

Stata 命令到 R 函数速查表

沈明宏 (mhshenaa@connect.ust.hk)
2024 年 6 月 3 日 更新

本表默认使用RStudio、tidyverse 和 statart 包。本表暂不包括文本变量处理、类别变量处理、时间变量处理、列表（list）处理、矩阵处理等更高级的议题。

```
在 RStudio 中下载 tidyverse 和 statart:  
install.packages("tidyverse")  
install.packages("devtools")  
devtools::install_github("socimh/statart")
```

所有人都有权基于教学、自习、科研等学术目的查看、编辑和分享本文档。**如非授权，禁止使用本文档中的内容进行商业活动。**

Stata 命令	R 函数	功能
基础功能和快捷键		
clear	一般不需要，需要时 rm(data)	清除掉数据
tab 键（需较高版本）	回车键；tab 键（Copilot）	补全当前代码
crtl + D (Win), ctrl + cmd + D (Mac)	cmd + 回车键	运行选中代码
crtl + D (Mac)	crtl + cmd + D (Mac)	复制选中代码
pwd 或直接看左下角	getwd() 或直接看 Console 处	查看工作文件夹
cd "path/to/folder"	setwd("path/to/folder") 或直接在 Files 处更改	更换工作文件夹
dir	dir() 或直接看 Files (Go to Working Directory)	查看工作文件夹中内容
global path "path/to/folder" cd \$path	path <- "path/to/folder" setwd(path)	存储全局宏变量
ssc install package	install.packages("package")	下载用户自定义的包
	devtools::install_github("user/package")	下载用户发布在 Github 的包
无需操作，自动加载	library(package)	加载用户自定义的包
Variables 窗口	vaariables(data) 或 names_as_column(data)	查看变量名列表
Variables 窗口	codebook(tb) variables(tb)	查看变量列表
Variables 窗口输入 x	variables(tb, "x")	根据关键词搜索变量
sysuse auto.dta	data("mtcars")	加载内置的示例数据

Stata 命令	R 函数	功能
use data.dta	很少有 .RData 或 .rds 格式的数据 load("data.RData") data <- readRDS("file.rds")	载入原生数据
菜单栏 File -> Import	菜单栏 File -> Import Dataset 更多内容, 详见我的代码 Week 1-3。	导入外部数据
help command	help(package) 或直接在 Help 窗口搜索	搜索帮助文档
探索数据		
br	view(data)	观察数据表
list	data	打印数据表
比较繁琐	print_headtail(data)	打印数据的开头和结尾几行
比较繁琐	print_interval(data)	打印数据的等间距的几行
des	glimpse(data)	描述数据结构
codebook	codebook_detail(data)	显示详细编码本
codebook, c	codebook(data)	显示简洁编码本
br x 或 list x (个人推荐 browse, 后面全部省略 list 命令)	select(data, x) 或 pull(data, x) 或 data\$x	选中、观察某个变量
br x1 x3-x5	select(data, x1, x3:x5)	选中、观察某些变量
br x* y*	select(data, s_match("x* y*"))	选中、观察某些变量 (批量选择)
br in 1/10	head(data, 10)	选中、观察数据前几行
br in 7/12	slice(data, 7:12)	选中、观察数据某几行
br in -10/-1	tail(data, 10)	选中、观察数据后几行
br if x1 == 1 & x2 == 2	filter(data, x1 == 1 & x2 == 2)	根据某些变量取值, 选中、观察数据特定行
br x*	select(data, starts_with("x"))	找到所有以某些字符开头的变量
br *x	select(data, ends_with("x"))	找到所有以某些字符结尾的变量
br *x*	select(data, contains("x"))	找到所有包含某些字符的变量
br *	select(data, everything())	找到所有变量
tab x fre x	tab(data, x) fre(data, x)	用表格描述单个类别变量取值分布
tab x1 x2	tab2(data, x1, x2) fre2(data, x1, x2)	用表格描述两个类别变量取值分布
table x1 x2 x3	tab(data, x1, x2, x3) (长表) fre(data, x1, x2, x3) (长表)	用表格描述多个类别变量取值分布

Stata 命令	R 函数	功能
tab1 x1 x2 x3	tab1(data, x1, x2, x3) frel(data, x1, x2, x3)	逐个描述变量
sum	summary(data) (不推荐) summ(data)	用统计量描述所有变量
sum x	summ(data, x)	用统计量描述单个变量
sum x, d dis r(p25)	summ(data, x, .stat = "q1")	用单个统计量描述单个变量
bys group: sum x	data %>% group_by(group) %>% summ(x)	用统计量分组描述所有变量
tabstat x1 x2, by(group) stat(mean sd)	data %>% group_by(group) %>% summ(x1, x2) 或 data %>% group_by(group) %>% tabstat(x1, x2, .stat = c("mean", "sd"))	用统计量分组描述若干变量
collapse (count) x1 x2 (mean) x1_avg = x1 x2_avg = x2, by(group)	data <- data %>% group_by(group) %>% summ(x1, x2)	用统计量分组描述若干变量, 并替换原数据
变量和数值标签		
lab def gender 1 male 2 female lab val gender gender	data <- data %>% mutate(x = factor(x, labels = c("male", "female"))	给数值变量的所有取值 贴上标签
lab def x 9 "Unknown" 10 "Refused" lab val x x	不推荐在 R 中使用这种数值标签。 data <- data %>% mutate(x = haven::labelled(x, c(Unknown = 9, Refused = 10))	给数值变量的部分取值 贴上标签
lab var x "Variable Name"	R 用户通常不看变量标签。 data <- data %>% mutate(x = haven::labelled(x, label = "Variable Name")	贴上变量标签
构造和修改变量		
clonevar x2 = x	data <- data %>% mutate(x2 = x)	克隆一个变量
gen x2 = x^2	data <- data %>% mutate(x2 = x^2)	生成新变量
gen x = _n	data <- data %>% mutate(x = row_number())	生成第几行变量

Stata 命令	R 函数	功能
gen x2 = x^2 if x > 0	data <- data %>% mutate(x2 = ifelse(x > 0, x^2, NA))	根据条件生成新变量
gen x2 = x^2 in 11/20	data <- data %>% mutate(x2 = ifelse(row_number() %in% 11:20, x^2, NA))	根据位置生成新变量
replace x = x^2	data <- data %>% mutate(x = x^2)	修改旧变量
replace x = . if x == 1	data <- data %>% mutate(x = na_if_value(x, 1))	把单个值变成缺失
replace x = 1 if x == .	data <- data %>% mutate(x = value_if_na(x, 1))	把缺失值变成某个值
replace x = x^2 if x > 0	data <- data %>% mutate(x = ifelse(x > 0, x^2, x))	根据条件修改旧变量
replace x = x^2 in 11/20	data <- data %>% mutate(x = ifelse(row_number() %in% 11:20, x^2, x))	根据位置修改旧变量
mark dummy if x1 == 1 & x2 == 2	data <- data %>% mutate(dummy = x1 == 1 & x2 == 2)	根据条件生成逻辑变量 / 虚拟变量
tab x, gen(x)	data <- data %>% fastDummies::dummy_cols("x")	生成（类别变量）x 所有取值的虚拟变量
recode x 1/5=1 6/10=2 11/15=3 else = ., g(x3g)	data <- data %>% mutate(x3g = cut_breaks(x, breaks = c(6, 11))) 或 data <- data %>% mutate(x3g = case_when(x %in% 1:5 ~ 1, x %in% 6:10 ~ 2, x %in% 11:15 ~ 3, TRUE ~ NA_real_))	重编码变量
encode str_x, g(x)	data <- data %>% mutate(x = factor(str_x))	把文本变量转换成带标签的数值变量
destring str_x, g(num_x)	starwars %>% mutate(num_x = as_numeric(str_x))	把文本变量转换成数值变量
tostring num_x, g(str_x)	starwars %>% mutate(str_x = as_character(num_x))	把数值变量转换成文本变量

Stata 命令	R 函数	功能
egen bins = cut(x), group(10)	data <- data %>% mutate(bins = cut_quantile(x, 10))	根据分位数, 将连续变量均等地分成类别变量
比较繁琐	data <- data %>% mutate(bins = cut_length(x, 10))	根据取值, 将连续变量分成若干类类别变量
egen x_sum = rowtotal(x1-x5)	data <- data %>% mutate(x_sum = row_sum(x1:x5))	通过多个变量生成新变量 (示例一)
egen x_mi = rowmiss(x1-x5)	data <- data %>% mutate(x_mi = row_miss(x1:x5))	通过多个变量生成新变量 (示例二)
egen x_avg = mean(x)	data <- data %>% mutate(x_avg = mean(x, na.rm = TRUE))	通过计算函数生成新变量
gen lag_x = x[_n-1]	data <- data %>% mutate(x = lag(x))	生成滞后变量
gen lead_x = x[_n+2]	data <- data %>% mutate(x = lead(x, 2))	生成前定变量
foreach v of var v1- v5 { gen `v'_sq = `v'^2 }	data <- data %>% mutate(across(v1:v5, ~ .x^2, .names = "{col}_sq"))	对一些变量跑循环 (示例一)
forval i in 1/5 { gen v`i'_sq = v`i'^2 if v`i' > 0 }	data <- data %>% mutate(across(v1:v5, ~ ifelse(.x > 0, .x^2, NA), .names = "{col}_sq"))	对一些变量跑循环 (示例二, 带有条件语句)
forval i = 6/10 { forval j = 3/5 { replace x_`i'`j' = x_`i'`j' + 1 } }	data <- data %>% mutate(across(s_match("x_[6- 10]_[3-5]"), ~ .x + 1))	对一些变量跑循环 (示例三, 变量名中有特定数字)
ren x x2	data <- data %>% rename(x2 = x)	重命名变量
ren *, lower	data <- data %>% rename_with(str_to_lower)	批量重命名变量到小写
筛选变量或个案		
keep x1 x3-x5	data <- data %>% select(x1, x3:x5)	保留多个变量
比较复杂	data <- data %>% select(1, 3:5)	根据变量位置保留变量
drop x	data <- data %>% select(-x) 或	删除单个变量

Stata 命令	R 函数	功能
	<pre>data <- data %>% mutate(x = NULL)</pre>	
drop x1 x3-x5	<pre>data <- data %>% select(-c(x1, x3:x5))</pre>	删除多个变量
比较复杂	<pre>data <- data %>% select(-c(1, 3:5))</pre>	根据变量位置删除变量
keep in 1/10	<pre>data <- data %>% head(10)</pre>	保留数据前几行
keep in 7/12	<pre>data <- data %>% slice(7:12)</pre>	保留数据某几行
keep in -10/-1	<pre>data <- data %>% tail(10)</pre>	保留数据后几行
keep if x1 == 1 & x2 == 2	<pre>data <- data %>% filter(x1 == 1 & x2 == 2)</pre>	根据某些变量取值，保留数据特定行
keep if !missing(x) 或 keep if x < .	<pre>data <- data %>% filter(!is.na(x))</pre>	保留 x 未缺失行
set seed 2024 sample 1000, c	<pre>data <- data %>% slice_sample(n = 1000) %>% set_seed(2024)</pre>	抽取 1000 条个案的随机样本，并确保可重复
set seed 2024 sample 10	<pre>data <- data %>% slice_sample(prop = .1) %>% set_seed(2024)</pre>	抽取 10% 的随机样本，并确保可重复
bys group: ...	<pre>data %>% group_by(group) %>% ...</pre>	分组使用某些命令/函数
bys x1 x2: gen n = _n tab n if n == 1	<pre>summarise(data, n_distinct(x))</pre>	查看某些变量未重复行的数目
duplicates drop x1 x2	<pre>data <- data %>% distinct(x, .keep_all = TRUE)</pre>	只保留某些变量未重复行
use data1.dta merge 1:1 id using data2_id.dta, keep(match) keepusing(id) nogen	<pre>data1 <- data1 %>% semi_join(data2, by = "id")</pre>	根据 id 合并两个数据，且只保留 data2 匹配成功的样本、data1 中的变量。即相当于在 data1 中，根据 id 筛选特定行。
use data1.dta merge 1:1 id using data2_id.dta, keep(master) keepusing(id) nogen	<pre>data1 <- data1 %>% anti_join(data2, by = "id")</pre>	根据 id 合并两个数据，且只保留 data2 中没有的个案、data1 中的变量。即相当于在 data1 中，根据 id 筛选特定行。

Stata 命令	R 函数	功能
排列变量或个案		
order x1-x4 *	data <- data %>% relocate(x1:x4)	把一些变量移到最前面
order x1-x4, b(x5)	data <- data %>% relocate(x1:x4, .before = x5)	把一些变量移到其他位置
sort x1 x2	data <- data %>% arrange(x1, x2)	根据 x1 和 x2 升序排列数据
gsort +x1 -x2	data <- data %>% arrange(x1, -x2)	根据 x1 和 x2 升序或降序排列数据
合并数据		
use data1.dta append using data2.dta	data <- data1 %>% bind_rows(data2)	垂直合并两个数据
use data1.dta merge 1:1 _n using data2.dta, nogen	data <- data1 %>% bind_cols(data2)	根据行的位置，水平合并两个数据
use data1.dta merge 1:1 id using data2.dta, nogen	data <- data1 %>% full_join(data2, by = "id")	根据 id，一对一水平合并两个数据
use data1.dta merge m:1 id using data2.dta, nogen	data <- data1 %>% full_join(data2, by = "id")	根据 id，多对一水平合并两个数据
use data1.dta merge 1:1 id using data2.dta, keep(master match)	data <- data1 %>% left_join(data2, by = "id")	根据 id 合并两个数据，且只保留 data1 中的样本
use data1.dta merge 1:1 id using data2.dta, keep(using match)	data <- data1 %>% right_join(data2, by = "id")	根据 id 合并两个数据，且只保留 data2 中的样本
use data1.dta merge 1:1 id using data2.dta, keep(match)	data <- data1 %>% inner_join(data2, by = "id")	根据 id 合并两个数据，且只保留 data1 和 data2 匹配成功的个案
merge 1:1 id1 id2 using data2.dta, nogen	data <- data1 %>% inner_join(data2, by = c("id1", "id2"))	根据 id1 和 id2 合并两个数据
use data1.dta ren * *_data1 ren id_data1 id save data1.dta, replace	data1 <- data1 %>% rename_with(~ paste0(., "_data1")) %>% rename(id = id_data1) data2 <- data2 %>%	根据 id 合并两个数据，并生成变量名后缀

Stata 命令	R 函数	功能
<pre>use data2.dta, clear ren * *_data2 ren id_data2 id save data2.dta, replace use data1.dta, clear merge 1:1 id using data2.dta, nogen</pre>	<pre>rename_with(~ paste0(., "_data1")) %>% rename(id = id_data1) data <- data1 %>% full_join(data2, by = "id")</pre>	
比较复杂	<pre>data <- data1 %>% full_join(data2, by = "id", suffix = c("_data1", "_data2"))</pre>	根据 id 合并两个数据，并仅对重复变量生成变量名后缀
转换长、宽数据		
<pre>reshape long var, i(id) j(year)</pre>	<pre>data <- data %>% pivot_longer(starts_with("var"), names_to = "year", names_prefix = "var")</pre>	宽数据到长数据（变量名开头相同）
比较复杂	<pre>data <- data %>% pivot_longer(-id)</pre>	宽数据到长数据（任意变量名）
<pre>reshape wide var, i(id) j(year)</pre>	<pre>data <- data %>% pivot_wider(names_from = "year", values_from = "var")</pre>	长数据到宽数据
<pre>xpouse, clear</pre>	<pre>data <- datawizard::data_rotate(data)</pre>	转置数据（调换长和宽）
构造数据		
比较复杂	<pre>newdata <- expand_grid(x1 = 1:5, x2 = c("a", "b", "c"))</pre>	生成一个数据，含有指定变量取值的所有组合方式
比较复杂	<pre>newdata <- data %>% complete(x1, x2)</pre>	生成一个数据，含有若干变量取值的所有组合方式，并保留原数据中其他变量
比较复杂	<pre>newdata <- data %>% expand(x1, x2)</pre>	生成一个数据，含有若干变量取值的所有组合方式，并删除原数据中其他变量
回归模型		
<pre>reg y x</pre>	<pre>data %>% regress(y ~ x)</pre>	OLS 回归

Stata 命令	R 函数	功能
logit y x	data %>% regress(y ~ x, "logit")	Logit 回归
ologit y x	data %>% regress(y ~ x, "ologit")	定序 Logit 回归
mlogit y x	data %>% regress(y ~ x, "mlogit")	多项 Logit 回归
poisson y x	data %>% regress(y ~ x, "poisson")	泊松回归
mixed y x grp:	data %>% regress(y ~ x + (1 grp), "mixed")	混合效应模型；多层次模型；分层线性模型
abc y x	fit <- abc(y ~ x, data) tidy_coef(fit)	更多小众模型
快速画图		
hist x	data %>% s_plot(x)	频数直方图（x 为连续变量）
scatter y x	data %>% s_plot(x, y)	散点图（x 和 y 均为连续变量）
gr bar x	data %>% s_plot(x)	条形图（x 为类别变量）
vioplot y, over(x)	data %>% s_plot(x, y)	小提琴图（x 为类别变量，y 为连续变量）
比较繁琐	data %>% s_plot(x, y)	热力图（x 和 y 均为类别变量）