Coding sample

Alonso Quijano

The following R code reproduces the statistical analysis presented in the writing sample titled *Maternal sexual empowerment and early sexual onset among female adolescents: Evidence from a prevalence study in Ecuador*. It shows how I typically handle data and how I share my code with research partners or other people who wish to replicate my results. The code can also be found on my github https://github.com/aquijanoruiz/R_projects/blob/master/early_sexual_activity/early_sexual_activity_code_sample.Rmd. A short chunk of stata code

This code takes raw data from the National Statistics Institute of Ecuador (INEC). It first downloads from the INEC website the zip that contains the .dta files. I also added some simple code that downloads the packages used in the study in case they have not been installed already. By doing this, I make sure anyone can easily run my code immediately without any wasted time.

Then, I describe the data wrangling process. You will see how I use the dplyr package. I merge the different datasets and select the information that is useful for the analysis. It involves matching subjects' data among various datasets and combining them into one data frame. I proceed by creating the variables that were applied in the models. Creating some of these variables involved using several logical operators and reassigning the levels of the factors. I finished by running the logistic regressions and creating a summary statistics table. I also added a simple CDF plot.

I do not like to write convoluted code. I instead prefer to create my own functions and apply loops to optimize the amount of code I write (which you may see in this sample). I am not yet the best at writing the most efficient code, but I believe my code is easy-to-read and concise. I hope you find my code runs smoothly, is easy to follow, and well-commented.

Before finishing, I would like to mention that I am keen on geospatial analysis. I like creating interactive maps using tmap and leaflet. At the beginning of the pandemic, I created a map illustrating the number of confirmed covid-19 cases in Ecuador. I was actually one of the first in my country to create an interactive map with detailed data for every province and city in my country. You can check the map at this url https://bit.ly/3k0LnJE.

```
################################################################################
#####                          1 Loading the data                          #####
#####                          1.1 The packages                            #####
################################################################################

# The following code automatically downloads the packages in case they are not
# installed in your computer already.

if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(readstata13)) install.packages("readstata13", repos = "http://cran.us.r-project.org")
if(!require(kableExtra)) install.packages("kableExtra", repos = "http://cran.us.r-project.org")
if(!require(stargazer)) install.packages("stargazer", repos = "http://cran.us.r-project.org")

################################################################################
#####                          1.2 Downloading the data                    #####
################################################################################
```

```r
# The data is downloaded directly from the permanent link that contains the zip file with
# all the datasets

# 1) The "people" dataset will contain the demographic and economic data for each of the
# members of the household, contained in the "1_BDD_ENS2018_f1_personas.dta" file

# 2) The "women" dataset will contain the data about the sexual health of women aged 10 to
# 49 years, contained in the "4_BDD_ENS2018_f2_mef.dta" file

# 3) The "behavior" dataset will contain the data about behavioral risk factors of people
# aged 5 to 18 years, contained in the "8_BDD_ENS2018_f4_fact_riesgo.dta" file

# 4) The "house" dataset will contain the data about the house the household lives in,
# contained in the "2_BDD_ENS2018_f1_hogar.dta" file

options(timeout=600) # we change the download timeout time to 600

# We give the url a name
url <- "https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/ENSANUT/ENSANUT_20
# We create a temporary directory
td <- tempdir()
# We create the placeholder file
tf <- tempfile(tmpdir=td, fileext = ".zip")
# We download the data into the placeholder file
download.file(url,tf)

# We get the name of the file inside the zip file that contains the demographic and
# economic data, unzip it, get the full path name of it, and finally load it

# We can use this code to look at the files contained inside the zip file
unzip(tf, list=TRUE)$Name
```

```
##  [1] "BDD_ENSANUT_2018_STATA_1/"
##  [2] "BDD_ENSANUT_2018_STATA_1/1_BDD_ENS2018_f1_personas.dta"
##  [3] "BDD_ENSANUT_2018_STATA_1/2_BDD_ENS2018_f1_hogar.dta"
##  [4] "BDD_ENSANUT_2018_STATA_1/3_BDD_ENS2018_f1_etiqueta.dta"
##  [5] "BDD_ENSANUT_2018_STATA_1/4_BDD_ENS2018_f2_mef.dta"
##  [6] "BDD_ENSANUT_2018_STATA_1/5_BDD_ENS2018_f2_lactancia.dta"
##  [7] "BDD_ENSANUT_2018_STATA_1/6_BDD_ENS2018_f2_salud_ninez.dta"
##  [8] "BDD_ENSANUT_2018_STATA_1/7_BDD_ENS2018_f3_ssrh.dta"
##  [9] "BDD_ENSANUT_2018_STATA_1/8_BDD_ENS2018_f4_fact_riesgo.dta"
## [10] "BDD_ENSANUT_2018_STATA_1/9_BDD_ENS2018_f5_des_inf.dta"
```

```r
# We get the name of the file, get its full path, unzip it, and then load it

people.f.name <- unzip(tf, list=TRUE)$Name[2] # The people dataset is number 2
women.f.name <- unzip(tf, list=TRUE)$Name[5] # The women dataset is number 5
behavior.f.name <- unzip(tf, list=TRUE)$Name[9] # The behavior dataset is number 9
house.f.name <- unzip(tf, list=TRUE)$Name[3] # The house dataset is the number 3

people.f.path <- file.path(td, people.f.name)
women.f.path <- file.path(td, women.f.name)
behavior.f.path <- file.path(td, behavior.f.name)
```

```
house.f.path <- file.path(td, house.f.name)

unzip(tf, files=c(people.f.name, women.f.name, behavior.f.name, house.f.name),
      exdir=td, overwrite=TRUE)

# Now, we can load the three files
people <- read.dta13(people.f.path)
women <- read.dta13(women.f.path)
behavior <- read.dta13(behavior.f.path)
house <- read.dta13(house.f.path)

################################################################################
#####                          1.3 Extracting the variable labels            #####
################################################################################

# As these are STATA files, the label of each of the variables is stored inside the
# datasets, we can extract them using the following code:
data.key.people <- data.frame(variable = names(people),
                              label = attr(people,"var.labels"))

data.key.women <- data.frame(variable = names(women),
                             label = attr(women,"var.labels"))

data.key.behavior <- data.frame(variable = names(behavior),
                                label = attr(behavior,"var.labels"))

data.key.house <- data.frame(variable = names(house),
                             label = attr(house,"var.labels"))

# Let's look at the first 12 variables of the people set and their labels
head(data.key.people, 12)
```

```
##     variable                                                label
## 1       area                                                Área
## 2       prov                                           Provincia
## 3        upm                               Indentificador de upm
## 4     id_viv                          Indentificador de vivienda
## 5   id_hogar                            Indentificador del hogar
## 6     id_per                         Indentificador de la persona
## 7    persona                                       Cód. Persona
## 8       sexo                                               Sexo
## 9  f1_s2_3_1       3.1  ¿Cuántos años cumplidos tiene?: años
## 10 f1_s2_3_2      3.2  ¿Cuántos años cumplidos tiene?: meses
## 11 f1_s2_4_1 4.1  ¿Cuál es la fecha de nacimiento de (...)? día
## 12 f1_s2_4_2 4.2  ¿Cuál es la fecha de nacimiento de (...)? mes
```

```
# The name of each variable is assigned according to its code in the survey. For example,
# whether a woman between 12 and 49 years old has ever had sexual intercourse can be found
# in the variable f2_s8_803 of the women set, which corresponds to form 2, section 8,
# question 803
summary(women$f2_s8_803)
```

```
##                 si                 no no desea contestar               NA's
```

```
##               8879              13002                258              26561
```

```
# We can see that the whole dataset is in Spanish. This should be no surprise since the
# official language in Ecuador is Spanish. For your better understanding, I will rename
# every variable of interest from Spanish into English

################################################################################
#####                        2 Data wrangling                              #####
#####                        2.1 Analyzing the structure of the data       #####
################################################################################

# We have to first see how the dataset is structured, I wil rename some variables first
people <- people %>% mutate(household_id = id_hogar,
                            subject_id = id_per,
                            person = as.integer(persona),
                            sex = sexo,
                            mother = f1_s2_15_1,
                            father = f1_s2_14_1,
                            age = f1_s2_3_1)

levels(people$sex) <- c("male", "female")

# Let's look at the first household in our dataset. This household has 6 members, one
# female and five males. Who is whose mother and father? We have to look at the "person",
# "mother", and "father" variables. Person 1 and person 2 are the mother and father of
# persons 3 to 6
people_id <- people %>% select(household_id, subject_id, person, sex, age, mother, father)
people_id %>% filter(household_id == "010150000201011")
```

```
##      household_id        subject_id person    sex age mother father
## 1 010150000201011 01015000020101101      1   male  28     NA     NA
## 2 010150000201011 01015000020101102      2 female  28     NA     NA
## 3 010150000201011 01015000020101103      3   male  13      2      1
## 4 010150000201011 01015000020101104      4   male  11      2      1
## 5 010150000201011 01015000020101105      5   male   6      2      1
## 6 010150000201011 01015000020101106      6   male   3      2      1
```

```
################################################################################
#####                        2.2 Extracting the mother's Ids               #####
################################################################################

# Because the final dataset will require the observation for each subject (female
# adolescent) and their respectivemother to be in one single row, we cannot work with
# this data set simply as it is. To use the left_join() funtion, we first need to add
# a column with the unique Id of each person's mother. We will then use then the
# mother's unique Id to merge the data

# We first create a separate data.frame with the mothers' Ids
mothers_id <- people_id %>% group_by(household_id) %>%
  slice(mother) %>%  # we take only the mothers
  distinct(person, .keep_all = TRUE) %>% # we eliminate repeated observations
  ungroup() %>% mutate(mother_id = subject_id) %>% select(household_id, mother_id, person)
```

```r
# We add the mothers' Id to the people_id data.frame we created
people_id <- left_join(people_id, mothers_id, by = c("household_id" = "household_id",
                                                     "mother" = "person"))

people_id <- people_id %>% select(subject_id, mother_id) # we select only what we need
head(people_id, 6) # we got what we wanted, an additional column with the Id of each
```

```
##          subject_id         mother_id
## 1 01015000020101101              <NA>
## 2 01015000020101102              <NA>
## 3 01015000020101103 01015000020101102
## 4 01015000020101104 01015000020101102
## 5 01015000020101105 01015000020101102
## 6 01015000020101106 01015000020101102
```

```r
# person's mother next to the Id of that person

################################################################################
#####                        2.3 Merging the datasets                      #####
#####                        2.3.1 The "daughters" set                     #####
################################################################################

# Now we can merge the four dataset and filter the girls between 12 and 18.
# We will call this new dataset "daugthers"
daughters <- people %>% # the demographic and economic data
  left_join(women, by = c("subject_id" = "id_per")) %>% # the data about sexual health
  left_join(behavior, by = c("subject_id" = "id_per")) %>% # the behavioral variables
  left_join(house, by = c("household_id" = "id_hogar")) %>% # the data about the house
  left_join(people_id, by = "subject_id") %>% # the mothers' Ids (we'll use this to
  # filter the "mothers" set)
  filter(sex == "female" & age == 16) # we filter the girls who are 16

nrow(daughters) # we have 11,446 girls in our dataset. This will not be final version
```

```
## [1] 1636
```

```r
# as we will continue cleaning the data (this includes eliminating NAs, errors, etc.)

n_distinct(daughters$mother_id) # we can also see we have data for 8,063 mothers. We have
```

```
## [1] 1342
```

```r
# more daughters than mothers because some are sisters, and there is missing data for
# some mothers whether because they do not live with their daughters, were not at home
# when the survey took place, etc.

################################################################################
#####                        2.3.2 The "mothers" set                       #####
################################################################################

# The data in the "mothers" set corresponds to the data of the mothers of those girls
```

```r
# included in the "daughters" set
mothers <- people %>%
  left_join(women, by = c("subject_id" = "id_per")) %>% # the data about sexual health
  semi_join(daughters, by = c("subject_id" = "mother_id")) # we filter only the mothers
# of those in the daugthers set. This is why we created the people_id data frame :)

################################################################################
#####                        3. Variables                                  #####
#####                        3.1 Creating some useful functions            #####
################################################################################

# Some of the answers are coded as 88 and 99 when respondents either do not remember or
# do not want to answer. We can create a funtion to get rid of those values

ninenineTOna <- function(x){
  y = ifelse(x == 77 | x == 88 | x == 99 , NA, x)
  return(y)
}

# As the survey was done in Ecuador, answers are coded in Spanish, we can create a
# function to translate the YES/NO questions and store them as factors
# We created two functions with the same purpose but with the levels inverted. We will
# apply different levels to different variables depending from what angle we want to
# look at the variable

sinoTOyesno <- function(x){
  x = as.integer(x)
  y = factor(x, levels = c(1, 2), labels = c("yes", "no"))
  return(y)
}

sinoTOnoyes <- function(x){
  x = as.integer(x)
  y = factor(x, levels = c(2, 1), labels = c("no", "yes"))
  return(y)
}

################################################################################
#####                        3.2 Dependent variable                        #####
################################################################################

# early sexual acrtivity ------------------------------
daughters$early_sexual_activity <- sinoTOnoyes(daughters$f2_s8_803)

# coercion at first intercourse ------------------------------
# (Not the dependent variable but something we will look at)
daughters$coercion_1st_intercourse <- factor(with(daughters,
  ifelse(as.integer(f2_s8_807) == 1 | as.integer(f2_s8_807) == 2, "no", "yes"),
  levels = c("yes", "no")))

################################################################################
#####                    3.3 Independent variables                         #####
#####                    3.3.1 Social, economic and demographic variabless #####
```

```r
##############################################################################

# a) income & number of members in the household

# We calculated the total income for each household. We need to sum up the different sources
# of income of each member (scattered in many variables/columns), and then we need total
# the income of each member to get the overall income of the whole household.

income <- select(people, household_id, f1_s3_15, f1_s3_16_2, f1_s3_17, f1_s3_18,
                 f1_s3_19, f1_s3_20_2, f1_s3_22_2)

nineninetozero <- function(x){ # We create a variable to change the 999999s for zero
  x = ifelse(x == 999999, 0, x) # and apply it to all the columns
  return(x)
}

income[, c(2:8)] <- sapply(income[, c(2:8)], FUN = nineninetozero)
income$f1_s3_17 <- income$f1_s3_17 * (-1) # We changed the sign of the reported expenses
income <- income %>% mutate(income = rowSums(.[, 2:8], na.rm = TRUE)) # we sum the columns

# We sum the income of each houlsehold member
income <- income %>% group_by(household_id) %>%
  summarize(h_income = sum(income, na.rm = TRUE),
            h_num_members = n())

# b) area (urban/rural)
daughters$rural <- daughters$area.x
levels(daughters$rural) <- c("no", "yes")

# c) internet access
daughters$h_internet <- sinoTOyesno(daughters$f1_s1_42)

# d) ethnic minority
daughters$minority <- with(daughters,
  factor(ifelse(!as.integer(f1_s2_9) == 6 & !as.integer(f1_s2_9) == 7, "yes", "no"),
         levels = c("no", "yes")))

# e) Misses school
daughters$attends_school <- sinoTOyesno(daughters$f1_s2_17)

##############################################################################
#####                  3.3.2 Knowledge of sexual education                #####
##############################################################################

# a) didn't know what was happening to her when she had their first period
daughters$period_knowledge <- sinoTOyesno(daughters$f2_s8_841)

# b) cannot answer correctly: can AIDS spread through handshake?
daughters$aids_knowledge <- with(daughters,
  factor(ifelse(as.integer(f2_s10_1011_1) == 1 | as.integer(f2_s10_1011_1) == 3, "no", "yes"),
         levels = c("yes", "no")))

# c) cannot answer correctly: can a women get pregnant the first time she has sex?
```

```r
daughters$pregnancy_knowledge <- sinoTOyesno(daughters$f2_s8_845)

# d) has ever received info about sexuality and primary source (school, home, other)
daughters$sexuality_knowledge <- factor(with(daughters,
  case_when(is.na(f2_s8_800d) ~ NA_character_, as.integer(f2_s8_800d) == 2 ~ "no info",
            as.integer(f2_s8_801d) == 1 ~ "family", as.integer(f2_s8_801d) == 2 ~ "school",
            TRUE ~ "other")), levels = c("no info", "family", "school", "other"))


################################################################################
#####                  3.3.3 Behavioral risk factors                       #####
################################################################################

# a) ever drunk alcohol -------------------------------
daughters$ever_drunk_alcohol <- sinoTOnoyes(daughters$f4_s5_500)

# b) ever smoked -------------------------------
daughters$ever_smoked <- sinoTOnoyes(daughters$f4_s6_600)

################################################################################
#####                  3.3.4 Characteristics of the mother                 #####
################################################################################

# a) mother's age at first birth
# We subtract the year of birth of the youngest child from the year of birth of the mother
mothers$m_age_1st_birth <-
  with(mothers, pmin(f2_s2_218_1_b3, f2_s2_218_2_b3, f2_s2_218_3_b3, f2_s2_218_4_b3,
                     f2_s2_218_5_b3, f2_s2_218_6_b3, f2_s2_218_7_b3, f2_s2_218_8_b3,
                     f2_s2_218_9_b3, f2_s2_218_10_b3, na.rm = TRUE)) - mothers$f1_s2_4_3

# b) mother had teenage birth
mothers$m_teenage_birth <-
  factor(ifelse(mothers$m_age_1st_birth <= 19, "yes", "no"), levels = c("no", "yes"))

# c) mother's age at first intercourse
mothers$m_age_1st_intercourse <- ninenineTOna(coalesce(mothers$f2_s8_804, mothers$f2_s8_831))

# d) mother's education attainment (no formal education, secondary, tertiary)
mothers$m_education <- mothers$f1_s2_19_1
levels(mothers$m_education) <- c("none", "none", "none", "primary", "primary", "secondary",
                                 "secondary", "tertiary", "tertiary", "tertiary")

mothers$m_finished_ps <- mothers$m_education # mother finished primary school
levels(mothers$m_finished_ps) <- c("no", "yes", "yes", "yes")

mothers$m_finished_hs <- mothers$m_education # mother finished high school
levels(mothers$m_finished_hs) <- c("no", "no", "yes", "yes")

mothers$m_finished_college <- mothers$m_education # mother finished college
levels(mothers$m_finished_college) <- c("no", "no", "no", "yes")

# e) empowerment & sexual decision making of the mother
# We measure empowerment as the ability of the mothers's to make their own sexual decisions
```

```r
# We classify unempowered women as those who aren't able to turn down sex
# We also classify unempowerment as the inability to demand the use of contraception

mothers$m_empowerment <- factor(with(mothers,
  case_when(is.na(f2_s6_604) ~ NA_character_,
            as.integer(f2_s6_613) == 8  ~ "no", # partner does not allow contraception
            as.integer(f2_s8_835) == 6 ~ "no", # has unprotected sex because partner
            # does not like contraception
            as.integer(f2_s8_834) == 2 & ! as.integer(f2_s8_835) == 4 &
            as.integer(f2_s8_836) == 1 & !as.integer(f2_s8_837) == 2 ~ "no",
            as.integer(f2_s8_839) == 2 ~ "no", # cannot turn down sex
            TRUE ~ "yes")), levels = c("yes", "no"))

# f) mother has a job
mothers$m_job <- with(mothers,
  factor(case_when(is.na(f1_s3_1) ~ NA_character_,
                   as.integer(f1_s3_1) == 2 & as.integer(f1_s3_2) == 12 ~ "no",
                   TRUE ~ "yes")), levels = c("yes", "no"))

##############################################################################
#####                        3.4 Merging the data frames                 #####
##############################################################################

daughters_tidy <- daughters %>%  select(household_id, subject_id, mother_id,
  early_sexual_activity, rural, minority, h_internet, attends_school,  period_knowledge,
  aids_knowledge, pregnancy_knowledge, sexuality_knowledge, ever_drunk_alcohol, ever_smoked,
  coercion_1st_intercourse) %>% left_join(income, by = c("household_id" = "household_id"))

mothers_tidy <- mothers %>% select(subject_id, m_teenage_birth, m_empowerment, m_job,
  m_age_1st_intercourse, m_education, m_finished_ps, m_finished_hs, m_finished_college)

data <- daughters_tidy %>% left_join(mothers_tidy, by = c("mother_id" = "subject_id")) %>%
  filter(!is.na(early_sexual_activity)) # we eliminate NAs
```

```r
##############################################################################
#####                        4 The logit models                          #####
##############################################################################

# We will run logistic regressions to see what variables are most correlated
# with early sexual activity.

# Model 1: ------------------------------
# m_empowerment + control variables

logit_m1 <- glm(early_sexual_activity ~ minority + rural + h_income + h_num_members + h_internet +
                  attends_school + period_knowledge + pregnancy_knowledge + aids_knowledge +
                  sexuality_knowledge + m_job + m_education +  m_empowerment,
                data = data, family = "binomial")

# Model 2: ------------------------------
# m_empowerment & m_teenage_birth + control variables

logit_m2 <- glm(early_sexual_activity ~ minority + rural + h_income + h_num_members + h_internet +
```

```
                  attends_school + period_knowledge + pregnancy_knowledge + aids_knowledge +
                  sexuality_knowledge + m_job + m_education +  m_empowerment + m_teenage_birth,
                data = data, family = "binomial")

# Model 3: ------------------------------
# m_empowerment & m_teenage_birth & m_age_1st_intercourse + control variables

logit_m3 <- glm(early_sexual_activity ~ minority + rural + h_income + h_num_members + h_internet +
                  attends_school + period_knowledge + pregnancy_knowledge + aids_knowledge +
                  sexuality_knowledge + m_job + m_education +  m_empowerment + m_teenage_birth +
                  m_age_1st_intercourse, data = data, family = "binomial")

stargazer(logit_m1, logit_m2, logit_m3, title="Logistic Regression Results",
          covariate.labels = c("Ethnic minority", "Lives in a rural area", "Household income",
          "Number of members in the household", "Does not have internet", "Misses school",
          "Lacks knowledge about period", "Lacks knowledge about pregnancy",
          "Lacks knowledge about AIDs", "Knows about sexuality from family",
          "Knows about sexuality from school", "Knows about sexuality from other sources",
          "Mother has a job", "Mother finished primary school", "Mother finished secondary school",
          "Mother finished college", "Mother lacks sexual bargaining", "Mother had a teenage birth",
          "Mother's age at first intercourse"),
          align=TRUE, header = FALSE, star.cutoffs = c(.05, .01, .001),
          dep.var.labels = c("Early sexual activity"), no.space = TRUE)
```

```
# Model 4/5/6 (includes drinking and smoking) ------------------------------
# m_empowerment & m_teenage_birth & m_age_1st_intercourse + control variables
logit_m4 <- glm(early_sexual_activity ~ minority + rural + h_income + h_num_members + h_internet +
                  attends_school + period_knowledge + pregnancy_knowledge + aids_knowledge +
                  sexuality_knowledge + m_job + m_education + ever_drunk_alcohol + ever_smoked +
                  m_empowerment,
                data = data, family = "binomial")

logit_m5 <- glm(early_sexual_activity ~ minority + rural + h_income + h_num_members + h_internet +
                  attends_school + period_knowledge + pregnancy_knowledge + aids_knowledge +
                  sexuality_knowledge + m_job + m_education + ever_drunk_alcohol + ever_smoked +
                  m_empowerment + m_teenage_birth,
                data = data, family = "binomial")

logit_m6 <- glm(early_sexual_activity ~ minority + rural + h_income + h_num_members + h_internet +
                  attends_school + period_knowledge + pregnancy_knowledge + aids_knowledge +
                  sexuality_knowledge + m_job + m_education + ever_drunk_alcohol + ever_smoked +
                  m_empowerment + m_teenage_birth + m_age_1st_intercourse,
                data = data, family = "binomial")

stargazer(logit_m4, logit_m5, logit_m6, title="Logistic Regression Results",
          covariate.labels = c("Ethnic minority", "Lives in a rural area", "Household income",
          "Number of members in the household", "Does not have internet", "Misses school",
          "Lacks knowledge about period", "Lacks knowledge about pregnancy",
          "Lacks knowledge about AIDs", "Knows about sexuality from family",
          "Knows about sexuality from school", "Knows about sexuality from other sources",
          "Mother has a job", "Mother finished primary school", "Mother finished secondary school",
          "Mother finished college", "Ever drunk alcohol", "Ever smoked",
          "Mother lacks sexual bargaining", "Mother had a teenage birth",
```

Table 1: Logistic Regression Results

	Dependent variable:		
	Early sexual activity		
	(1)	(2)	(3)
Ethnic minority	0.314	0.324	0.325
	(0.255)	(0.259)	(0.270)
Lives in a rural area	−0.053	−0.025	0.058
	(0.231)	(0.233)	(0.241)
Household income	0.0001	0.0001	0.0001
	(0.0001)	(0.0001)	(0.0001)
Number of members in the household	0.098	0.070	0.037
	(0.056)	(0.058)	(0.059)
Does not have internet	0.430	0.447	0.404
	(0.246)	(0.249)	(0.257)
Misses school	2.127***	2.139***	2.163***
	(0.319)	(0.322)	(0.340)
Lacks knowledge about period	0.251	0.242	0.350
	(0.252)	(0.254)	(0.264)
Lacks knowledge about pregnancy	0.290	0.323	0.424
	(0.260)	(0.261)	(0.272)
Lacks knowledge about AIDs	−0.024	0.033	−0.120
	(0.308)	(0.311)	(0.337)
Knows about sexuality from family	2.631*	2.578*	2.564*
	(1.141)	(1.205)	(1.204)
Knows about sexuality from school	2.069	2.125	2.056
	(1.109)	(1.173)	(1.170)
Knows about sexuality from other sources	3.000*	2.976*	3.144*
	(1.191)	(1.250)	(1.248)
Mother has a job	0.631**	0.575*	0.595*
	(0.225)	(0.228)	(0.236)
Mother finished primary school	−1.187	−1.332	−1.161
	(0.683)	(0.695)	(0.707)
Mother finished secondary school	−0.693	−0.790	−0.584
	(0.685)	(0.697)	(0.710)
Mother finished college	−0.943	−0.961	−0.554
	(0.734)	(0.744)	(0.761)
Mother lacks sexual bargaining	0.628*	0.600*	0.478
	(0.297)	(0.300)	(0.316)
Mother had a teenage birth		0.778***	0.294
		(0.220)	(0.258)
Mother's age at first intercourse			−0.188***
			(0.055)
Constant	−4.583***	−4.829***	−1.389
	(1.287)	(1.334)	(1.649)
Observations	828	824	783
Log Likelihood	-321.446	-314.520	-291.948
Akaike Inf. Crit.	678.891	667.040	623.895
Note:			*p<0.05; **p<0.01; ***p<0.001

```
          "Mother's age at first intercourse"), align=TRUE, header = FALSE,
          star.cutoffs = c(.05, .01, .001), dep.var.labels = c("Early sexual activity"),
          no.space = TRUE)
```

```
##############################################################################
#####              5 Summary statistics                                  #####
#####              5.1 Percentages and means of each variable by gruop   #####
##############################################################################

# We are going to create a table with the summary statistics of each variable
# We first create a copy of the data, which we will use to create the table
data_copy <- data

# The variable sexuality_knowledge has several levels in it. We will split this
# variable into different columns. Because this is a factor, whose levels are 1 for the
# first level, 2 for the seconde, etc., we will code as 2 those who learned about
# sexuality from X source and 1 otherwise

data_copy$value <- 2
data_copy <- spread(data_copy, sexuality_knowledge, value, fill = 1, sep = "_")

# We will use the chi square test and t test to compare variables within groups (early
# sexual activity and no early sexual activity). For that we will apply some loops

# chi square test (categorical variables) ------------------------------

cat_var <- c("minority", "rural", "h_internet", "attends_school", "period_knowledge",
             "pregnancy_knowledge", "aids_knowledge", "sexuality_knowledge_no info",
             "sexuality_knowledge_family", "sexuality_knowledge_school",
             "sexuality_knowledge_other", "ever_drunk_alcohol", "ever_smoked", "m_job",
             "m_finished_ps", "m_finished_hs", "m_finished_college", "m_teenage_birth",
             "m_empowerment")


chi_sq_test  <- sapply(cat_var, function(x){
  chi_sq <- chisq.test(data_copy[,"early_sexual_activity"], data_copy[,x])
  return(chi_sq$p.value)
})

mean_early_sex  <- sapply(cat_var, function(x){
  mean <- mean(as.integer(data_copy[data_copy$early_sexual_activity == "yes", x]) == 2,
               na.rm = TRUE)
  return(mean)
})

mean_no_early_sex  <- sapply(cat_var, function(x){
  mean <- mean(as.integer(data_copy[data_copy$early_sexual_activity == "no", x]) == 2,
               na.rm = TRUE)
  return(mean)
})

# We put the percentages and p values everything in one table
summary_statistics <- tibble(variable = cat_var, mean_early_sex = mean_early_sex,
```

Table 2: Logistic Regression Results

	Dependent variable:		
	Early sexual activity		
	(1)	(2)	(3)
Ethnic minority	0.671	0.695	0.667
	(0.409)	(0.411)	(0.414)
Lives in a rural area	−0.411	−0.372	−0.363
	(0.354)	(0.354)	(0.358)
Household income	0.0002	0.0002	0.0002
	(0.0001)	(0.0001)	(0.0001)
Number of members in the household	0.118	0.095	0.087
	(0.091)	(0.093)	(0.093)
Does not have internet	0.775*	0.774*	0.675
	(0.357)	(0.356)	(0.358)
Misses school	1.554**	1.532**	1.568**
	(0.495)	(0.498)	(0.513)
Lacks knowledge about period	−0.232	−0.199	−0.168
	(0.426)	(0.427)	(0.434)
Lacks knowledge about pregnancy	−0.165	−0.174	−0.122
	(0.434)	(0.435)	(0.437)
Lacks knowledge about AIDs	−0.007	−0.037	−0.208
	(0.505)	(0.510)	(0.524)
Knows about sexuality from family	1.288	1.329	1.244
	(1.585)	(1.623)	(1.623)
Knows about sexuality from school	0.659	0.771	0.636
	(1.537)	(1.578)	(1.580)
Knows about sexuality from other sources	1.656	1.755	1.623
	(1.648)	(1.684)	(1.683)
Mother has a job	0.488	0.436	0.455
	(0.325)	(0.327)	(0.333)
Mother finished primary school	−2.610**	−2.748**	−2.460*
	(0.971)	(0.979)	(0.981)
Mother finished secondary school	−2.018*	−2.128*	−1.782
	(0.980)	(0.984)	(0.987)
Mother finished college	−2.215*	−2.275*	−1.792
	(1.047)	(1.046)	(1.057)
Ever drunk alcohol	0.872**	0.868**	0.785*
	(0.328)	(0.328)	(0.333)
Ever smoked	1.415**	1.358**	1.404**
	(0.497)	(0.500)	(0.512)
Mother lacks sexual bargaining	0.685	0.704	0.587
	(0.438)	(0.440)	(0.450)
Mother had a teenage birth		0.353	−0.131
		(0.311)	(0.358)
Mother's age at first intercourse			−0.191*
			(0.080)
Constant	−2.446	−2.483	0.998
	(1.749)	(1.782)	(2.272)
Observations	417	415	401
Log Likelihood	-155.455	-154.608	-150.149
Akaike Inf. Crit.	350.910	351.216	344.298

Note: *p<0.05; **p<0.01; ***p<0.001

```
                              mean_no_early_sex = mean_no_early_sex, p_value = chi_sq_test)

# t test (continuous variables) ------------------------------

cont_var <- c("h_income", "h_num_members", "m_age_1st_intercourse")

t_test  <- sapply(cont_var, function(x){
  t_test <- t.test(data_copy[,x] ~ data_copy[,"early_sexual_activity"], var.equal = TRUE)
  return(t_test$p.value)
})

mean_early_sex  <- sapply(cont_var, function(x){
  mean <- mean(data_copy[data_copy$early_sexual_activity == "yes", x], na.rm = TRUE)
  return(mean)
})

mean_no_early_sex  <- sapply(cont_var, function(x){
  mean <- mean(data_copy[data_copy$early_sexual_activity == "no", x], na.rm = TRUE)
  return(mean)
})

# We add the new means and p values to the table we already made
summary_statistics <- rbind(summary_statistics, tibble(variable = cont_var,
  mean_early_sex = mean_early_sex, mean_no_early_sex = mean_no_early_sex,
  p_value = t_test))

# Adding some format to the table ------------------------------

summary_statistics[, 2:3] <- round(summary_statistics[, 2:3], digits = 2)
summary_statistics[, 4] <- round(summary_statistics[, 4], digits = 3)

summary_statistics_copy <- summary_statistics

summary_statistics_copy$p_value <-
  with(summary_statistics_copy, case_when(p_value < 0.0005 ~ paste("0.000", "***"),
                                          p_value <= 0.001 ~ paste(as.character(p_value), "***"),
                                          p_value <= 0.01 ~ paste(p_value, "***"),
                                          p_value <= 0.05 ~ paste(p_value, "**"),
                                          p_value <= 0.1 ~ paste(as.numeric(p_value), " *"),
                                          TRUE ~ as.character(p_value)))

names(summary_statistics_copy) <- c("Variables", "Early sexual activity",
                                    "No early sexual activity", "p value")

summary_statistics_copy$Variables <- c("Ethnic minority", "Lives in a rural area",
  "Does not have internet", "Misses school", "Lacks knowledge about period",
  "Lacks knowledge about pregnancy", "Lacks knowledge about AIDs",
  "Does not know about sexuality", "Knows about sexuality from family",
  "Knows about sexuality from school", "Knows about sexuality from other sources",
  "Has ever drunk alcohol", "Has ever smoked", "Mother has a job", "Mother finished primary school",
  "Mother finished high school", "Mother finished college", "Mother had a teenage birth",
  "Mother lacks sexual bargaining", "Household income", "Number of members in the household",
  "Mother's age at first intercourse")
```

Table 3: Percentage and mean levels of explanatory variables by group

Variables	Early sexual activity	No early sexual activity	p value
Ethnic minority	0.30	0.21	0.001 ***
Lives in a rural area	0.48	0.41	0.033 **
Does not have internet	0.75	0.55	0.000 ***
Misses school	0.44	0.05	0.000 ***
Lacks knowledge about period	0.28	0.19	0.000 ***
Lacks knowledge about pregnancy	0.20	0.17	0.273
Lacks knowledge about AIDs	0.18	0.12	0.014 **
Does not know about sexuality	0.09	0.07	0.273
Knows about sexuality from family	0.15	0.10	0.015 **
Knows about sexuality from school	0.68	0.80	0.000 ***
Knows about sexuality from other sources	0.08	0.03	0.000 ***
Has ever drunk alcohol	0.62	0.44	0.000 ***
Has ever smoked	0.09	0.03	0.000 ***
Mother has a job	0.64	0.60	0.282
Mother finished primary school	0.91	0.96	0.003 ***
Mother finished high school	0.44	0.51	0.1 *
Mother finished college	0.09	0.16	0.021 **
Mother had a teenage birth	0.68	0.48	0.000 ***
Mother lacks sexual bargaining	0.15	0.11	0.108
Household income	646.79	630.03	0.944
Number of members in the household	5.64	5.42	0.08 *
Mother's age at first intercourse	16.11	17.64	0.000 ***

Note: makecell[l]p values for comparison of percentagges using chi-square. p values for comparison of means using t-test. Ns = 401–828. *p < .1; **p < .05; ***p < .01

```
kable(summary_statistics_copy, format = "latex", booktabs = TRUE,
      caption = "Percentage and mean levels of explanatory variables by group") %>%
  footnote(general = "p values for comparison of percentagges using chi-square. p values for
           comparison of means using t-test. Ns = 401-828. *p < .1; **p < .05; ***p < .01",
           threeparttable = T, footnote_as_chunk = T)
```

```
################################################################################
#####       5.2 CDF of the mother's age at first intercourse by gruop       #####
################################################################################

plot(ecdf(data_copy[data_copy$early_sexual_activity == "yes", "m_age_1st_intercourse"]),
          col = "cadetblue1", main = "", xlab = "Age at first intercourse", ylab = "Density",
     cex.axis=0.75, cex.lab=0.75)
plot(ecdf(data_copy[data_copy$early_sexual_activity == "no", "m_age_1st_intercourse"]),
     col = "palegreen", add = TRUE)
legend("left", c("Early sexual activity", "No early sexual activity"),
       col = c("cadetblue1", "palegreen"), lwd = 5, bty = "n", cex = 0.6)
```

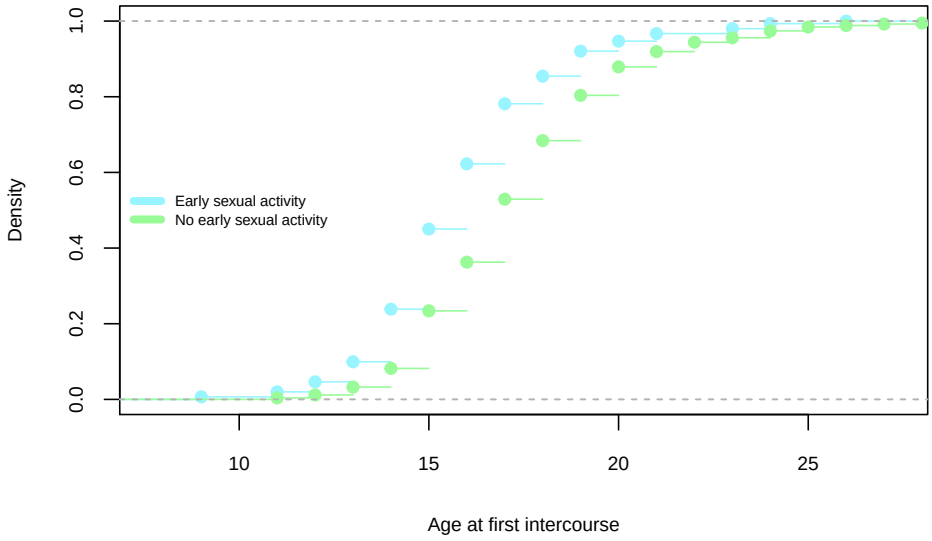

Figure 1: Cumulative histogram of age at first intercourse of mothers by group

```
##############################################################################
#####                              5.3 Q & A                             #####
##############################################################################

# 1)  What percent of girls in our sample had a coerced first sex?
mean(data$coercion_1st_intercourse == "yes", na.rm = TRUE)
```

```
## [1] 0.3441176
```

```
# 2) What percent of girls in our sample had their first intercourse before age 16?
index <- as.integer(names(logit_m1$fitted.values)) # these are the observations used in our model
data_copy <- data[index,]

prop.table(table(data_copy$early_sexual_activity))
```

```
##
##        no       yes
## 0.8357488 0.1642512
```

```
# 3) What is the mean age and sd of the mothers' age at first intercourse by group?

data_copy %>% group_by(early_sexual_activity) %>%
  summarize(mean = mean(m_age_1st_intercourse, na.rm = TRUE),
            sd = sd(m_age_1st_intercourse, na.rm = TRUE))
```

```
## # A tibble: 2 x 3
```

```
##   early_sexual_activity  mean    sd
##   <fct>                 <dbl> <dbl>
## 1 no                     17.7  2.79
## 2 yes                    16.1  2.48
```