Persistence of incomes

Grattan Insitute Working Paper

HP

Grattan Institute Support

Founding members

Program support

Higher Education

TMF_logo_green.pdf

Affiliate Partners

Google

Origin Foundation

Senior Affiliates

EY

PwC

Stockland

The Scanlon Foundation

Wesfarmers

Affiliates

Ashurst

Corrs

Deloitte

Jacobs

Mercy Health

Sinclair Knight Merz

Urbis

Westpac

Grattan Institute Support

Table of contents

List of Figures

List of boxes

```
library(foreign)
library(ggplot2)
library(scales)
library(grattan)
library(data.table)
library(tidyr)
library(dplyr)
library(magrittr)

weighted.var.se <- function(x, w, na.rm=FALSE){
  # Computes the variance of a weighted mean following Cochran
  # http://stats.stackexchange.com/questions/25895/computing-st
  if (na.rm) { w <- w[i <- !is.na(x)]; x <- x[i] }
  n = length(w)
  xWbar = weighted.mean(x,w,na.rm=na.rm)
  wbar = mean(w)
  out = n/((n-1)*sum(w)^2)*(sum((w*x-wbar*xWbar)^2)-2*xWbar*sum
  return(out)
}
```

```
read_hilda_strip_first_letter_add_column_id <-
  function(filename){
    temp <- fread(filename)
    nms <- names(temp)
    # detect the prefix if it is the wave id.
    nms.prefixes <- unique(gsub("^(.).*$", "\\1", nms))
    yearid <- nms.prefixes[nms.prefixes %in% letters[1:14]]
    # we're interested in stripping the names that specify the
    setnames(temp, old = nms, new = gsub("^[a-n]", "", nms))

    make_negative_NA <- function(x){
      if (is.numeric(x)){
        x[x < 0] <- NA
      }

      x
    }

    wave.decoder <-
      data.table(
        wave = letters[1:14],
        Year = 2000 + 1:14
      ) %>%
      setkey(wave)

    temp <-
      temp %>%
      mutate_each(funs(make_negative_NA)) %>%
      mutate(wave = yearid) %>%
      as.data.table %>%
      setkey(wave) %>%
      merge(wave.decoder) %>%
      mutate(income = tifefp) %>%
      as.data.table

    tmp.svy.quantiles <-
      survey::svydesign(ids = ~xwaveid, strata = ~xhhstrat, weights =
                        data = temp) %>%
      survey::svyquantile(x = ~income, design = ., quantiles = (0:10)/

    quantile.index <- c(1,4:11)

    temp %<>% mutate(tot_inc_percentile_contemporaneous =
                       as.numeric(factor(cut(income,
                                             # deciles.  So choose the
                                             # the eighth, the ninth,
```

```
                                        # This corresponds
                                        # 80-90th percenti
                                        # entiles.
                                        breaks = tmp.svy.q
                                        include.lowest = T
                    tot_inc_percentile_contemporaneous = ifels


                    top_decile = tot_inc_percentile_contempora
                    second_dec = tot_inc_percentile_contempora
                    third_decile = tot_inc_percentile_contempo
                    fourth_decile = tot_inc_percentile_contemp
                    fifth_decile = tot_inc_percentile_contempo
                    sixth_decile = tot_inc_percentile_contempo
                    bottom_deciles = tot_inc_percentile_contem
                    top_quintile = top_decile | second_dec) %>
      as.data.table


    return(temp)
  }

hilda_list <-
  lapply(list.files(path = "../../../Data/HILDA/Wave14/", full.
                    pattern = "^E.*csv$"),  # enumerated person
         read_hilda_strip_first_letter_add_column_id)
```

```
all_hilda <- rbindlist(hilda_list, fill = TRUE)

prop_stay_by_id <-
  all_hilda %>%
  select(
```

```
    xwaveid
    ,wave
    ,lnwte
    ,hhwte
    ,tifefp
    ,tifefn
    ,top_decile
    ,top_quintile                                          le.index)),
  ) %>%
  filter(lnwte != 0) %>%
  tbl_df %>%
  group_by(xwaveid) %>%
  filter(any(top_decile)) %>%
  arrange(wave) %>%
  mutate(cum_topdecile = cumsum(top_decile)) %>%
  group_by(xwaveid) %>%
  mutate(first_enters_decile = lag(cum_topdecile) == 0 & top_decile,
         first_enters_decile_at = as.character(ifelse(first_enters_dec
         has_entered_top_decile = as.logical(cumsum(top_decile)),
         top_quintile_after_top_decile = top_decile | (as.logical(cums
         leaves_decile = top_decile > lead(top_decile),
         leaves_quintile = top_quintile > lead(top_quintile)) %>%  #on
  group_by(xwaveid) %>%
  summarise(potential_time_in_top_decile = sum(has_entered_top_decile)
            years_in_sample = n(),
            time_in_top_decile = sum(top_decile),
            time_in_top_quintile = sum(top_quintile_after_top_decile),
            last_lnwte = last(lnwte))

## Error in .f(.x[[i]], ...): object 'xwaveid' not found

# Note that quite a lot have na valued weights  -
prop_stay_by_id.narm <-
```

```
  prop_stay_by_id %>%
  filter(complete.cases(.))

## Error in eval(lhs, parent, parent): object
'prop_stay_by_id' not found

weighted.mean(prop_stay_by_id.narm$time_in_top_quintile/12,# /
              prop_stay_by_id.narm$last_lnwte)

## Error in weighted.mean(prop_stay_by_id.narm$time_in_top_quin
prop_stay_by_id.narm$last_lnwte): object
'prop_stay_by_id.narm' not found
```

```
all_hilda %>%
  dplyr::select(
    xwaveid
    ,wave
    ,Year
    ,lnwte
    ,hhwte
    ,tifefp
    ,tifefn
    ,income
    ,top_decile
    ,top_quintile
  ) %>%
  filter(lnwte != 0) %>%
  group_by(xwaveid) %>%
  arrange(wave) %>%
  mutate(cum_topdecile = cumsum(top_decile)) %>%
  group_by(xwaveid) %>%
  mutate(first_enters_decile = lag(cum_topdecile) == 0 & top_de
         first_enters_decile_at = as.character(ifelse(first_enters_dec
         has_entered_top_decile = as.logical(cumsum(top_decile)),
         top_quintile_after_top_decile = top_decile | (as.logical(cums
         leaves_decile = top_decile > lead(top_decile),
         leaves_quintile = top_quintile > lead(top_quintile),
         top_deciler = as.logical(max(top_decile)),
         not_top_deciler_but_top_quintile = !top_deciler & max(top_qui
         last_lnwte = last(lnwte)) %>%
  mutate(real_income = cpi_inflator(income, from_fy = yr2fy(Year), to_
  group_by(top_deciler, not_top_deciler_but_top_quintile) %>%
  summarise(mean.income = weighted.mean(real_income, last_lnwte),
            sum(top_decile),
            sd.income = weighted.var.se(real_income, last_lnwte))

## Error in .f(.x[[i]], ...): object 'xwaveid' not found
```

```
all_hilda %>%
  filter(lnwte != 0 | is.na(lnwte)) %>%
  group_by(xwaveid) %>%
  mutate(ever_in_top_decile = sum(top_decile) > 0) %>%
  ungroup %>%
  filter(ever_in_top_decile) %>%
  group_by(xwaveid) %>%
  summarise(years_in_top_decile = sum(top_decile),
            years_in_top_quinti = sum(top_quintile),
            years_in_sample = n(),
            last_lnwte = last(lnwte),
            last_age = last(hgage)) %>%
  ungroup %>%
  filter(!is.na(last_lnwte)) %>%

  summarise(prop_time_top_quintile = weighted.mean(years_in_top_decile
```

```
                                                    last_lnwte))
```

```
## Error in grouped_df_impl(data, unname(vars), drop):
Column `xwaveid` is unknown
```

```
all_hilda %>%
  group_by(xwaveid) %>%
  filter(sum(top_decile) > 0) %>%
  filter(lnwte != 0) %>%
  #mutate(inc_quantile = paste0("Q", tot_inc_percentile_contemp
  as.data.table %>%
  setkey(tot_inc_percentile_contemporaneous) %>%
  merge(data.table(tot_inc_percentile_contemporaneous = 1:10),
  ggplot(aes(x = factor(tot_inc_percentile_contemporaneous))) +
  geom_bar(aes(y = ..count../sum(..count..)))
```

```
## Error in grouped_df_impl(data, unname(vars), drop):
Column `xwaveid` is unknown
```

```
all_hilda %>%
  group_by(xwaveid) %>%
  filter(max(hgage) <= 60,
         min(hgage) >= 30) %>%

  # Exclude xwaveids who never have top_decile
  filter(sum(top_decile) > 0) %>%
  filter(lnwte != 0) %>%
  as.data.table %>%

  # Forces 2:3 to be present in the chart:
  setkey(tot_inc_percentile_contemporaneous) %>%
```

```
  merge(data.table(tot_inc_percentile_contemporaneous = 1:10), all.y =

  grplot(aes(x = factor(tot_inc_percentile_contemporaneous))) +
  geom_bar(aes(y = ..count../sum(..count..),
               weight = lnwte))
```

```
## Error in grouped_df_impl(data, unname(vars), drop):
Column `xwaveid` is unknown
```

```
all_hilda %>%
  group_by(xwaveid) %>%
  filter(max(hgage) <= 60,
         min(hgage) >= 30) %>%
  filter(any(top_decile)) %>%
  filter(lnwte != 0) %>%
  mutate(lnwte_last = last(lnwte)) %>%
  #mutate(inc_quantile = paste0("Q", tot_inc_percentile_contemporaneou
  as.data.table %>% # select(xwaveid, Year, lnwte, top_decile, tot_inc
  ungroup %>%
  group_by(tot_inc_percentile_contemporaneous) %>%
  summarise(time.in = sum(lnwte_last)) %>%
  ungroup %>%
  mutate(time.in = time.in/sum(time.in)) %>%
  arrange(tot_inc_percentile_contemporaneous) %>%
  as.data.table %>%
  setkey(tot_inc_percentile_contemporaneous) %>%
  merge(data.table(tot_inc_percentile_contemporaneous = 1:10), all.y =
  mutate(time.in = ifelse(tot_inc_percentile_contemporaneous %in% 1:3,
                          first(time.in)/3,
                          time.in)) %>%
  # tbl_df
  grplot(aes(x = factor(tot_inc_percentile_contemporaneous),
```

```
            y = time.in)) +
  geom_bar(stat = "identity") +
  scale_y_continuous("Proportion of time spent",
                     label=percent,
                     expand = c(0,0),
                     limits = c(0,0.5)) +
  scale_x_discrete("Contemporaneous total income decile") +
  theme(axis.title.y = element_text(angle = 90, margin = margin(7,7,7,7, "pt")))
```

```
## Error in grouped_df_impl(data, unname(vars), drop):
Column ‘xwaveid‘ is unknown
```

Figure 0.1: For those who reach the top decile, most will stay in the top two deciles

Proportion of years spent in given decile by those respondents aged no more than 60 and no less than 30 who were in the top decile at least once in the HILDA survey period

0.0.1 How many hit top decile only once

```
all_hilda %>%
  group_by(xwaveid) %>%
  filter(max(hgage) <= 60,
         min(hgage) >= 30) %>%
  filter(sum(top_decile) > 0) %>%
  filter(lnwte != 0) %>%
  mutate(lnwte_last = last(lnwte)) %>%
  #mutate(inc_quantile = paste0("Q", tot_inc_percentile_contemporaneous)) %>%
  as.data.table %>% # select(xwaveid, Year, lnwte, top_decile, tot_inc_percentile_contemporaneous)
  ungroup %>%
  group_by(tot_inc_percentile_contemporaneous) %>%
  summarise(time.in = sum(lnwte_last)) %>%
  ungroup %>%
  mutate(time.in = time.in/sum(time.in)) %>%
  arrange(tot_inc_percentile_contemporaneous)
```

```
## Error in grouped_df_impl(data, unname(vars), drop):
Column ‘xwaveid‘ is unknown
```

```
all_hilda %>%
  group_by(xwaveid) %>%
  filter(lnwte != 0) %>%
  mutate(hits_top_decile_only_once = sum(top_decile) == 1,
         never_hits_top_decile = sum(top_decile) == 0,
         prop_in_top_decile = mean(top_decile),
         last_lnwte = last(lnwte)) %>%
  as.data.table %$%
  weighted.mean(hits_top_decile_only_once, last_lnwte)
```

```
## Error in grouped_df_impl(data, unname(vars), drop):
Column ‘xwaveid‘ is unknown
```

0.0.2 Number of times in top decile

```
number_times_top_decile <-
  all_hilda %>%
  filter(lnwte != 0 | is.na(lnwte)) %>%
  group_by(xwaveid) %>%
  mutate(lnwte_last = last(lnwte),
         z = n(),
         sum_topdecile = sum(top_decile)) %>%
  ungroup %>%
  filter(!is.na(lnwte_last)) %>%
```

```
  group_by(sum_topdecile) %>%
  dplyr::summarise(mean.time = sum(lnwte_last)) %>%
  ungroup %>%
  mutate(mean.time = mean.time/sum(mean.time)) %>%
  arrange(sum_topdecile)

## Error in grouped_df_impl(data, unname(vars), drop):
Column ‘xwaveid‘ is unknown

number_times_top_decile

## Error in eval(expr, envir, enclos): object
’number_times_top_decile’ not found
```

```
number_times_top_decile %>%
  grplot(aes(x = factor(sum_topdecile),
             y = mean.time)) +
  geom_bar(stat = "identity") +
  scale_y_continuous("Prop of respondents",
                     label=percent,
                     expand = c(0,0),
                     limits = c(0,1)) +
  xlab("Years in top decile")

## Error in eval(lhs, parent, parent): object
’number_times_top_decile’ not found
```

```
decile_presence_by_top_decile_longevity <-
  all_hilda %>%
  group_by(xwaveid) %>%
```

Figure 0.2
Proportion of respondents in HILDA

Notes:

Source: HILDA

```
  filter(max(hgage) <= 60,
         min(hgage) >= 30) %>%
  group_by(xwaveid) %>%
  mutate(top_deciler = sum(top_decile) > 0,
         sum_topdecile = sum(top_decile),
         last_lnwte = last(lnwte)) %>%
  ungroup %>%
  filter(last_lnwte > 0, !is.na(last_lnwte),
         top_deciler) %>%
  group_by(sum_topdecile) %>%
  summarise(Q10 = weighted.mean(top_decile, last_lnwte),
            #Q10nw = mean(top_decile),
            Q09 = weighted.mean(second_dec,last_lnwte),
            Q08 = weighted.mean(third_decile,last_lnwte),
            Q07 = weighted.mean(fourth_decile,last_lnwte),
            Q06 = weighted.mean(fifth_decile,last_lnwte),
            Q05 = weighted.mean(sixth_decile, last_lnwte),
            QX = weighted.mean(bottom_deciles,last_lnwte),
            n = sum(last_lnwte)) %>%
  arrange(sum_topdecile)

## Error in grouped_df_impl(data, unname(vars), drop):
Column ‘xwaveid‘ is unknown

decile_presence_by_top_decile_longevity %>%
  gather(decile, presence, Q10:QX, factor_key = TRUE) %>%
  arrange(sum_topdecile, desc(decile)) %>%
```

```
  mutate(decile.text = ifelse(sum_topdecile == 1,
                              as.character(decile),
                              NA_character_)) %>%
  mutate(presence = 13 * presence) %>%
  group_by(sum_topdecile) %>%
  arrange(sum_topdecile, decile) %>%
  mutate(text.y = cumsum(presence),
         text.color = ifelse(decile == "Q10", "white", "black")
  ungroup %>%
  arrange(sum_topdecile, decile) %>%
  grplot(aes(x = factor(sum_topdecile),
             y = presence,
             fill = decile)) +
  geom_bar(stat = "identity") +
  xlab("Years in top decile") +
  scale_y_continuous("Years", breaks = 0:13) +
  geom_text(aes(y = text.y, label = decile.text,
                color = text.color),
            vjust = 1.2) +
  scale_color_manual(values = c("white" = "white", "black" = "b
```

```
## Error in eval(lhs, parent, parent): object
'decile_presence_by_top_decile_longevity' not found
```

```
marimekko_data_decile_presence_by_top_decile_longevity <-
  decile_presence_by_top_decile_longevity %>%
  gather(decile, presence, Q10:QX, factor_key = TRUE) %>%
  mutate(subpresence = presence * n) %>%
  arrange(sum_topdecile, decile) %>%
  group_by(decile) %>%
  mutate(cumsum_n = cumsum(n),
         cumsum_lag = lag(cumsum_n, default = 0),
```

7

```
         cumsum_lead = lead(cumsum_n, default = Inf)) %>%
  ungroup %>%
  group_by(sum_topdecile) %>%
  mutate(cumsubpresence = cumsum(presence),
         cumsubpresence_lag = lag(cumsubpresence, default=0),
         cumsubpresence_lead = lead(cumsubpresence, default = 0)) %>%
  ungroup %>%
  mutate(xmin = cumsum_lag/max(cumsum_n),
         xmax = cumsum_n/max(cumsum_n),
         xcenter = (xmin + xmax)/2) %>%
  group_by(decile) %>%
  mutate(cumsubpresence_decile_lag = lag(cumsubpresence, default=0),
         cumsubpresence_decile_lead = lead(cumsubpresence, default = 0
         y_center = cumsubpresence_lag + presence / 2) %>%
  ungroup
```

```
## Error in eval(lhs, parent, parent): object
'decile_presence_by_top_decile_longevity' not found
```

```
marimekko_data_decile_presence_by_top_decile_longevity %>%
{
  grplot(., aes(xmin = xmin,
                xmax = xmax,
                ymin = cumsubpresence_lag,
                ymax = cumsubpresence,
                fill = decile)) +
  geom_rect() +
  geom_segment(aes(x = xmin, xend = xmax,
                   y = cumsubpresence_lag, yend = cumsubpresence_lag),
               color = "black") +
  geom_segment(aes(x = xmax, xend = xmax,
```

```
                  y = cumsubpresence, yend = cumsubpresence_decile_lead),
              color = "black") +
  geom_hline(yintercept = 1) +
  scale_color_manual(values = c("white" = "white", "black" = "black")) +
  scale_y_continuous(expand = c(0,0),
                     label = percent) +
  scale_x_continuous("Years in top decile",
                     expand = c(0,0),
                     breaks = c(unique((.$xmin + .$xmax)/2)),
                     labels = c(paste0(1:14))) +
  scale_y_continuous("Prop time in each decile",
                     label = percent,
                     expand = c(0,0),
                     breaks = c(0:10)/10) +
  coord_equal() +
  annotate("text",
           x = c(rep(min(.$xmax), 5) / 2, 0.38, 0.75),
           y = c(sort(unique(filter(., sum_topdecile == min(sum_topdecile))$y_center), decreasing = TRUE)[1:5], 0.52, 0.3),
           label = c("Bottom 40% of incomes", "5th income decile", paste0(6:8, "th\nincome\ndecile"), paste0(9:10, "th income decile"),
           hjust = c(0.3, rep(0.5, 4), 0.3, 0.3),
           size = 20/(14/5),
           fontface = "bold",
           lineheight = 0.75,
           color = c(rep("black", 5), rep("white", 2)))
}

## Error in eval(lhs, parent, parent): object
'marimekko_data_decile_presence_by_top_decile_longevity'
not found
```

Figure 0.3
Percentage of time cohort was in decile

```
marimekko.data %>%
  mutate(xmax.ppt = 100 * xmax/max(xmin)) %>%
  group_by(sum_topdecile) %>%
  mutate(uncum.y = c(first(ymax), diff(ymax))) %>%
  select(sum_topdecile, decile, xmax.ppt, uncum.y) %>%
  spread(decile, uncum.y) %>%
  write_csv("hilda_marimekko.csv")

## Error in eval(lhs, parent, parent): object
'marimekko.data' not found
```

0.1 Responding person