



BOLD  
THINKERS  
DRIVING  
REAL-WORLD  
IMPACT

# Weight Calibration across Packages

Stas Kolenikov

9/23/2019

# Weight calibration

- Last step in creating analysis weights in survey data files
- Adjusting the weights so that they sum to known population totals in different subgroups (age, sex, race, ethnicity, geography, etc.)
- Desirable to minimize changes from the input weights (probability of selection, nonresponse adjustments, frame integration, etc.)

Deville & Sarndal (1992)

# Contenders

## Stata

- `ipfraking` (Kolenikov 2014, 2019)
- `svyca1` (official Stata)
- `survwgt` (Winter 2002)
- `sreweight` (Pacifico 2014)

## R

- `survey::calibrate()` (Lumley 2010)

## SAS

- `rake_and_trim()` (Izrael, Battaglia, Hoaglin, Frankel, Ball, 2017)

# Out of scope

- SUDAAN PROC WGTADJ, PROC WGTADJX
- Stata ipfweight (Bergmann 2011)
- R library(ReGenesees) (Zardetto 2015)
- R library(ipfr) (Ward, Macfarlane 2019)

# Expectations

- Produce usable results
- Provide weight diagnostics
- Speed
- Fool proof

# Running example



# Running example

CPS 2018 March ASEC data

- estimate control totals based on 13353 adults in CA
- calibrate 8403 adults in TX on
  - sex
  - age (14 categories)
  - race/ethnicity (6 categories)
  - education (5 categories)
  - HH income (9 categories)
  - nativity (3 categories)
  - marital status (6 categories)
  - own vs. rent
  - metro area of TX (23 categories)



BOLD  
THINKERS  
DRIVING  
REAL-WORD  
IMPACT

# Tasks and tests

1. Straight raking
2. Raking with divergent population control totals
3. Raking with bounded weight adjustment ratios [0.3,3]
4. Raking with bounded weight values (2nd and 98th percentile of unrestricted distribution)
5. Linear calibration
6. Linear calibration with trimming
7. (Informative error expected) incorrect specification of control totals



# Performance summary

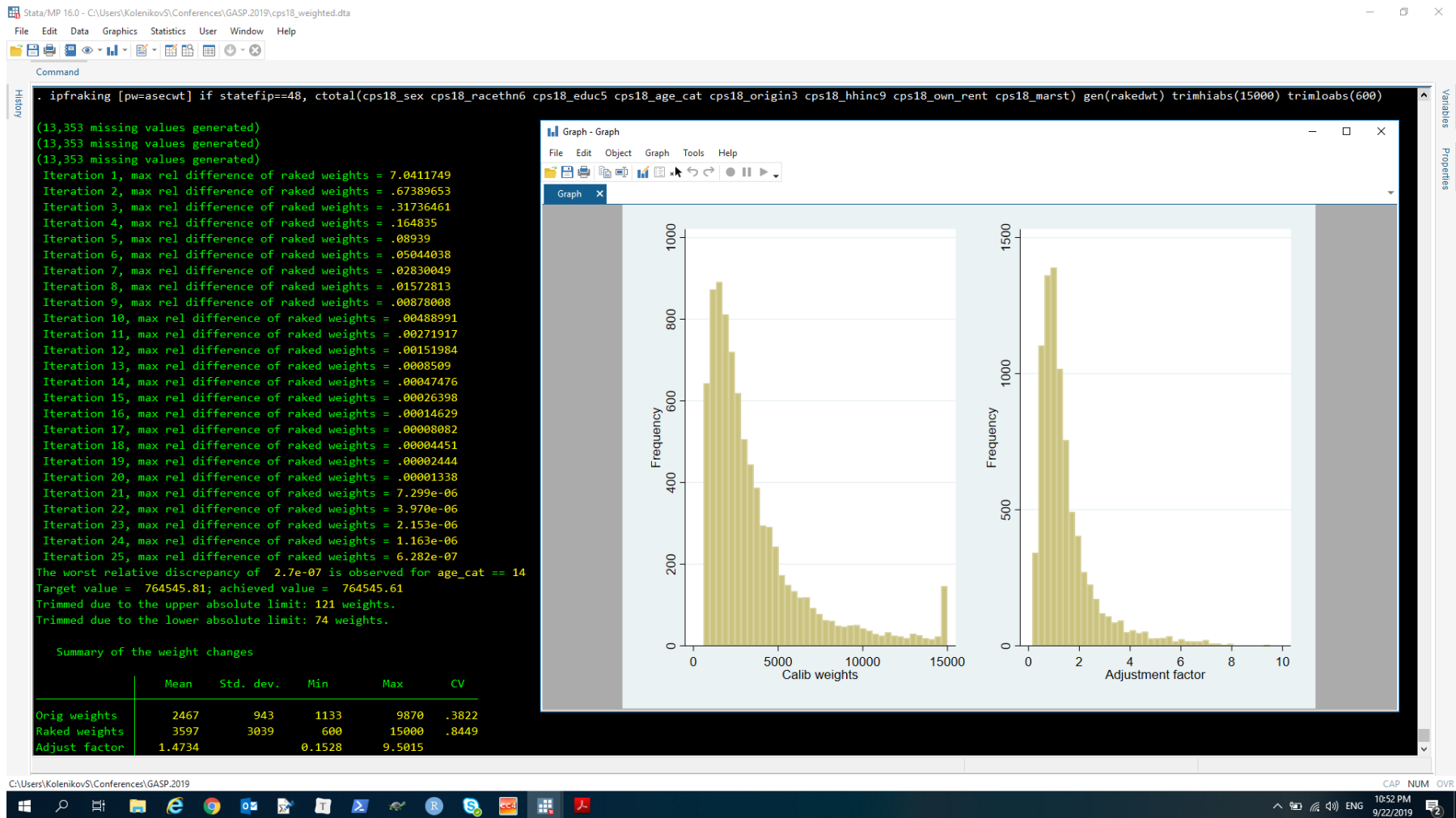
package	TOT	DIV	REL	ABS	LIN	LIN+TR	time
ipfraking	name	+W	+	+	+	N	7.14 sec
svycal	name/=	F	F	N	+	+	0.18 sec
survwgt	order	NW	N	N	N	N	0.80 sec
sreweight	order	F	F	N	+	N	0.19 sec
calibrate	name	-W	..	+	+	+	0.35 sec
rake_and_trim	name+magic	F	-W	+	N	N	61 sec

N: no documented functionality exists

W: issued reasonable warnings

F: failed with cryptic error message / no message

# Stata ipfraking



# Stata svyca1

The image shows a Stata 16.0 command window. The title bar at the top reads "Stata/MP 16.0 - C:\Users\Kolenikov\Conferences\GASP.2019\cps18\_weighted.dta". The menu bar includes File, Edit, Data, Graphics, Statistics, User, Window, and Help. The command window itself has a blue header labeled "Command" and a scrollable area below it. The command entered is: 

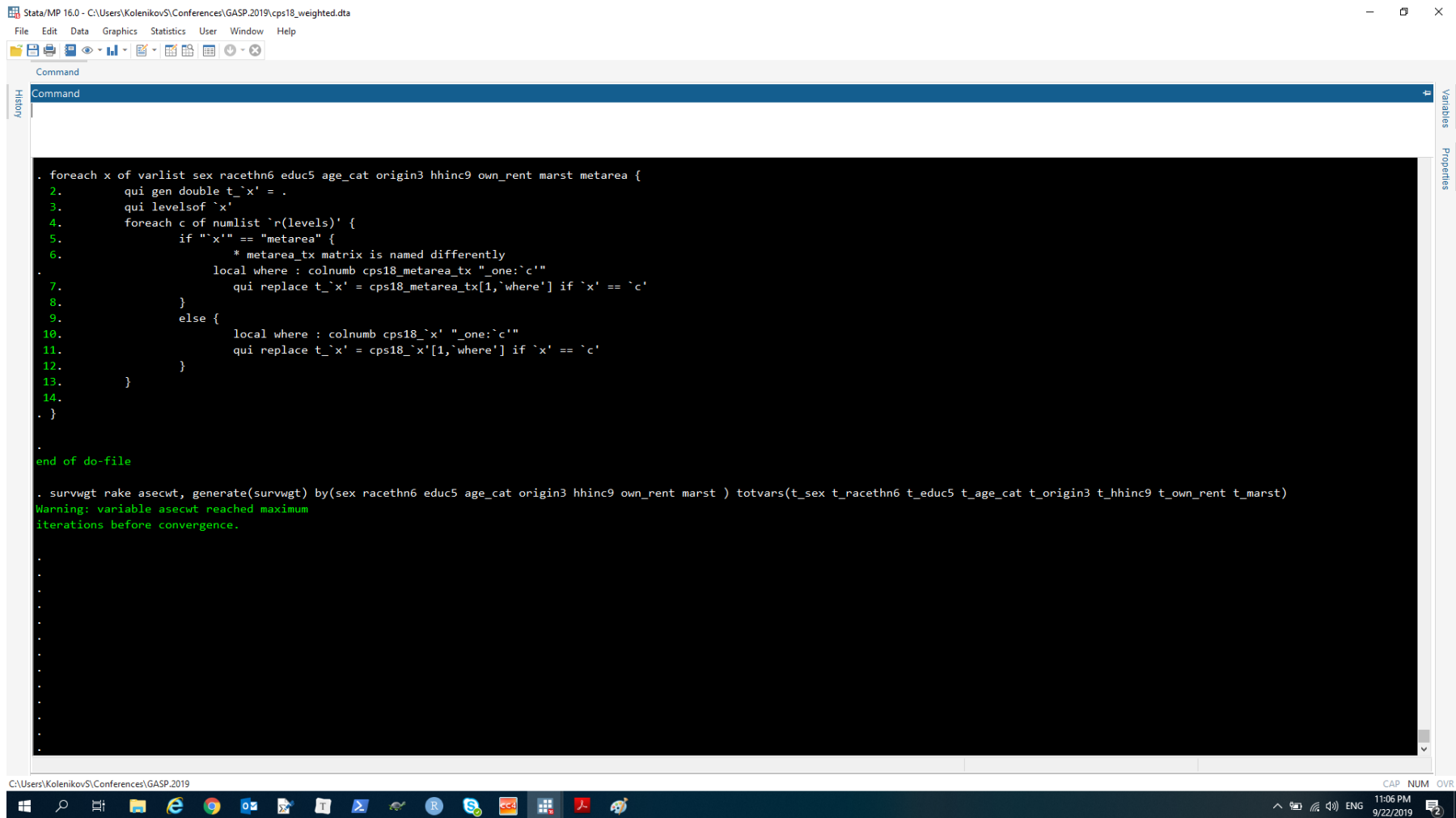
```
. svyset rake ibn.sex ibn.racethn6 ibn.educ5 ibn.age_cat ibn.origin3 ibn.hhinc9 ibn.own_rent ibn.marst ///  
> [pw=asecwt] if statefip==48, generate(svycalwt) nocons totals( ///  
> 1.sex = 14813330.68 2.sex = 15411378.64 ///  
> 1.racethn6 = 12400052.47 2.racethn6 = 1665649.68 3.racethn6 = 4731138.97 ///  
> 4.racethn6 = 335014.83 5.racethn6 = 439771.25 6.racethn6 = 10653082.12 ///  
> 1.educ5 = 4237156.58 2.educ5 = 7062603.95 3.educ5 = 8783383.20 ///  
> 4.educ5 = 6651846.95 5.educ5 = 3489718.64 ///  
> 1.age_cat = 1427812.21 2.age_cat = 2692454.42 3.age_cat = 3071987.71 ///  
> 4.age_cat = 2859233.95 5.age_cat = 2856726.84 6.age_cat = 2557485.26 ///  
> 7.age_cat = 2508571.32 8.age_cat = 2435574.39 9.age_cat = 2394585.24 ///  
> 10.age_cat = 2236585.55 11.age_cat = 1802990.20 ///  
> 12.age_cat = 1268298.67 13.age_cat = 1347857.75 14.age_cat = 764545.81 ///  
> 1.origin3 = 14214474.58 2.origin3 = 5584485.04 3.origin3 = 10425749.7 ///  
> 1.hhinc9 = 2910673.86 2.hhinc9 = 3957861.36 3.hhinc9 = 4113413.54 ///  
> 4.hhinc9 = 3855284.67 5.hhinc9 = 3079011.69 6.hhinc9 = 3056394.29 ///  
> 7.hhinc9 = 2296345.18 8.hhinc9 = 3083671.77 9.hhinc9 = 3872052.96 ///  
> 1.own_rent = 17591393.55 2.own_rent = 12633315.77 ///  
> 1.marst = 14849097.16 2.marst = 487877.77 3.marst = 667292.81 ///  
> 4.marst = 2621033.03 5.marst = 1628860.57 6.marst = 9970547.98 ///  
> )
```

 The output shows several collinearity warnings: 

```
note: 6.racethn6 omitted because of collinearity  
note: 5.educ5 omitted because of collinearity  
note: 14.age_cat omitted because of collinearity  
note: 3.origin3 omitted because of collinearity  
note: 9.hhinc9 omitted because of collinearity  
note: 2.own_rent omitted because of collinearity  
note: 6.marst omitted because of collinearity
```

 The command window also shows the end of the do-file. The taskbar at the bottom shows the Windows Start button, search icon, and several open applications including Stata, Chrome, and Word. The system clock in the bottom right corner shows 11:02 PM on 9/22/2019.

# Stata survwgt



The screenshot shows the Stata 16.0 MP command window with a do-file for creating weights. The code defines a loop over variables sex, race, education, age, origin, income, own rent, marital status, and metarea. It uses the `levelsof` command to get the number of levels for each variable and then uses the `replace` command to create a weight variable `t_x` based on the metarea matrix. The `survwgt` command is used to generate the final weights, and the `rake` command is used to adjust the weights. A warning message indicates that the variable `asecwt` reached the maximum iterations before convergence.

```
. Stata/MP 16.0 - C:\Users\KolenikovS\Conferences\GASP.2019\cps18_weighted.dta
File Edit Data Graphics Statistics User Window Help

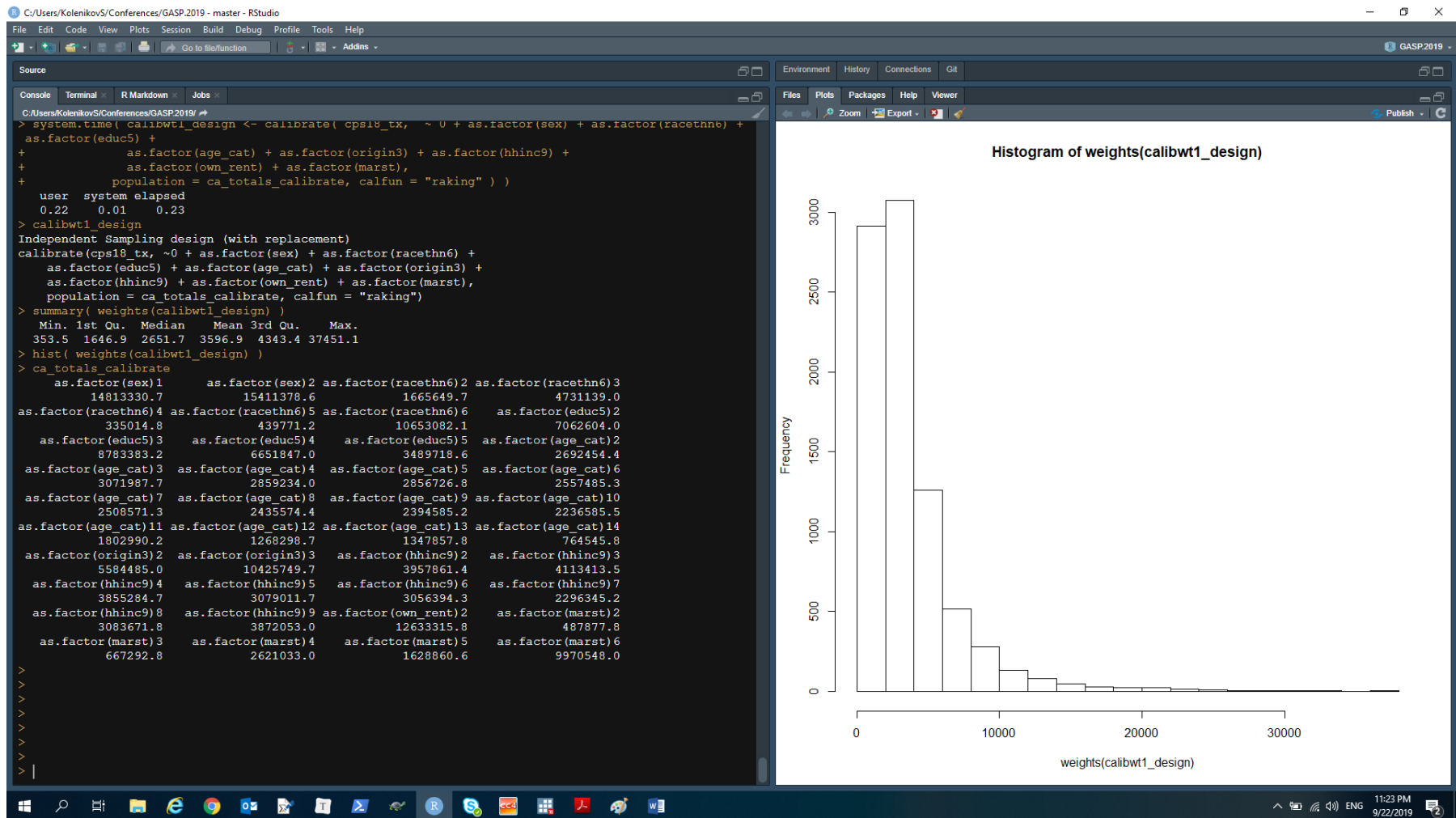
Command
History Variables Properties

. foreach x of varlist sex race6 educ5 age_cat origin3 hhinc9 own_rent marst metarea {
2.     qui gen double t_`x' = .
3.     qui levelsof `x'
4.     foreach c of numlist `r(levels)' {
5.         if "`x'" == "metarea" {
6.             * metarea_tx matrix is named differently
7.             local where : colnumb cps18_metarea_tx "_one:'c'"
8.             qui replace t_`x' = cps18_metarea_tx[1,`where'] if `x' == `c'
9.         }
10.        else {
11.            local where : colnumb cps18_`x' "_one:'c'"
12.            qui replace t_`x' = cps18_`x'[1,`where'] if `x' == `c'
13.        }
14.    }
. }

end of do-file

. survwgt rake asewt, generate(survwgt) by(sex race6 educ5 age_cat origin3 hhinc9 own_rent marst ) totvars(t_sex t_race6 t_educ5 t_age_cat t_origin3 t_hhinc9 t_own_rent t_marst)
Warning: variable asewt reached maximum
iterations before convergence.
```

# R survey::calibrate()



# SAS rake\_and\_trim()

\*\*\*\* Program terminated at iteration 11 because raking converged \*\*\*\*

*The FREQ Procedure*

*Weighted Distribution After Raking*

Sex	Output Weight Sum of Weights	Target Total	Sum of Weights Difference	% of Output Weights	Target % of Weights	Difference in %	Marginal Category Difference in %
1	14813715.70	14813331	385.02	49.012	49.011	0.001	0.003
2	15410993.62	15411379	-385.02	50.988	50.989	-0.001	-0.002

*Weighted Distribution After Raking*

# SAS rake\_and\_trim()

File Explorer window showing the directory structure and file properties for 'rntwt4.rtf, ...'.

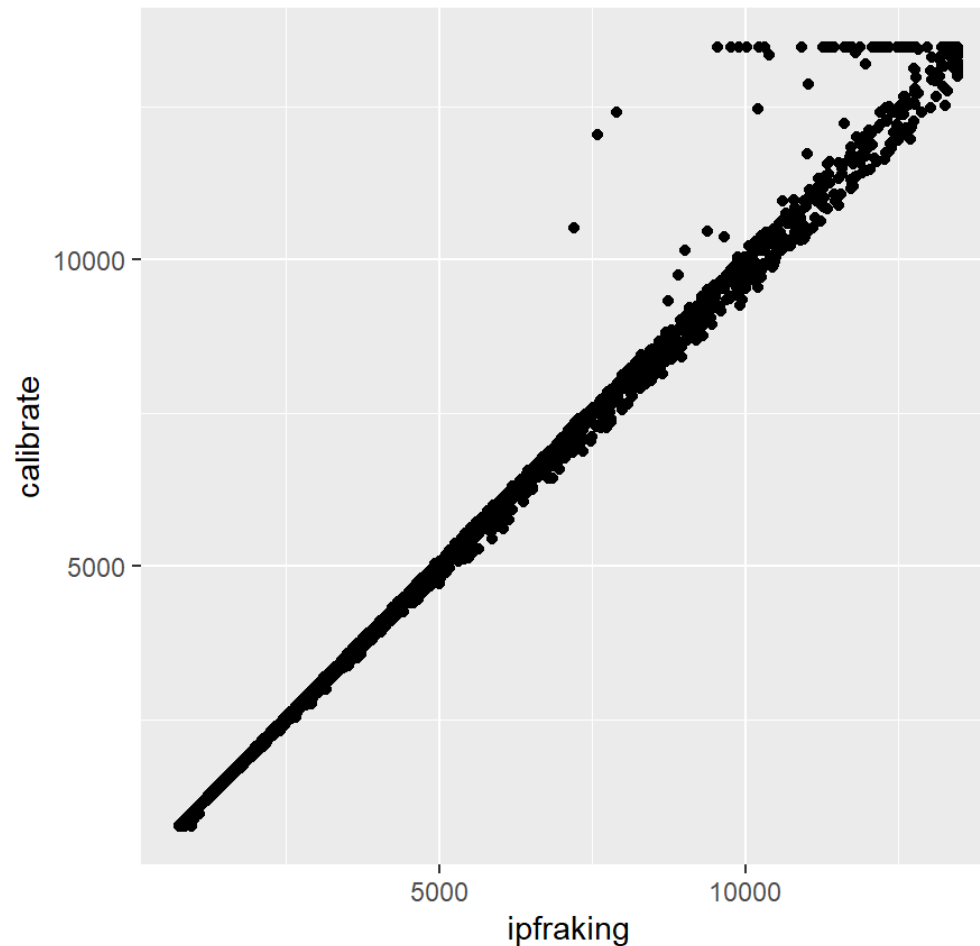
Directory path: Stas Kolenikov > Conferences > GASP.2019 > SAS

Name	Date modified	Type	Size
cps18_tx_wt1.sas7bdat	9/20/2019 6:20 PM	SAS7BDAT File	13,268 KB
cps18_tx_wt2.sas7bdat	9/22/2019 4:31 PM	SAS7BDAT File	13,268 KB
cps18_tx_wt3.sas7bdat	9/22/2019 5:27 PM	SAS7BDAT File	13,696 KB
cps18_tx_wt4.sas7bdat	9/22/2019 5:58 PM	SAS7BDAT File	13,696 KB
rntwt1	9/20/2019 6:20 PM	Log File	2,680 KB
rntwt1.rtf	9/20/2019 6:12 PM	Rich Text Format	1,765 KB
rntwt2	9/22/2019 4:31 PM	Log File	2,308 KB
rntwt2.rtf	9/22/2019 4:31 PM	Rich Text Format	2,698 KB
rntwt3	9/22/2019 5:27 PM	Log File	18,413 KB
rntwt3.rtf	9/22/2019 5:27 PM	Rich Text Format	7,799 KB
rntwt3a	9/22/2019 4:44 PM	Log File	18,413 KB
rntwt3a.rtf	9/22/2019 4:44 PM	Rich Text Format	7,799 KB
rntwt4	9/22/2019 5:58 PM	Log File	6,230 KB
rntwt4.rtf	9/22/2019 5:58 PM	Rich Text Format	2,918 KB

Properties dialog box for 'rntwt4.rtf, ...':

- General tab: 3 Files, 0 Folders
- Type: Multiple Types
- Location: All in C:\Users\KolenikovS\Conferences\GASP.2019
- Size: 22.3 MB (23,391,759 bytes)
- Size on disk: 22.3 MB (23,396,352 bytes)
- Attributes: ☐ Read-only, ☐ Hidden

# Weight trimming $\neq$ methodology



BOLD  
THINKERS  
DRIVING  
REAL-WORLD  
IMPACT



# Misspecified control totals

package	Extra in pop	Extra in data	Wrong order
ipfraking	E	E	
svycal	E	!!!	
survwgt	N/A	N/A	!!!
sreweight	E	E	!!!
calibrate	E	E	
rake_and_trim	!?!?	!!!	

E: issued an error and stopped

!!!: did not issue an error – results highly suspect!

# Thanks and out

Questions?

- [stas\\_kolenikov@abtassoc.com](mailto:stas_kolenikov@abtassoc.com)
- @StatStas on Twitter