What will affect the number of total children in a family?

Zijian wang

2020/10/13

Abstract

Most of family have children, but some of them do not have; I want to figure out what will affect the number of total children in a family. I explore the date of the 2017 General Social Survey to find what will affect the total number of children in a family. I choose six variables to discover whether these variables affect the number of children and how they affect.

Introduction

This paper uses data from the 2017 General Social Survey: Families Cycle 31(GSS). This data set contains too many variables. First of all, I use code(gss_cleaning, authors: Rohan Alexander and Sam Caetano) to clean the data, then get 81 variables. In these 81 variables, I found that many of them may affect each other, and some may affect the same variable. Moreover, I choose the total number of the child in a family as the variable that I want to figure out how different variables affect it.

The primary purpose of the research is to find out what will affect the total number of children in a family, and I choose five variables: age, sex, age_first_child,age_youngest_child_under_6,children_in_household . First, I try to figure out how this single variable affects the total number of children. I will plot different types of graphs and analyze which variable has a significant influence. Then I will use those variables to build a model to find out the relationship between them and the total number of the children.

Through the model, I found that total number of children can not be explained by single variable.

Data set

The dataset is the 2017 General Special Survey: Families Cycle 31(GSS), and I use code(gss_cleaning, authors: Rohan Alexander and Sam Caetano) to choose 81 variables.

This data was collected with computer-assisted telephone interviews(CATI). And respondents were interviewed in the official language of their choice. All interviewing took place using centralized telephone facilities in five of Statistics Canada's regional offices and with calls being made from approximately 9:00 a.m. to 9:30 p.m. Mondays to Fridays. Interviewing was also scheduled from 10:00 a.m. to 5:00 p.m. on Saturdays and 1:00 p.m. to 9:00 p.m. on Sundays. (Gss31_Use_Guide)

The target population for the 2017 GSS included all persons 15 years of age and older in Canada, excluding:

1. Residents of the Yukon, Northwest Territories, and Nunavut; and
2. Full-time residents of institutions.(Gss31_Use_Guide)

The survey frame was created using two different components:

1. ts of telephone numbers in use (both landline and cellular) available to Statistics Canada from various sources (telephone companies, Census of population, etc.);

2.The Address Register (AR): List of all dwellings within the ten provinces. (Gss31_Use_Guide)

The target sample size (i.e., the desired number of respondents) for the 2017 GSS was 20,000, while the actual number of respondents was 20,602. (Gss31_Use_Guide)

Non-response: Those who refused to participate at first were re-contacted up to two more times to explain the importance of the survey and encourage their participation. For cases in which the timing of the interviewer's call was inconvenient, an appointment was arranged to call back at a more convenient time. For cases in which there was no one home, numerous callbacks were made. The overall response rate for the 2017 GSS was 52.4%. (Gss31_Use_Guide)

This dataset's key features are to gather data on social trends to monitor changes in Canadians' living conditions and well-being over time; provide information on specific social policy issues of current or emerging interest(Gss31_Use_Guide). The strengths of this dataset are it gathers many classification variables, and these variables are helpful in the analysis of data. This data's weakness is that some variables gather too many missing values, so these kinds of variables are not useful and need more time to clean data.

I select six classification variables, such as the number of total children, age, Age_first_child , number_total_children_intention,Age_youngest_child_under_6, Children_in_household.

The number of total children: This data is the total number of children reported by respondents, and this variable is capped at seven children and more. This is the data that I want to figure out how other variables affect it. And I found that 30% of respondents have no children, and 30% of respondents have two children, 15% have one child, and 15% have three children. 10% of respondents have four or more.

Age: This data is the age of respondent with decimal at the time of the survey interview, and this variable is capped at 80 years and older.

Number_total_children_intention: This data is the total number of children intending to have, and these variables are capped at five children and more. This data is similar to future_children_intention. And also the the data I have chosen are numerical variables and have more information.

Age_first_child: This data is the respondent's first child's age, and this variable is capped at age 60 and older. This data looks

like data(age)

Age_youngest_child_under_6: This is the age of the youngest child under 6.

Children_in_household: This data is children that live in the respondent's household. 80% of children do not live in the respondent's household.

clean data(authors: Rohan Alexander and Sam Caetano)

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test

## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ----------------------------------------------------------------------------------- tid
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
##
## -- Column specification ----------------------------------------------------------------------------
## cols(
##   .default = col_double()
## )
## i Use `spec()` for the full column specifications.
```

```
## Rows: 20,602
## Columns: 81
## Rowwise:
## $ caseid                            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,...
## $ age                               <dbl> 52.7, 51.1, 63.6, 80.0, 28.0, 63.0...
## $ age_first_child                   <dbl> 27, 33, 40, 56, NA, 37, 40, 59, NA...
## $ age_youngest_child_under_6        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ total_children                    <dbl> 1, 5, 5, 1, 0, 2, 2, 7, 0, 1, 0, 0...
## $ age_start_relationship            <dbl> NA, NA, NA, NA, 25.3, NA, NA, NA, ...
## $ age_at_first_marriage             <dbl> NA, NA, NA, NA, NA, NA, NA, 22.1, ...
## $ age_at_first_birth                <dbl> 25.9, NA, 23.2, 27.3, NA, 25.8, 18...
## $ distance_between_houses           <dbl> 30, NA, NA, NA, NA, NA, NA, NA, NA...
## $ age_youngest_child_returned_work <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ feelings_life                     <dbl> 8, 10, 8, 10, 8, 9, 4, 10, 8, 5, 1...
## $ sex                               <chr> "Female", "Male", "Female", "Femal...
## $ place_birth_canada                <chr> "Born in Canada", "Born in Canada"...
## $ place_birth_father                <chr> "Born in Canada", "Born in Canada"...
## $ place_birth_mother                <chr> "Born in Canada", "Born in Canada"...
## $ place_birth_macro_region          <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ place_birth_province              <chr> "Quebec", "Ontario", "Ontario", "A...
## $ year_arrived_canada               <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ province                          <chr> "Quebec", "Manitoba", "Ontario", "...
## $ region                            <chr> "Quebec", "Prairie region", "Ontar...
## $ pop_center                        <chr> "Larger urban population centres (...
## $ marital_status                    <chr> "Single, never married", "Married"...
## $ aboriginal                        <chr> "No", "No", "No", "No", "No", "No"...
## $ vis_minority                      <chr> "Not a visible minority", "Not a v...
## $ age_immigration                   <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ landed_immigrant                  <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ citizenship_status                <chr> "By birth", "By birth", "By birth"...
## $ education                         <chr> "High school diploma or a high sch...
## $ own_rent                          <chr> "Owned by you or a member of this ...
## $ living_arrangement                <chr> "Alone", "Spouse only", "Spouse on...
## $ hh_type                           <chr> "Low-rise apartment (less than 5 s...
## $ hh_size                           <dbl> 1, 2, 2, 2, 2, 2, 1, 1, 1, 6, 5, 1...
## $ partner_birth_country             <chr> "Canada", "Canada", "Canada", "Can...
## $ partner_birth_province            <chr> "Quebec", "Manitoba", "Ontario", "...
## $ partner_vis_minority             <chr> "Not a visible minority", "Not a v...
## $ partner_sex                      <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ partner_education                <chr> "Trade certificate or diploma", "B...
## $ average_hours_worked             <chr> "30.0 to 40.0 hours", "50.1 hours ...
## $ worked_last_week                 <chr> "Yes", "Yes", "No", "No", "No", "N...
## $ partner_main_activity            <chr> "Working at a paid job or business...
## $ self_rated_health                <chr> "Excellent", "Good", "Very good", ...
## $ self_rated_mental_health         <chr> "Excellent", "Good", "Good", "Very...
## $ religion_has_affiliation         <chr> "Has religious affiliation", "Don’...
## $ regilion_importance              <chr> "Somewhat important", "Don’t know"...
## $ language_home                    <chr> "French", "English", "French", "En...
## $ language_knowledge               <chr> "French only", "English only", "Bo...
## $ income_family                    <chr> "$25,000 to $49,999", "$75,000 to ...
## $ income_respondent                <chr> "$25,000 to $49,999", "Less than $...
## $ occupation                       <chr> "Sales and service occupations", "...
## $ childcare_regular                <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ childcare_type                   <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ childcare_monthly_cost           <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ ever_fathered_child              <chr> "NA", "Yes", "NA", "NA", "No", "NA...
## $ ever_given_birth                 <chr> "Yes", "NA", "Yes", "Yes", "NA", "...
## $ number_of_current_union          <chr> "NA", "NA", "NA", "NA", "Second un...
## $ lives_with_partner               <chr> "No", "No", "No", "No", "Yes", "No...
## $ children_in_household            <chr> "No child", "No child", "No child"...
## $ number_total_children_intention  <dbl> NA, NA, NA, NA, 2, NA, NA, NA, NA,...
## $ has_grandchildren                <chr> "No", "Yes", "Yes", "No", "No", "Y...
## $ grandparents_still_living        <chr> "No", "No", "No", "No", "Yes", "No...
## $ ever_married                     <chr> "No", "Yes", "Yes", "Yes", "No", "...
## $ current_marriage_is_first        <chr> "NA", "Yes", "Yes", "Yes", "NA", "...
## $ number_marriages                 <dbl> 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0...
## $ religion_participation           <chr> "Once or twice a year", "Don’t kno...
## $ partner_location_residence       <chr> "In the same province", "NA", "NA"...
## $ full_part_time_work              <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ time_off_work_birth              <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ reason_no_time_off_birth         <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ returned_same_job                <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ satisfied_time_children          <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ provide_or_receive_fin_supp      <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ fin_supp_child_supp              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_child_exp               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_lump                    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_other                   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_agreement               <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ future_children_intention        <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ is_male                          <dbl> 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0...
## $ main_activity                    <chr> "NA", "NA", "NA", "NA", "NA", "NA"...
## $ age_diff                         <chr> "NA", "Respondent is 4 years older...
## $ number_total_children_known      <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1...
```

Plot for each variable

Fig 1.1 total children

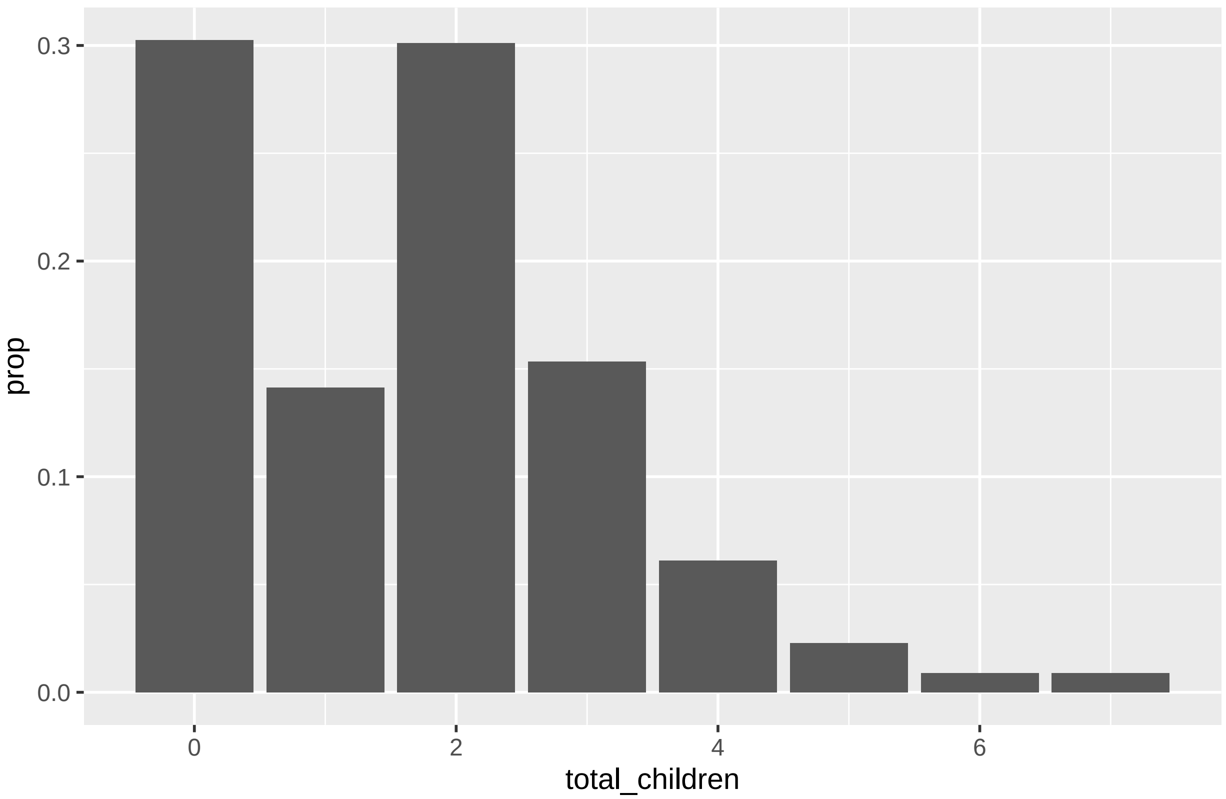


Fig 1.2 Age

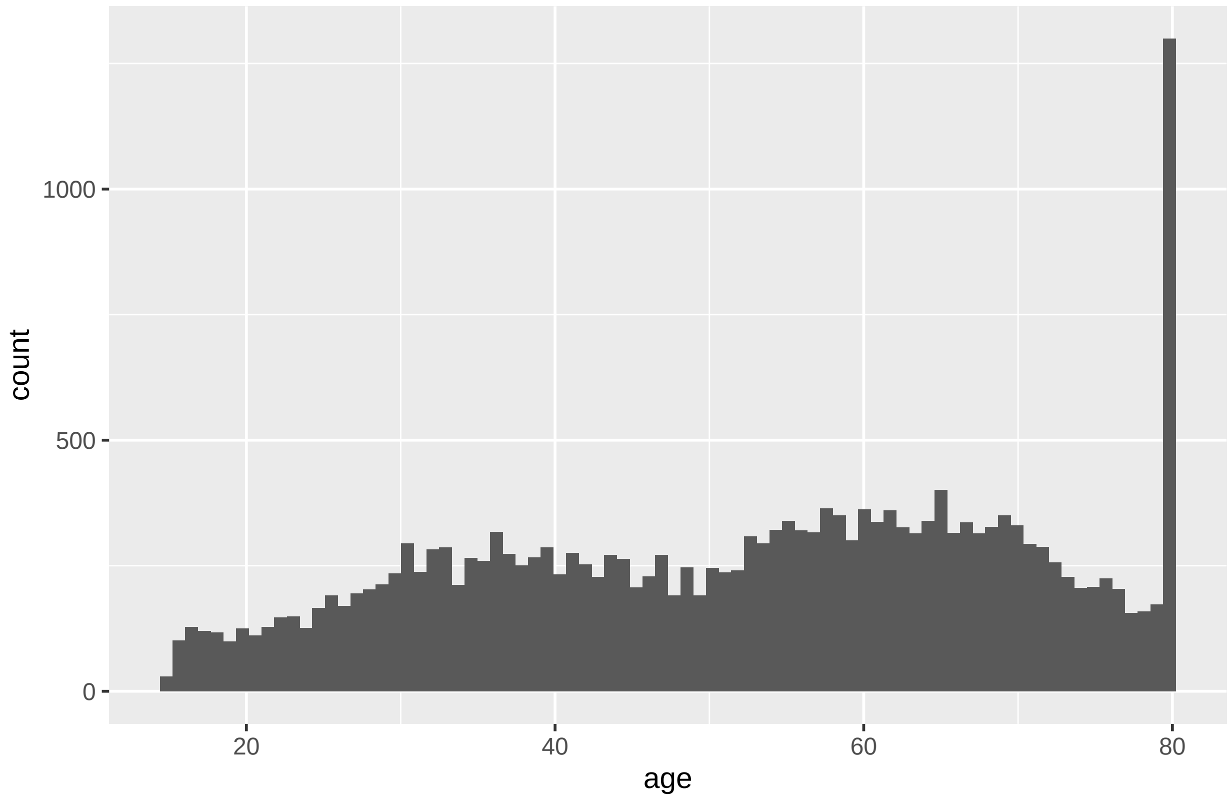


Fig 1.3 age of first child

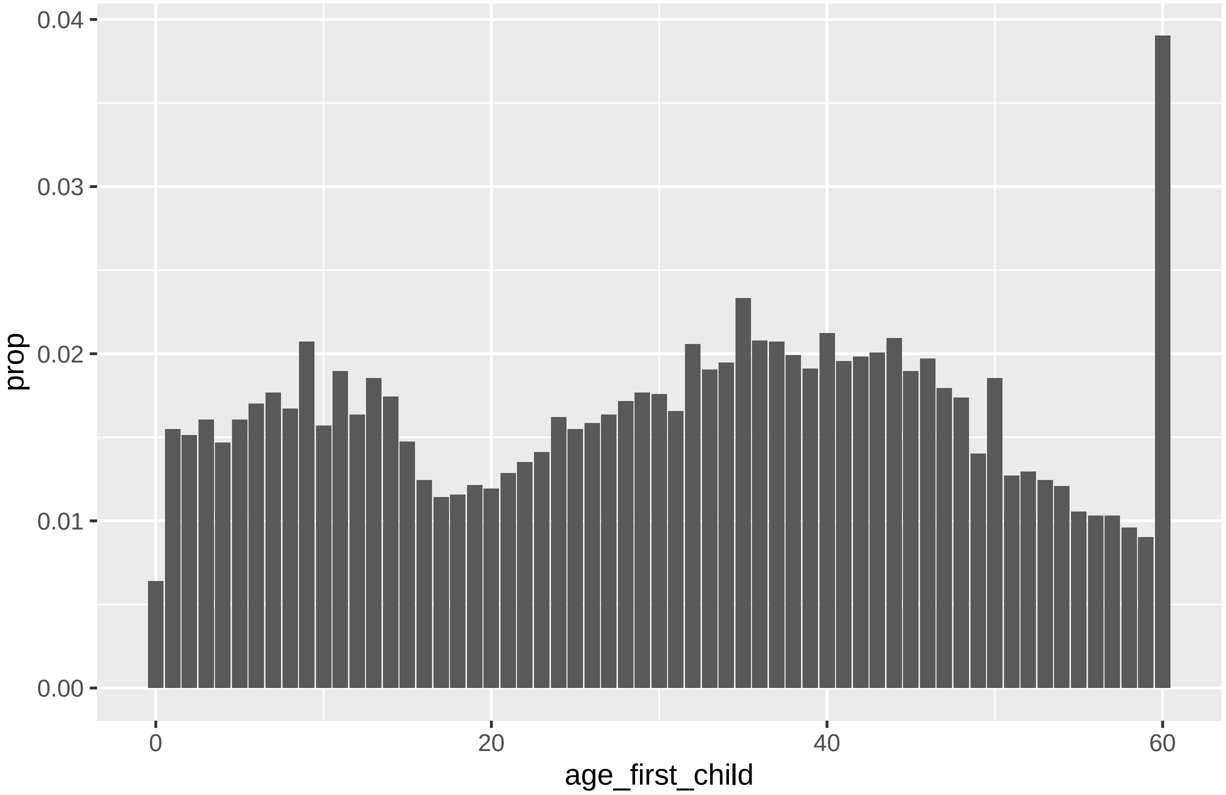


Fig 1.4 age of youngest child under 6

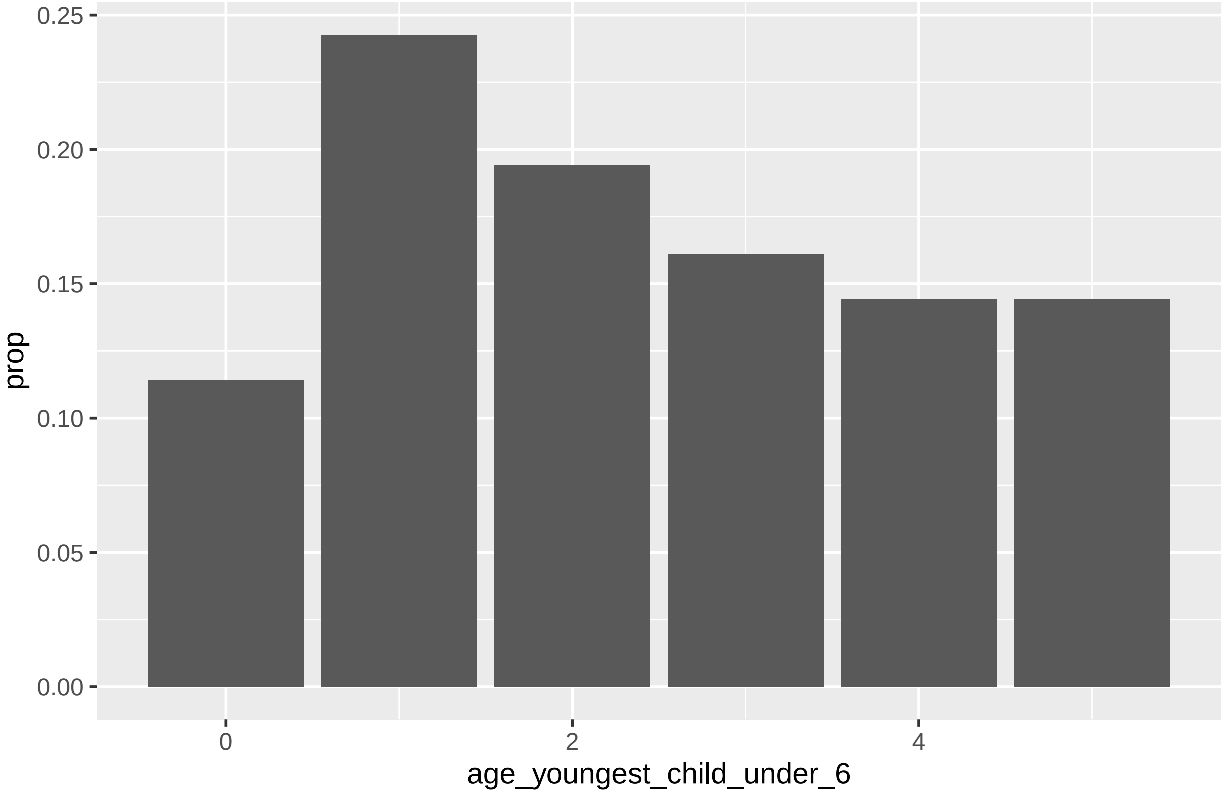


Fig 1.5 number_total_children_intention

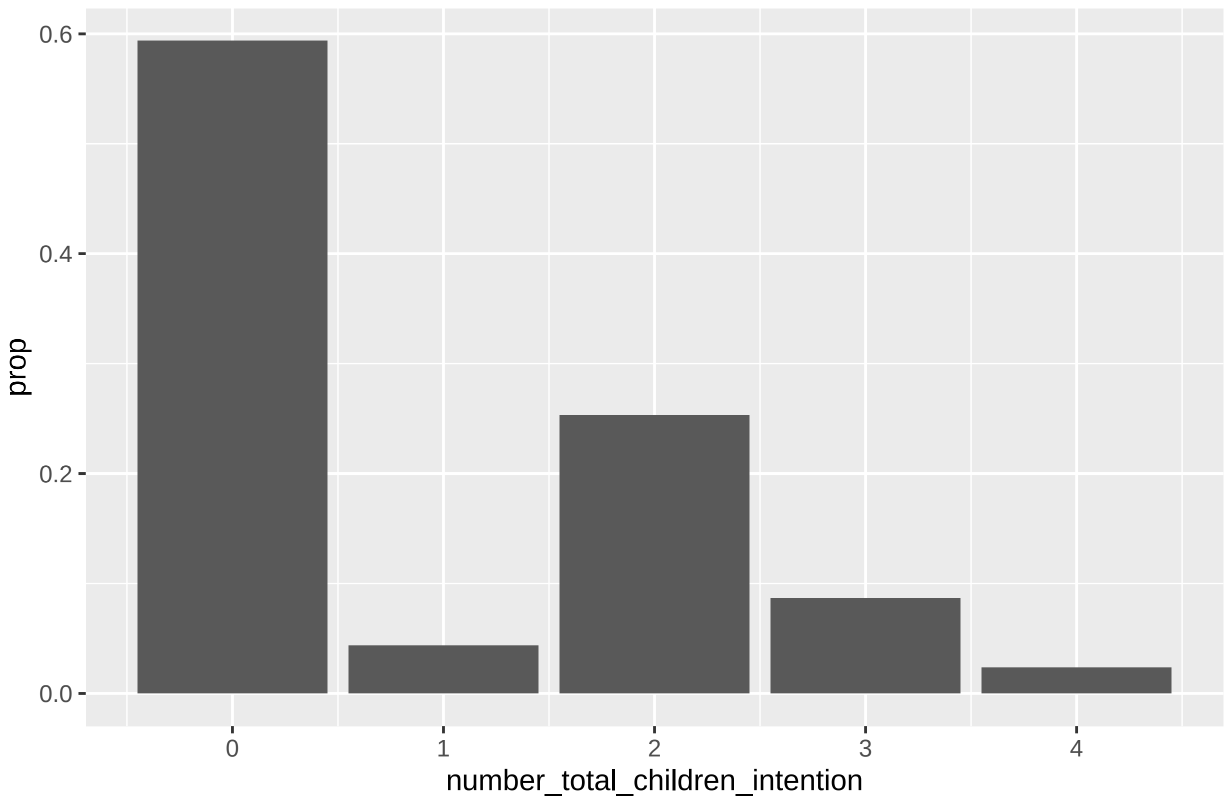
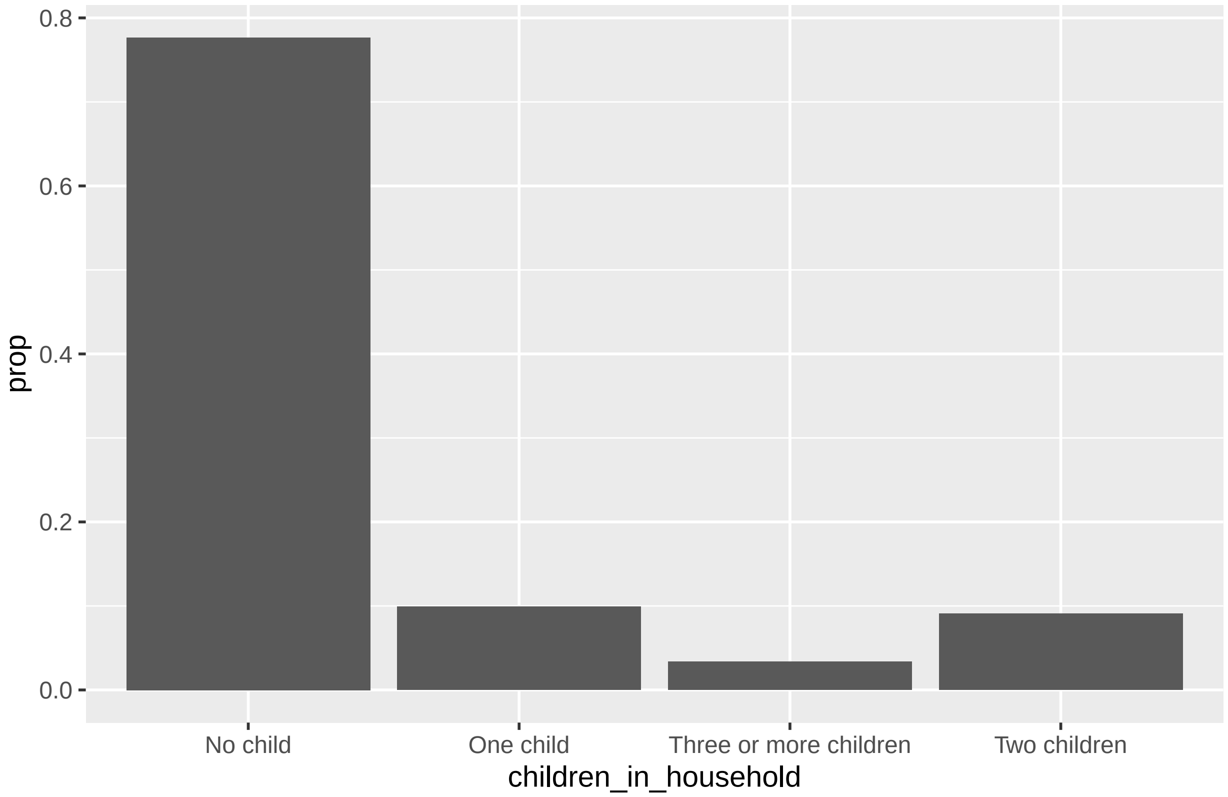


Fig 1.6 total children in household



Total number of children and other variables

Fig 2.1 Age of first child and total children

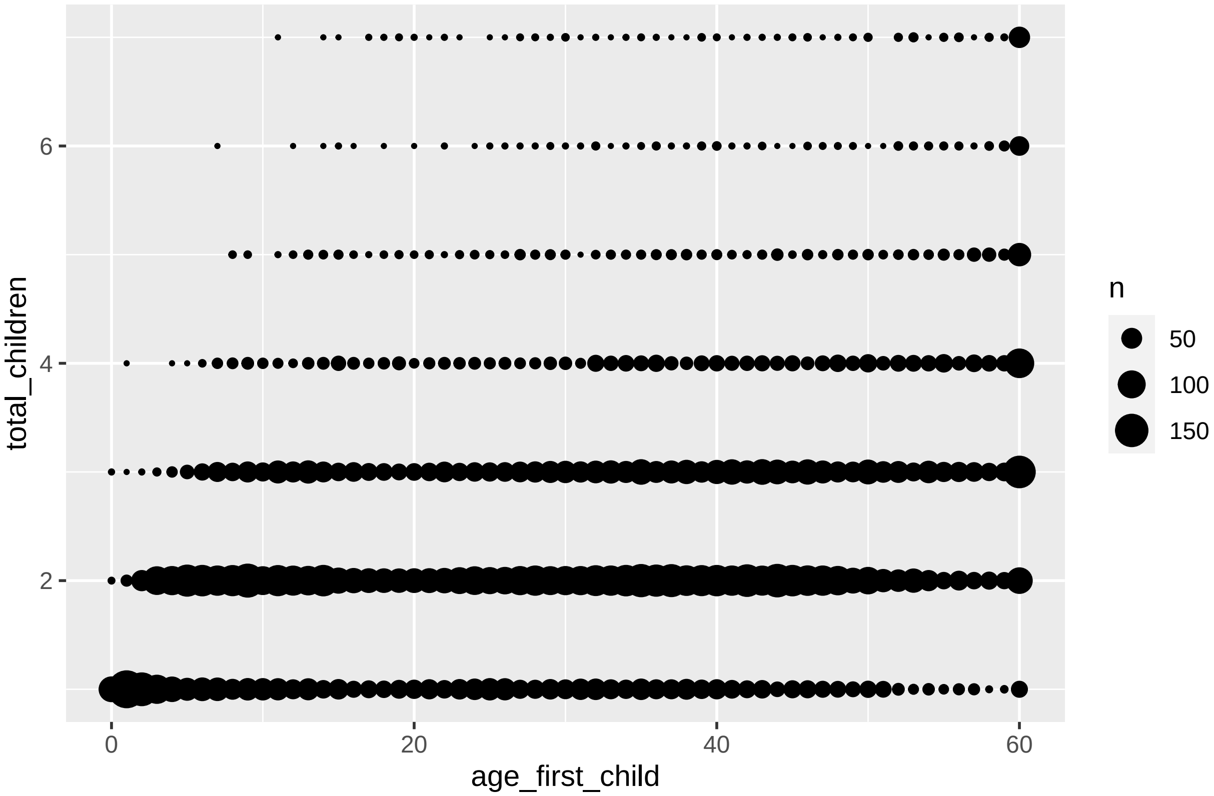


Fig 2.2 Age and total children

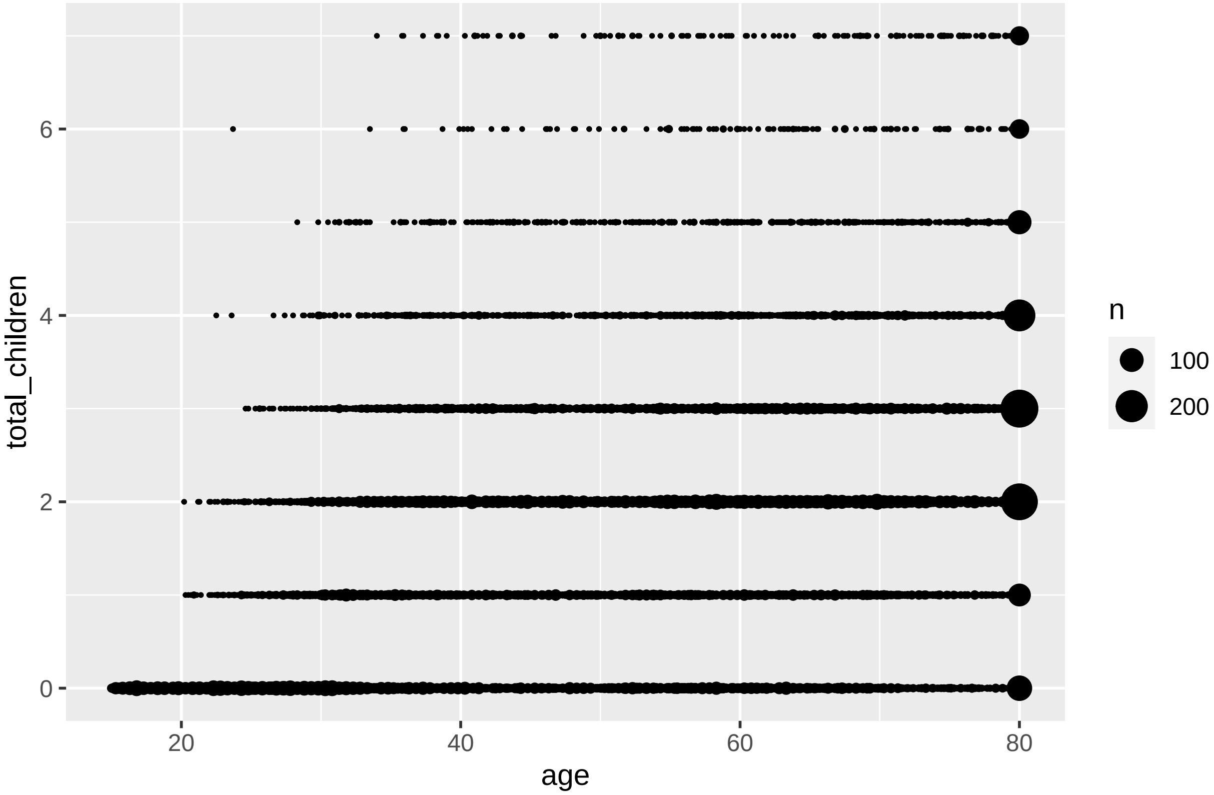


Fig 2.3 total children and age_youngest_child_under_6

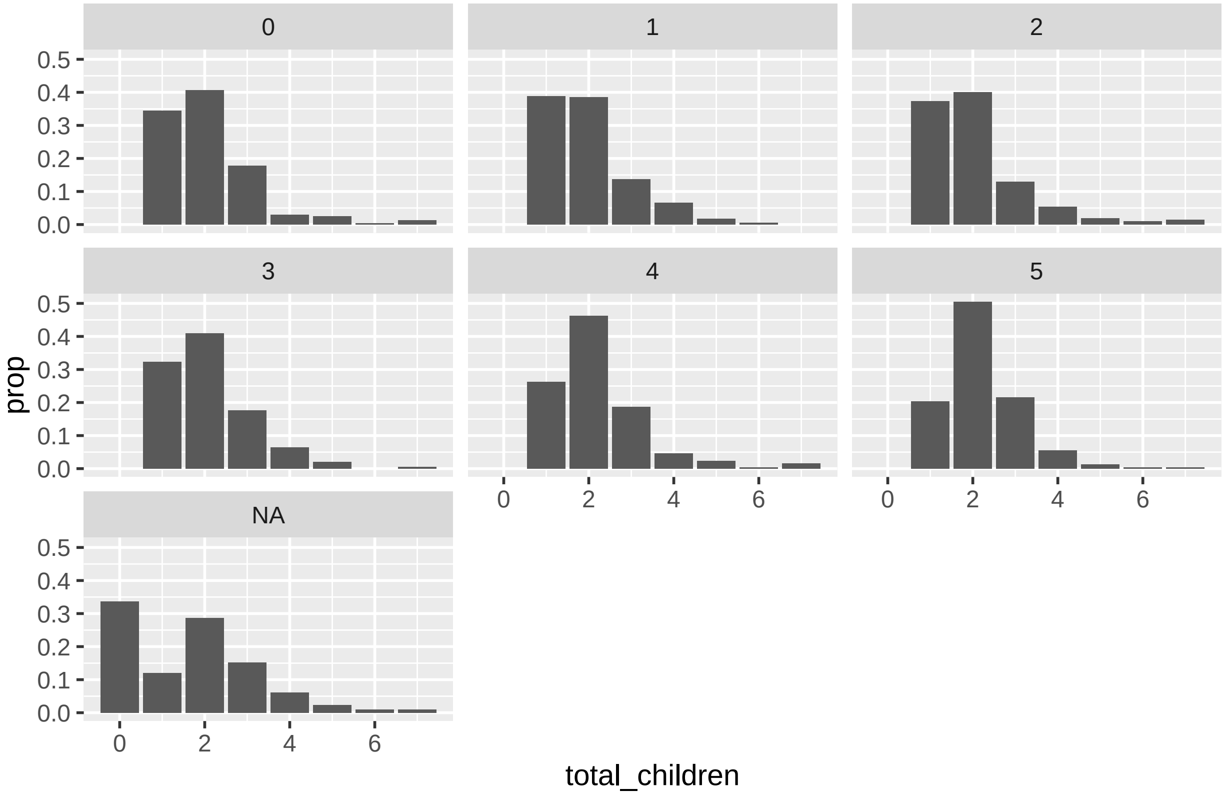


Fig 2.4 total children and number_total_children_intention

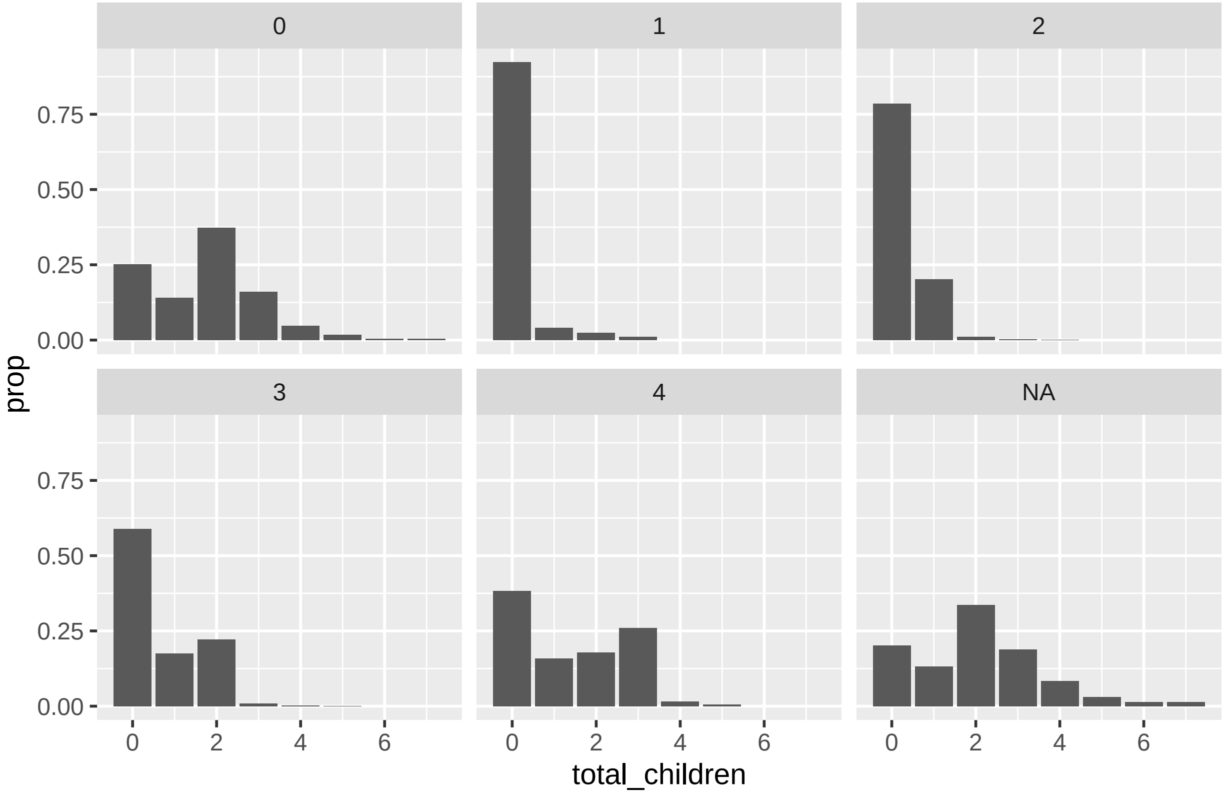
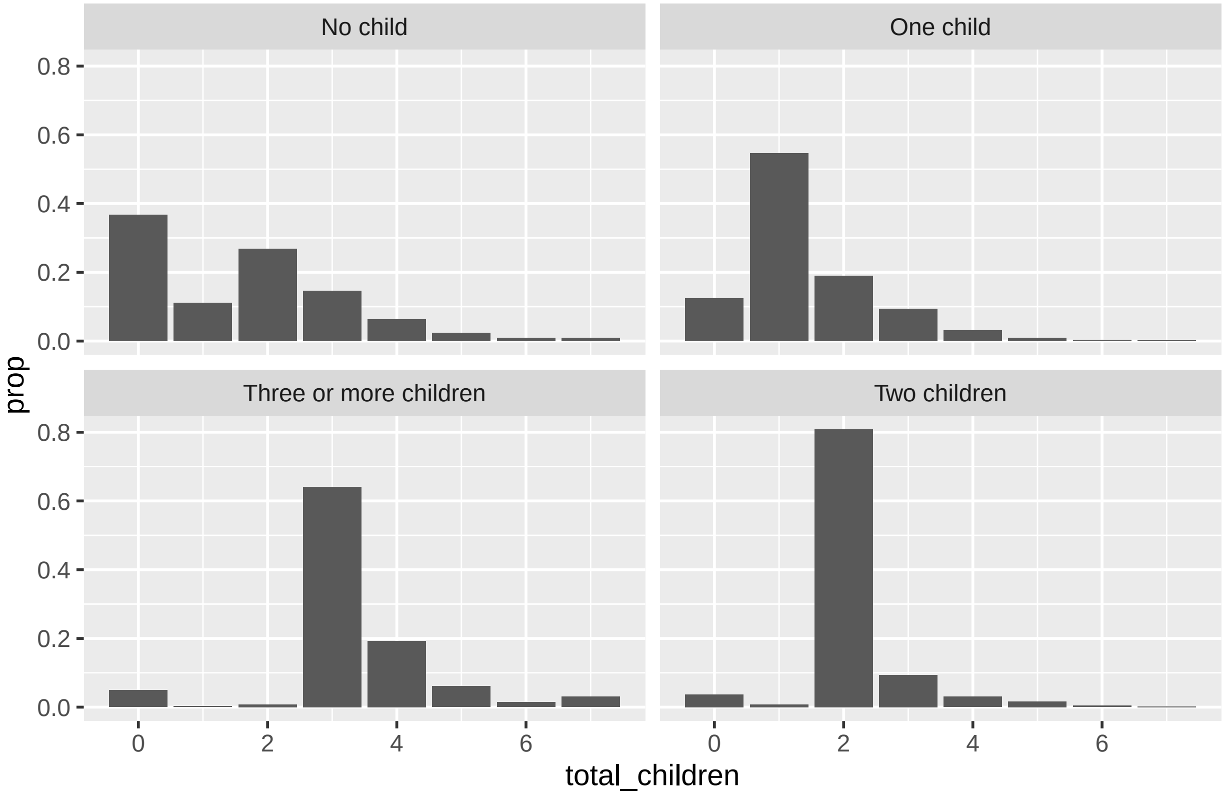


Fig 2.5 total children and children_in_household



Model

The total number of children is a numerical variable, so I choose to use linear regression to predict this variable.

Linear regression is a linear approach to modeling the relationship between a dependent variable and one independent variable.

First, I select age to build a mod.

But R-square of this model is only 0.1992,

I need to add more variables to let my R-square increase.

So, I choose to use multivariate regression,

Multivariate regression is an extension of linear regression to multivariate outcomes.

I think everyone may have an intention before they have a child,

so number_total_children_intention is the second variables that I selected,

and future_children_intention is a similar variable, but it is not a numerical variable, so I did not

select it.

And then I choose age_first_child and age_youngest_child_under_6,

These two variables are based on one they have a child already.

Then, I selected children_in_household as my last variable; this is how many children live in the respondent's house. This variable is based on they have children.

I use those variables to build my second mod.

The R-squared of this mod is 0.8232.

And all variables’ p-value is very small.

Moreover, I think this model is a good model

I have thought about an alternative model, and the variables are age, sex, life feeling score, and income. However, the r-squared of this mod is only 0.22, it must miss some main variables, but I can not find that variables, so I did not use this one.

Software

The software that I used is “lm”, and I can use the p-value and f-value to check my model.

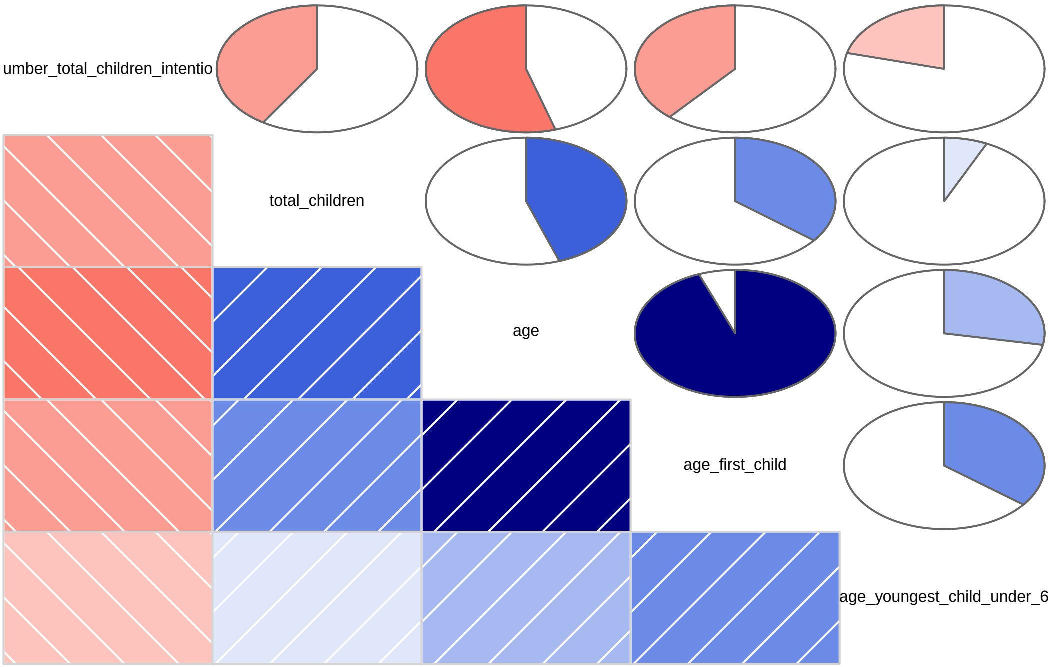
The P-value of the intercept is $< 2.2e-16$ and all variables’ p-value all very small; this means this model can consider being statistically significant.

And the f-value of the variables are large enough to pass the model check.

This is my first try

```
##
## Call:
## lm(formula = total_children ~ age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7205 -0.8271 -0.1662  0.7814  6.0025
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.2760540  0.0288599  -9.565   <2e-16 ***
## age          0.0374574  0.0005235  71.547   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.333 on 20581 degrees of freedom
## Multiple R-squared:  0.1992, Adjusted R-squared:  0.1991
## F-statistic:  5119 on 1 and 20581 DF,  p-value: < 2.2e-16
```

This is the correlation of each variable



This is my second model

```
##
## Call:
## lm(formula = total_children ~ age + age_first_child + age_youngest_child_under_6 +
##     number_total_children_intention + children_in_household,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00406 -0.11295 -0.01119  0.08750  2.84839
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                  1.654933   0.083543  19.809
## age                                         -0.005517   0.001967  -2.805
## age_first_child                              0.108136   0.002557  42.289
## age_youngest_child_under_6                  -0.098306   0.006676 -14.724
## number_total_children_intention             -0.020226   0.008545  -2.367
## children_in_householdOne child              -0.439943   0.049036  -8.972
## children_in_householdThree or more children  1.164339   0.051814  22.472
## children_in_householdTwo children            0.192487   0.048281   3.987
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## age                                          0.00508 **
## age_first_child                              < 2e-16 ***
## age_youngest_child_under_6                   < 2e-16 ***
## number_total_children_intention              0.01804 *
## children_in_householdOne child               < 2e-16 ***
## children_in_householdThree or more children  < 2e-16 ***
## children_in_householdTwo children           6.95e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4221 on 1864 degrees of freedom
##   (18711 observations deleted due to missingness)
## Multiple R-squared:  0.8232, Adjusted R-squared:  0.8225
## F-statistic:  1240 on 7 and 1864 DF,  p-value: < 2.2e-16
```

Mean of total children

```
## [1] 1.678813
```

I give each variable a abbreviation

afc = age_first_child

aycu6 = age_youngest_child_under_6

ntci = number_total_children_intention

1c = children_in_householdOne child

2c = children_in_householdThree or more children

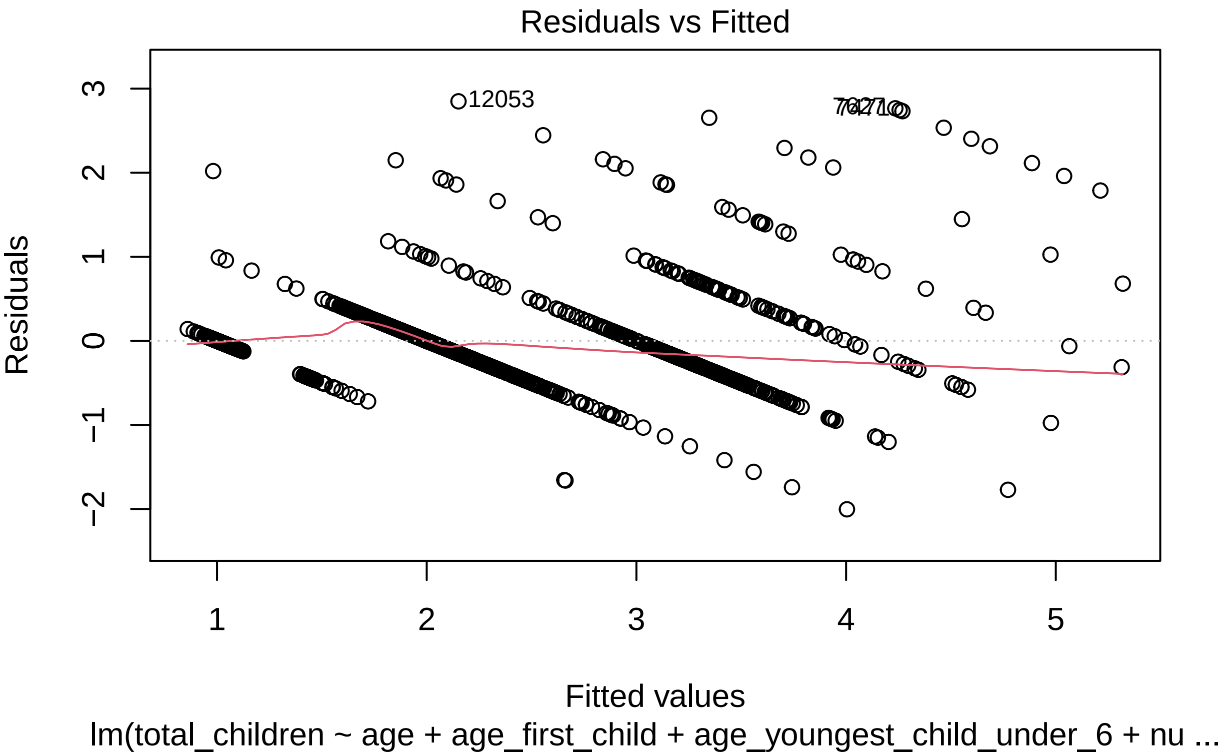
3c = children_in_householdTwo children

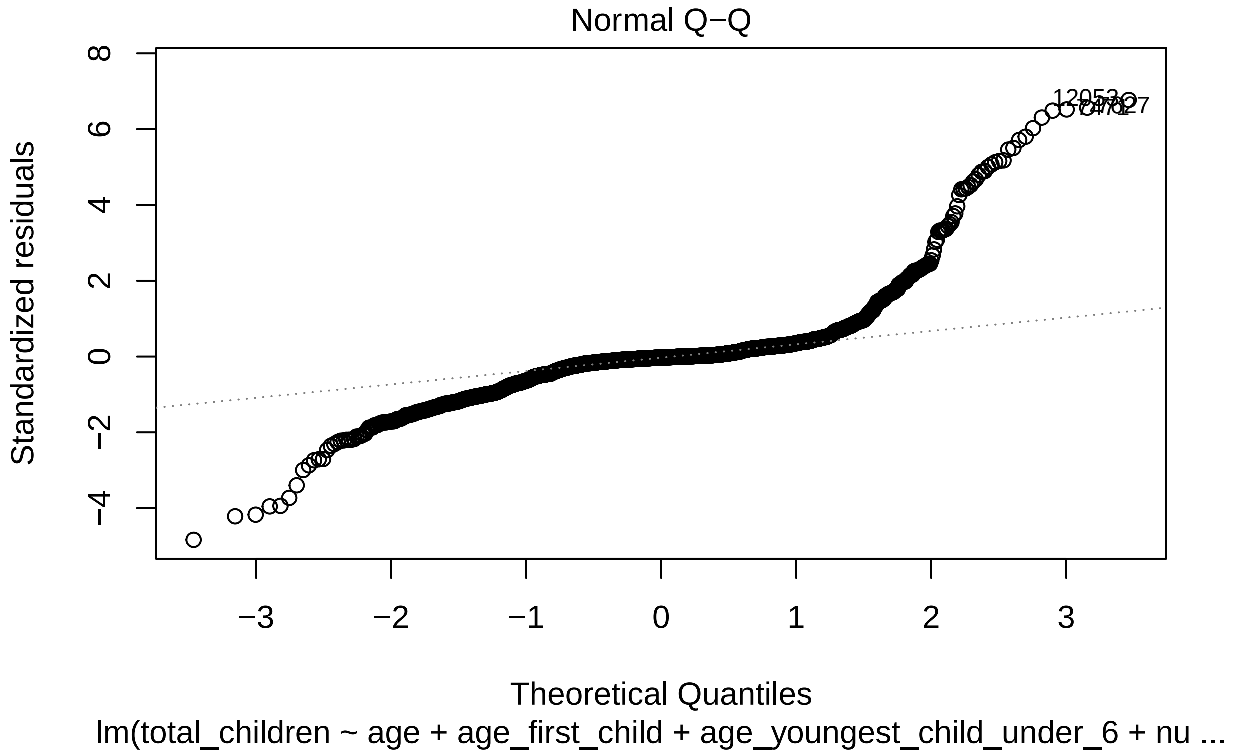
$$totalchildren = \beta_0 + \beta 1 \times age + \beta 2 \times afc + \beta 3 \times aycu6 + \beta 4 \times ntci + \beta 5 \times 1c + \beta 6 \times 3c + \beta 7 \times 2c + \epsilon$$

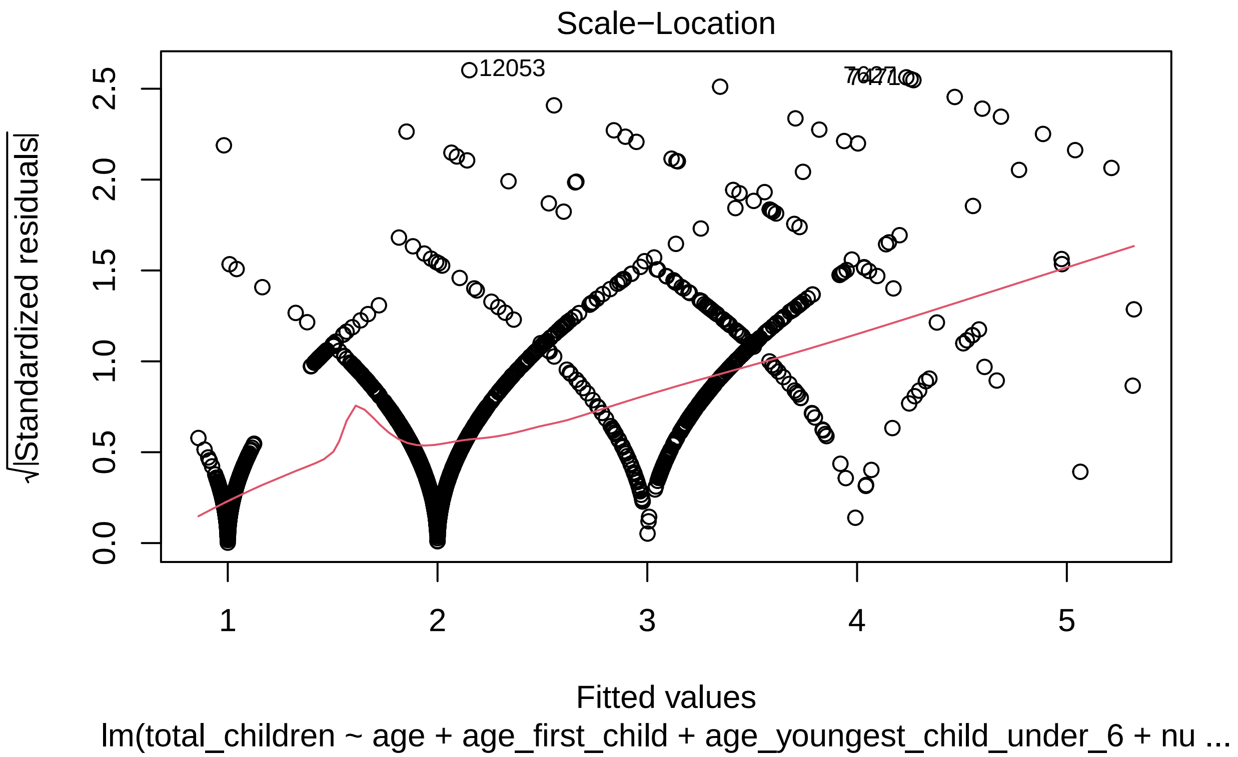
$$\hat{totalchildren} = 1.66 - 0.01 \times age + 0.11 \times afc - 0.1 \times aytcu6 - 0.02 \times ntci - 0.44 \times 1c + 1.16 \times 3c + 0.19 \times 2c$$

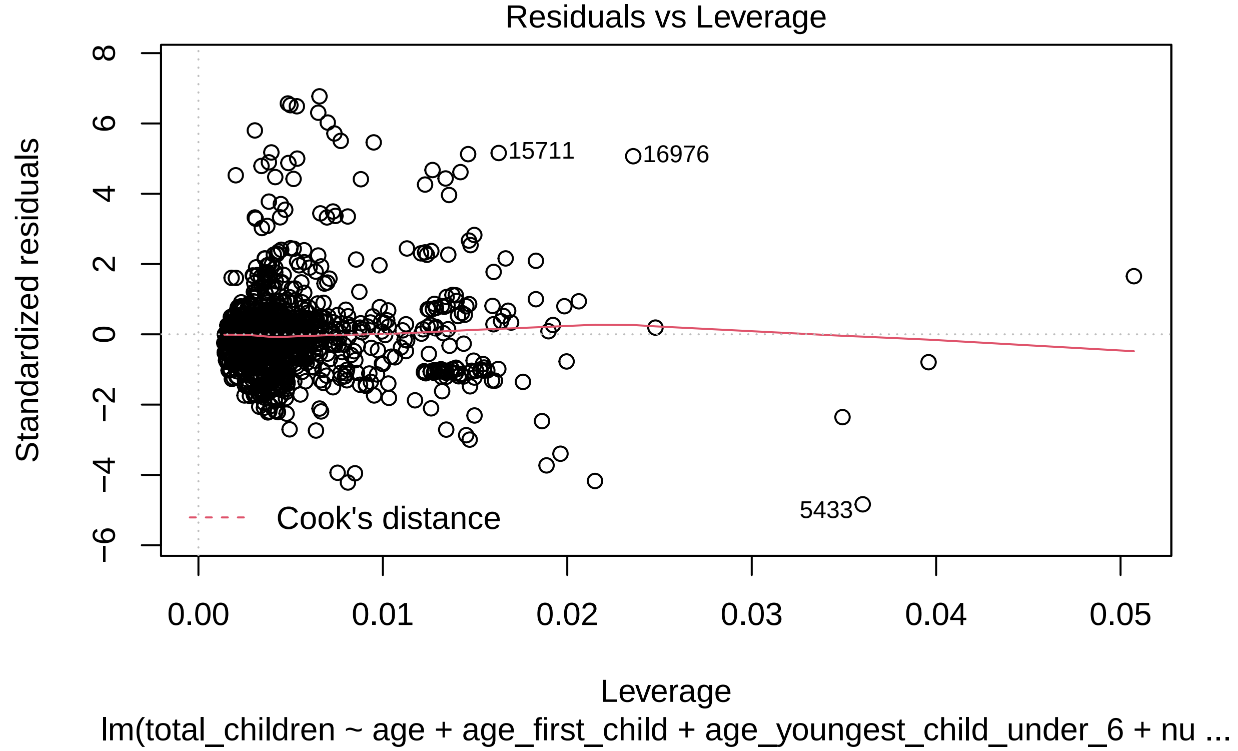
```
## Analysis of Variance Table
##
## Response: total_children
##                                   Df Sum Sq Mean Sq F value    Pr(>F)
## age                                1 145.17  145.17  814.89 < 2.2e-16 ***
## age_first_child                    1 863.81  863.81 4848.92 < 2.2e-16 ***
## age_youngest_child_under_6         1  63.62   63.62  357.15 < 2.2e-16 ***
## number_total_children_intention    1  39.95   39.95  224.23 < 2.2e-16 ***
## children_in_household              3 433.66  144.55  811.43 < 2.2e-16 ***
## Residuals                       1864 332.06    0.18
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Plot of the model









Results

Age: age and total_children are negative correlation, and while age increase one, total_children decrease 0.01.

Age_first_child: age_first_child and total_children are positive correlation, and while age_first_child increase one, total_children increase 0.11.

Age_youngest_child_under_6: age_youngest_child_under_6 and total_children are negative correlation, and while age_youngest_child_under_6 increase one, total_children decrease 0.1.

number_total_children_intention: number_total_children_intention and total_children are negative correlation, and while number_total_children_intention increase one, total_children decrease 0.02.

Children_in_householdOne child: children_in_householdOne child and total_children are negative correlation, and while children_in_householdOne child increase one, total_children decrease 0.44.

Children_in_householdThree or more children: children_in_householdThree or more children and total_children are positive correlation, and while children_in_householdThree or more children increase one, total_children decrease 1.16.

Children_in_householdTwo children: children_in_householdTwochildren and total_children are positive correlation, and while children_in_householdTwochildren increase one, total_children decrease 0.19.

Discussion

Fig2.1 shows that when the age of the first child increase, the total number of children increase. This is the law of nature; when respondents have ten children, their first child's age can not be two years old.

Fig2.2 shows that when age increase, the total number of children increases. This is the law of nature; the respondent can not have ten children when they are ten years old.

However, age and total children negatively correlate in my model; it may be because my intercept is 1.66, and when the respondent is too old, they can not have a new child.

Fig2.3 shows that while the youngest child's age increases, the percentage of two children increases, but the percent of three or more does not increase. However, this variable has a negative correlation; when the youngest child's age increases, the first child's age increases. For example, when the first child is five years old, and in this situation, the first child's age and youngest child's age are the same, it almost does not influence the total number of children because the respondent only has one. However, if the respondent's first child is 40 years old, and the youngest child is one year old, the respondent may have a new child in a few years. If the respondent's youngest child is five years old, by the law of nature, it is almost impossible that the respondent has a new child because the respondent is at least 65 years old; in this situation, the youngest child's age increase, total children decrease.

Fig2.4 shows that while the number of total children intent increases, the total number of children increases. However, this variable has a negative correlation. Then I notice that about 95% do not have a child when respondents intend to have one child. Even with the intent to have four children, still, 40% of respondents have no child. This may be the reason this variable is a negative correlation.

Fig2.5 shows that when the respondent has one child at home, they probably only have one child, but the intercept of this model is 1.66, so this variable is not surprising to have a negative correlation. When they have two children at home, they have at least two children, so this variable has a positive correlation. Moreover, when they have more than three children at home, they at least have three children.

Overall, I found that number of total children intent is not influence the number of total children too much. I can delete this variable from my model.

Weakness

The model that I build is only considered when you already have a child due to variables that I selected, I should select some variables that can show why people choose not to have a child.

And this size of this data is too small and have too much missing value.

Appendix

how to download the file

1.go http://dc.chass.utoronto.ca/myaccess.html

2. Click SDA @ chass and login
3. Continue in English
4. Find GSS
5. Click "Data" on General social survey on Family (cycle 31), 2017.
6. Then download
7. Click file, stata and select all and then click continues
8. create and download

link: https://github.com/nerowangg/-What-will-affect-the-number-of-total-children-in-a-family.git

all code are in github post package.

References

General Social Survey Cycle 31: Families Public Use Microdata File Documentation and User's Guide

2017 General Social Survey: Families Cycle 31 Public Use Microdata File PUMF

gss_cleaning (authors: Rohan Alexander and Sam Caetano) gss_dict.txt gss_labeel.txt

2017 General Social Survey: Families Cycle 31(GSS)

package:

ggplot, corrgram, janitor, tidyverse