

R By within Individual Groups Variables, Averages

Fan Wang

2020-04-01

Contents

1	Nested within Group Stats	1
1.1	Build Function	1
1.2	Test Program	3
1.2.1	Generate Within Individual Groups	3
1.2.2	Within Group Averages	4

1 Nested within Group Stats

Go to the [RMD](#), [R](#), [PDF](#), or [HTML](#) version of this file. Go back to [fan's REconTools Package](#), [R Code Examples Repository \(bookdown site\)](#), or [Intro Stats with R Repository \(bookdown site\)](#).

By Multiple within Individual Groups Variables, Averages for All Numeric Variables within All Groups of All Group Variables (Long to very Wide). Suppose you have an individual level final outcome. The individual is observed for N periods, where each period the inputs differ. What inputs impacted the final outcome?

Suppose we can divide N periods in which the individual is in the data into a number of years, a number of semi-years, a number of quarters, or uneven-staggered lengths. We might want to generate averages across individuals and within each of these different possible groups averages of inputs.

Then we want to version of the data where each row is an individual, one of the variables is the final outcome, and the other variables are these different averages: averages for the 1st, 2nd, 3rd year in which individual is in data, averages for 1st, ..., final quarter in which individual is in data.

1.1 Build Function

This function takes as inputs:

1. **vars.not.groups2avg**: a list of variables that are not the within-individual or across-individual grouping variables, but the variables we want to average over. Within individual grouping averages will be calculated for these variables using the not-listed variables as within individual groups (excluding vars.indi.grp groups).
2. **vars.indi.grp**: a list or individual variables, and also perhaps villages, province, etc id variables that are higher than individual ID. Note the groups are across individual higher level group variables.
3. the remaining variables are all within individual grouping variables.

the function output is a dataframe:

1. each row is an individual
2. initial variables individual ID and across individual groups from *vars.indi.grp*.
3. other variables are all averages for the variables in *vars.not.groups2avg*
 - if there are 2 within individual group variables, and the first has 3 groups (years), the second has 6 groups (semi-years), then there would be 9 average variables.

- each average variables has the original variable name from vars.not.groups2avg plus the name of the within individual grouping variable, and at the end 'c_x', where x is a integer representing the category within the group (if 3 years, x=1, 2, 3)

```
# Data Function
# https://fanwangecon.github.io/R4Econ/summarize/summ/ByGroupsSummWide.html
f.by.groups.summ.wide <- function(df.groups.to.average,
                                vars.not.groups2avg,
                                vars.indi.grp = c('S.country','ID'),
                                display=TRUE) {

  # 1. generate categoricals for full year (m.12), half year (m.6), quarter year (m.4)
  # 2. generate categoricals also for uneven years (m12t14) using
  # stagger (+2 rather than -1)
  # 3. reshape wide to long, so that all categorical date groups appear in var=value,
  # and categories in var=variable
  # 4. calculate mean for all numeric variables for all date groups
  # 5. combine date categorical variable and value, single var:
  # m.12.c1= first year average from m.12 averaging

  #####
  # Step 1
  #####
  # 1. generate categoricals for full year (m.12), half year (m.6), quarter year (m.4)
  # 2. generate categoricals also for uneven years (m12t14) using stagger
  # (+2 rather than -1)

  #####
  # S2: reshape wide to long, so that all categorical date groups appear in var=value,
  # and categories in var=variable; calculate mean for all
  # numeric variables for all date groups
  #####
  df.avg.long <- df.groups.to.average %>%
    gather(variable, value, -one_of(c(vars.indi.grp,
                                      vars.not.groups2avg))) %>%
    group_by(!!!syms(vars.indi.grp), variable, value) %>%
    summarise_if(is.numeric, funs(mean(., na.rm = TRUE)))

  if (display){
    dim(df.avg.long)
    options(repr.matrix.max.rows=10, repr.matrix.max.cols=20)
    print(df.avg.long)
  }

  #####
  # S3 combine date categorical variable and value, single var:
  # m.12.c1= first year average from m.12 averaging; to do this make
  # data even longer first
  #####

  # We already have the averages, but we want them to show up as variables,
  # mean for each group of each variable.
  df.avg.allvars.wide <- df.avg.long %>%
    ungroup() %>%

```

```

mutate(all_m_cate = paste0(variable, '_c', value)) %>%
select(all_m_cate, everything(), -variable, -value) %>%
gather(variable, value, -one_of(vars.indi.grp), -all_m_cate) %>%
unite('var_mcate', variable, all_m_cate) %>%
spread(var_mcate, value)

if (display){
  dim(df.avg.allvars.wide)
  options(repr.matrix.max.rows=10, repr.matrix.max.cols=10)
  print(df.avg.allvars.wide)
}

return(df.avg.allvars.wide)
}

```

1.2 Test Program

In our sample dataset, the number of nutrition/height/income etc information observed within each country and month of age group are different. We have a panel dataset for children observed over different months of age.

We have two key grouping variables: 1. country: data are observed for guatemala and cebu 2. month-age (survey month round=svymthRound): different months of age at which each individual child is observed

A child could be observed for many months, or just a few months. A child's height information could be observed for more months-of-age than nutritional intake information. We eventually want to run regressions where the outcome is height/weight and the input is nutrition. The regressions will be at the month-of-age level. We need to know how many times different variables are observed at the month-of-age level.

```

# Library
library(tidyverse)

# Load Sample Data
setwd('C:/Users/fan/R4Econ/_data/')
df <- read_csv('height_weight.csv')

```

1.2.1 Generate Within Individual Groups

In the data, children are observed for different number of months since birth. We want to calculate quarterly, semi-year, annual, etc average nutritional intakes. First generate these within-individual grouping variables. We can also generate uneven-staggered calendar groups as shown below.

```

mth.var <- 'svymthRound'
df.groups.to.average<- df %>%
  filter(!sym(mth.var) >= 0 & !sym(mth.var) <= 24) %>%
  mutate(m12t24=(floor((!sym(mth.var) - 12) %/% 14) + 1),
         m8t24=(floor((!sym(mth.var) - 8) %/% 18) + 1),
         m12 = pmax((floor((!sym(mth.var)-1) %/% 12) + 1), 1),
         m6 = pmax((floor((!sym(mth.var)-1) %/% 6) + 1), 1),
         m3 = pmax((floor((!sym(mth.var)-1) %/% 3) + 1), 1))

# Show Results
options(repr.matrix.max.rows=30, repr.matrix.max.cols=20)
vars.arrange <- c('S.country', 'indi.id', 'svymthRound')
vars.groups.within.indi <- c('m12t24', 'm8t24', 'm12', 'm6', 'm3')
as.tibble(df.groups.to.average %>%

```

```
group_by(!!!syms(vars.arrange)) %>%
arrange(!!!syms(vars.arrange)) %>%
select(!!!syms(vars.arrange), !!!syms(vars.groups.within.indi)))
```

1.2.2 Within Group Averages

With the within-group averages created, we can generate averages for all variables within these groups.

[illegible]

This is the tabular version of results

```
dim(df.avg.allvars.wide)
```

[1] 2023 38

```
names(df.avg.allvars.wide)
```

```
## [1] "S.country"      "indi.id"        "cal_m12_c1"     "cal_m12_c2"     "cal_m12t24_c0"  "cal_m12t24_c1"
## [9] "cal_m3_c3"      "cal_m3_c4"      "cal_m3_c5"      "cal_m3_c6"      "cal_m3_c7"      "cal_m3_c8"
## [17] "cal_m6_c3"      "cal_m6_c4"      "cal_m8t24_c0"   "cal_m8t24_c1"   "prot_m12_c1"     "prot_m12_c2"
## [25] "prot_m3_c1"     "prot_m3_c2"     "prot_m3_c3"     "prot_m3_c4"     "prot_m3_c5"     "prot_m3_c6"
## [33] "prot_m6_c1"     "prot_m6_c2"     "prot_m6_c3"     "prot_m6_c4"     "prot_m8t24_c0"  "prot_m8t24_c1"
```

```
df.avg.allvars.wide[1:20,] %>% kable() %>% kable_styling_fc_wide()
```

Year	Country	Population (millions)	GDP (billion USD)	GDP per capita (USD)	Life expectancy (years)	Fertility rate (children per woman)	Infant mortality rate (per 1,000 live births)	HDI	Economic Indicators										Social Indicators										Environmental Indicators									
									Export (billion USD)	Import (billion USD)	Trade balance (billion USD)	Government expenditure (billion USD)	Private expenditure (billion USD)	Public debt (billion USD)	Unemployment rate (%)	Urban population (%)	Rural population (%)	Population growth rate (%)	Population density (per sq km)	Urban population (millions)	Rural population (millions)	Population growth rate (%)	Population density (per sq km)	Urban population (millions)	Rural population (millions)	Population growth rate (%)	Population density (per sq km)	Urban population (millions)	Rural population (millions)	Population growth rate (%)	Population density (per sq km)	Urban population (millions)	Rural population (millions)	Population growth rate (%)	Population density (per sq km)	Urban population (millions)	Rural population (millions)	
2019	USA	331.9	21.4	64,560	78.1	1.3	7.1	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2018	USA	328.2	21.0	64,000	77.9	1.3	7.2	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2017	USA	324.5	20.6	63,500	77.7	1.3	7.3	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2016	USA	320.8	20.2	62,800	77.5	1.3	7.4	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2015	USA	317.1	19.8	62,300	77.3	1.3	7.5	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2014	USA	313.4	19.4	61,800	77.1	1.3	7.6	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2013	USA	309.7	19.0	61,300	76.9	1.3	7.7	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2012	USA	306.0	18.6	60,800	76.7	1.3	7.8	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2011	USA	302.3	18.2	60,300	76.5	1.3	7.9	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2010	USA	298.6	17.8	59,800	76.3	1.3	8.0	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2009	USA	295.0	17.4	59,300	76.1	1.3	8.1	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2008	USA	291.4	17.0	58,800	75.9	1.3	8.2	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2007	USA	287.8	16.6	58,300	75.7	1.3	8.3	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2006	USA	284.2	16.2	57,800	75.5	1.3	8.4	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2005	USA	280.6	15.8	57,300	75.3	1.3	8.5	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2004	USA	277.0	15.4	56,800	75.1	1.3	8.6	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2003	USA	273.4	15.0	56,300	74.9	1.3	8.7	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2002	USA	269.8	14.6	55,800	74.7	1.3	8.8	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2001	USA	266.2	14.2	55,300	74.5	1.3	8.9	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
2000	USA	262.6	13.8	54,800	74.3	1.3	9.0	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1999	USA	259.0	13.4	54,300	74.1	1.3	9.1	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1998	USA	255.4	13.0	53,800	73.9	1.3	9.2	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1997	USA	251.8	12.6	53,300	73.7	1.3	9.3	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1996	USA	248.2	12.2	52,800	73.5	1.3	9.4	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1995	USA	244.6	11.8	52,300	73.3	1.3	9.5	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1994	USA	241.0	11.4	51,800	73.1	1.3	9.6	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1993	USA	237.4	11.0	51,300	72.9	1.3	9.7	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1992	USA	233.8	10.6	50,800	72.7	1.3	9.8	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1991	USA	230.2	10.2	50,300	72.5	1.3	9.9	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1990	USA	226.6	9.8	49,800	72.3	1.3	10.0	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1989	USA	223.0	9.4	49,300	72.1	1.3	10.1	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1988	USA	219.4	9.0	48,800	71.9	1.3	10.2	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1987	USA	215.8	8.6	48,300	71.7	1.3	10.3	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1986	USA	212.2	8.2	47,800	71.5	1.3	10.4	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1985	USA	208.6	7.8	47,300	71.3	1.3	10.5	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1984	USA	205.0	7.4	46,800	71.1	1.3	10.6	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1983	USA	201.4	7.0	46,300	70.9	1.3	10.7	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1982	USA	197.8	6.6	45,800	70.7	1.3	10.8	0.92	1,200	1,500	300	4,500	5,000	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500	1,000	1,200	1,500							
1981	USA	194.2	6.2	45,300	70.5	1.3	10.9	0.92	1,200	1,500	300	4,																										