Data Preparation and Preliminary Descriptive Analysis

Lemi Daba

1/19/2022

The data and documentations can be downloaded from here.

Data Preparation

```r
# Clear working environment and set working directory
rm(list = ls())
# setwd("D:/Doc/paper/Census_2011")

# Load essential libraries
library(haven)
library(tidyverse)
library(janitor)
library(lubridate)
library(AER)
library(stargazer)
library(scales)
library(broom)

theme_set(theme_light())
```

Read in the data:

```r
person_tot <- read_dta("sa-census-2011-person-all-v1.2.dta")
```

```r
person_tot <- person_tot %>%
  clean_names() %>%
  # retain household questionnaires only
  filter(qn_type == 1) %>%
  mutate(dob = make_date(p01_year, p01_month, p01_day)) %>%
  # the census was conducted on 9-10 October, 2011
  mutate(
    age_month = interval(dob, ymd(20111010)) %/% months(1),
    age_year = interval(dob, ymd(20111010)) %/% years(1)
  ) %>%
  select(sn, f00_nr, dob, f03_sex, age_month, age_year, everything())
```

We still have 4,337,697 observations.

Prepare kids' and parents' data separately and join them later:

```r
kids <- person_tot %>%
  # filter those kids whose parents are alive
  filter(p14_motheralive == 1 & p15_fatheralive == 1) %>%
  mutate(
    moth_no = str_c(sn, p14a_motherpnr),
    fath_no = str_c(sn, p15a_fatherpnr)
  ) %>%
  select(
    # other variables need to be added (not final list)
    moth_no,
    fath_no,
    child_dob = dob,
    child_sex = f03_sex,
    child_age_year = age_year,
    child_age_month = age_month,
    child_sch_attend = p17_schoolattend,
    child_educ = p20_edulevel,
    child_private = p19_edupubpriv,
    child_pop_group = p05_pop_group
  )

mothers <- person_tot %>%
  mutate(moth_no = str_c(sn, f00_nr)) %>%
  semi_join(kids, by = "moth_no") %>%
  # other variables need to be added (not final list)
  select(
    moth_no,
    moth_age_year = age_year,
    moth_age_month = age_month,
    moth_dob = dob,
    moth_marital = p03_marital_st,
    moth_pp_group = p05_pop_group,
    moth_educ = p20_edulevel,
    moth_ceb = p32_childeverborn,
    moth_age_fstbr = p33_agefirstbirth,
    moth_employ = derp_employ_status,
    moth_employ_official = derp_employ_status_official,
    moth_employ_extended = derp_employ_status_expanded
  )

fathers <- person_tot %>%
  mutate(fath_no = str_c(sn, f00_nr)) %>%
  semi_join(kids, by = "fath_no") %>%
  # other variables need to be added (not final list)
  select(
    fath_no,
    fath_age_year = age_year,
    fath_age_month = age_month,
    fath_dob = dob,
    fath_marital = p03_marital_st,
    fath_pp_group = p05_pop_group,
    fath_educ = p20_edulevel,
    fath_employ = derp_employ_status,
    fath_employ_official = derp_employ_status_official,
    fath_employ_extended = derp_employ_status_expanded
  )
```

Merge kids' and parents' data:

```
data <- kids %>%
  left_join(mothers, by = "moth_no") %>%
  left_join(fathers, by = "fath_no")
```

Take a look at `data`:

```
data
```

```
## # A tibble: 2,093,472 x 30
##          moth_no     fath_no child_dob  child_sex child_age_year child_age_month
##            <dbl>       <dbl> <date>         <dbl>          <dbl>           <dbl>
##  1  118080376073     1.18e11 2007-04-28         2              4              53
##  2  118080376073     1.18e11 2010-11-09         1              0              11
##  3 1180803770498     1.18e12 1964-01-01         1             47             573
##  4 1180803925698     1.18e12 1988-10-03         1             23             276
##  5 1180803925698     1.18e12 1992-06-16         2             19             231
##  6  118080392562     1.18e11 2011-08-13         2              0               1
##  7  118080388682     1.18e11 2007-06-14         1              4              51
##  8  118080388682     1.18e11 2008-09-03         1              3              37
##  9  118083483952     1.18e11 1997-10-10         1             14             168
## 10  118083483952     1.18e11 1999-07-22         2             12             146
## # ... with 2,093,462 more rows, and 24 more variables: child_sch_attend <dbl>,
## #   child_educ <dbl>, child_private <dbl>, child_pop_group <dbl>,
## #   moth_age_year <dbl>, moth_age_month <dbl>, moth_dob <date>,
## #   moth_marital <dbl>, moth_pp_group <dbl>, moth_educ <dbl>, moth_ceb <dbl>,
## #   moth_age_fstbr <dbl>, moth_employ <dbl>, moth_employ_official <dbl>,
## #   moth_employ_extended <dbl>, fath_age_year <dbl>, fath_age_month <dbl>,
## #   fath_dob <date>, fath_marital <dbl>, fath_pp_group <dbl>, ...
```

Next, get the dob of first-born and second-born kids

```
firstborn_dob <- data %>%
  select(moth_no, child_dob) %>%
  group_by(moth_no) %>%
  arrange(child_dob, .by_group = TRUE) %>%
  slice(1) %>%
  ungroup() %>%
  rename(firstborn_dob = "child_dob") %>%
  mutate(firstborn_age = interval(firstborn_dob, ymd(20111010)) %/% years(1))

secondborn_dob <- data %>%
  select(moth_no, child_dob) %>%
  group_by(moth_no) %>%
  arrange(child_dob, .by_group = TRUE) %>%
  slice(2) %>%
  ungroup() %>%
  rename(secondborn_dob = "child_dob") %>%
  mutate(secondborn_age = interval(secondborn_dob, ymd(20111010)) %/% years(1))
```

Merge these with `data`:

```
data <- data %>%
  left_join(firstborn_dob, by = "moth_no") %>%
  left_join(secondborn_dob, by = "moth_no") %>%
  filter(firstborn_age <= 18)
```

Print the updated `data`:

```r
data
```

```
## # A tibble: 979,233 x 34
##         moth_no      fath_no child_dob  child_sex child_age_year child_age_month
##           <dbl>        <dbl> <date>         <dbl>          <dbl>           <dbl>
##  1 118080376073      1.18e11 2007-04-28         2              4              53
##  2 118080376073      1.18e11 2010-11-09         1              0              11
##  3 118080392562      1.18e11 2011-08-13         2              0               1
##  4 118080388682      1.18e11 2007-06-14         1              4              51
##  5 118080388682      1.18e11 2008-09-03         1              3              37
##  6 118083483952      1.18e11 1997-10-10         1             14             168
##  7 118083483952      1.18e11 1999-07-22         2             12             146
##  8 118097426731      1.18e12 1998-06-16         1             13             159
##  9 118097426731      1.18e12 2003-12-25         2              7              93
## 10 118097426731      1.18e12 2006-09-22         1              5              60
## # ... with 979,223 more rows, and 28 more variables: child_sch_attend <dbl>,
## #   child_educ <dbl>, child_private <dbl>, child_pop_group <dbl>,
## #   moth_age_year <dbl>, moth_age_month <dbl>, moth_dob <date>,
## #   moth_marital <dbl>, moth_pp_group <dbl>, moth_educ <dbl>, moth_ceb <dbl>,
## #   moth_age_fstbr <dbl>, moth_employ <dbl>, moth_employ_official <dbl>,
## #   moth_employ_extended <dbl>, fath_age_year <dbl>, fath_age_month <dbl>,
## #   fath_dob <date>, fath_marital <dbl>, fath_pp_group <dbl>, ...
```

So, data now has 979233 observations.

Now, we proceed to getting the 2+ and 3+ samples.

```r
# Get the 2+ and 3+ sample
gt2_sample0 <- data %>%
  filter(firstborn_age %>% between(6, 18)) %>%
  group_by(moth_no) %>%
  arrange(child_dob, .by_group = TRUE) %>%
  mutate(
    birth_order = row_number(child_dob),
    no_kids = n()
  ) %>%
  ungroup() %>%
  filter(no_kids >= 2)

gt3_sample0 <- gt2_sample0 %>%
  filter(no_kids >= 3) %>%
  filter(secondborn_age %>% between(6, 18))
```

At this stage, gt2_sample0 has 567077 obs. and gt3_sample0 has 255207 observations.

Next, make a data set with the relevant instruments for the 2+ sample:

```r
parity_gt2 <- gt2_sample0 %>%
  select(moth_no, child_dob, birth_order, no_kids, child_sex) %>%
  filter(birth_order %in% c(1, 2, 3)) %>%
  pivot_wider(
    id_cols = moth_no,
    names_from = birth_order,
    values_from = c(child_sex, child_dob)
  ) %>%
  mutate(
    boy_1 = case_when(
      child_sex_1 == 1 ~ 1,
      TRUE ~ 0
    ),
    boy_2 = case_when(
      child_sex_2 == 1 ~ 1,
      TRUE ~ 0
    ),
    boy_12 = case_when(
      (child_sex_1 == 1 & child_sex_2 == 1) ~ 1,
      TRUE ~ 0
    ),
    girl_12 = case_when(
      (child_sex_1 == 2 & child_sex_2 == 2) ~ 1,
      TRUE ~ 0
    ),
    same_sex_12 = boy_12 + girl_12,
    twins_1 = case_when(
      (child_dob_1 == child_dob_2) ~ 1,
      TRUE ~ 0
    ),
    twins_2 = case_when(
      (child_dob_2 == child_dob_3) ~ 1,
      TRUE ~ 0
    )
  ) %>%
  select(-contains(c("child_sex", "child_dob")))
```

Do the same for the 3+ sample:

```
parity_gt3 <- gt3_sample0 %>%
  select(moth_no, child_dob, birth_order, no_kids, child_sex) %>%
  # extend the vector up to 4 to get the 3+ sample
  filter(birth_order %in% c(1, 2, 3, 4)) %>%
  pivot_wider(
    id_cols = moth_no,
    names_from = birth_order,
    values_from = c(child_sex, child_dob)
  ) %>%
  mutate(
    boy_1 = case_when(
      child_sex_1 == 1 ~ 1,
      TRUE ~ 0
    ),
    boy_2 = case_when(
      child_sex_2 == 1 ~ 1,
      TRUE ~ 0
    ),
    boy_3 = case_when(
      child_sex_3 == 1 ~ 1,
      TRUE ~ 0
    ),
    boy_12 = case_when(
      (child_sex_1 == 1 & child_sex_2 == 1) ~ 1,
      TRUE ~ 0
    ),
    girl_12 = case_when(
      (child_sex_1 == 2 & child_sex_2 == 2) ~ 1,
      TRUE ~ 0
    ),
    same_sex_12 = boy_12 + girl_12,
    # The following part is unique to the 3+ sample
    boy_123 = case_when(
      (child_sex_1 == 1 & child_sex_2 == 1 & child_sex_3 == 1) ~ 1,
      TRUE ~ 0
    ),
    girl_123 = case_when(
      (child_sex_1 == 2 & child_sex_2 == 2 & child_sex_3 == 2) ~ 1,
      TRUE ~ 0
    ),
    same_sex_123 = boy_123 + girl_123,
    # and for the twins:
    twins_1 = case_when(
      (child_dob_1 == child_dob_2) ~ 1,
      TRUE ~ 0
    ),
    twins_2 = case_when(
      (child_dob_2 == child_dob_3) ~ 1,
      TRUE ~ 0
    ),
    twins_3 = case_when(
      (child_dob_3 == child_dob_4) ~ 1,
      TRUE ~ 0
    )
  ) %>%
  select(-contains(c("child_sex", "child_dob")))
```

Merge the initial 2+ and 3+ samples with the parity data sets and get rid of observations with missing values to get our analysis samples:

```
gt2_analysis_sample <- gt2_sample0 %>%
  left_join(parity_gt2, by = "moth_no") %>%
  filter(birth_order == 1) %>%
  # Filter out twins at first birth (and unrealistic obs.)
  filter(!(twins_1 == 1), no_kids < 10) %>%
  mutate(boy = case_when(child_sex == 1 ~ 1, child_sex == 2 ~ 0)) %>%
  select(
    moth_no:child_sex, boy, birth_order:twins_2, no_kids,
    everything(), -(firstborn_dob:secondborn_age)
  ) %>%
  filter(!is.na(moth_dob), !is.na(fath_dob), !is.na(child_private)) %>%
  select(-fath_employ, -moth_employ)

# Do the same for the 3+ sample
gt3_analysis_sample <- gt3_sample0 %>%
  left_join(parity_gt3, by = "moth_no") %>%
  filter(birth_order %in% c(1, 2)) %>%
  filter(!(twins_1 == 1 | twins_2 == 1)) %>%
  filter(no_kids < 10) %>%
  mutate(boy = case_when(child_sex == 1 ~ 1, child_sex == 2 ~ 0)) %>%
  select(
    moth_no:child_sex, boy, birth_order:twins_2, twins_3, no_kids,
    everything(), -(firstborn_dob:secondborn_age)
  ) %>%
  filter(!is.na(moth_dob), !is.na(fath_dob), !is.na(child_private)) %>%
  select(-fath_employ, -moth_employ)
```

Let's take a look at the analysis samples:

```
gt2_analysis_sample
```

```
## # A tibble: 81,348 x 38
##      moth_no  fath_no child_dob  child_sex   boy birth_order no_kids boy_1 boy_2
##        <dbl>    <dbl> <date>         <dbl> <dbl>       <int>   <int> <dbl> <dbl>
##  1   1.00e11  1.00e11 2001-05-23         1     1           1       4     1     1
##  2   1.00e11  1.00e11 2003-07-12         2     0           1       2     0     0
##  3   1.00e11  1.00e11 1999-09-22         1     1           1       3     1     0
##  4   1.00e11  1.00e11 2002-03-21         2     0           1       3     0     1
##  5   1.00e11  1.00e11 2003-09-26         2     0           1       2     0     1
##  6   1.00e11  1.00e11 1997-08-04         1     1           1       3     1     1
##  7   1.00e11  1.00e11 2000-05-10         1     1           1       3     1     1
##  8   1.00e11  1.00e11 2000-11-05         1     1           1       2     1     0
##  9   1.00e11  1.00e11 2001-01-23         1     1           1       3     1     0
## 10   1.00e11  1.00e11 1999-04-08         1     1           1       5     1     0
## # ... with 81,338 more rows, and 29 more variables: boy_12 <dbl>,
## #   girl_12 <dbl>, same_sex_12 <dbl>, twins_1 <dbl>, twins_2 <dbl>,
## #   child_age_year <dbl>, child_age_month <dbl>, child_sch_attend <dbl>,
## #   child_educ <dbl>, child_private <dbl>, child_pop_group <dbl>,
## #   moth_age_year <dbl>, moth_age_month <dbl>, moth_dob <date>,
## #   moth_marital <dbl>, moth_pp_group <dbl>, moth_educ <dbl>, moth_ceb <dbl>,
## #   moth_age_fstbr <dbl>, moth_employ_official <dbl>, ...
```

```
gt3_analysis_sample
```

```
## # A tibble: 56,662 x 43
##      moth_no  fath_no child_dob  child_sex   boy birth_order no_kids boy_1 boy_2
##        <dbl>    <dbl> <date>         <dbl> <dbl>       <int>   <int> <dbl> <dbl>
##  1   1.00e11  1.00e11 2001-05-23         1     1           1       4     1     1
##  2   1.00e11  1.00e11 2004-02-03         1     1           2       4     1     1
##  3   1.00e11  1.00e11 1994-03-07         1     1           2       3     0     1
##  4   1.00e11  1.00e11 1999-09-22         1     1           1       3     1     0
##  5   1.00e11  1.00e11 2001-04-01         1     1           2       3     0     1
##  6   1.00e11  1.00e11 2002-03-21         2     0           1       3     0     1
##  7   1.00e11  1.00e11 2003-07-13         1     1           2       3     0     1
##  8   1.00e11  1.00e11 1997-08-04         1     1           1       3     1     1
##  9   1.00e11  1.00e11 2003-05-05         1     1           2       3     1     1
## 10   1.00e11  1.00e11 2000-05-10         1     1           1       3     1     1
## # ... with 56,652 more rows, and 34 more variables: boy_3 <dbl>, boy_12 <dbl>,
## #   girl_12 <dbl>, same_sex_12 <dbl>, boy_123 <dbl>, girl_123 <dbl>,
## #   same_sex_123 <dbl>, twins_1 <dbl>, twins_2 <dbl>, twins_3 <dbl>,
## #   child_age_year <dbl>, child_age_month <dbl>, child_sch_attend <dbl>,
## #   child_educ <dbl>, child_private <dbl>, child_pop_group <dbl>,
## #   moth_age_year <dbl>, moth_age_month <dbl>, moth_dob <date>,
## #   moth_marital <dbl>, moth_pp_group <dbl>, moth_educ <dbl>, ...
```

At this stage, `gt2_analysis_sample` has 81348 observations and `gt3_analysis_sample` has 56662 observations.

Do further cleaning and generate some outcome variabes:

```r
# Generate dummy for private school attendance and mother's LFP status
gt2_analysis_sample <- gt2_analysis_sample %>%
  filter(child_private %in% c(1, 2, 9)) %>%
  mutate(
    private_school = case_when(
      child_private == 2 ~ 1, TRUE ~ 0
    ),

    moth_lfp_offic = case_when(
      moth_employ_official %in% c(1, 2) ~ 1,
      moth_employ_official %in% c(3, 4) ~ 0
    ),

    moth_lfp_ext = case_when(
      moth_employ_extended %in% c(1, 2) ~ 1,
      moth_employ_extended == 3 ~ 0
    )
  ) %>%
  filter(!is.na(moth_employ_official)) %>%
  mutate(
    moth_pp_group_fct =
      case_when(
        moth_pp_group == 1 ~ "Black African",
        moth_pp_group == 2 ~ "Coloured",
        moth_pp_group == 3 ~ "Indian or Asian",
        moth_pp_group == 4 ~ "White",
        moth_pp_group == 5 ~ "Other",
      ) %>% factor()
  ) %>%
  mutate( child_sex = case_when(
      child_sex == 1 ~ "Male", child_sex == 2 ~ "Female"
  ) %>% factor() )

# Construct educational attainment variable
gt2_analysis_sample <- gt2_analysis_sample %>%
  filter(child_educ %in% 0:12 | child_educ == 98) %>%
  mutate(
    child_educ_gen = case_when(
      as.numeric(child_educ) == 98 ~ -1,
      TRUE ~ as.numeric(child_educ)
    )
  ) %>%
  group_by(child_age_year, boy) %>%
  mutate(mean_educ_age_sex = mean(child_educ_gen)) %>%
  ungroup() %>%
  mutate(educ_attain = child_educ_gen / mean_educ_age_sex)
```

Display the final analysis sample:

```r
gt2_analysis_sample
```

```
## # A tibble: 80,993 x 45
##       moth_no  fath_no child_dob  child_sex   boy birth_order no_kids boy_1 boy_2
##         <dbl>    <dbl> <date>     <fct>     <dbl>       <int>   <int> <dbl> <dbl>
##  1    1.00e11  1.00e11 2001-05-23 Male          1           1       4     1     1
##  2    1.00e11  1.00e11 2003-07-12 Female        0           1       2     0     0
##  3    1.00e11  1.00e11 1999-09-22 Male          1           1       3     1     0
##  4    1.00e11  1.00e11 2002-03-21 Female        0           1       3     0     1
##  5    1.00e11  1.00e11 2003-09-26 Female        0           1       2     0     1
##  6    1.00e11  1.00e11 1997-08-04 Male          1           1       3     1     1
##  7    1.00e11  1.00e11 2000-05-10 Male          1           1       3     1     1
##  8    1.00e11  1.00e11 2000-11-05 Male          1           1       2     1     0
##  9    1.00e11  1.00e11 2001-01-23 Male          1           1       3     1     0
## 10    1.00e11  1.00e11 1999-04-08 Male          1           1       5     1     0
## # ... with 80,983 more rows, and 36 more variables: boy_12 <dbl>,
## #   girl_12 <dbl>, same_sex_12 <dbl>, twins_1 <dbl>, twins_2 <dbl>,
## #   child_age_year <dbl>, child_age_month <dbl>, child_sch_attend <dbl>,
## #   child_educ <dbl>, child_private <dbl>, child_pop_group <dbl>,
## #   moth_age_year <dbl>, moth_age_month <dbl>, moth_dob <date>,
## #   moth_marital <dbl>, moth_pp_group <dbl>, moth_educ <dbl>, moth_ceb <dbl>,
## #   moth_age_fstbr <dbl>, moth_employ_official <dbl>, ...
```

We can see that `gt2_analysis_sample` has 80993 observations.

Preliminary Analysis

Only data from the `gt2_analysis_sample` is used in this preliminary analysis. Since the `gt3_analysis_sample` is a subset of this sample, a symmetric type of analysis will apply.

Table of summary statistics

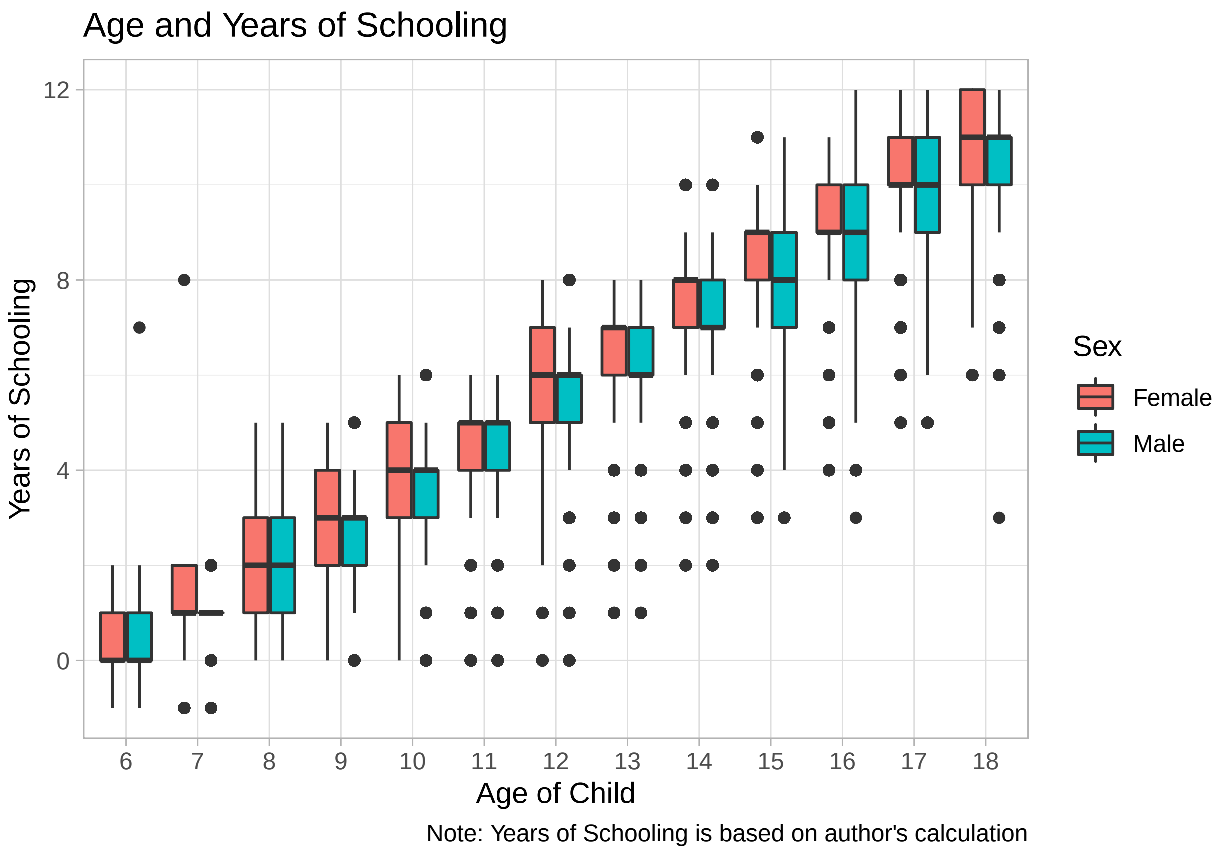
```
gt2_analysis_sample %>%
  select(
    boy, no_kids:same_sex_12, twins_2, child_age_year, child_educ_gen,
    educ_attain, private_school, moth_age_year, fath_age_year
  ) %>%
  as.data.frame() %>%
  stargazer(type = "text")
```

```
##
## =====================================================================
## Statistic        N     Mean  St. Dev.  Min   Pctl(25) Pctl(75)  Max
## ---------------------------------------------------------------------
## boy            80,993 0.503   0.500     0     0        1       1
## no_kids        80,993 2.606   0.863     2     2        3       9
## boy_1          80,993 0.503   0.500     0     0        1       1
## boy_2          80,993 0.507   0.500     0     0        1       1
## boy_12         80,993 0.262   0.440     0     0        1       1
## girl_12        80,993 0.252   0.434     0     0        1       1
## same_sex_12    80,993 0.514   0.500     0     0        1       1
## twins_2        80,993 0.010   0.097     0     0        0       1
## child_age_year 80,993 12.178  3.647     6     9        15      18
## child_educ_gen 80,993 5.674   3.383     -1    3        8       12
## educ_attain    80,993 1.000   0.531   -2.650  0.871    1.118   18.548
## private_school 80,993 0.104   0.305     0     0        0       1
## moth_age_year  80,993 35.997  6.195     18    32       40      64
## fath_age_year  80,993 40.950  7.734     13      36        45      107
## ----------------------------------------------------------------------
```

Some Graphs

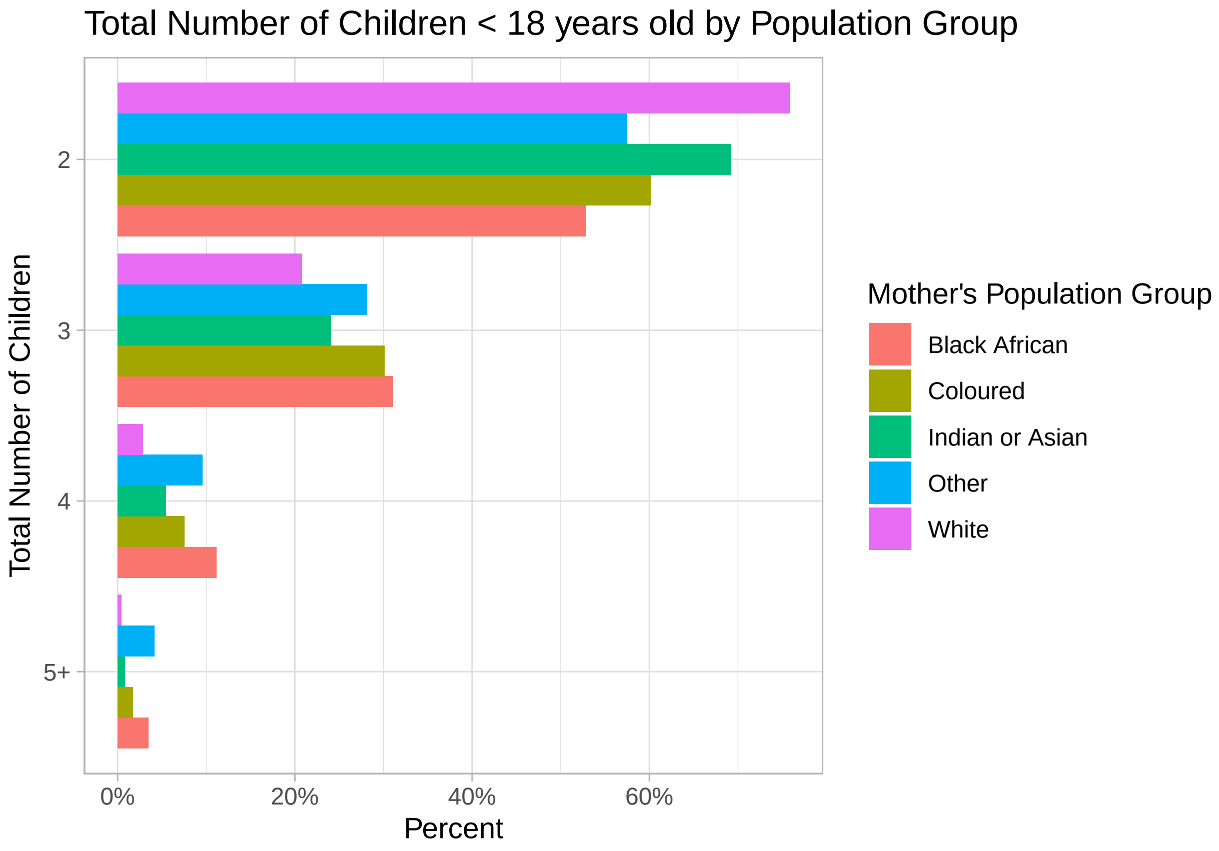
The following is a box plot of years of schooling for all kids in the sample by age and gender.

```
gt2_analysis_sample %>%
  mutate(child_age = factor(child_age_year)) %>%
  select(child_age, child_sex, child_educ_gen) %>%
  ggplot(aes(child_age, child_educ_gen)) +
  geom_boxplot(aes(fill = child_sex)) +
  labs(
    title = "Age and Years of Schooling",
    x = "Age of Child",
    y = "Years of Schooling",
    fill = "Sex",
    caption = "Note: Years of Schooling is based on author's calculation"
  )
```



As a prelude to a hetrogeneity analysis, let's look at the proportion of families having a certain number of children by population group.

```
gt2_analysis_sample %>%
  count(moth_pp_group_fct, no_kids) %>%
  # filter(moth_pp_group_fct != "Other") %>%
  mutate(
    no_kids = if_else(no_kids %in% 5:9, "5+", as.character(no_kids)),
    no_kids = factor(no_kids) %>% fct_rev()
  ) %>%
  group_by(moth_pp_group_fct) %>%
  mutate(prop = n / sum(n)) %>%
  ggplot(aes(no_kids, prop)) +
  geom_col(mapping = aes(fill = moth_pp_group_fct), position = "dodge") +
  coord_flip() +
  scale_y_continuous(label = percent) +
  labs(
    title = "Total Number of Children < 18 years old by Population Group",
    x = "Total Number of Children",
    y = "Percent",
    fill = "Mother's Population Group"
  )
```

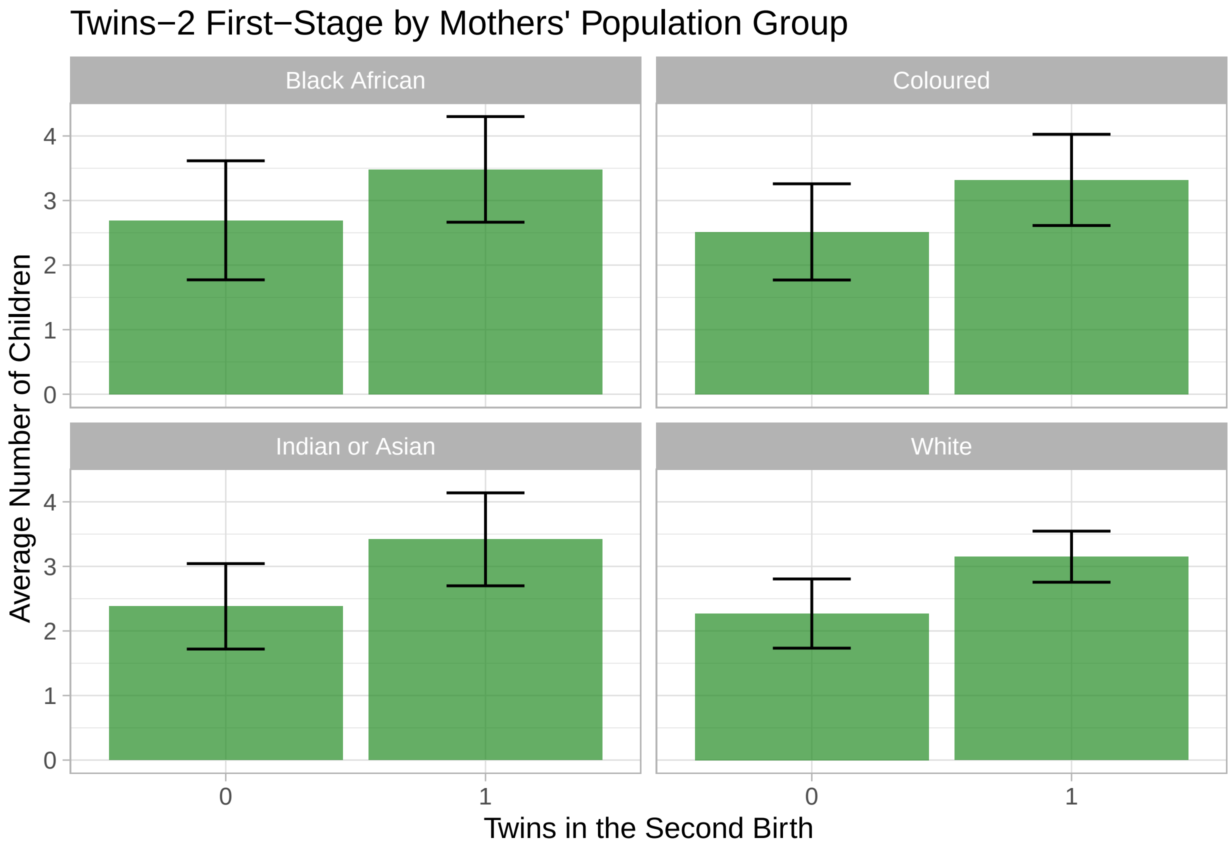


We can also see the first stage effect of the twins instrumnet diaggregated by the mother's population group:

```
my_sum <- gt2_analysis_sample %>%
  mutate(twins_2 = factor(twins_2)) %>%
  group_by(twins_2, moth_pp_group_fct) %>%
  summarise(
    n = n(),
    mean_no_kids = mean(no_kids),
    sd=sd(no_kids)
  ) %>%
  mutate(se = sd/sqrt(n)) %>%
  mutate(ic = se*qt(.975, n-1))
```

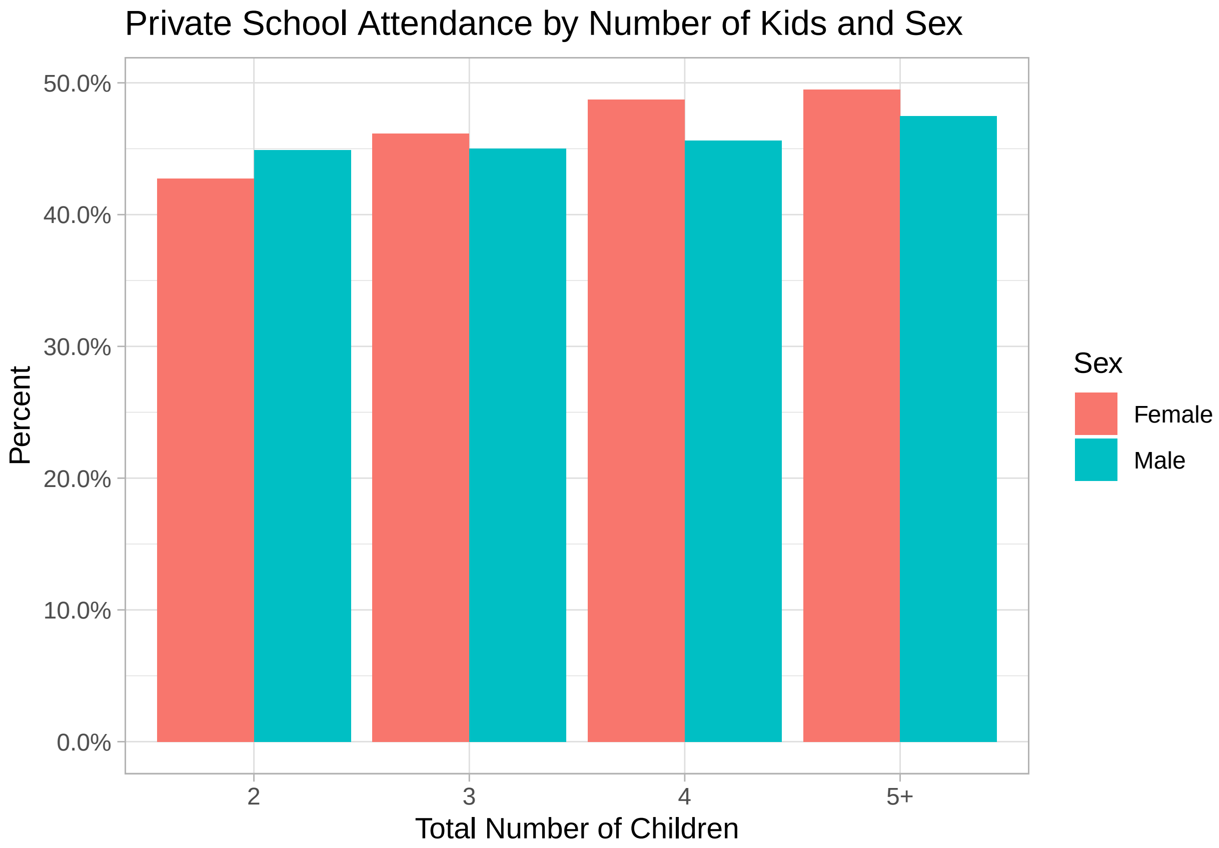
```
## `summarise()` has grouped output by 'twins_2'. You can override using the `.groups` argument.
```

```
my_sum %>%
  filter(moth_pp_group_fct != "Other") %>%
  ggplot() +
  geom_col(aes(twins_2, mean_no_kids), fill = "forestgreen", alpha = 0.7) +
  geom_errorbar(aes(x = twins_2, ymin = mean_no_kids - sd,
                    ymax = mean_no_kids + sd), size = 0.5, width = 0.3) +
  facet_wrap(~moth_pp_group_fct) +
  labs(
    title = "Twins-2 First-Stage by Mothers' Population Group",
    x = "Twins in the Second Birth",
    y = "Average Number of Children"
  )
```



The other outcome variable I intend to explore is private school attendance. The following graph shows the proportion of firstborns attending private school:

```
gt2_analysis_sample %>%
  mutate(no_kids = if_else(no_kids %in% 5:9, "5+", as.character(no_kids)),
         no_kids = factor(no_kids) ) %>%
  count(no_kids, child_sex, private_school) %>%
  group_by(no_kids) %>%
  mutate(prop = n/sum(n)) %>%
  ggplot(aes(no_kids, prop)) +
  geom_col(mapping = aes(fill = child_sex), position = "dodge") +
  # coord_flip() +
  scale_y_continuous(label = percent) +
  labs(
    title = "Private School Attendance by Number of Kids and Sex",
    x = "Total Number of Children",
    y = "Percent",
    fill = "Sex"
  )
```



I also run a simple linear regression of private school attendance dummy on the same sex dummy, disaggregated by number of children in the family and sex of the firstborn child. This is graphically shown below:

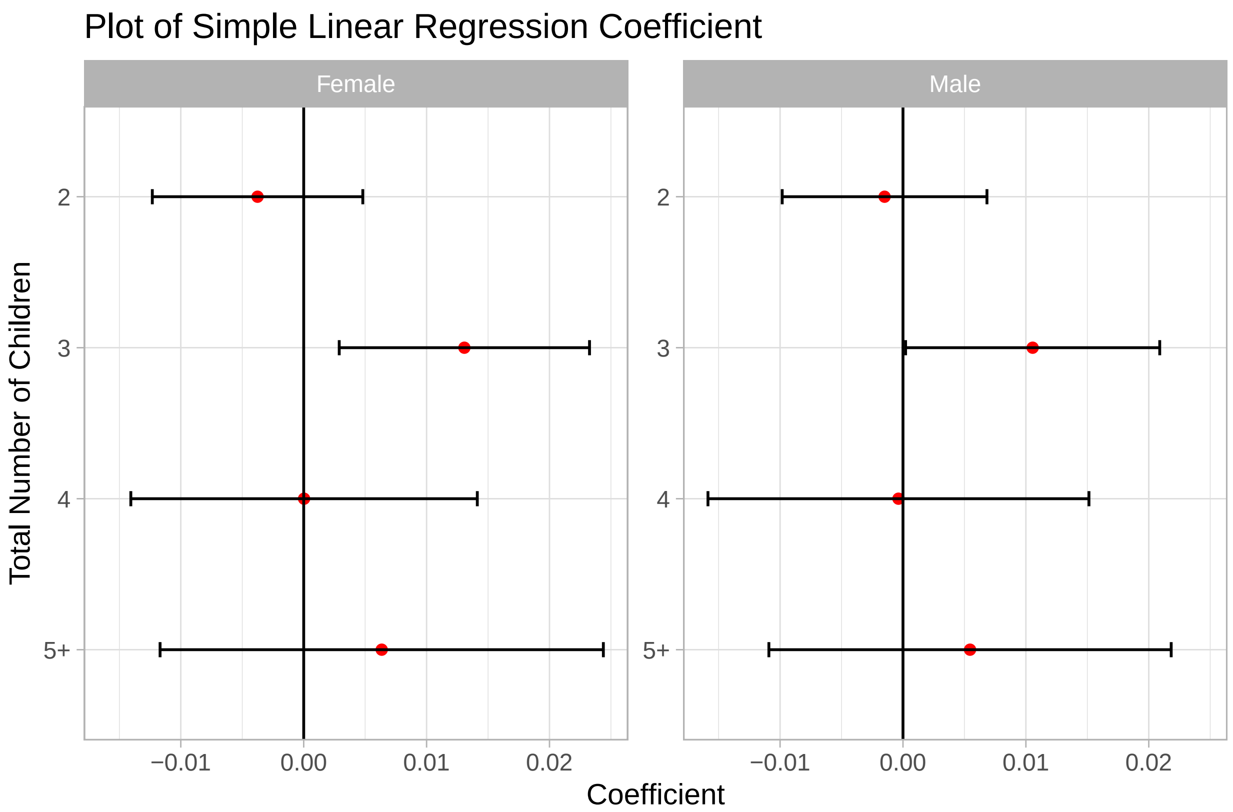
```
simple_model_twins <- function(tbl) {
  lm(private_school ~ twins_2, data = tbl)
}

simple_model_samesex <- function(tbl) {
  lm(private_school ~ same_sex_12, data = tbl)
}

get_confs <- function(mod) {
  confint(mod) %>%
    as_tibble()
}
```

```
simple_mod_data_samesex <- gt2_analysis_sample %>%
  mutate(no_kids = if_else(no_kids %in% 5:9, "5+", as.character(no_kids)),
         no_kids = factor(no_kids, levels = c("2", "3", "4", "5+")) ) %>%
  select(private_school, no_kids, child_sex, twins_2, same_sex_12) %>%
  group_by(no_kids, child_sex) %>%
  nest() %>%
  mutate(
    model = map(data, simple_model_samesex ),
    summaries = map(model, tidy),
    conf_ints = map(model, get_confs)
  ) %>%
  unnest(c(summaries, conf_ints)) %>%
  rename("conf_low" = `2.5 %`, "conf_high" = `97.5 %`) %>%
  arrange(no_kids, child_sex)

simple_mod_data_samesex %>%
  filter(term == "same_sex_12") %>%
  ggplot(aes(fct_rev(no_kids), estimate)) +
  geom_point(color = "red") +
  facet_wrap(~ child_sex, scales = "free_y") +
  geom_errorbar(aes(ymin = conf_low, ymax = conf_high), width = .1 ) +
  geom_hline(aes(yintercept = 0)) +
  coord_flip() +
  labs(
    title = "Plot of Simple Linear Regression Coefficient",
    x = "Total Number of Children",
    y = "Coefficient",
    caption = "The red dots indicate coefficient estimate and the lines are 95% confidence intervals"
      )
```



The red dots indicate coefficient estimate and the lines are 95% confidence intervals

First-Stage Regression

```
m1 <- lm(no_kids ~ same_sex_12, data = gt2_analysis_sample)
m2 <- lm(no_kids ~ boy_1 + boy_2 + same_sex_12, data = gt2_analysis_sample)
m3 <- lm(no_kids ~ boy_1 + boy_12 + girl_12, data = gt2_analysis_sample)
m4 <- lm(no_kids ~ twins_2, data = gt2_analysis_sample)
stargazer(m1, m2, m3, m4, type = "text", keep.stat = c("n", "rsq"))
```

```
##
## =================================================
##                       Dependent variable:
##               ------------------------------------
##                             no_kids
##                 (1)       (2)      (3)      (4)
## -------------------------------------------------
## boy_1                 -0.033*** -0.020**
##                         (0.006)  (0.009)
##
## boy_2                 -0.013**
##                         (0.006)
##
## same_sex_12  0.046*** 0.046***
##              (0.006)   (0.006)
##
## boy_12                          0.034***
##                                 (0.009)
```

```
##
## girl_12                         0.059***
##                                 (0.009)
##
## twins_2                                  0.805***
##                                          (0.031)
##
## Constant     2.583*** 2.606***  2.593*** 2.599***
##              (0.004)   (0.006)  (0.006)  (0.003)
##
## -------------------------------------------------
## Observations  80,993   80,993    80,993   80,993
## R2            0.001     0.001    0.001    0.008
## =================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

In column 1, the same sex dummy is used (i.e., whether the first two births are both boys or both girls). In column 2, the regression controls for the gender of the first two kids. The result is identical to column 1. In column 3, the same sex instrument is disaggregated by gender. As shown, parents are more likely to have an additional child if the first two births are females than if they are males, suggesting sex preference in favor of boys. Finally, column 4 shows that having a twin in the second birth is strongly correlated with a larger family size. But the standard error is relatively larger compared to the other instruments, possibly because of the relatively small number of twin births in the sample (< 1%).

2SLS Regression

Here, I run preliminary 2SLS regressions with different outcome variables. First, let's see the effect on educational attainment.

```r
OLS1 <- lm(educ_attain ~ no_kids, data = gt2_analysis_sample)
IV1 <- ivreg(educ_attain ~ no_kids | twins_2, data = gt2_analysis_sample)
IV2 <- ivreg(educ_attain ~ no_kids | same_sex_12, data = gt2_analysis_sample)
IV3 <- ivreg(educ_attain ~ no_kids | boy_12 + girl_12, data = gt2_analysis_sample)
stargazer(OLS1, IV1, IV2, IV3, type = "text", keep.stat = c("n", "rsq"))
```

```
##
## =================================================
##                       Dependent variable:
##              ------------------------------------
##                          educ_attain
##                 OLS           instrumental
##                                 variable
##                 (1)      (2)      (3)      (4)
## -------------------------------------------------
## no_kids      -0.016***  -0.015   0.135    0.089
##               (0.002)  (0.024)  (0.084)  (0.067)
##
## Constant     1.041***  1.039*** 0.647*** 0.769***
##               (0.006)  (0.062)  (0.219)  (0.175)
##
## -------------------------------------------------
## Observations  80,993    80,993   80,993   80,993
## R2             0.001    0.001    -0.060   -0.028
## =================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

The first column is obtained using OLS, which tells the usual story, that number of kids is negatively related with educational attainment. Comumns 2-4 are IV estimates, where 2 and 3 use variations of the same sex instrument and column 4 uses the twins instrument. All show a consistent result; the number of children does NOT affect educational attainment.

What about private school attendance? Here is the 2SLS result using private school attendance dummy as an outcome variable.

```r
OLS2 <- lm(private_school ~ no_kids, data = gt2_analysis_sample)
IV4 <- ivreg(private_school ~ no_kids | twins_2, data = gt2_analysis_sample)
IV5 <- ivreg(private_school ~ no_kids | same_sex_12, data = gt2_analysis_sample)
IV6 <- ivreg(private_school ~ no_kids | boy_12 + girl_12, data = gt2_analysis_sample)
stargazer(OLS2, IV4, IV5, IV6, type = "text", keep.stat = c("n", "rsq"))
```

```
##
## ===============================================
##                      Dependent variable:
##              ----------------------------------
##                        private_school
##                 OLS          instrumental
##                                variable
##                 (1)      (2)      (3)     (4)
## -----------------------------------------------
## no_kids      -0.030***  0.003    0.014  -0.002
##               (0.001)  (0.014)  (0.047) (0.038)
##
## Constant     0.182***  0.097***  0.067   0.108
##               (0.003)  (0.036)  (0.123) (0.099)
##
## -----------------------------------------------
## Observations  80,993    80,993  80,993  80,993
## R2             0.007    -0.001  -0.008   0.001
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```