Easy Data.Frame

Smart Labels for R

Huashan Chen

2015年12月19日

第1节 Easy Data.Frame

ezdf 包的目的是使 R 支持类似 SPSS 或 Stata 那样对用户友好的标签输出。ezdf 包并不是要定义一套新的制表函数，而是控制相关制表函数（如 pander）在输出时，能够自动带上对应的标签。除此之外，ezdf 也封装了几个常用的制表方法。

众所周知，在 R 的体系当中，并无变量标签或者数值标签的定义。对于类别变量，在 R 中使用 factor 类型可起到部分标签的功能。对于变量标签，在data.frame 中尽管可以直接使用标签来命名变量，例如 df$年龄，但是实际使用中多有不便。

在 R 中导入 SPSS 或 Stata 等传统统计软件的数据格式可有多个包来实现，例如 foreign、readStata13、haven、sas7bdat 等等。这些包在导入数据时，都能保持原数据中所定义的标签。然而所有这些包目前来说各有优缺点，即使对同一个格式也做不到支持各个版本的导入，因此难以提供一揽子解决方案。更重要的，各个包导入数据之后所定义的标签属性各不相同，导致对标签的使用难以统一。更不用说，在制作表格或者统计结果输出时，能够让 R 做到标签友好。

ezdf 包目前支持采用 foreign 或 haven 导入 Stata 以及 SPSS 格式数据。在 haven 包中，导入数据后，变量标签的定义位于数据库每列字段的 label 属性当中，数值标签的定义在每列字段的 labels 属性当中。在 foreign 包中，导入数据后，变量标签的定义在数据框的 variable.labels 属性当中，数值标签则由参数 use.value.labels 控制是否转化为 factor 类型。

第2节 导入数据

2.1 导入 Stata 数据

导入 Stata 数据使用 `readStata()` 函数:

```
library(ezdf)
dat <- readStata('CGSS2013（居民问卷）发布版_2014.dta', encoding
= 'GB2312')
# View(dat)
```

参数 `encoding` 设置 Stata 标签的编码，该参数默认值为 UTF-8。有的 Stata 数据对变量名以及字符变量（string）的值都采用不同编码，对于这种情况，需分别设置 `varNameEncoding` 和 `charEncoding`。

2.2 导入 SPSS 数据

导入 SPSS 数据使用 `readSPSS()` 函数:

```
readSPSS(file, lib = "foreign", ...)
```

参数 `lib` 设置导入所使用的 R 包，目前支持 `foreign` 和 `haven`。

2.3 将 data.frame 转换为 ezdf

若要创建一个新的 `ez.data.frame` 对象，可用 `as.ez(dt, meta)` ，例如:

```
data(iris)
library(ezdf)
d1 = as.ez(iris)
class(d1)
## [1] "ez.data.frame" "data.table"    "data.frame"
```

第3节 标签的存储

3.1 变量标签

Easy data.frame 定义了一个继承自 `data.frame` 的类 `ez.data.frame`，

并将变量标签存储在 meta 属性当中。meta 可为 data.frame 或 matrix 类，且至少包括两列字段，其中第一列为变量名，第二列为变量标签。

示例的 iris 数据集并不包含任何标签信息，可以用 setmeta() 方法设置自定义的标签。 meta 参数必须是至少包括两列的 data.frame 对象，其第一列被当作变量名，第二列被当作变量标签。setmeta() 将清除数据框中原有的变量标签设置，并返回一个 ezdf 对象。

```
d1$test = sample(5, size = nrow(iris), replace = T)
setmeta(d1, data.frame(var= 'test', lbl = 'Test VAR'))
attr(d1, 'meta')
##     var      lbl
## 1: test Test VAR
```

若要获取数据框中某些变量的标签，可用 varLabels() 方法。

```
varLabels(d1, c('Species', 'test'))
## [1] ""         "Test VAR"

# 用 default = "var" 控制空标签的输出为变量名。该参数默认值为 "",
即输出空标签。
varLabels(d1, c('Species', 'test'), default = "var")
## [1] "Species"  "Test VAR"

# 返回所有已定义的变量标签
varLabels(d1)
##     var      lbl
## 1: test Test VAR

# 设置变量标签
varLabels(d1, "test") <- "New Label"
varLabels(d1, "test")
## [1] "New Label"
```

3.2 数值标签

数值标签的存储采用 haven 的格式，即数据框中的数值型字段可有一个属性 labels，其数据类型为命名的数值型，例如

```
c(V1 = 1, V2 = 2, V3 = 3, MI = 9)
```

注意，对非数值型变量（例如：字符型、factor、逻辑型等）设置 `labels` 属性，在表格输出时`ezdf` 不会对其进行自动输出。

数值标签的获取和设置使用 `valueLabels()` 方法，例如：

```
vl1 = valueLabels(d1, 'test')

# 数值标签可以"加减"
vl2 = vl1 + c("Class1"=1, "Class2"=2, "Class3"=3, MI=9, MM=8)
valueLabels(d1, 'test') = vl2
```

第4节 制表函数

4.1 tbl()

生成长表统计量，命令为 `tbl(ez, expr, func = 'mean', N = FALSE, sort = TRUE)`，其中 `expr` 为 `formula` 对象，当公式右端只有一个自变量时，忽略 `func` 参数，输出频次统计量；`func` 为需要采用的统计量，例如 `mean`、`sum`、`var`。

```
tbl(d1, ~test)
##    test\nNew Label  N
## 1:       1++Class1 25
## 2:       2++Class2 27
## 3:       3++Class3 34
## 4:            4++  36
## 5:            5++  28

# 分组求均值，添加样本数
tbl(d1, Sepal.Length ~ Species + test, 'mean', N = T)
##        Species test\nNew Label Sepal.Length  N
##  1:     setosa       1++Class1     5.069231 13
##  2:     setosa       2++Class2     4.830000 10
##  3:     setosa       3++Class3     5.050000 10
##  4:     setosa            4++      5.155556  9
##  5:     setosa            5++      4.900000  8
##  6: versicolor       1++Class1     6.100000  8
##  7: versicolor       2++Class2     6.030000 10
```

```
##  8: versicolor       3++Class3     5.880000 10
##  9: versicolor            4++      5.812500 16
## 10: versicolor            5++      5.983333  6
## 11:  virginica       1++Class1     6.150000  4
## 12:  virginica       2++Class2     7.000000  7
## 13:  virginica       3++Class3     6.735714 14
## 14:  virginica            4++      6.272727 11
## 15:  virginica            5++      6.607143 14

# tbl() 默认按照公式右端 x 的值排序, 如果取消排序
tbl(d1, Sepal.Length ~ Species + test, 'mean', N = T, sort = F)
##        Species test\nNew Label Sepal.Length  N
##  1:     setosa            5++      4.900000  8
##  2:     setosa       3++Class3     5.050000 10
##  3:     setosa       2++Class2     4.830000 10
##  4:     setosa       1++Class1     5.069231 13
##  5:     setosa            4++      5.155556  9
##  6: versicolor       1++Class1     6.100000  8
##  7: versicolor            4++      5.812500 16
##  8: versicolor       3++Class3     5.880000 10
##  9: versicolor       2++Class2     6.030000 10
## 10: versicolor            5++      5.983333  6
## 11:  virginica            5++      6.607143 14
## 12:  virginica       1++Class1     6.150000  4
## 13:  virginica            4++      6.272727 11
## 14:  virginica       2++Class2     7.000000  7
## 15:  virginica       3++Class3     6.735714 14
```

4.2 ctbl()

ctbl() 是对 table() 的封装, 采用 ctbl(ez, expr) 的调用方式。

```
ctbl(d1, Sepal.Length ~ Species + test)
## , , test = 1
##
##              Species
## Sepal.Length setosa versicolor virginica
##          4.3      0          0         0
```

```
1       ##          4.4      0          0          0
        ##          4.5      0          0          0
        ##          4.6      3          0          0
        ##          4.7      0          0          0
        ##          4.8      2          0          0
        ##          4.9      1          0          0
        ##          5        2          0          0
        ##          5.1      0          0          0
        ##          5.2      0          0          0
        ##          5.3      1          0          0
        ##          5.4      2          1          0
        ##          5.5      0          2          0
        ##          5.6      0          0          0
        ##          5.7      1          0          0
        ##          5.8      1          0          1
        ##          5.9      0          0          0
        ##          6        0          0          0
        ##          6.1      0          1          1
        ##          6.2      0          1          0
        ##          6.3      0          0          1
        ##          6.4      0          1          1
        ##          6.5      0          0          0
        ##          6.6      0          0          0
        ##          6.7      0          1          0
        ##          6.8      0          0          0
        ##          6.9      0          0          0
        ##          7        0          1          0
        ##          7.1      0          0          0
        ##          7.2      0          0          0
        ##          7.3      0          0          0
        ##          7.4      0          0          0
        ##          7.6      0          0          0
        ##          7.7      0          0          0
        ##          7.9      0          0          0
        ##
        ## , , test = 2
        ##
        ##             Species
```

```
2       ## Sepal.Length setosa versicolor virginica
        ##          4.3      1          0         0
        ##          4.4      2          0         0
        ##          4.5      0          0         0
        ##          4.6      0          0         0
        ##          4.7      1          0         0
        ##          4.8      1          0         0
        ##          4.9      1          0         0
        ##          5        2          0         0
        ##          5.1      0          0         0
        ##          5.2      0          0         0
        ##          5.3      0          0         0
        ##          5.4      2          0         0
        ##          5.5      0          0         0
        ##          5.6      0          2         0
        ##          5.7      0          2         0
        ##          5.8      0          0         0
        ##          5.9      0          1         1
        ##          6        0          0         0
        ##          6.1      0          2         0
        ##          6.2      0          1         0
        ##          6.3      0          0         1
        ##          6.4      0          0         0
        ##          6.5      0          0         0
        ##          6.6      0          0         0
        ##          6.7      0          2         1
        ##          6.8      0          0         0
        ##          6.9      0          0         0
        ##          7        0          0         0
        ##          7.1      0          0         0
        ##          7.2      0          0         1
        ##          7.3      0          0         1
        ##          7.4      0          0         0
        ##          7.6      0          0         0
        ##          7.7      0          0         1
        ##          7.9      0          0         1
        ##
        ## , , test = 3
```

```
##
##              Species
## Sepal.Length setosa versicolor virginica
##          4.3      0          0         0
##          4.4      0          0         0
##          4.5      0          0         0
##          4.6      0          0         0
##          4.7      1          0         0
##          4.8      0          0         0
##          4.9      1          0         0
##          5        3          1         0
##          5.1      3          1         0
##          5.2      1          0         0
##          5.3      0          0         0
##          5.4      1          0         0
##          5.5      0          2         0
##          5.6      0          0         0
##          5.7      0          1         1
##          5.8      0          0         0
##          5.9      0          0         0
##          6        0          2         0
##          6.1      0          0         1
##          6.2      0          0         1
##          6.3      0          0         2
##          6.4      0          0         1
##          6.5      0          1         1
##          6.6      0          1         0
##          6.7      0          0         0
##          6.8      0          0         1
##          6.9      0          1         2
##          7        0          0         0
##          7.1      0          0         0
##          7.2      0          0         1
##          7.3      0          0         0
##          7.4      0          0         0
##          7.6      0          0         1
##          7.7      0          0         2
##          7.9      0          0         0
```

```
4       ##
        ## , , test = 4
        ##
        ##              Species
        ## Sepal.Length setosa versicolor virginica
        ##          4.3      0          0         0
        ##          4.4      0          0         0
        ##          4.5      1          0         0
        ##          4.6      0          0         0
        ##          4.7      0          0         0
        ##          4.8      0          0         0
        ##          4.9      0          1         1
        ##          5        1          1         0
        ##          5.1      3          0         0
        ##          5.2      2          1         0
        ##          5.3      0          0         0
        ##          5.4      0          0         0
        ##          5.5      1          1         0
        ##          5.6      0          2         0
        ##          5.7      1          2         0
        ##          5.8      0          1         1
        ##          5.9      0          1         0
        ##          6        0          1         2
        ##          6.1      0          0         0
        ##          6.2      0          0         1
        ##          6.3      0          3         1
        ##          6.4      0          1         0
        ##          6.5      0          0         1
        ##          6.6      0          0         0
        ##          6.7      0          0         2
        ##          6.8      0          1         1
        ##          6.9      0          0         0
        ##          7        0          0         0
        ##          7.1      0          0         1
        ##          7.2      0          0         0
        ##          7.3      0          0         0
        ##          7.4      0          0         0
        ##          7.6      0          0         0
```

```
5       ##          7.7      0          0         0
        ##          7.9      0          0         0
        ##
        ## , , test = 5
        ##
        ##             Species
        ## Sepal.Length setosa versicolor virginica
        ##          4.3      0          0         0
        ##          4.4      1          0         0
        ##          4.5      0          0         0
        ##          4.6      1          0         0
        ##          4.7      0          0         0
        ##          4.8      2          0         0
        ##          4.9      1          0         0
        ##          5        0          0         0
        ##          5.1      2          0         0
        ##          5.2      0          0         0
        ##          5.3      0          0         0
        ##          5.4      0          0         0
        ##          5.5      1          0         0
        ##          5.6      0          1         1
        ##          5.7      0          0         0
        ##          5.8      0          2         1
        ##          5.9      0          0         0
        ##          6        0          1         0
        ##          6.1      0          1         0
        ##          6.2      0          0         0
        ##          6.3      0          0         1
        ##          6.4      0          0         3
        ##          6.5      0          0         2
        ##          6.6      0          1         0
        ##          6.7      0          0         2
        ##          6.8      0          0         0
        ##          6.9      0          0         1
        ##          7        0          0         0
        ##          7.1      0          0         0
        ##          7.2      0          0         1
        ##          7.3      0          0         0
```

```
6       ##          7.4      0          0         1
        ##          7.6      0          0         0
        ##          7.7      0          0         1
        ##          7.9      0          0         0
        # 等价于
        table(d1$Sepal.Length, d1$Species, d1$test)
        ## , , d1$test = 1
        ##
        ##                d1$Species
        ## d1$Sepal.Length setosa versicolor virginica
        ##             4.3      0          0         0
        ##             4.4      0          0         0
        ##             4.5      0          0         0
        ##             4.6      3          0         0
        ##             4.7      0          0         0
        ##             4.8      2          0         0
        ##             4.9      1          0         0
        ##             5        2          0         0
        ##             5.1      0          0         0
        ##             5.2      0          0         0
        ##             5.3      1          0         0
        ##             5.4      2          1         0
        ##             5.5      0          2         0
        ##             5.6      0          0         0
        ##             5.7      1          0         0
        ##             5.8      1          0         1
        ##             5.9      0          0         0
        ##             6        0          0         0
        ##             6.1      0          1         1
        ##             6.2      0          1         0
        ##             6.3      0          0         1
        ##             6.4      0          1         1
        ##             6.5      0          0         0
        ##             6.6      0          0         0
        ##             6.7      0          1         0
        ##             6.8      0          0         0
        ##             6.9      0          0         0
        ##             7        0          1         0
```

```
7       ##             7.1      0          0         0
        ##             7.2      0          0         0
        ##             7.3      0          0         0
        ##             7.4      0          0         0
        ##             7.6      0          0         0
        ##             7.7      0          0         0
        ##             7.9      0          0         0
        ##
        ## , , d1$test = 2
        ##
        ##                d1$Species
        ## d1$Sepal.Length setosa versicolor virginica
        ##             4.3      1          0         0
        ##             4.4      2          0         0
        ##             4.5      0          0         0
        ##             4.6      0          0         0
        ##             4.7      1          0         0
        ##             4.8      1          0         0
        ##             4.9      1          0         0
        ##             5        2          0         0
        ##             5.1      0          0         0
        ##             5.2      0          0         0
        ##             5.3      0          0         0
        ##             5.4      2          0         0
        ##             5.5      0          0         0
        ##             5.6      0          2         0
        ##             5.7      0          2         0
        ##             5.8      0          0         0
        ##             5.9      0          1         1
        ##             6        0          0         0
        ##             6.1      0          2         0
        ##             6.2      0          1         0
        ##             6.3      0          0         1
        ##             6.4      0          0         0
        ##             6.5      0          0         0
        ##             6.6      0          0         0
        ##             6.7      0          2         1
        ##             6.8      0          0         0
```

```
8       ##             6.9      0          0         0
        ##             7        0          0         0
        ##             7.1      0          0         0
        ##             7.2      0          0         1
        ##             7.3      0          0         1
        ##             7.4      0          0         0
        ##             7.6      0          0         0
        ##             7.7      0          0         1
        ##             7.9      0          0         1
        ##
        ## , , d1$test = 3
        ##
        ##                d1$Species
        ## d1$Sepal.Length setosa versicolor virginica
        ##             4.3      0          0         0
        ##             4.4      0          0         0
        ##             4.5      0          0         0
        ##             4.6      0          0         0
        ##             4.7      1          0         0
        ##             4.8      0          0         0
        ##             4.9      1          0         0
        ##             5        3          1         0
        ##             5.1      3          1         0
        ##             5.2      1          0         0
        ##             5.3      0          0         0
        ##             5.4      1          0         0
        ##             5.5      0          2         0
        ##             5.6      0          0         0
        ##             5.7      0          1         1
        ##             5.8      0          0         0
        ##             5.9      0          0         0
        ##             6        0          2         0
        ##             6.1      0          0         1
        ##             6.2      0          0         1
        ##             6.3      0          0         2
        ##             6.4      0          0         1
        ##             6.5      0          1         1
        ##             6.6      0          1         0
```

```
9       ##             6.7      0          0         0
        ##             6.8      0          0         1
        ##             6.9      0          1         2
        ##             7        0          0         0
        ##             7.1      0          0         0
        ##             7.2      0          0         1
        ##             7.3      0          0         0
        ##             7.4      0          0         0
        ##             7.6      0          0         1
        ##             7.7      0          0         2
        ##             7.9      0          0         0
        ##
        ## , , d1$test = 4
        ##
        ##                d1$Species
        ## d1$Sepal.Length setosa versicolor virginica
        ##             4.3      0          0         0
        ##             4.4      0          0         0
        ##             4.5      1          0         0
        ##             4.6      0          0         0
        ##             4.7      0          0         0
        ##             4.8      0          0         0
        ##             4.9      0          1         1
        ##             5        1          1         0
        ##             5.1      3          0         0
        ##             5.2      2          1         0
        ##             5.3      0          0         0
        ##             5.4      0          0         0
        ##             5.5      1          1         0
        ##             5.6      0          2         0
        ##             5.7      1          2         0
        ##             5.8      0          1         1
        ##             5.9      0          1         0
        ##             6        0          1         2
        ##             6.1      0          0         0
        ##             6.2      0          0         1
        ##             6.3      0          3         1
        ##             6.4      0          1         0
```

```
##             6.5      0          0         1
##             6.6      0          0         0
##             6.7      0          0         2
##             6.8      0          1         1
##             6.9      0          0         0
##             7        0          0         0
##             7.1      0          0         1
##             7.2      0          0         0
##             7.3      0          0         0
##             7.4      0          0         0
##             7.6      0          0         0
##             7.7      0          0         0
##             7.9      0          0         0
##
## , , d1$test = 5
##
##                d1$Species
## d1$Sepal.Length setosa versicolor virginica
##             4.3      0          0         0
##             4.4      1          0         0
##             4.5      0          0         0
##             4.6      1          0         0
##             4.7      0          0         0
##             4.8      2          0         0
##             4.9      1          0         0
##             5        0          0         0
##             5.1      2          0         0
##             5.2      0          0         0
##             5.3      0          0         0
##             5.4      0          0         0
##             5.5      1          0         0
##             5.6      0          1         1
##             5.7      0          0         0
##             5.8      0          2         1
##             5.9      0          0         0
##             6        0          1         0
##             6.1      0          1         0
##             6.2      0          0         0
```

```
##             6.3      0          0         1
##             6.4      0          0         3
##             6.5      0          0         2
##             6.6      0          1         0
##             6.7      0          0         2
##             6.8      0          0         0
##             6.9      0          0         1
##             7        0          0         0
##             7.1      0          0         0
##             7.2      0          0         1
##             7.3      0          0         0
##             7.4      0          0         1
##             7.6      0          0         0
##             7.7      0          0         1
##             7.9      0          0         0
```

4.3 ftable()

ftable.ez.data.frame 方法是对 ftable() 的封装，其调用方式为 ftable(ez, formula, style = 1, prop_margin = 1, ...)。其中 style = 1 输出频次; style = 2 输出百分比，由 prop_margin 参数指定行百分比还是列百分比; style = 3 输出百分比和行加总频次。

```
ftable(d1, Species~test)
##
##             setosa versicolor virginica
##   1++Class1     13          8         4
##   2++Class2     10         10         7
##   3++Class3     10         10        14
##   4++            9         16        11
##   5++            8          6        14
ftable(d1, Species~test, style = 2)
##
##                setosa versicolor virginica
##   1++Class1 0.5200000  0.3200000 0.1600000
##   2++Class2 0.3703704  0.3703704 0.2592593
##   3++Class3 0.2941176  0.2941176 0.4117647
##   4++       0.2500000  0.4444444 0.3055556
```

```
##   5++       0.2857143  0.2142857 0.5000000
(t1 = ftable(d1, Species~test, style = 3))
##              setosa versicolor virginica  N
## 1++Class1 0.5200000  0.3200000 0.1600000 25
## 2++Class2 0.3703704  0.3703704 0.2592593 27
## 3++Class3 0.2941176  0.2941176 0.4117647 34
## 4++       0.2500000  0.4444444 0.3055556 36
## 5++       0.2857143  0.2142857 0.5000000 28
```

第5节 与 pander 包自动结合

pander 是用于 markdown 格式输出的 R 包，提供了非常丰富的表格输出功能。在加载 ezdf 包之后，会自动与 pander 包结合，实现自动标签输出。

```
# pander 输出
library(pander)

pander(t1, ez = d1)
```

	setosa	versicolor	virginica	N
1++Class1	0.52	0.32	0.16	25
2++Class2	0.3704	0.3704	0.2593	27
3++Class3	0.2941	0.2941	0.4118	34
4++	0.25	0.4444	0.3056	36
5++	0.2857	0.2143	0.5	28

pander 与回归结果输出：

```
# 加上数值标签
options('ezdfKeepVal' = T)
pander(tbl(dat, a66 ~ s5a, 'mean'))

# 数值与标签之间分隔符
options('ezdfValueLabelSep' = '=')
pander(tbl(dat, a66 ~ s5a, 'mean'))

# 跑个线性回归
```

```
m1 = lm(a6 ~ a2 + a10, dat)
pander(m1)
```

第6节 选项

目前支持下述三个选项:

- `options('ezdfKeepVal' = T)`
- `options('ezdfValueLabelSep' = '=')`
- `options('ezdfKeepVarName' =  T)`

```
options('ezdfKeepVal' = T)
options('ezdfValueLabelSep' = '=')
options('ezdfKeepVarName' =  F)

tbl(d1, ~test)
##    New Label  N
## 1:  1=Class1 25
## 2:  2=Class2 27
## 3:  3=Class3 34
## 4:       4=  36
## 5:       5=  28
pander(tbl(d1, ~test))
```

Warning in tbl.ez.data.frame(d1, ~test) : No Y variables in expr, output frequencies. Function 'mean' is omitted

New Label	N
1=Class1	25
2=Class2	27
3=Class3	34
4=	36
5=	28

```
options('ezdfKeepVarName' =  T)
options('ezdfValueLabelSep' = '++')
pander(tbl(d1, ~test))
```

Warning in tbl.ez.data.frame(d1, ~test) : No Y variables in expr, output frequencies. Function 'mean' is omitted

test New Label	N
1++Class1	25
2++Class2	27
3++Class3	34
4++	36
5++	28