

统计分析与建模

高珍

gaozhen@tongji.edu.cn

数据预处理

- **数据过滤**

- 缺失值处理
- 异常值处理
- 数据去重
- 数据规范
- 数据采样
- 数据排序
- 数据向量化
- 列联表
- 分组汇总

数据过滤

- which()、subset()

```
``{r subset-1}  
df=data.frame(year=c(2000,2001,2000,2003,2001),  
  month=c(1,2,3,4,5))  
df  
``
```

year <dbl>	month <dbl>
2000	1
2001	2
2000	3
2003	4
2001	5

5 rows

```
``{r subset-2}  
subset(df,year=='2000')  
df[df$year==2000,]  
df[which(df$year==2000),]  
``
```

year <dbl>	month <dbl>
2000	1
2000	3

data.frame
2 x 2

year <dbl>	month <dbl>
2000	1
2000	3

data.frame
2 x 2

year <dbl>	month <dbl>
2000	1
2000	3

data.frame
2 x 2

	year <dbl>	month <dbl>
1	2000	1
3	2000	3

2 rows

数据预处理

- 数据过滤
- **缺失值处理**
- 异常值处理
- 数据去重
- 数据规范
- 数据采样
- 数据排序
- 数据向量化
- 列联表
- 分组汇总

缺失值处理

一些特殊的数:

(1) FALSE (假): 以0计算

(2) NA (缺失值): 参与计算

(3) NULL: 不参与计算

(4) NaN: 无意义的数, 比如 $\sqrt{-2}$

```
NA == NA
```

```
## [1] NA
```

```
NA+8
```

```
## [1] NA
```

```
NA^0
```

```
## [1] 1
```

```
1/NA
```

```
## [1] NA
```

```
1/0
```

```
## [1] Inf
```

```
1/0-1/0
```

```
## [1] NaN
```

```
```{r NA}
x<-c(1,2,3,NA,4);
mean(x)
mean(x,na.rm=T)
```
```

```
[1] NA
[1] 2.5
```

缺失值检测

- ①判断x是否缺失值的函数是`is.na(x)`，是返回TRUE,否则返回FALSE。
- ②判断x是否完整的函数是`complete.cases(x)`。
- ③ `summary`可以显示每个变量的缺失值数量。
- ④返回数据缺失模式使用mice包中的`md.pattern(x)`函数。

缺失值处理

对于缺失数据通常有三种方法：

方法1：当缺失数据较少时直接删除相应样本

方法2：对缺失数据进行插补

方法3：使用对缺失数据不敏感的分析方法，如决策树

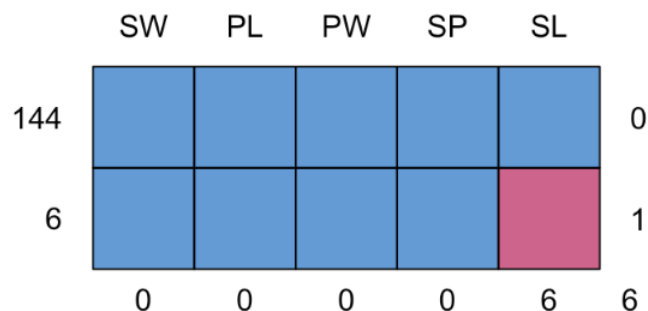
```
```{r NA}
v=c(1,2,3,NA,4,NA,5)
print(v)
v[is.na(v)]=0
print(v)
```
```

```
[1] 1 2 3 NA 4 NA 5
[1] 1 2 3 0 4 0 5
```

缺失值检测

```
``{r kurtosis/skewness}  
library(mice)  
x=iris  
x[sample(1:nrow(x), 6),1] <- NA  
#随机在iris数据集第1列生成6行NA  
colnames(x)<-c("SL", "SW", "PL", "PW", "SP")  
md.pattern(x)  
``
```

R Console



数据预处理

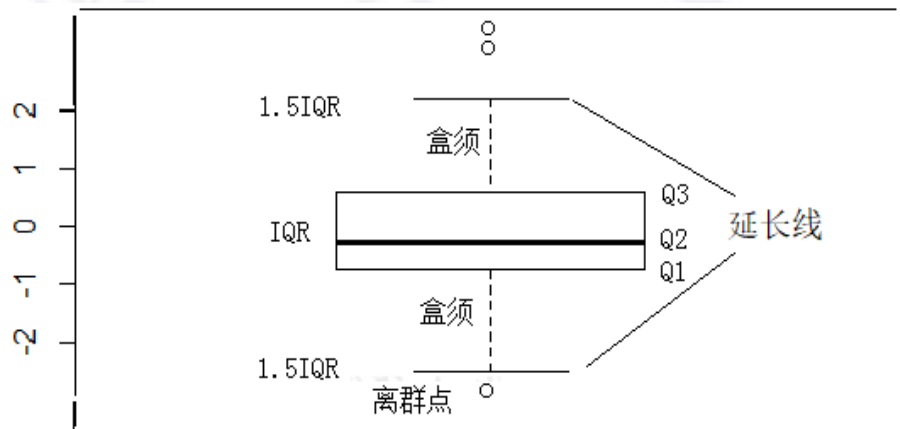
- 数据过滤
- 缺失值处理
- **异常值处理**
- 数据去重
- 数据规范
- 数据采样
- 数据排序
- 数据向量化
- 列联表
- 分组汇总

异常值检测

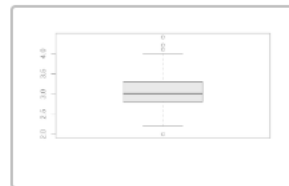
- 异常值（离群点）是指测量数据中的随机错误或偏差，包括错误值或偏离均值的孤立点值。在数据处理中，异常值会极大的影响回归或分类的效果。
- 为了避免异常值造成的损失，需要在数据预处理阶段进行异常值检测。另外，某些情况下，异常值检测也可能是研究的目的，如数据造假的发现、电脑入侵检测等。

箱线图检测离群点(boxplot.stats)

在一条数轴上，以数据的上下四分位数（Q1-Q3）为界画一个矩形盒子（中间50%的数据落在盒内）；在数据的中位数位置画一条线段为中位线；默认延长线不超过盒长的1.5倍，延长线之外的点认为是异常值（用○标记）



```
``{r noise}
y<-boxplot.stats(x[,2], coef=1.5, do.conf=TRUE, do.out=TRUE)
boxplot(x[,2]) #绘制箱线图
y$stats
y$out
``
```



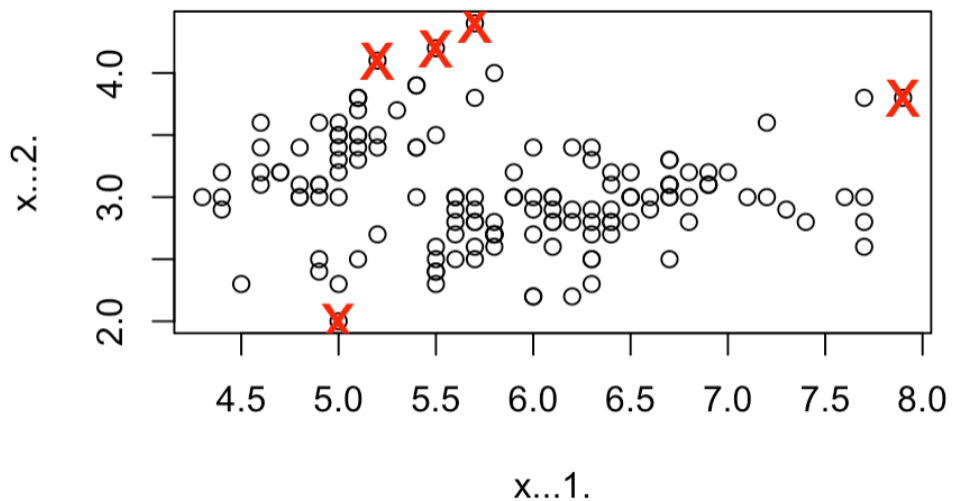
R Console

```
[1] 2.2 2.8 3.0 3.3 4.0
[1] 4.4 4.1 4.2 2.0
```

```
[1] 2.2 2.8 3.0 3.3 4.0
[1] 4.4 4.1 4.2 2.0
```

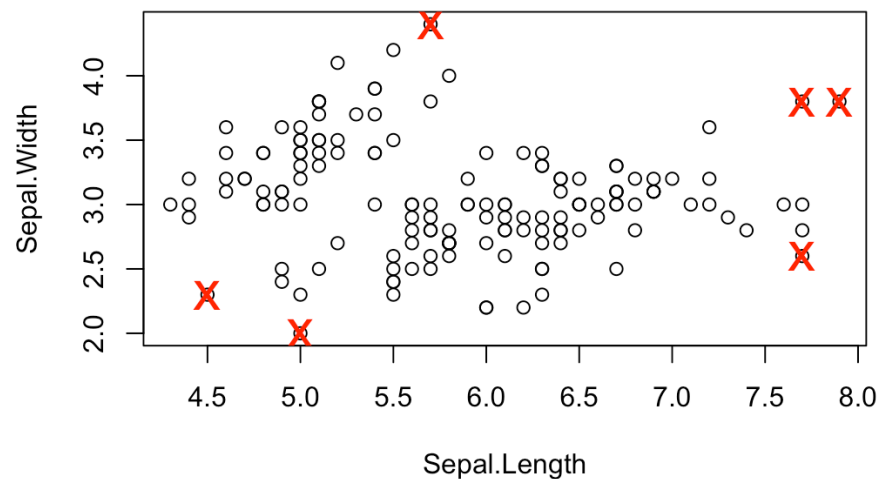
散点图检测离群点(boxplot.stats)

```
`{r outliers-2}  
#寻找a为异常值的坐标位置  
a<-which(x[,2] %in% boxplot.stats(x[,2])$out)  
#寻找b为异常值的坐标位置  
b<-which(x[,1] %in% boxplot.stats(x[,1],coef=1.0)$out)  
df<-data.frame(x[,1], x[,2])  
plot(df) #绘制x, y的散点图  
p2<-union(a,b) #寻找变量x或y为异常值的坐标位置  
points(df[p2,],col="red",pch="x",cex=2) #标记异常值  
`
```



聚类方法检测异常值

```
``{r outliers-3}  
k<-kmeans(iris[,c(1,2)],centers=3) #kmeans聚类为3类  
#k$cluster #输出聚类结果  
#centers返回每个样本对应的聚类中心样本  
centers <- k$centers[k$cluster, ]  
#计算每个样本到其聚类中心的距离  
distances<-sqrt(rowSums((iris[,1:2]-centers)^2))  
#找到距离最大的6个样本，认为是异常值  
out<-order(distances,decreasing=TRUE)[1:6]  
plot(iris[,c(1,2)])  
points(iris[out,c(1,2)],col="red",pch="x",cex=2) #标记异常值  
``
```



数据预处理

- 数据过滤
- 缺失值处理
- 异常值处理
- **数据去重**
- 数据规范
- 数据采样
- 数据排序
- 数据向量化
- 列联表
- 分组汇总

数据去重

- `unique()`
- `duplicated()`

(1) 建立是否重复索引

```
index<-duplicated(data.set$Ensembl)
```

```
index
```

```
[1] FALSE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE
```

(2) 去掉重复行

```
data.set2<-data.set[!index,]
```

数据预处理

- 数据过滤
- 缺失值处理
- 异常值处理
- 数据去重
- **数据规范**
- 数据采样
- 数据排序
- 数据向量化
- 列联表
- 分组汇总

数据规范

(1)数据的中心化

`scale(data, center=T, scale=F)`

$$x = x - \mu$$

(2)数据的标准化

`scale(data, center=T, scale=T)`

$$x = (x - \mu) / \sigma$$

`scale(data, center=F, scale=T)`

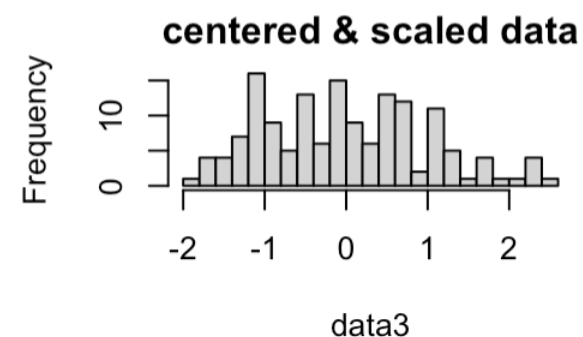
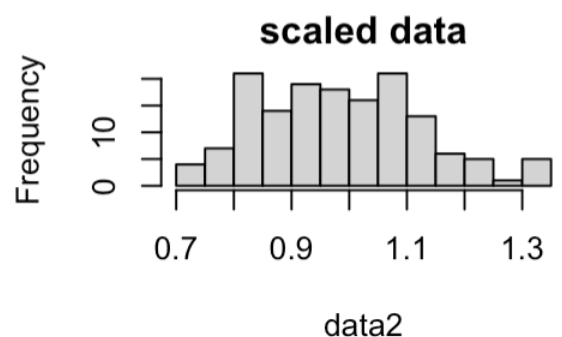
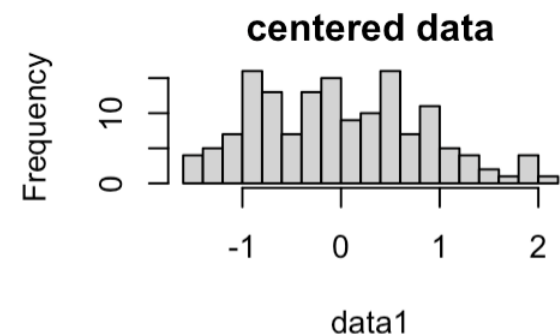
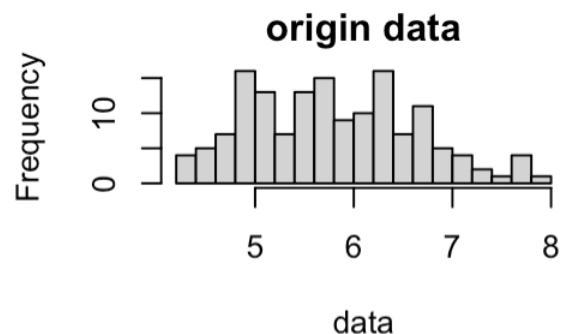
$$x = \frac{x}{\sqrt{\sum x^2 / (n - 1)}}$$

(3)小数定标规范化

移动变量的小数点位置来将变量映射到[-1,1]

`options(digits = 4)` #控制输出结果的有效位数

```
{r scale}
data=iris[,1]
#summary(data)
data1=scale(data,center=T,scale=F)
#summary(data1)
data2=scale(data,center=F,scale=T)
#summary(data2)
data3=scale(data,center=T,scale=T)
#summary(data3)
par(mfrow=c(2,2))
hist(data,main="origin data",breaks = 20)
hist(data1,main="centered data",breaks = 20)
hist(data2,main="scaled data",breaks = 20)
hist(data3,main="centered & scaled data",breaks = 20)
}
```



数据预处理

- 数据过滤
- 缺失值处理
- 异常值处理
- 数据去重
- 数据规范
- **数据采样**
- 数据排序
- 数据向量化
- 列联表
- 分组汇总

数据采样(sample)

```
{r sample}  
set.seed(1234)  
data=data.frame(c1=c(1:10),c2=c(11:20))  
k=nrow(data)  
idx=sample(k,round(0.8*k))  
train=data[idx,]  
(test=data[-idx,])  
`
```

| | c1
<int> | c2
<int> |
|---|--------------------|--------------------|
| 3 | 3 | 13 |
| 9 | 9 | 19 |

2 rows

数据预处理

- 数据过滤
- 缺失值处理
- 异常值处理
- 数据去重
- 数据规范
- 数据采样
- **数据排序**
- 数据向量化
- 列联表
- 分组汇总

数据排序

- `sort()`

```
``{r srot}  
data=data.frame(v1=c(1,3,2,5,4),v2=c('a','c','b','e','d'))  
#data frame对象 含有v1,v2两列  
data[sort(data$v1,index.return=TRUE)$ix,]  
#对data的数据按v1排列,v1须为numeric  
``
```

| | v1
<dbl> | v2
<chr> |
|---|--------------------|--------------------|
| 1 | 1 | a |
| 3 | 2 | b |
| 2 | 3 | c |
| 5 | 4 | d |
| 4 | 5 | e |

5 rows

数据预处理

- 数据过滤
- 缺失值处理
- 异常值处理
- 数据去重
- 数据规范
- 数据采样
- 数据排序
- **数据向量化**
- 列联表
- 分组汇总

数据向量化

- `as.vector()`

1. matrix
2. array

```
``{r as.vector}
(m=matrix(c(1,2,3,4),nrow=2))
as.vector(m)
``
```

```
      [,1] [,2]
[1,]     1     3
[2,]     2     4
[1] 1 2 3 4
```

- `unlist()`

1. list
2. data frame

`unlist` simplifies it to produce a vector which contains all the atomic components which occur in `x`

```
``{r unlist}
l=list(c1=c(1,2,3),c2=c('a','b'),c3=data.frame(c31=c(4,5),c32=c('c','d')))
print(l)
unlist(l)
``
```

```
$c1
[1] 1 2 3
```

```
$c2
[1] "a" "b"
```

```
$c3
  c31 c32
1   4   c
2   5   d
```

| c11 | c12 | c13 | c21 | c22 | c3.c311 | c3.c312 | c3.c321 | c3.c322 |
|-----|-----|-----|-----|-----|---------|---------|---------|---------|
| "1" | "2" | "3" | "a" | "b" | "4" | "5" | "c" | "d" |

数据预处理

- 数据过滤
- 缺失值处理
- 异常值处理
- 数据去重
- 数据规范
- 数据采样
- 数据排序
- 数据向量化
- **列联表**
- 分组汇总

列联表

- table()

| 列联表 | | | |
|-----|-----|-----|-----|
| 员工 | 产品X | 产品Y | 总计 |
| 甲 | 23 | 27 | 50 |
| 乙 | 22 | 33 | 55 |
| 总计 | 45 | 60 | 105 |

```
```{r contingency table}
(v=sample(letters[1:5],10,replace=TRUE))
(t=table(v))
length(t);names(t)
```
```

```
[1] "a" "c" "c" "c" "e" "a" "d" "d" "c" "b"
v
a b c d e
2 1 4 2 1
[1] 5
[1] "a" "b" "c" "d" "e"
```

```
```{r contingency_table_2}
(a<-rep(letters[1:3],each=4))
(b<-sample(LETTERS[1:3],12,replace=T))
(t=table(a,b))
nrow(t);colnames(t);rownames(t)
```
```

```
[1] "a" "a" "a" "a" "b" "b" "b" "b" "c" "c" "c" "c"
[1] "A" "C" "A" "A" "A" "A" "C" "C" "B" "A" "B" "B"
b
a   A B C
a 3 0 1
b 2 0 2
c 1 3 0
[1] 3
[1] "A" "B" "C"
[1] "a" "b" "c"
```

列联表

```
```{r contingency_table_3}
library(vcd)
#风湿性关节炎新疗法的双盲临床实验
#Improved:
#none: 无改善
#some: 一定程度改善
#marked: 显著改善
#Treatment
#Treated: 药物治疗
#Placebo: 安慰剂治疗
head(Arthritis)
```
```

| | ID
<int> | Treatment
<fctr> | Sex
<fctr> | Age
<int> | Improved
<ord> |
|---|-------------|---------------------|---------------|--------------|-------------------|
| 1 | 57 | Treated | Male | 27 | Some |
| 2 | 46 | Treated | Male | 29 | None |
| 3 | 77 | Treated | Male | 30 | None |
| 4 | 17 | Treated | Male | 32 | Marked |
| 5 | 36 | Treated | Male | 46 | Marked |
| 6 | 23 | Treated | Male | 58 | Marked |

```
```{r contingency_table_4}
library(vcd)
(mytable=with(Arthritis,table(Improved)))
prop.table(mytable)
```
```

```
Improved
  None  Some Marked
    42    14    28
Improved
  None      Some   Marked
0.5000000 0.1666667 0.3333333
```

列联表

```
``{r contingency_table_5}  
(mytable=xtabs(~Treatment+Improved,data=Arthritis))  
margin.table(mytable,1)  
margin.table(mytable,2)|  
``
```

| | | Improved | | |
|-----------|---------|----------|--------|--------|
| Treatment | | None | Some | Marked |
| Placebo | | 29 | 7 | 7 |
| Treated | | 13 | 7 | 21 |
| Treatment | | | | |
| Placebo | Treated | | | |
| | | 43 | 41 | |
| Improved | | | | |
| | None | Some | Marked | |
| | 42 | 14 | 28 | |

```
``{r contingency_table_6}  
(mytable=xtabs(~Treatment+Improved,data=Arthritis))  
prop.table(mytable,1)  
prop.table(mytable,2)  
prop.table(mytable)  
``
```

| | | Improved | | |
|-----------|--|----------|------|--------|
| Treatment | | None | Some | Marked |
| Placebo | | 29 | 7 | 7 |
| Treated | | 13 | 7 | 21 |

| | | Improved | | |
|-----------|--|-----------|-----------|-----------|
| Treatment | | None | Some | Marked |
| Placebo | | 0.6744186 | 0.1627907 | 0.1627907 |
| Treated | | 0.3170732 | 0.1707317 | 0.5121951 |

| | | Improved | | |
|-----------|--|-----------|-----------|-----------|
| Treatment | | None | Some | Marked |
| Placebo | | 0.6904762 | 0.5000000 | 0.2500000 |
| Treated | | 0.3095238 | 0.5000000 | 0.7500000 |

| | | Improved | | |
|-----------|--|------------|------------|------------|
| Treatment | | None | Some | Marked |
| Placebo | | 0.34523810 | 0.08333333 | 0.08333333 |
| Treated | | 0.15476190 | 0.08333333 | 0.25000000 |

列联表

```
```{r contingency_table_7}  
(mytable=xtabs(~Treatment+Improved,data=Arthritis))
addmargins(mytable)
addmargins(prop.table(mytable))
```
```

| | Improved | | |
|-----------|----------|------|--------|
| Treatment | None | Some | Marked |
| Placebo | 29 | 7 | 7 |
| Treated | 13 | 7 | 21 |

| | Improved | | | |
|-----------|----------|------|--------|-----|
| Treatment | None | Some | Marked | Sum |
| Placebo | 29 | 7 | 7 | 43 |
| Treated | 13 | 7 | 21 | 41 |
| Sum | 42 | 14 | 28 | 84 |

| | Improved | | | |
|-----------|------------|------------|------------|------------|
| Treatment | None | Some | Marked | Sum |
| Placebo | 0.34523810 | 0.08333333 | 0.08333333 | 0.51190476 |
| Treated | 0.15476190 | 0.08333333 | 0.25000000 | 0.48809524 |
| Sum | 0.50000000 | 0.16666667 | 0.33333333 | 1.00000000 |

```
```{r contingency_table_8}  
(mytable=xtabs(~Treatment+Improved,data=Arthritis))
addmargins(prop.table(mytable,1),2)
addmargins(prop.table(mytable,2),1)
```
```

| | Improved | | |
|-----------|----------|------|--------|
| Treatment | None | Some | Marked |
| Placebo | 29 | 7 | 7 |
| Treated | 13 | 7 | 21 |

| | Improved | | | |
|-----------|-----------|-----------|-----------|-----------|
| Treatment | None | Some | Marked | Sum |
| Placebo | 0.6744186 | 0.1627907 | 0.1627907 | 1.0000000 |
| Treated | 0.3170732 | 0.1707317 | 0.5121951 | 1.0000000 |

| | Improved | | |
|-----------|-----------|-----------|-----------|
| Treatment | None | Some | Marked |
| Placebo | 0.6904762 | 0.5000000 | 0.2500000 |
| Treated | 0.3095238 | 0.5000000 | 0.7500000 |
| Sum | 1.0000000 | 1.0000000 | 1.0000000 |

数据预处理

- 数据过滤
- 缺失值处理
- 异常值处理
- 数据去重
- 数据规范
- 数据采样
- 数据排序
- 数据向量化
- 列联表
- **分组汇总**

分组汇总

- `aggregate(x, by, FUN, ...)`

```
{r aggregate-1}  
table(state.region)  
}
```

```
state.region  
 Northeast      South North Central      West  
           9           16           12           13
```

```
{r aggregate-2}  
aggregate(state.x77, list(Region = state.region), mean)  
}
```

| Region
<fctr> | Population
<dbl> | Income
<dbl> | Illiteracy
<dbl> | Life Exp
<dbl> | Murder
<dbl> | HS Grad
<dbl> | Frost
<dbl> | Area
<dbl> |
|------------------|---------------------|-----------------|---------------------|-------------------|-----------------|------------------|----------------|---------------|
| Northeast | 5495.111 | 4570.222 | 1.000000 | 71.26444 | 4.722222 | 53.96667 | 132.7778 | 18141.00 |
| South | 4208.125 | 4011.938 | 1.737500 | 69.70625 | 10.581250 | 44.34375 | 64.6250 | 54605.12 |
| North Central | 4803.000 | 4611.083 | 0.700000 | 71.76667 | 5.275000 | 54.51667 | 138.8333 | 62652.00 |
| West | 2915.308 | 4702.615 | 1.023077 | 71.23462 | 7.215385 | 62.00000 | 102.1538 | 134463.00 |

4 rows

分组汇总

- `aggregate(x, by, FUN, ...)`

```
{r aggregate-3}  
aggregate(state.x77,  
  list(Region = state.region,  
        Cold = state.x77[, "Frost"] > 130),  
  mean)  
...
```

| Region
<fctr> | Cold
<lgl> | Population
<dbl> | Income
<dbl> | Illiteracy
<dbl> | Life Exp
<dbl> | Murder
<dbl> | HS Grad
<dbl> | Frost
<dbl> | Area
<dbl> |
|------------------|---------------|---------------------|-----------------|---------------------|-------------------|-----------------|------------------|----------------|---------------|
| Northeast | FALSE | 8802.8000 | 4780.400 | 1.1800000 | 71.12800 | 5.580000 | 52.06000 | 110.6000 | 21838.60 |
| South | FALSE | 4208.1250 | 4011.938 | 1.7375000 | 69.70625 | 10.581250 | 44.34375 | 64.6250 | 54605.12 |
| North Central | FALSE | 7233.8333 | 4633.333 | 0.7833333 | 70.95667 | 8.283333 | 53.36667 | 120.0000 | 56736.50 |
| West | FALSE | 4582.5714 | 4550.143 | 1.2571429 | 71.70000 | 6.828571 | 60.11429 | 51.0000 | 91863.71 |
| Northeast | TRUE | 1360.5000 | 4307.500 | 0.7750000 | 71.43500 | 3.650000 | 56.35000 | 160.5000 | 13519.00 |
| North Central | TRUE | 2372.1667 | 4588.833 | 0.6166667 | 72.57667 | 2.266667 | 55.66667 | 157.6667 | 68567.50 |
| West | TRUE | 970.1667 | 4880.500 | 0.7500000 | 70.69167 | 7.666667 | 64.20000 | 161.8333 | 184162.17 |

7 rows

分组汇总

- `aggregate(x, by, FUN, ...)`

```
``{r aggregate-4}
testDF <- data.frame(v1 = c(1,3,5,7,8,3,5,NA,4,5,7,9),
                     v2 = c(11,33,55,77,88,33,55,NA,44,55,77,99) )
by1 <- c("red", "blue", 1, 2, NA, "big", 1, 2, "red", 1, NA, 12)
by2 <- c("wet", "dry", 99, 95, NA, "damp", 95, 99, "red", 99, NA, NA)
aggregate(x = testDF, by = list(by1, by2), FUN = "mean")
``
```

| Group.1
<chr> | Group.2
<chr> | v1
<dbl> | v2
<dbl> |
|------------------|------------------|-------------|-------------|
| 1 | 95 | 5 | 55 |
| 2 | 95 | 7 | 77 |
| 1 | 99 | 5 | 55 |
| 2 | 99 | NA | NA |
| big | damp | 3 | 33 |
| blue | dry | 3 | 33 |
| red | red | 4 | 44 |
| red | wet | 1 | 11 |

8 rows

```
``{r aggregate-5}
fby1 <- factor(by1, exclude = "")
fby2 <- factor(by2, exclude = "")
aggregate(x = testDF, by = list(fby1, fby2), FUN = "mean")
``
```

| Group.1
<fctr> | Group.2
<fctr> | v1
<dbl> | v2
<dbl> |
|-------------------|-------------------|-------------|-------------|
| 1 | 95 | 5.0 | 55.0 |
| 2 | 95 | 7.0 | 77.0 |
| 1 | 99 | 5.0 | 55.0 |
| 2 | 99 | NA | NA |
| big | damp | 3.0 | 33.0 |
| blue | dry | 3.0 | 33.0 |
| red | red | 4.0 | 44.0 |
| red | wet | 1.0 | 11.0 |
| 12 | NA | 9.0 | 99.0 |
| NA | NA | 7.5 | 82.5 |

1-10 of 10 rows

分组汇总

- `aggregate(formula, data, FUN, ...)`

```
{r aggregate-one~one}  
aggregate(weight ~ feed, data = chickwts, mean)
```

| feed
<fctr> | weight
<dbl> |
|----------------|-----------------|
| casein | 323.5833 |
| horsebean | 160.2000 |
| linseed | 218.7500 |
| meatmeal | 276.9091 |
| soybean | 246.4286 |
| sunflower | 328.9167 |

6 rows

```
{r aggregate-many~one}  
aggregate(cbind(Ozone, Temp) ~ Month, data = airquality, mean)
```

| Month
<int> | Ozone
<dbl> | Temp
<dbl> |
|----------------|----------------|---------------|
| 5 | 23.61538 | 66.73077 |
| 6 | 29.44444 | 78.22222 |
| 7 | 59.11538 | 83.88462 |
| 8 | 59.96154 | 83.96154 |
| 9 | 31.44828 | 76.89655 |

5 rows

```
{r aggregate-one~many}  
aggregate(breaks ~ wool + tension, data = warpbreaks, mean)
```

| wool
<fctr> | tension
<fctr> | breaks
<dbl> |
|----------------|-------------------|-----------------|
| A | L | 44.55556 |
| B | L | 28.22222 |
| A | M | 24.00000 |
| B | M | 28.77778 |
| A | H | 24.55556 |
| B | H | 18.77778 |

6 rows

```
{r aggregate-many~many}  
aggregate(cbind(ncases, ncontrols) ~ alcgp + tobgp, data = esoph, sum)
```

| alcgp
<ord> | tobgp
<ord> | ncases
<dbl> | ncontrols
<dbl> |
|----------------|----------------|-----------------|--------------------|
| 0-39g/day | 0-9g/day | 9 | 261 |
| 40-79 | 0-9g/day | 34 | 179 |
| 80-119 | 0-9g/day | 19 | 61 |
| 120+ | 0-9g/day | 16 | 24 |
| 0-39g/day | 10-19 | 10 | 84 |
| 40-79 | 10-19 | 17 | 85 |
| 80-119 | 10-19 | 19 | 49 |

数据预处理

- 数据过滤
- 缺失值处理
- 异常值处理
- 数据去重
- 数据规范
- 数据采样
- 数据排序
- 数据向量化
- 列联表
- 分组汇总



Thanks



高 珍
同济大学
计算机科学与技术学院
(gaozhen@tongji.edu.cn)

