
  - 回归系数：X每变化一个单位，Y变化的单位量
- 自变量(X)为因子变量
  - 线性模型要求预测变量是数值型
  - 用一系列与因子水平相对应的数值型的**对照变量**来代替因子
  - 对照变量定义了对比方式/矩阵

Q5.1 对单个因子变量，对比方式产生的规律是什么

单因素两水平的变量如何产生对比方式

以R中的mtcars数据为例

mtcars数据收集了若干种汽车的若干参数，这里想考察汽车的变速器(am, Transmission, 0 = automatic, 1 = manual)对耗油量(mpg, miles per gallon)的影响。

```
> mtcarsNew = within(mtcars, {Transmission = NA;
                              Transmission[am == 0] ='Auto';
                              Transmission[am == 1] = 'Manual'})
> mtcarsNew$Transmission = factor(mtcarsNew$Transmission)
> mtcarsNew = mtcarsNew[c('mpg', 'Transmission')]
> head(mtcarsNew, 3)
                  mpg  Transmission
Mazda RX4         21     Manual
Mazda RX4 Wag     21     Manual
Datsun 710        23     Manual
> Model = lm(data = mtcarsNew, mpg~Transmission)
```

单因素两水平变量如何产生对比方式

```
> summary(Model)
.....
Coefficients:
                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         17.15     1.12        15.25    1.1e-15 ***
Transmission Manual 7.24      1.76        4.11     0.00029 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
.....
```

- 模型回归系数部分中显示的是当汽车为手动挡(Manual)时的信息；
- 模型对自变量生成了一个新变量，设置了其对比方式

```
> contrasts(mtcarsNew$Transmission)
       Manual
Auto        0
Manual      1
```

单因素多水平变量的对比方式

```
> ThreeLevel = factor(c('A','B','C'))
> contrasts(ThreeLevel)
  B C
A 0 0
B 1 0
C 0 1
```

```
> FourLevel = factor(c('A','B','C','D'))
> contrasts(FourLevel)
  B C D
A 0 0 0
B 1 0 0
C 0 1 0
D 0 0 1
```

Q5.2 模型是如何利用对比方式来获取回归系数的

Q5.1 对单个因子变量，对比方式产生的规律是什么

1. 对于某个含有 k 个水平的因子，会产生 $k-1$ 个对比变量；
2. 对比变量以一个 $k \times (k-1)$ 的矩阵的形式存在；
3. 矩阵中每一列表示一个对比变量或对比方式；
4. 矩阵中的每一行代表某个水平在不同对比方式下的编码；
5. 默认的对比方式总是将某个水平作为基线（比如 A），每种对比方式表示用其他的某个水平和基线进行比较。

单因素多水平变量的对比方式

```
> FourLevel = factor(c('A','B','C','D'))
> contrasts(FourLevel)
  B C D
A 0 0 0
B 1 0 0
C 0 1 0
D 0 0 1
```

$$Y = b_1X_B + b_2X_C + b_3X_D + d + \sigma$$

A水平： $Y_A = d + \sigma$

B水平： $Y_B = b_1 + d + \sigma$

C水平： $Y_C = b_2 + d + \sigma$

D水平： $Y_D = b_3 + d + \sigma$

$$Y_B - Y_A = b_1$$

$$Y_C - YA = b_2$$

$$Y_D - YA = b_3$$

单因素 k 水平默认的对比方式下，回归系数等于真实效应差异

treatment coding VS sum coding

```
> contr.treatment(2)
  2
1 0
2 1
> contr.treatment(3)
  2 3
1 0 0
2 1 0
3 0 1
> contr.treatment(4)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

```
> contr.sum(2)
  [,1]
1    1
2   -1
> contr.sum(3)
  [,1] [,2]
1    1    0
2    0    1
3   -1   -1
> contr.sum(4)
  [,1] [,2] [,3]
1    1    0    0
2    0    1    0
3    0    0    1
4   -1   -1   -1
```

可改变线性模型中因子的对比方式

设置因子对比方式

通过全局参数设置

```
> options(contrasts = c('contr.sum', 'contr.poly'))

 # 需要同时依次定义无序因子和有序因子的对比方式，后者为定义有序因子的，不常用，这里不展开说了。
```

在模型中设置(局部)

```
> ComMatrix = contr.sum(2)
> rownames(ComMatrix) = levels(mtcarsNew$Transmission)
> colnames(ComMatrix) = 'Auto'
> ComMatrix
          Auto
Auto       1
Manual    -1

> Model = lm(data = mtcarsNew, mpg~Transmission,
                  contrasts = list(Transmission = ComMatrix))
```

设置因子对比方式

```
> summary(Model)
.....................................................................................
Coefficients:
                      Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)           20.7698    0.8822       23.543      < 2e-16 ***
Transmission Auto     -3.6225    0.8822       -4.106      0.000285 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
.....................................................................................
```

What?! 回归系数成了原来的一半了!

单因素两水平 sum coding 下的对比

```
> contrasts(mtcarsNew$Transmission)
          [,1]
Auto       1
Manual    -1
```

$$Y = b_1 X_{\text{Transmission}} + d + \sigma$$

Manual水平： $Y_{Manual} = -b_1 + d + \sigma$

Auto水平： $Y_{Auto} = b_1 + d + \sigma$

$$Y_{Auto} - Y_{Manual} = 2b_1$$

$$b_1 = \frac{Y_{Auto} - YManual}{2}$$

单因素两水平sum coding的对比下，回归系数等于真实效应的一半

单因素三水平 sum coding 下的对比

```
> contrasts(ThreeLevel)
   [,1] [,2]
A    1    0
B    0    1
C   -1   -1
```

$$Y = b_1X_1 + b_2X_2 + d + \sigma$$

A 水平：$Y_A = b_1 + d + \sigma$

B 水平：$Y_B = b_2 + d + \sigma$

C 水平：$Y_C = -b_1 - b_2 + d + \sigma$

$$Y_A - Y_C = 2b_1 + b_2$$

$$Y_B - YC = b_1 + 2b_2$$

单因素多水平sum coding的对比下，回归系数不等于真实效应

因子水平超过2时，采用sum coding的对比方式存在问题

两因素两水平无交互的情况

treatment coding

$$\boldsymbol{Y = b_1 X_A + b_2 X_B + d + \sigma}$$

$$X_A = \begin{cases} 0, & X_A = A1 \\ 1, & X_A = A2 \end{cases}$$

$$X_B = \begin{cases} 0, & X_B = B1 \\ 1, & X_B = B2 \end{cases}$$

$$Y_{A1} = b_2 + d + \sigma$$

$$Y_{A2} = b_1 + b_2 + d + \sigma$$

$$\boldsymbol{Y_{A2} - YA_1 = b_1}$$

同理

$$\boldsymbol{Y_{B2} - YB_1 = b_2}$$

采用treatment coding编码的对比下，各因子的回归系数等于其真实效应

两因素两水平无交互的情况

sum coding

$$\boldsymbol{Y = b_1 X_A + b_2 X_B + d + \sigma}$$

$$X_A = \begin{cases} -1, & X_A = A1 \\ 1, & X_A = A2 \end{cases}$$

$$X_B = \begin{cases} -1, & X_B = B1 \\ 1, & X_B = B2 \end{cases}$$

$Y_{A1} = -b_1 + d + \sigma$

$Y_{A2} = b_1 + d + \sigma$

$\boldsymbol{Y_{A2} - YA_1 = 2b_1}$

同理

$\boldsymbol{Y_{B2} - YB_1 = 2b_2}$

采用sum coding编码的对比下，各因子的回归系数等于其真实效应的一半

两因素两水平有交互的情况

treatment coding

$$\boldsymbol{Y = b_1 X_A + b_2 X_B + b_3 X_A X_B + d + \sigma}$$

$$X_A = \begin{cases} 0, & X_A = A1 \\ 1, & X_A = A2 \end{cases}$$

$$X_B = \begin{cases} 0, & X_B = B1 \\ 1, & X_B = B2 \end{cases}$$

$$Y_{A1} = b_2 + d + \sigma$$

$$Y_{A2} = b_1 + b_2 + b_3 + d + \sigma$$

$$\boldsymbol{Y_{A2} - YA_1 = b_1 + b_3}$$

同理

$$\boldsymbol{Y_{B2} - YB_1 = b_2 + b_3}$$

$$Y_{A1B1} = d + \sigma$$

$$Y_{A1B2} = b_2 + d + \sigma$$

$$Y_{A2B1} = b_1 + d + \sigma$$

$$Y_{A2B2} = b_1 + b_2 + b_3 + d + \sigma$$

$$\boldsymbol{(Y_{A2B2} - Y_{A2B1}) - (Y_{A1B2} - Y_{A1B1}) = b_3}$$

treatment coding的对比下，交互作用的回归系数等于真实的效应，主效应的回归系数不是真实的效应

两因素两水平有交互的情况

sum coding

$$\boldsymbol{Y = b_1X_A + b_2X_B + b_3X_AX_B + d + \sigma}$$

$$X_A = \begin{cases} -1, & X_A = A1 \\ 1, & X_A = A2 \end{cases}$$

$$X_B = \begin{cases} -1, & X_B = B1 \\ 1, & X_B = B2 \end{cases}$$

$$Y_{A1} = -b_1 - b_2 + d + \sigma$$

$$Y_{A2} = b_1 - b_2 + d + \sigma$$

$$\boldsymbol{Y_{A2} - YA_1 = 2b_1}$$

同理

$$\boldsymbol{Y_{B2} - YB_1 = 2b_2}$$

$$Y_{A1B1} = -b_1 - b_2 + b_3 + d + \sigma$$

$$Y_{A1B2} = -b_1 + b_2 - b_3 + d + \sigma$$

$$Y_{A2B1} = b_1 - b_2 - b_3 + d + \sigma$$

$$Y_{A2B2} = b_1 + b_2 + b_3 + d + \sigma$$

$$\boldsymbol{(Y_{A2B2} - Y_{A2B1}) - (Y_{A1B2} - Y_{A1B1}) = 4b_3}$$

即sum coding的对比下，交互作用的回归系数等于真实的效应的$\frac{1}{4}$，主效应的回归系数等于真实的效应的$\frac{1}{2}$

[-1, 1] → [-0.5, 0.5]　回归系数 = 真实效应

对比方式对回归系数的影响

Q5.3 treatment coding 与 sum coding下，回归系数与真实效应的关系？

- **treatment coding 适用于无交互的主效应 & 有交互的交互作用**
- **一旦因素有超过两个水平，sum coding 不适用**

	treatment coding		sum coding	
	主效应	交互作用	主效应	交互作用
单因素两水平	相等		真实效应的一半	
单因素多水平	相等		不相等	
两因素两水平无交互	相等		真实效应的一半	
两因素两水平有交互	不相等	相等	真实效应的一半	真实效应的1/4
多因素多水平无交互	相等		不相等	
多因素多水平有交互	不相等	相等	不相等	不相等

对比方式对回归系数的影响

Q5.4 k 因素 n 水平的情况下($k \geq 1, n \geq 2$)，如何保证回归系数等于真实的效应值？

- **simple coding**

But simple coding is not simple……

R 里面没有相关函数！ DIY！！

```
> contr.simple(2)
  [,1]
1 -0.5
2  0.5
```

```
> contr.simple(3)
     [,1]  [,2]
[1,] -0.33 -0.33
[2,]  0.67 -0.33
[3,] -0.33  0.67
```

张光耀

心理学小雪僧

线性模型中无序因子变量的对比方式与回归系数的关系：原理，困境及其解决

来自专栏中国R语言社区

simple coding 编码规律

Q5.5 simple coding 的编码规律?

- **每一列代表一个对比方式；**
- **每一行代表一个水平；**
- **第一个水平为基线；**
- **对于某一列 C，表示基线之外的某个水平 L 与基线的对比；**
- **在 C 列中，水平 L 所在行的编码为$1-\frac{1}{n}$，其余行编码为$-\frac{1}{n}$。**

```
> contr.simple(2)
  [,1]
1 -0.5
2  0.5
```

```
> contr.simple(3)
      [,1]  [,2]
[1,] -0.33 -0.33
[2,]  0.67 -0.33
[3,] -0.33  0.67
```

```
> contr.simple(4)
      [,1]  [,2]  [,3]
[1,] -0.25 -0.25 -0.25
[2,]  0.75 -0.25 -0.25
[3,] -0.25  0.75 -0.25
[4,] -0.25 -0.25  0.75
```

Back to LMM

```
> ModelNew = lmer(data = Data,
                  Y~A*B+(B|Sub)+(1|Item),
                  contrasts = list(A = contr.simple(2),
                                   B = contr.simple(2)))
```

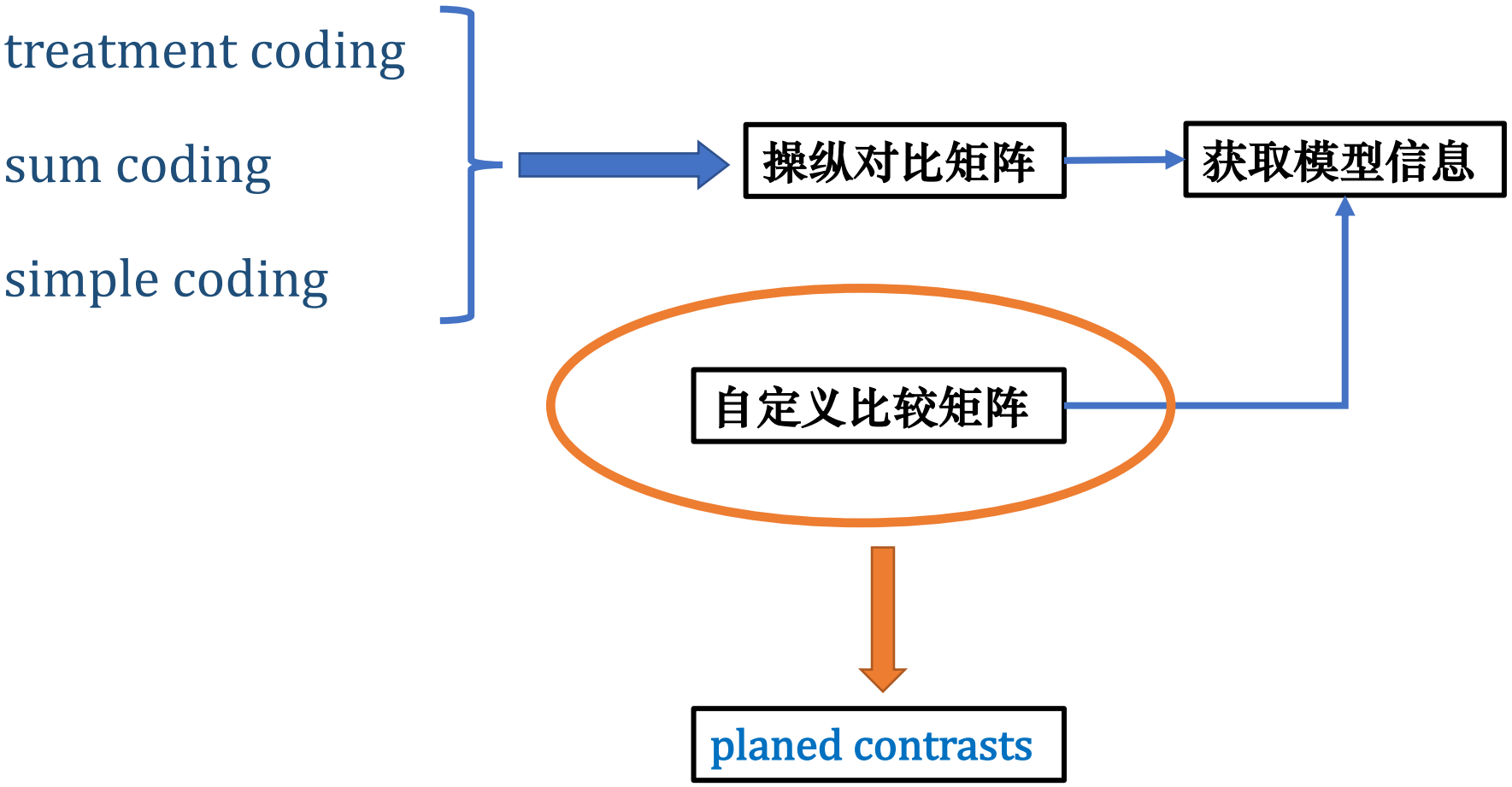
```
> summary(ModelNew)$coef
             Estimate   Std. Error
(Intercept)  248.44     10.09
A1           6.34       5.09
B1           6.32       7.20
A1:B1        -9.49      12.17
```

JAMOVI

Fixed Effects Parameter Estimates

Names	Effect	Estimate	SE
(Intercept)	(Intercept)	248.439	10.094
A1	A2 - A1	6.344	5.085
B1	B2 - B1	6.319	7.197
A1 ✻ B1	A2 - A1 ✻ B2 - B1	−9.492	12.169

结果一致！！！


treatment coding
sum coding
simple coding
操纵对比矩阵
获取模型信息
自定义比较矩阵
planed contrasts

planed contrasts

Q6 planed contrasts 时如何设置比较矩阵（简单的对比矩阵）？

- **每个contrast的和为0；**
- **每个contrast的绝对值的和为1；**
- **任意两个contrast的对应乘积和为0**

错误！！

	Cont1
A	0.5
B	-0.5
C	0.5
D	0

错误！！

	Cont1	Cont2	Cont3
A	0.5	0	0
B	-0.5	0.5	0
C	0	-0.5	0.5
D	0	0	-0.5

正确！！

	Cont1	Cont2
A	0.5	0
B	-0.5	0
C	0	0.5
D	0	-0.5

小结

Q5.1 对单个因子变量，对比方式产生的规律是什么

1. 对比变量以一个 $k \times (k-1)$ 的矩阵的形式存在；
2. 列表示对比变量/对比方式；
3. 行代表某个水平的编码；
4. 总是将某个水平作为基线。

Q5.2 模型是如何利用对比方式来获取回归系数的

Q6 如何自定义对比矩阵?

- 每个contrast的和为0；
- 每个contrast的绝对值的和为1；
- 任意两个contrast的对应乘积和为0

Q5.3 treatment coding 与 sum coding下，回归系数与真实效应的关系?

- treatment coding 适用于无交互的主效应 & 有交互的交互作用
- 一旦因素有超过两个水平，sum coding 不适用

Q5.4 如何保证回归系数等于真实的效应值？

- simple coding

Q5.5 simple coding 的编码规律?

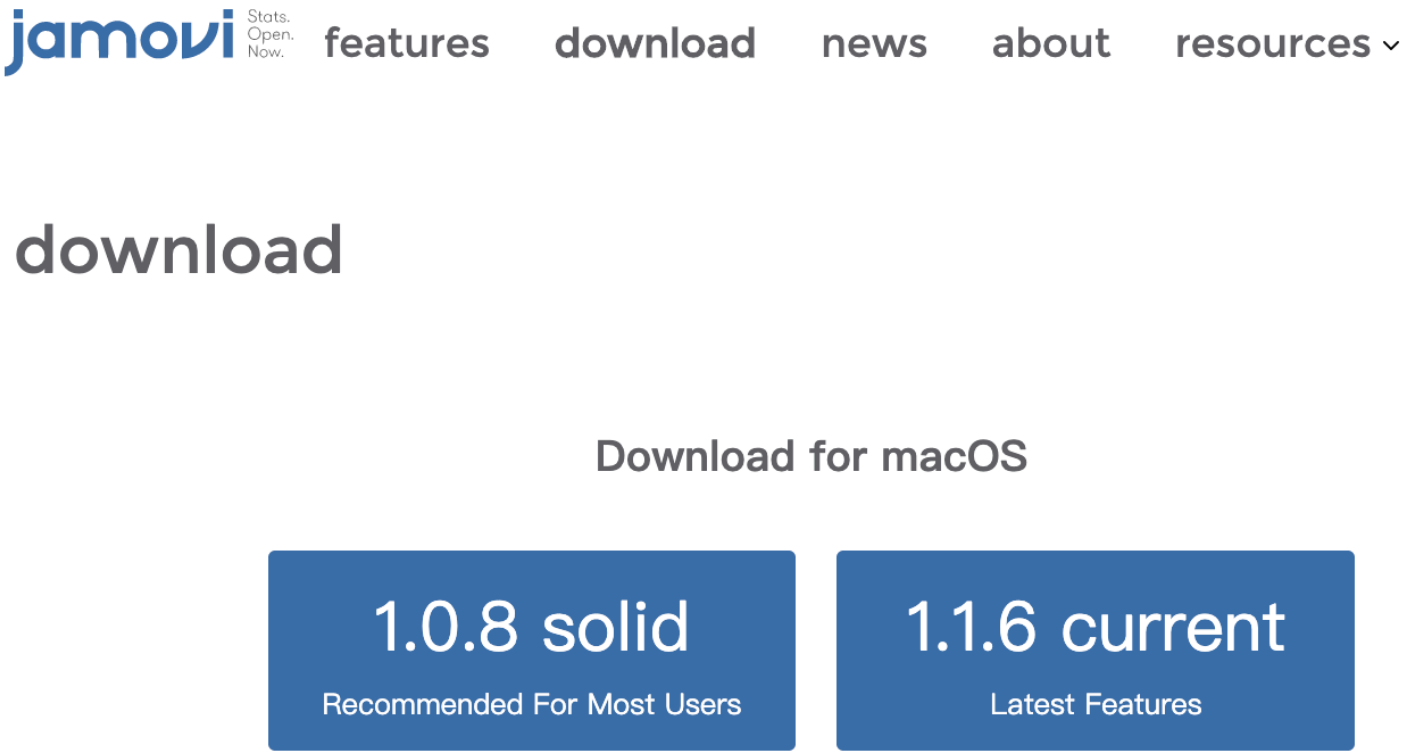
jamovi中的实现

Q1 jamovi 是啥?

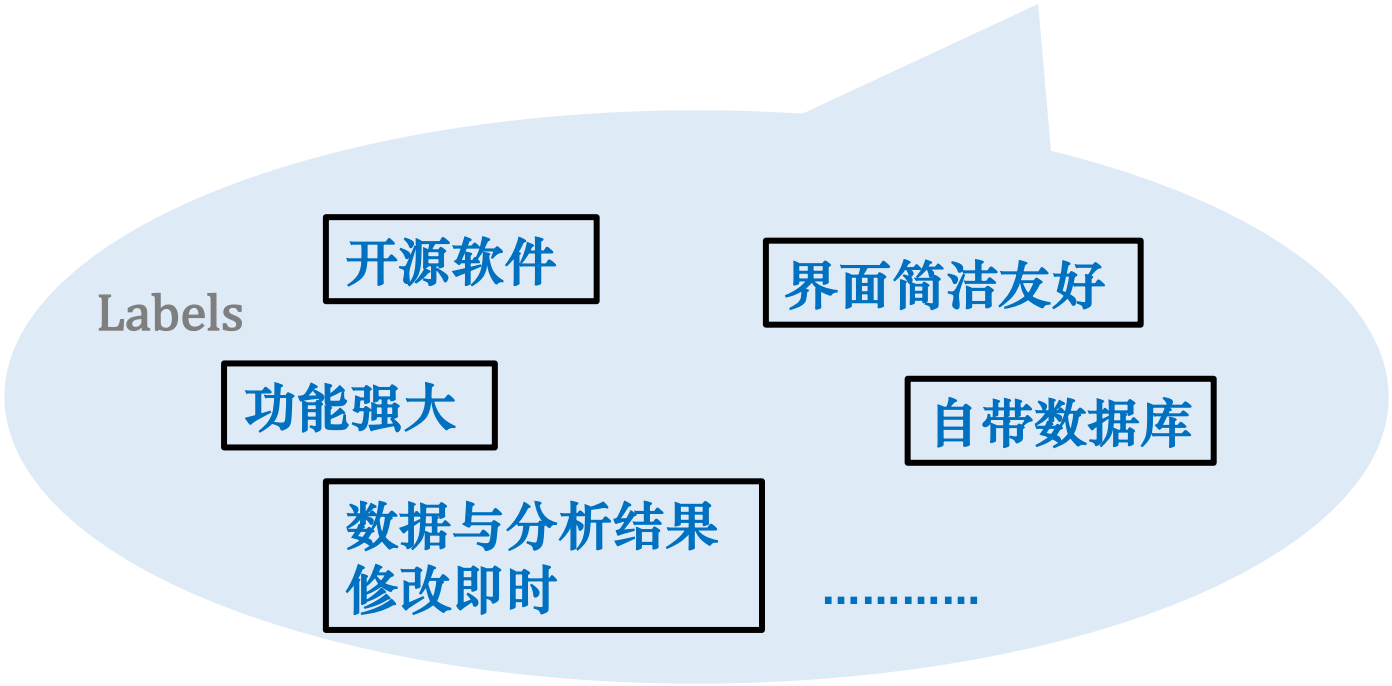
Q2 为啥用jamovi?

Q3 怎么用jamovi?

https://www.jamovi.org/download.html



基于R语言建立的统计分析软件



为啥用？谁用谁知道

怎么用？一用就知道

R or JAMOVI

不要拘泥于一个软件：

编辑数据： Excel is more powerful;

筛选模型： R is more powerful;

数据可视化： R is more powerful;

一般统计分析： R is more powerful；

初学者： JAMOVI is more powerful;

单一指标/“小数据”： JAMOVI is more powerful;