GSS Analysis Report

Qiyu Huang& Yuhan Zhu

10/19/2020

## Abstract

We are going to take a analysis on th Canadian General Social Survey. We have used logistic regression and linear regression to see relationships. Deetails are shown later in the discussion part.

## Introduction

We want to find if people are older, will they have more kids, more relationships, and have early kids when they are young? We also want to find if the income level is related with gender, with total number of children, and with satisfaction of their lives. Will the family be richer if they have more kids or poorer is what we want to find out as well. ## Data

## Installing package into '/home/rstudio-user/R/x86\_64-pc-linux-gnu-library/4.0'  
## (as 'lib' is unspecified)  
## Installing package into '/home/rstudio-user/R/x86\_64-pc-linux-gnu-library/4.0'  
## (as 'lib' is unspecified)  
## Installing package into '/home/rstudio-user/R/x86\_64-pc-linux-gnu-library/4.0'  
## (as 'lib' is unspecified)

## Warning in install.packages("readr", repos = "http://cran.us.r-project.org"):  
## installation of package 'readr' had non-zero exit status

## Installing package into '/home/rstudio-user/R/x86\_64-pc-linux-gnu-library/4.0'  
## (as 'lib' is unspecified)

##   
## Attaching package: 'janitor'

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

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.4 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# Data cleaning Part

### importing raw data

raw\_data <- read\_csv(“/cloud/project/AAGe4G0U.csv”) dict <- read\_lines(“gss\_dict.txt”, skip = 18) labels\_raw <- read\_file(“gss\_labels.txt”)

### set up dictionary

variable\_descriptions <- as\_tibble(dict) %>% filter(value!=“}”) %>% mutate(value = str\_replace(value, ".+%[0-9].\*f[ ]{2,}“,”“)) %>% mutate(value = str\_remove\_all(value,”"")) %>% rename(variable\_description = value) %>% bind\_cols(tibble(variable\_name = colnames(raw\_data)[-1]))

### variable names and values

labels\_raw\_tibble <- as\_tibble(str\_split(labels\_raw, “;”)[[1]]) %>% filter(row\_number()!=1) %>% mutate(value = str\_remove(value, “define”)) %>% mutate(value = str\_replace(value, “[ ]{2,}”, “XXX”)) %>% mutate(splits = str\_split(value, “XXX”)) %>% rowwise() %>% mutate(variable\_name = splits[1], cases = splits[2]) %>% mutate(cases = str\_replace\_all(cases, “”, "")) %>% select(variable\_name, cases) %>% drop\_na()

### variable name

labels\_raw\_tibble <- labels\_raw\_tibble %>% mutate(splits = str\_split(cases, “[ ]{0,}"[ ]{0,}”))

### creating a function

add\_cw\_text <- function(x, y){ if(!is.na(as.numeric(x))){ x\_new <- paste0(y, “==”, x,“~”) } else{ x\_new <- paste0(“"”,x,“",”) } return(x\_new) } ### Another function cw\_statements <- labels\_raw\_tibble %>% rowwise() %>% mutate(splits\_with\_cw\_text = list(modify(splits, add\_cw\_text, y = variable\_name))) %>% mutate(cw\_statement = paste(splits\_with\_cw\_text, collapse = "“)) %>% mutate(cw\_statement = paste0(”case\_when(“, cw\_statement,”TRUE~"NA")“)) %>% mutate(cw\_statement = str\_replace(cw\_statement,”,"",“,”,")) %>% select(variable\_name, cw\_statement)

### Do some final cleans of this function

cw\_statements <- cw\_statements %>% mutate(variable\_name = str\_remove\_all(variable\_name, “\r”)) %>% mutate(cw\_statement = str\_remove\_all(cw\_statement, “\r”))

### Apply that dictionary to the raw data

gss <- raw\_data %>% select(CASEID, agedc, achd\_1c, achdmpl, totchdc, acu0c, agema1c, achb1c, rsh\_131a, arretwk, slm\_01, sex, brthcan, brthfcan, brthmcan, brthmacr, brthprvc, yrarri, prv, region, luc\_rst, marstat, amb\_01, vismin, alndimmg, bpr\_16, bpr\_19, ehg3\_01b, odr\_10, livarr12, dwelc, hsdsizec, brthpcan, brtpprvc, visminpr, rsh\_125a, eop\_200, uhw\_16gr, lmam\_01, acmpryr, srh\_110, srh\_115, religflg, rlr\_110, lanhome, lan\_01, famincg2, ttlincg2, noc1610, cc\_20\_1, cc\_30\_1, ccmoc1c, cor\_031, cor\_041, cu0rnkc, pr\_cl, chh0014c, nochricc, grndpa, gparliv, evermar, ma0\_220, nmarevrc, ree\_02, rsh\_131b, rto\_101, rto\_110, rto\_120, rtw\_300, sts\_410, csp\_105, csp\_110a, csp\_110b, csp\_110c, csp\_110d, csp\_160, fi\_110) %>% mutate\_at(vars(agedc:fi\_110), .funs = funs(ifelse(.>=96, NA, .))) %>% mutate\_at(.vars = vars(sex:fi\_110), .funs = funs(eval(parse(text = cw\_statements %>% filter(variable\_name==deparse(substitute(.))) %>% select(cw\_statement) %>% pull()))))

### Change the attributes name

gss <- gss %>% clean\_names() %>% rename(age = agedc, age\_first\_child = achd\_1c, age\_youngest\_child\_under\_6 = achdmpl, total\_children = totchdc, age\_start\_relationship = acu0c, age\_at\_first\_marriage = agema1c, age\_at\_first\_birth = achb1c, distance\_between\_houses = rsh\_131a, age\_youngest\_child\_returned\_work = arretwk, feelings\_life = slm\_01, sex = sex, place\_birth\_canada = brthcan, place\_birth\_father = brthfcan, place\_birth\_mother = brthmcan, place\_birth\_macro\_region = brthmacr, place\_birth\_province = brthprvc, year\_arrived\_canada = yrarri, province = prv, region = region, pop\_center = luc\_rst, marital\_status = marstat, aboriginal = amb\_01, vis\_minority = vismin, age\_immigration = alndimmg, landed\_immigrant = bpr\_16, citizenship\_status = bpr\_19, education = ehg3\_01b, own\_rent = odr\_10, living\_arrangement = livarr12, hh\_type = dwelc, hh\_size = hsdsizec, partner\_birth\_country = brthpcan, partner\_birth\_province = brtpprvc, partner\_vis\_minority = visminpr, partner\_sex = rsh\_125a, partner\_education = eop\_200, average\_hours\_worked = uhw\_16gr, worked\_last\_week = lmam\_01, partner\_main\_activity = acmpryr, self\_rated\_health = srh\_110, self\_rated\_mental\_health = srh\_115, religion\_has\_affiliation = religflg, regilion\_importance = rlr\_110, language\_home = lanhome, language\_knowledge = lan\_01, income\_family = famincg2, income\_respondent = ttlincg2, occupation = noc1610, childcare\_regular = cc\_20\_1, childcare\_type = cc\_30\_1, childcare\_monthly\_cost = ccmoc1c, ever\_fathered\_child = cor\_031, ever\_given\_birth = cor\_041, number\_of\_current\_union = cu0rnkc, lives\_with\_partner = pr\_cl, children\_in\_household = chh0014c, number\_total\_children\_intention = nochricc, has\_grandchildren = grndpa, grandparents\_still\_living = gparliv, ever\_married = evermar, current\_marriage\_is\_first = ma0\_220, number\_marriages = nmarevrc, religion\_participation = ree\_02, partner\_location\_residence = rsh\_131b, full\_part\_time\_work = rto\_101, time\_off\_work\_birth = rto\_110, reason\_no\_time\_off\_birth = rto\_120, returned\_same\_job = rtw\_300, satisfied\_time\_children = sts\_410, provide\_or\_receive\_fin\_supp = csp\_105, fin\_supp\_child\_supp = csp\_110a, fin\_supp\_child\_exp = csp\_110b, fin\_supp\_lump = csp\_110c, fin\_supp\_other = csp\_110d, fin\_supp\_agreement = csp\_160, future\_children\_intention = fi\_110)

### Clean up

gss <- gss %>% mutate\_at(vars(age:future\_children\_intention), .funs = funs(ifelse(.==“Valid skip”|.==“Refusal”|.==“Not stated”, “NA”, .))) gss <- gss %>% mutate(is\_male = ifelse(sex==“Male”, 1, 0)) gss <- gss %>% mutate\_at(vars(fin\_supp\_child\_supp:fin\_supp\_other), .funs = funs(case\_when( .==“Yes”1, .==“No”0, .==“NA”~as.numeric(NA) ))) main\_act <- raw\_data %>% mutate(main\_activity = case\_when( mpl\_105a==“Yes”~ “Working at a paid job/business”, mpl\_105b==“Yes” ~ “Looking for paid work”, mpl\_105c==“Yes” ~ “Going to school”, mpl\_105d==“Yes” ~ “Caring for children”, mpl\_105e==“Yes” ~ “Household work”, mpl\_105i==“Yes” ~ “Other”, TRUE~ “NA”)) %>% select(main\_activity) %>% pull() age\_diff <- raw\_data %>% select(marstat, aprcu0c, adfgrma0) %>% mutate\_at(.vars = vars(aprcu0c:adfgrma0), .funs = funs(eval(parse(text = cw\_statements %>% filter(variable\_name==deparse(substitute(.))) %>% select(cw\_statement) %>% pull())))) %>% mutate(age\_diff = ifelse(marstat==“Living common-law”, aprcu0c, adfgrma0)) %>% mutate\_at(vars(age\_diff), .funs = funs(ifelse(.==“Valid skip”|.==“Refusal”|.==“Not stated”, “NA”, .))) %>% select(age\_diff) %>% pull() gss <- gss %>% mutate(main\_activity = main\_act, age\_diff = age\_diff) gss <- gss %>% rowwise() %>% mutate(hh\_size = str\_remove(string = hh\_size, pattern = “\ .*")) %>% mutate(hh\_size = case\_when( hh\_size=="One" ~ 1, hh\_size=="Two" ~ 2, hh\_size=="Three" ~ 3, hh\_size=="Four" ~ 4, hh\_size=="Five" ~ 5, hh\_size=="Six" ~ 6 )) gss <- gss %>% rowwise() %>% mutate(number\_marriages = str\_remove(string = number\_marriages, pattern = "\ .*”)) %>% mutate(number\_marriages = case\_when( number\_marriages==“No” ~ 0, number\_marriages==“One” ~ 1, number\_marriages==“Two” ~ 2, number\_marriages==“Three” ~ 3, number\_marriages==“Four” ~ 4 )) gss <- gss %>% rowwise() %>% mutate(number\_total\_children\_known = ifelse(number\_total\_children\_intention==“Don’t know”|number\_total\_children\_intention==“NA”, 0, 1)) %>% mutate(number\_total\_children\_intention = str\_remove(string = number\_total\_children\_intention, pattern = "\ .\*“)) %>% mutate(number\_total\_children\_intention = case\_when( number\_total\_children\_intention==”None" ~ 0, number\_total\_children\_intention==“One” ~ 1, number\_total\_children\_intention==“Two” ~ 2, number\_total\_children\_intention==“Three” ~ 3, number\_total\_children\_intention==“Four” ~ 4, number\_total\_children\_intention==“Don’t” ~ as.numeric(NA) ))

### save to a new csv file and finish cleaning

write\_csv(gss, “gss.csv”)

## Model

I have used ggplot, histogram, logistic regression, linear regression to analaysis these datasets.

This is the mathmetical notation for linear regression :y = β0 + β1x + e, where β0 is the intercept, there could be β2,β3 and so on. Y must be numerical. Predictors can be both numerical and categorical. In this case, I have picked some numerical variables as predictors and age as Y.

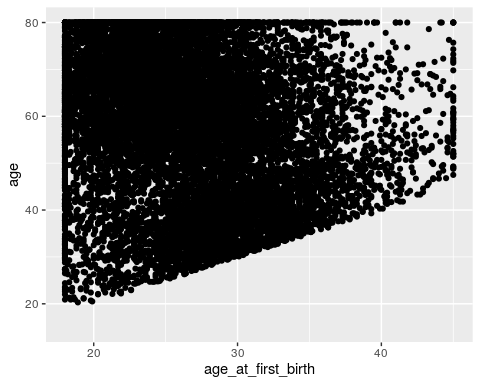
This is the mathmetical notation for logistic regression : log(p/(1-p)) = β0 + β1x,where β0 is the intercept, there could be β2,β3 and so on. P is the probability of an event that is going to occur. β1 is a coefficient represents changes in log adds for every one unit increase in x. Y can either be 1 or 0.

## Results

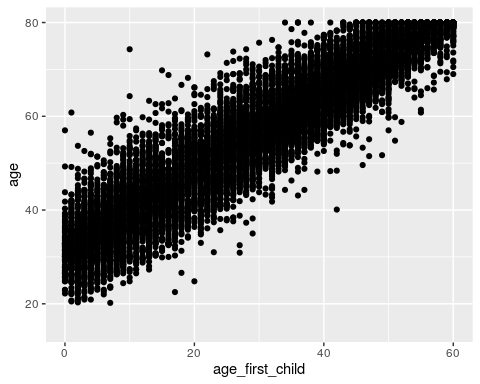
##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## .default = col\_character(),  
## caseid = col\_double(),  
## age = col\_double(),  
## age\_first\_child = col\_double(),  
## age\_youngest\_child\_under\_6 = col\_double(),  
## total\_children = col\_double(),  
## age\_start\_relationship = col\_double(),  
## age\_at\_first\_marriage = col\_double(),  
## age\_at\_first\_birth = col\_double(),  
## distance\_between\_houses = col\_double(),  
## age\_youngest\_child\_returned\_work = col\_double(),  
## feelings\_life = col\_double(),  
## hh\_size = col\_double(),  
## number\_total\_children\_intention = col\_double(),  
## number\_marriages = col\_double(),  
## fin\_supp\_child\_supp = col\_double(),  
## fin\_supp\_child\_exp = col\_double(),  
## fin\_supp\_lump = col\_double(),  
## fin\_supp\_other = col\_double(),  
## is\_male = col\_double(),  
## main\_activity = col\_logical()  
## # ... with 1 more columns  
## )  
## ℹ Use `spec()` for the full column specifications.

## caseid age age\_first\_child age\_youngest\_child\_under\_6  
## Min. : 1 Min. :15.00 Min. : 0.00 Min. :0.000   
## 1st Qu.: 5151 1st Qu.:37.30 1st Qu.:15.00 1st Qu.:1.000   
## Median :10302 Median :54.20 Median :32.00 Median :2.000   
## Mean :10302 Mean :52.19 Mean :30.57 Mean :2.412   
## 3rd Qu.:15452 3rd Qu.:66.78 3rd Qu.:44.00 3rd Qu.:4.000   
## Max. :20602 Max. :80.00 Max. :60.00 Max. :5.000   
## NA's :6835 NA's :18488   
## total\_children age\_start\_relationship age\_at\_first\_marriage  
## Min. :0.000 Min. :18.00 Min. :15.0   
## 1st Qu.:0.000 1st Qu.:25.00 1st Qu.:20.5   
## Median :2.000 Median :30.50 Median :22.8   
## Mean :1.679 Mean :33.63 Mean :24.1   
## 3rd Qu.:3.000 3rd Qu.:40.62 3rd Qu.:26.4   
## Max. :7.000 Max. :60.00 Max. :50.0   
## NA's :19 NA's :18566 NA's :15248   
## age\_at\_first\_birth distance\_between\_houses age\_youngest\_child\_returned\_work  
## Min. :18.00 Min. : 0.00 Min. : 0.200   
## 1st Qu.:22.80 1st Qu.: 4.00 1st Qu.: 0.500   
## Median :26.40 Median :10.00 Median : 6.000   
## Mean :26.86 Mean :17.13 Mean : 6.589   
## 3rd Qu.:30.30 3rd Qu.:24.75 3rd Qu.:12.000   
## Max. :45.00 Max. :90.00 Max. :48.000   
## NA's :7865 NA's :19476 NA's :19466   
## feelings\_life sex place\_birth\_canada place\_birth\_father  
## Min. : 0.000 Length:20602 Length:20602 Length:20602   
## 1st Qu.: 7.000 Class :character Class :character Class :character   
## Median : 8.000 Mode :character Mode :character Mode :character   
## Mean : 8.094   
## 3rd Qu.: 9.000   
## Max. :10.000   
## NA's :271   
## place\_birth\_mother place\_birth\_macro\_region place\_birth\_province  
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## year\_arrived\_canada province region pop\_center   
## Length:20602 Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## marital\_status aboriginal vis\_minority age\_immigration   
## Length:20602 Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## landed\_immigrant citizenship\_status education own\_rent   
## Length:20602 Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## living\_arrangement hh\_type hh\_size partner\_birth\_country  
## Length:20602 Length:20602 Min. :1.000 Length:20602   
## Class :character Class :character 1st Qu.:1.000 Class :character   
## Mode :character Mode :character Median :2.000 Mode :character   
## Mean :2.347   
## 3rd Qu.:3.000   
## Max. :6.000   
##   
## partner\_birth\_province partner\_vis\_minority partner\_sex   
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## partner\_education average\_hours\_worked worked\_last\_week   
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## partner\_main\_activity self\_rated\_health self\_rated\_mental\_health  
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## religion\_has\_affiliation regilion\_importance language\_home   
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## language\_knowledge income\_family income\_respondent occupation   
## Length:20602 Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## childcare\_regular childcare\_type childcare\_monthly\_cost  
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## ever\_fathered\_child ever\_given\_birth number\_of\_current\_union  
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## lives\_with\_partner children\_in\_household number\_total\_children\_intention  
## Length:20602 Length:20602 Min. :0.000   
## Class :character Class :character 1st Qu.:0.000   
## Mode :character Mode :character Median :0.000   
## Mean :0.903   
## 3rd Qu.:2.000   
## Max. :4.000   
## NA's :12202   
## has\_grandchildren grandparents\_still\_living ever\_married   
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## current\_marriage\_is\_first number\_marriages religion\_participation  
## Length:20602 Min. :0.0000 Length:20602   
## Class :character 1st Qu.:0.0000 Class :character   
## Mode :character Median :1.0000 Mode :character   
## Mean :0.7989   
## 3rd Qu.:1.0000   
## Max. :4.0000   
##   
## partner\_location\_residence full\_part\_time\_work time\_off\_work\_birth  
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## reason\_no\_time\_off\_birth returned\_same\_job satisfied\_time\_children  
## Length:20602 Length:20602 Length:20602   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## provide\_or\_receive\_fin\_supp fin\_supp\_child\_supp fin\_supp\_child\_exp  
## Length:20602 Min. :0.000 Min. :0.000   
## Class :character 1st Qu.:1.000 1st Qu.:0.000   
## Mode :character Median :1.000 Median :0.000   
## Mean :0.765 Mean :0.339   
## 3rd Qu.:1.000 3rd Qu.:1.000   
## Max. :1.000 Max. :1.000   
## NA's :20057 NA's :20057   
## fin\_supp\_lump fin\_supp\_other fin\_supp\_agreement future\_children\_intention  
## Min. :0.000 Min. :0.000 Length:20602 Length:20602   
## 1st Qu.:0.000 1st Qu.:0.000 Class :character Class :character   
## Median :0.000 Median :0.000 Mode :character Mode :character   
## Mean :0.055 Mean :0.055   
## 3rd Qu.:0.000 3rd Qu.:0.000   
## Max. :1.000 Max. :1.000   
## NA's :20057 NA's :20057   
## is\_male main\_activity age\_diff number\_total\_children\_known  
## Min. :0.0000 Mode:logical Length:20602 Min. :0.0000   
## 1st Qu.:0.0000 NA's:20602 Class :character 1st Qu.:0.0000   
## Median :0.0000 Mode :character Median :0.0000   
## Mean :0.4562 Mean :0.4123   
## 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000   
##

## Warning: Removed 7865 rows containing missing values (geom\_point).



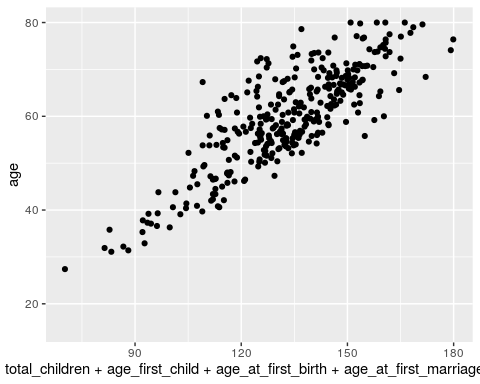
## Warning: Removed 6835 rows containing missing values (geom\_point).



## Warning: Removed 19 rows containing missing values (geom\_point).



## Warning: Removed 20272 rows containing missing values (geom\_point).



## linear regression

##   
## Call:  
## lm(formula = age ~ total\_children + age\_first\_child + age\_at\_first\_birth +   
## age\_at\_first\_marriage + age\_start\_relationship, data = gss)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7577 -0.3441 -0.0071 0.2994 13.8028   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.541116 0.470578 -1.150 0.2510   
## total\_children -0.010812 0.040020 -0.270 0.7872   
## age\_first\_child 0.985121 0.005762 170.956 <2e-16 \*\*\*  
## age\_at\_first\_birth 1.018111 0.013588 74.930 <2e-16 \*\*\*  
## age\_at\_first\_marriage -0.005123 0.012937 -0.396 0.6924   
## age\_start\_relationship 0.015489 0.006343 2.442 0.0151 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9223 on 324 degrees of freedom  
## (20272 observations deleted due to missingness)  
## Multiple R-squared: 0.9925, Adjusted R-squared: 0.9924   
## F-statistic: 8565 on 5 and 324 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = age ~ total\_children + income\_family, data = gss)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55.970 -12.472 1.057 12.067 41.785   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.75129 0.35396 112.304 < 2e-16 \*\*\*  
## total\_children 5.30882 0.07226 73.473 < 2e-16 \*\*\*  
## income\_family$125,000 and more -1.53603 0.40167 -3.824 0.000132 \*\*\*  
## income\_family$25,000 to $49,999 8.06564 0.40688 19.823 < 2e-16 \*\*\*  
## income\_family$50,000 to $74,999 5.46642 0.41852 13.061 < 2e-16 \*\*\*  
## income\_family$75,000 to $99,999 1.89703 0.43850 4.326 1.52e-05 \*\*\*  
## income\_familyLess than $25,000 6.86210 0.44340 15.476 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.44 on 20576 degrees of freedom  
## (19 observations deleted due to missingness)  
## Multiple R-squared: 0.2438, Adjusted R-squared: 0.2436   
## F-statistic: 1106 on 6 and 20576 DF, p-value: < 2.2e-16

## Logistic Regression

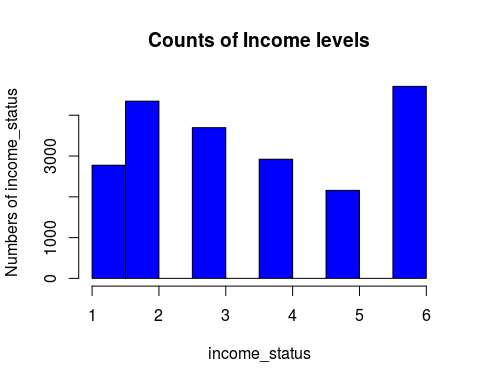
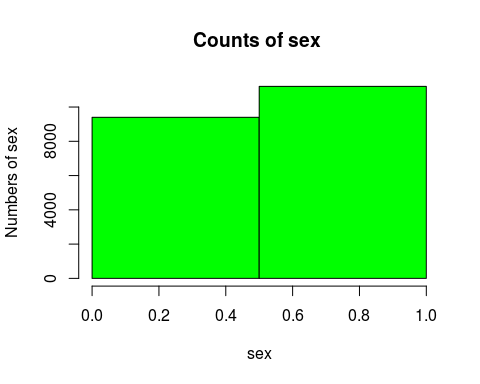
## # A tibble: 6 x 81  
## caseid age age\_first\_child age\_youngest\_ch… total\_children age\_start\_relat…  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 52.7 27 NA 1 NA   
## 2 2 51.1 33 NA 5 NA   
## 3 3 63.6 40 NA 5 NA   
## 4 4 80 56 NA 1 NA   
## 5 5 28 NA NA 0 25.3  
## 6 6 63 37 NA 2 NA   
## # … with 75 more variables: age\_at\_first\_marriage <dbl>,  
## # age\_at\_first\_birth <dbl>, distance\_between\_houses <dbl>,  
## # age\_youngest\_child\_returned\_work <dbl>, feelings\_life <dbl>, sex <chr>,  
## # place\_birth\_canada <chr>, place\_birth\_father <chr>,  
## # place\_birth\_mother <chr>, place\_birth\_macro\_region <chr>,  
## # place\_birth\_province <chr>, year\_arrived\_canada <chr>, province <chr>,  
## # region <chr>, pop\_center <chr>, marital\_status <chr>, aboriginal <chr>,  
## # vis\_minority <chr>, age\_immigration <chr>, landed\_immigrant <chr>,  
## # citizenship\_status <chr>, education <chr>, own\_rent <chr>,  
## # living\_arrangement <chr>, hh\_type <chr>, hh\_size <dbl>,  
## # partner\_birth\_country <chr>, partner\_birth\_province <chr>,  
## # partner\_vis\_minority <chr>, partner\_sex <chr>, partner\_education <chr>,  
## # average\_hours\_worked <chr>, worked\_last\_week <chr>,  
## # partner\_main\_activity <chr>, self\_rated\_health <chr>,  
## # self\_rated\_mental\_health <chr>, religion\_has\_affiliation <chr>,  
## # regilion\_importance <chr>, language\_home <chr>, language\_knowledge <chr>,  
## # income\_family <chr>, income\_respondent <chr>, occupation <chr>,  
## # childcare\_regular <chr>, childcare\_type <chr>,  
## # childcare\_monthly\_cost <chr>, ever\_fathered\_child <chr>,  
## # ever\_given\_birth <chr>, number\_of\_current\_union <chr>,  
## # lives\_with\_partner <chr>, children\_in\_household <chr>,  
## # number\_total\_children\_intention <dbl>, has\_grandchildren <chr>,  
## # grandparents\_still\_living <chr>, ever\_married <chr>,  
## # current\_marriage\_is\_first <chr>, number\_marriages <dbl>,  
## # religion\_participation <chr>, partner\_location\_residence <chr>,  
## # full\_part\_time\_work <chr>, time\_off\_work\_birth <chr>,  
## # reason\_no\_time\_off\_birth <chr>, returned\_same\_job <chr>,  
## # satisfied\_time\_children <chr>, provide\_or\_receive\_fin\_supp <chr>,  
## # fin\_supp\_child\_supp <dbl>, fin\_supp\_child\_exp <dbl>, fin\_supp\_lump <dbl>,  
## # fin\_supp\_other <dbl>, fin\_supp\_agreement <chr>,  
## # future\_children\_intention <chr>, is\_male <dbl>, main\_activity <lgl>,  
## # age\_diff <chr>, number\_total\_children\_known <dbl>

## # A tibble: 6 x 82  
## caseid age age\_first\_child age\_youngest\_ch… total\_children age\_start\_relat…  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 52.7 27 NA 1 NA   
## 2 2 51.1 33 NA 5 NA   
## 3 3 63.6 40 NA 5 NA   
## 4 4 80 56 NA 1 NA   
## 5 5 28 NA NA 0 25.3  
## 6 6 63 37 NA 2 NA   
## # … with 76 more variables: age\_at\_first\_marriage <dbl>,  
## # age\_at\_first\_birth <dbl>, distance\_between\_houses <dbl>,  
## # age\_youngest\_child\_returned\_work <dbl>, feelings\_life <dbl>, sex <chr>,  
## # place\_birth\_canada <chr>, place\_birth\_father <chr>,  
## # place\_birth\_mother <chr>, place\_birth\_macro\_region <chr>,  
## # place\_birth\_province <chr>, year\_arrived\_canada <chr>, province <chr>,  
## # region <chr>, pop\_center <chr>, marital\_status <chr>, aboriginal <chr>,  
## # vis\_minority <chr>, age\_immigration <chr>, landed\_immigrant <chr>,  
## # citizenship\_status <chr>, education <chr>, own\_rent <chr>,  
## # living\_arrangement <chr>, hh\_type <chr>, hh\_size <dbl>,  
## # partner\_birth\_country <chr>, partner\_birth\_province <chr>,  
## # partner\_vis\_minority <chr>, partner\_sex <chr>, partner\_education <chr>,  
## # average\_hours\_worked <chr>, worked\_last\_week <chr>,  
## # partner\_main\_activity <chr>, self\_rated\_health <chr>,  
## # self\_rated\_mental\_health <chr>, religion\_has\_affiliation <chr>,  
## # regilion\_importance <chr>, language\_home <chr>, language\_knowledge <chr>,  
## # income\_family <chr>, income\_respondent <chr>, occupation <chr>,  
## # childcare\_regular <chr>, childcare\_type <chr>,  
## # childcare\_monthly\_cost <chr>, ever\_fathered\_child <chr>,  
## # ever\_given\_birth <chr>, number\_of\_current\_union <chr>,  
## # lives\_with\_partner <chr>, children\_in\_household <chr>,  
## # number\_total\_children\_intention <dbl>, has\_grandchildren <chr>,  
## # grandparents\_still\_living <chr>, ever\_married <chr>,  
## # current\_marriage\_is\_first <chr>, number\_marriages <dbl>,  
## # religion\_participation <chr>, partner\_location\_residence <chr>,  
## # full\_part\_time\_work <chr>, time\_off\_work\_birth <chr>,  
## # reason\_no\_time\_off\_birth <chr>, returned\_same\_job <chr>,  
## # satisfied\_time\_children <chr>, provide\_or\_receive\_fin\_supp <chr>,  
## # fin\_supp\_child\_supp <dbl>, fin\_supp\_child\_exp <dbl>, fin\_supp\_lump <dbl>,  
## # fin\_supp\_other <dbl>, fin\_supp\_agreement <chr>,  
## # future\_children\_intention <chr>, is\_male <dbl>, main\_activity <lgl>,  
## # age\_diff <chr>, number\_total\_children\_known <dbl>, income\_status <chr>

##   
## 1 2 3 4 5 6   
## 2775 4345 3696 2921 2158 4707

## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion  
  
## Warning in apply(gss, 2, as.numeric): NAs introduced by coercion

## caseid age age\_first\_child age\_youngest\_child\_under\_6 total\_children  
## 1 1 52.7 27 NA 1  
## 2 2 51.1 33 NA 5  
## 3 3 63.6 40 NA 5  
## 4 4 80.0 56 NA 1  
## 5 5 28.0 NA NA 0  
## 6 6 63.0 37 NA 2  
## age\_start\_relationship age\_at\_first\_marriage age\_at\_first\_birth  
## 1 NA NA 25.9  
## 2 NA NA NA  
## 3 NA NA 23.2  
## 4 NA NA 27.3  
## 5 25.3 NA NA  
## 6 NA NA 25.8  
## distance\_between\_houses age\_youngest\_child\_returned\_work feelings\_life sex  
## 1 30 NA 8 1  
## 2 NA NA 10 0  
## 3 NA NA 8 1  
## 4 NA NA 10 1  
## 5 NA NA 8 0  
## 6 NA NA 9 1  
## place\_birth\_canada place\_birth\_father place\_birth\_mother  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## place\_birth\_macro\_region place\_birth\_province year\_arrived\_canada province  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## region pop\_center marital\_status aboriginal vis\_minority age\_immigration  
## 1 NA NA NA NA NA NA  
## 2 NA NA NA NA NA NA  
## 3 NA NA NA NA NA NA  
## 4 NA NA NA NA NA NA  
## 5 NA NA NA NA NA NA  
## 6 NA NA NA NA NA NA  
## landed\_immigrant citizenship\_status education own\_rent living\_arrangement  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## hh\_type hh\_size partner\_birth\_country partner\_birth\_province  
## 1 NA 1 NA NA  
## 2 NA 2 NA NA  
## 3 NA 2 NA NA  
## 4 NA 2 NA NA  
## 5 NA 2 NA NA  
## 6 NA 2 NA NA  
## partner\_vis\_minority partner\_sex partner\_education average\_hours\_worked  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## worked\_last\_week partner\_main\_activity self\_rated\_health  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## self\_rated\_mental\_health religion\_has\_affiliation regilion\_importance  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## language\_home language\_knowledge income\_family income\_respondent occupation  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## childcare\_regular childcare\_type childcare\_monthly\_cost ever\_fathered\_child  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## ever\_given\_birth number\_of\_current\_union lives\_with\_partner  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## children\_in\_household number\_total\_children\_intention has\_grandchildren  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA 2 NA  
## 6 NA NA NA  
## grandparents\_still\_living ever\_married current\_marriage\_is\_first  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## number\_marriages religion\_participation partner\_location\_residence  
## 1 0 NA NA  
## 2 1 NA NA  
## 3 1 NA NA  
## 4 1 NA NA  
## 5 0 NA NA  
## 6 1 NA NA  
## full\_part\_time\_work time\_off\_work\_birth reason\_no\_time\_off\_birth  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## returned\_same\_job satisfied\_time\_children provide\_or\_receive\_fin\_supp  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## fin\_supp\_child\_supp fin\_supp\_child\_exp fin\_supp\_lump fin\_supp\_other  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## fin\_supp\_agreement future\_children\_intention is\_male main\_activity age\_diff  
## 1 NA NA 0 NA NA  
## 2 NA NA 1 NA NA  
## 3 NA NA 0 NA NA  
## 4 NA NA 0 NA NA  
## 5 NA NA 1 NA NA  
## 6 NA NA 0 NA NA  
## number\_total\_children\_known income\_status  
## 1 0 2  
## 2 0 4  
## 3 0 4  
## 4 0 5  
## 5 1 3  
## 6 0 3

  ## Standard Logistic Regression

##   
## Call:  
## glm(formula = sex ~ age + age\_first\_child + total\_children +   
## feelings\_life + as.factor(income\_status), family = "binomial",   
## data = gss\_numeric)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4244 -1.1885 0.7674 1.0340 1.9379   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.786023 0.170562 22.197 < 2e-16 \*\*\*  
## age -0.097040 0.003906 -24.841 < 2e-16 \*\*\*  
## age\_first\_child 0.090983 0.003602 25.260 < 2e-16 \*\*\*  
## total\_children -0.189921 0.017017 -11.161 < 2e-16 \*\*\*  
## feelings\_life 0.025987 0.011501 2.260 0.0238 \*   
## as.factor(income\_status)2 -0.293959 0.070921 -4.145 3.40e-05 \*\*\*  
## as.factor(income\_status)3 -0.582232 0.072268 -8.057 7.85e-16 \*\*\*  
## as.factor(income\_status)4 -0.632205 0.075822 -8.338 < 2e-16 \*\*\*  
## as.factor(income\_status)5 -0.733849 0.080806 -9.082 < 2e-16 \*\*\*  
## as.factor(income\_status)6 -0.633354 0.071529 -8.855 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 18560 on 13584 degrees of freedom  
## Residual deviance: 17585 on 13575 degrees of freedom  
## (7017 observations deleted due to missingness)  
## AIC: 17605  
##   
## Number of Fisher Scoring iterations: 4

## Survey Estimation for Logistic Regression

install.packages(“survey”) library(survey)

## Using the Survey Library

## Installing package into '/home/rstudio-user/R/x86\_64-pc-linux-gnu-library/4.0'  
## (as 'lib' is unspecified)

## Loading required package: grid

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loading required package: survival

##   
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':  
##   
## dotchart

##   
## Call:  
## svyglm(formula = sex ~ age + age\_first\_child + total\_children +   
## feelings\_life + as.factor(income\_status), design = a, family = "binomial")  
##   
## Survey design:  
## svydesign(id = ~1, data = gss\_numeric, fpc = b)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.786023 0.140641 26.920 < 2e-16 \*\*\*  
## age -0.097040 0.003257 -29.794 < 2e-16 \*\*\*  
## age\_first\_child 0.090983 0.002992 30.410 < 2e-16 \*\*\*  
## total\_children -0.189921 0.013716 -13.847 < 2e-16 \*\*\*  
## feelings\_life 0.025987 0.009315 2.790 0.00528 \*\*   
## as.factor(income\_status)2 -0.293959 0.056767 -5.178 2.27e-07 \*\*\*  
## as.factor(income\_status)3 -0.582232 0.058178 -10.008 < 2e-16 \*\*\*  
## as.factor(income\_status)4 -0.632205 0.061268 -10.319 < 2e-16 \*\*\*  
## as.factor(income\_status)5 -0.733849 0.065778 -11.157 < 2e-16 \*\*\*  
## as.factor(income\_status)6 -0.633354 0.058028 -10.915 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1.002454)  
##   
## Number of Fisher Scoring iterations: 4

## Discussion

By using summary, we can see the details of each variable.

Then we created several ggplots to determine relationships between factors.

There is a positive linear relationship between age\_first\_child and age.

There is a postive linear relationship between age and total\_children, age\_first\_children, age\_at\_first\_birth, age\_at\_first\_marriage and age\_start\_relationship.

Then we run linear regression since we foud some relationship.

The function will be Y = (-0.54 + -0.01\* total\_children + 0.99\* age\_first\_child +1.02\* age\_at\_first\_birth-0.01\* age\_at\_first\_marriage+0.02\*age\_start\_relationship)

Since there are many missing values that has been removed, about 20272 rows. So we recreated a linear regression, this time, I added categorical variables income\_family to test if age and income are related.

The function will be Y = 39.75 + 5.31\* total\_children - 1.54\* ‘125,000 and more’ + 8.07\* ‘25,000 to $49,999’ + 5.47*‘50,000 to $74,999’ + 1.90* ‘75,000 to $99,999’ + 6.86\*‘Less than $25,000’

In order to generate logistic regression. We created a binary variable as our Y value which is sex. We set female = 1 and male = 0.

Then we sepertaed the income\_family by 1,2,3,4,5,6 levels. ‘1’ is the poorest family and ‘6’ is richest family as the rank.

Since logistic regression Y is a binary.Current Y value is a character. So we changed GSS file to numerical and named it gss\_numeric.

We created a histogram to show the income status from 1 through 6. According to the histogram, there are many income\_family that have income over $125,000 and between $25,000 to $49,000.

We also created a histogram of female and male. According to the histogram, there are more females than males in the dataset.

The logistic regression function we got is Y = 3.79-0.097\* age+0.091\* age\_first\_child-0.19\* total\_children+0.026\*feelings\_life-0.294x2-0.58x3-0.63x4-0.73x5-0.63x6 There is about 1000 difference between null deviance and residual deviance, the larger the difference is, the better the model is.

## Weaknesses

There are many missing values even after data cleaning. If by removing all the NAs,there will be 0 rows left. So in order to run some analysis, we have to obtain some missing values. But due to this is a questionaire, it’s hard to obtain perfect answers. There are too many columns, factors in the dataset, which can create mislead when doing analysis.

## Next Steps

Since we have produced predictive model, which is the logistic regression. We can try to predict some values. We also would like to create other algorithms such as random forest, or decision tree. Because there are too many factors in the dataset random forest and decision tree can deal with large dataset with multidimensions. we think GSS dataset is multidimensional.And of course, we should also set up a following survey to fill up these missing values by doing our best. Obtaining good dataset can lead to a better result.

## References

Technology, A. (n.d.). Data Centre. Retrieved October 20, 2020, from <http://dc.chass.utoronto.ca/myaccess.html> Tyagi, P. (2018, December 25). Decision Tree. Retrieved from <https://medium.com/@pytyagi/decision-tree-ac0c9e3b8258>