An {arsenal} of R Functions for Statistical Summaries

https://github.com/mayoverse/arsenal

Ethan Heinzen, Beth Atkinson, Jason Sinnwell

2021-08-26

MAYO CLINIC

Motivation

- Mayo Clinic: 3-site research hospital
- Historically a SAS shop
- Now ~500 R users
- A SAS license negotiation revealed our dependency on SAS
- How to port in-house macros and other SAS procedures to R?

MAYO CLINIC

rlocal

- First started with an internal-only package: `rlocal`
- Contained some "private" functions, some "public" functions
- Decided to separate "public" into a CRAN package

MAYO CLINIC

arsenal

- Goal: mimic and improve SAS functionality "easily"
- v0.1.2 on CRAN Dec 30, 2016
- Major releases Feb 2018, Jan 2019, and March 2019
- Minor releases every couple of months

MAYO CLINIC

Functions

- `tableby()` $\equiv$ `%table()`: create a table 1
- `paired()` $\equiv$ `%paired()`: create a table 1 for paired data
- `modelsum()` $\equiv$ `%modelsum()`: fit models over a set of independent variables
- `freqlist()` $\equiv$ `PROC FREQ`: compile frequency table
- `comparedf()` $\equiv$ `PROC COMPARE`: compare two datasets
- `write2*()` $\approx$ `ODS OUTPUT`: output results to a file

MAYO CLINIC

The dataset

```
library(arsenal)
data(mockstudy)
str(mockstudy)
```

```
## 'data.frame':    1499 obs. of  14 variables:
##  $ case       : int  110754 99706 105271 105001 112263 86205 99508 90158 88989 90515 ...
##  $ age        : int  67 74 50 71 69 56 50 57 51 63 ...
##   ..- attr(*, "label")= chr "Age in Years"
##  $ arm        : chr  "F: FOLFOX" "A: IFL" "A: IFL" "G: IROX" ...
##   ..- attr(*, "label")= chr "Treatment Arm"
##  $ sex        : Factor w/ 2 levels "Male","Female": 1 2 2 2 2 1 1 1 2 1 ...
##  $ race       : chr  "Caucasian" "Caucasian" "Caucasian" "Caucasian" ...
##   ..- attr(*, "label")= chr "Race"
##  $ fu.time    : int  922 270 175 128 233 120 369 421 387 363 ...
##  $ fu.stat    : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ ps         : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ hgb        : num  11.5 10.7 11.1 12.6 13 10.2 13.3 12.1 13.8 12.1 ...
##  $ bmi        : num  25.1 19.5 NA 29.4 26.4 ...
##   ..- attr(*, "label")= chr "Body Mass Index (kg/m^2)"
##  $ alk.phos   : int  160 290 700 771 350 569 162 152 231 492 ...
##  $ ast        : int  35 52 100 68 35 27 16 12 25 18 ...
##  $ mdquality.s: int  NA 1 1 1 NA 1 1 1 1 1 ...
##  $ age.ord    : Ord.factor w/ 8 levels "10-19"<"20-29"<..: 6 7 4 7 6 5 4 5 5 6 ...
```

MAYO CLINIC

tableby

A simple pipe table:

```
tb <- tableby(arm ~ sex + age, data = mockstudy)
summary(tb, text = TRUE)
```

```
##
##
## |             | A: IFL (N=428)   | F: FOLFOX (N=691) | G: IROX (N=380) | Total (N=1499)   | p value|
## |:------------|:---------------:|:-----------------:|:---------------:|:----------------:|-------:|
## |sex          |                  |                   |                 |                  |   0.190|
## |-  Male      |   277 (64.7%)    |    411 (59.5%)    |   228 (60.0%)   |   916 (61.1%)    |        |
## |-  Female    |   151 (35.3%)    |    280 (40.5%)    |   152 (40.0%)   |   583 (38.9%)    |        |
## |Age in Years |                  |                   |                 |                  |   0.614|
## |-  Mean (SD) | 59.673 (11.365)  |  60.301 (11.632)  | 59.763 (11.499) | 59.985 (11.519)  |        |
## |-  Range     | 27.000 - 88.000  |  19.000 - 88.000  | 26.000 - 85.000 | 19.000 - 88.000  |        |
```

MAYO CLINIC

The markdown equivalent:

```
tb <- tableby(arm ~ sex + age, data = mockstudy)
summary(tb)
```

	A: IFL (N=428)	F: FOLFOX (N=691)	G: IROX (N=380)	Total (N=1499)	p value
sex					0.190
Male	277 (64.7%)	411 (59.5%)	228 (60.0%)	916 (61.1%)	
Female	151 (35.3%)	280 (40.5%)	152 (40.0%)	583 (38.9%)	
Age in Years					0.614
Mean (SD)	59.673 (11.365)	60.301 (11.632)	59.763 (11.499)	59.985 (11.519)	
Range	27.000 - 88.000	19.000 - 88.000	26.000 - 85.000	19.000 - 88.000	

Notice that sex (a categorical) is treated differently from age (continuous). `tableby()` supports categoricals (character, logical, factor), numerics, ordered factors, `survival::Surv()` objects, dates, and results from `arsenal::selectall()`.

MAYO CLINIC

tableby

Common requests:

- Change labels: `set_labels()`, `labels()<-`, `labelTranslations=`
- Change summary statistics: `tableby.control()` or inline
- Change statistical test (p-value): `tableby.control()` or inline, or `modpval.tableby()`.
- Change decimal points: `tableby.control()` or inline

MAYO CLINIC

tableby

```
tb <- tableby(arm ~ fe(sex, digits.pct = 0) + notest(age, digits = 1, "median", "q1q3"),
              data = mockstudy)
summary(tb, pfootnote = TRUE)
```

	A: IFL (N=428)	F: FOLFOX (N=691)	G: IROX (N=380)	Total (N=1499)	p value
sex					0.190[1]
Male	277 (65%)	411 (59%)	228 (60%)	916 (61%)	
Female	151 (35%)	280 (41%)	152 (40%)	583 (39%)	
Age in Years					
Median	61.0	61.0	61.0	61.0	
Q1, Q3	53.0, 68.0	52.0, 69.0	52.0, 68.0	52.0, 68.0	

1. Fisher's Exact Test for Count Data

MAYO CLINIC

tableby

Without a by-variable:

```
tb <- tableby( ~ sex + age, data = mockstudy)
summary(tb)
```

	Overall (N=1499)
sex	
Male	916 (61.1%)
Female	583 (38.9%)
Age in Years	
Mean (SD)	59.985 (11.519)
Range	19.000 - 88.000

MAYO CLINIC

tableby

A stratified, subsetted, and multiple endpoint summary:

```
tb <- tableby(list(arm, sex) ~ age, strata = ps, data = mockstudy, subset = ps %in% 0:1)
summary(tb)
```

ps		A: IFL (N=340)	F: FOLFOX (N=521)	G: IROX (N=305)	Total (N=1166)	p value
0	**Age in Years**					0.740
	Mean (SD)	60.101 (10.948)	60.173 (11.096)	59.361 (11.904)	59.935 (11.261)	
	Range	27.000 - 81.000	22.000 - 82.000	26.000 - 85.000	22.000 - 85.000	
1	**Age in Years**					0.582
	Mean (SD)	60.579 (12.026)	61.342 (11.918)	60.081 (11.037)	60.800 (11.721)	
	Range	28.000 - 88.000	26.000 - 88.000	28.000 - 84.000	26.000 - 88.000	

ps		Male (N=720)	Female (N=446)	Total (N=1166)	p value
0	**Age in Years**				0.614
	Mean (SD)	59.757 (11.031)	60.221 (11.637)	59.935 (11.261)	
	Range	27.000 - 85.000	22.000 - 82.000	22.000 - 85.000	
1	**Age in Years**				0.045
	Mean (SD)	61.599 (11.748)	59.500 (11.588)	60.800 (11.721)	
	Range	26.000 - 88.000	28.000 - 88.000	26.000 - 88.000	

MAYO CLINIC

tableby

Other features:

- `as.data.frame()`, `as.data.frame(summary())`
- Subset variables, change the order, delete variable: `[`, `head()`, `tail()`
- Sort by p-value: `sort()`
- Filter by p-value: <, <=, >, >=, etc.
- Merge two tables: `merge()`
- Custom p-values and user statistics

MAYO CLINIC

modelsum

The basic table (modeling `alk.phos ~ arm` and `alk.phos ~ ps`)

```
ms <- modelsum(alk.phos ~ arm + ps, data = mockstudy)
summary(ms)
```

	estimate	std.error	p.value	adj.r.squared	Nmiss
(Intercept)	175.577	6.779	< 0.001	0.001	266
Treatment Arm F: FOLFOX	-13.593	8.715	0.119		
Treatment Arm G: IROX	-2.070	9.842	0.833		
(Intercept)	143.772	4.813	< 0.001	0.046	266
ps	46.719	5.979	< 0.001		

MAYO CLINIC

Add common adjusters:

```
ms <- modelsum(alk.phos ~ arm + ps, data = mockstudy, adjust = ~ sex + age)
summary(ms)
```

	estimate	std.error	p.value	adj.r.squared	Nmiss
(Intercept)	175.548	20.587	< 0.001	-0.001	266
Treatment Arm F: FOLFOX	-13.701	8.730	0.117		
Treatment Arm G: IROX	-2.245	9.860	0.820		
sex Female	3.016	7.521	0.688		
Age in Years	-0.017	0.319	0.956		
(Intercept)	148.391	19.585	< 0.001	0.045	266
ps	46.721	5.987	< 0.001		
sex Female	1.169	7.343	0.874		
Age in Years	-0.084	0.311	0.787		

MAYO CLINIC

Don't show common adjusters:

```
ms <- modelsum(alk.phos ~ arm + ps, data = mockstudy, adjust = ~ sex + age)
summary(ms, show.adjust = FALSE, show.intercept = FALSE)
```

	estimate	std.error	p.value	adj.r.squared	Nmiss
Treatment Arm F: FOLFOX	-13.701	8.730	0.117	-0.001	266
Treatment Arm G: IROX	-2.245	9.860	0.820		
ps	46.721	5.987	< 0.001	0.045	266

MAYO CLINIC

modelsum

Other options:

- Change model "family": Poisson, Binomial, Survival, Negative Binomial, Ordinal, Conditional Logistic, Relative Risk
- Change labels: `labels()<-`, `set_labels()`, `labelTranslations=`
- Change decimal places: `modelsum.control()`
- Change summary statistics: `modelsum.control()`
- `as.data.frame()`, `as.data.frame(summary())`
- Subset variables, change the order, delete variable: `[`, `head()`, `tail()`
- Merge two tables: `merge()`

MAYO CLINIC

freqlist

```
fl <- freqlist(~ sex + arm + ps, data = mockstudy)
summary(fl)
```

sex	Treatment Arm	ps	Freq	Cumulative Freq	Percent	Cumulative Percent
Male	A: IFL	0	122	122	8.14	8.14
		1	101	223	6.74	14.88
		2	8	231	0.53	15.41
		NA	46	277	3.07	18.48
	F: FOLFOX	0	168	445	11.21	29.69
		1	148	593	9.87	39.56
		2	16	609	1.07	40.63
		NA	79	688	5.27	45.90
	G: IROX	0	101	789	6.74	52.64
		1	80	869	5.34	57.97
		2	10	879	0.67	58.64
		NA	37	916	2.47	61.11
Female	A: IFL	0	66	982	4.40	65.51
		1	51	1033	3.40	68.91
		2	11	1044	0.73	69.65
		NA	23	1067	1.53	71.18
	F: FOLFOX	0	110	1177	7.34	78.52
		1	95	1272	6.34	84.86
		2	13	1285	0.87	85.72
		NA	62	1347	4.14	89.86
	G: IROX	0	68	1415	4.54	94.40
		1	56	1471	3.74	98.13
		2	9	1480	0.60	98.73
		NA	19	1499	1.27	100.00

MAYO CLINIC

freqlist

Sorted by frequency:

```
fl <- freqlist(~ sex + arm + ps, data = mockstudy)
summary(sort(fl), dupLabels = TRUE)
```

sex	Treatment Arm	ps	Freq	Cumulative Freq	Percent	Cumulative Percent
Male	A: IFL	2	8	8	0.53	0.53
Female	G: IROX	2	9	17	0.60	1.13
Male	G: IROX	2	10	27	0.67	1.80
Female	A: IFL	2	11	38	0.73	2.54
Female	F: FOLFOX	2	13	51	0.87	3.40
Male	F: FOLFOX	2	16	67	1.07	4.47
Female	G: IROX	NA	19	86	1.27	5.74
Female	A: IFL	NA	23	109	1.53	7.27
Male	G: IROX	NA	37	146	2.47	9.74
Male	A: IFL	NA	46	192	3.07	12.81
Female	A: IFL	1	51	243	3.40	16.21
Female	G: IROX	1	56	299	3.74	19.95
Female	F: FOLFOX	NA	62	361	4.14	24.08
Female	A: IFL	0	66	427	4.40	28.49
Female	G: IROX	0	68	495	4.54	33.02
Male	F: FOLFOX	NA	79	574	5.27	38.29
Male	G: IROX	1	80	654	5.34	43.63
Female	F: FOLFOX	1	95	749	6.34	49.97
Male	A: IFL	1	101	850	6.74	56.70
Male	G: IROX	0	101	951	6.74	63.44
Female	F: FOLFOX	0	110	1061	7.34	70.78
Male	A: IFL	0	122	1183	8.14	78.92
Male	F: FOLFOX	1	148	1331	9.87	88.79
Male	F: FOLFOX	0	168	1499	11.21	100.00

MAYO CLINIC

freqlist

Other options:

- `as.data.frame()`, `as.data.frame(summary())`
- Change labels: `labels()<-`, `set_labels()`, `labelTranslations=`
- Subset variables, change the order, delete variable: `[`, `head()`, `tail()`
- Merge two tables: `merge()`

MAYO CLINIC

comparedf

```
mockstudy2 <- muck_up_mockstudy()
cdf <- comparedf(mockstudy, mockstudy2, by = "case")
cdf
```

```
## Compare Object
##
## Function Call:
## comparedf(x = mockstudy, y = mockstudy2, by = "case")
##
## Shared: 9 non-by variables and 1495 observations.
## Not shared: 7 variables and 4 observations.
##
## Differences found in 3/7 variables compared.
## 3 variables compared have non-identical attributes.
```

```
# summary(cdf)
```

MAYO CLINIC

comparedf

```
cdf <- comparedf(mockstudy, mockstudy2, by = "case", int.as.num = TRUE,
                 factor.as.char = TRUE, tol.vars = "case")
cdf
```

```
## Compare Object
##
## Function Call:
## comparedf(x = mockstudy, y = mockstudy2, by = "case", int.as.num = TRUE,
##     factor.as.char = TRUE, tol.vars = "case")
##
## Shared: 10 non-by variables and 1495 observations.
## Not shared: 5 variables and 4 observations.
##
## Differences found in 5/10 variables compared.
## 4 variables compared have non-identical attributes.
```

MAYO CLINIC

comparedf

```
tail(diffs(cdf))
```

```
##      var.x var.y   case values.x values.y row.x row.y
## 3045   hgb   hgb 112460       NA       -9  1493  1487
## 3046   hgb   hgb 112463       NA       -9  1494  1488
## 3047   hgb   hgb 112484       NA       -9  1497  1491
## 3048   ast   ast  86205       27       36     6     3
## 3049   ast   ast 105271      100       36     3     2
## 3050   ast   ast 110754       35       36     1     1
```

```
diffs(cdf, by.var = TRUE)
```

```
##          var.x       var.y    n NAs
## 1          arm         Arm    0   0
## 2          sex         sex 1495   0
## 3         race        race 1285   0
## 4           ps          ps    1   1
## 5          hgb         hgb  266 266
## 6          bmi         bmi    0   0
## 7     alk.phos    alk.phos    0   0
## 8          ast         ast    3   0
## 9  mdquality.s mdquality.s    0   0
## 10     age.ord     age.ord    0   0
```

MAYO CLINIC

write2

Three main functions: `write2word()`, `write2pdf()`, `write2html()`

Can use other output formats supported by R Markdown.

```
write2pdf(list(
  tb,
  summary(lm(age ~ sex, data = mockstudy)),
  "\\newpage",
  "# My modelsum table",
  ms,
  code.chunk(1 + 1)
), file = "test.pdf")
```

MAYO CLINIC

Resources

Docs: `https://mayoverse.github.io/arsenal/`

Issues: `https://github.com/mayoverse/arsenal/issues/`

This presentation: `https://github.com/eheinzen/2021_arsenal_RMedicine/tree/Recording`

Connect with us on Github: `@eheinzen`, `@bethatkinson`, `@sinnweja`, `https://github.com/mayoverse/`

MAYO CLINIC