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 large datasets and how I share my code with colleagues or other people who wish to replicate my results. The code can also be found on github.

This study uses data from the National Statistics Institute of Ecuador (INEC). It first downloads the zip that contains the .dta files. I added some code that automatically downloads the packages used in case they are not installed already. Then, I describe the data wrangling process, for which I use the dplyr package. I merge different datasets and select the information that is useful for the analysis. This involves matching subjects' data among various datasets. I proceed by creating the variables, for which I use several logical operators and reassign the levels of the factors. I finish by creating the summary statistics table and run the logistic regression.

I like to create my own functions and apply loops to optimize the amount of code I write (which you may see in this sample).This makes my code easy-to-read and concise. Before finishing, I would like to mention that I am keen on geospatial analysis. I like creating interactive maps using tmap and leaflet. You can see some of the geospatial work I do at this url.

```
################################################################################
#####                              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")

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

# 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
```

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

```r
# 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
```

```r
################################################################################
#####                        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)

# 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
```

```
# 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
```

```
# 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
```

```
# 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
# 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
```

```r
# 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 and order the levels
# as

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 activity ------------------------------
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("no", "yes")))

# used contraception at 1st intercourse ----------------------------
daughters$contraception_1st_intercourse <-
  factor(ifelse(as.integer(daughters$f2_s8_808) == 1, "yes", "no"), levels = c("no", "yes"))

# has ever been pregnant ----------------------------
daughters$teenage_pregnancy <-
  (factor(ifelse(as.integer(daughters$f2_s2_200) == 1 | as.integer(daughters$f2_s2_207) == 1,
                 "yes", "no"), levels = c("no", "yes")))

################################################################################
#####                    3.3 Independent variables                         #####
#####                    3.3.1 Mother-related variables                    #####
################################################################################

# a) empowerment & sexual decision making of the mother
# We measure empowerment as the ability of the mothers's to make their own sexual decisions
# 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_lack_empowerment <-
  factor(with(mothers,
              case_when(is.na(f2_s6_604) ~ NA_character_,
                        as.integer(f2_s6_613) == 8 ~ "yes",
```

```
                        # partner does not allow contraception
                        as.integer(f2_s8_835) == 6 ~ "yes",
                        # 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) == 1 ~ "yes",
                        as.integer(f2_s8_839) == 2 ~ "yes",
                        # cannot turn down sex
                        TRUE ~ "no")), levels = c("no", "yes"))

# b) mother had teenage birth
# We first calculate the year in which she had her first birth and then subtract that year
# 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

mothers$m_teenage_birth <-
  factor(ifelse(mothers$m_age_1st_birth <= 19, "yes", "no"), levels = c("no", "yes"))

# c) mother's age
mothers$m_age <- mothers$edadanios.x

# d) mother's marital status
mothers$m_marital_status <-
  factor(with(mothers,
              case_when(is.na(f2_s9_900) ~ NA_character_,
                        as.integer(f2_s9_900) == 1 ~ "married",
                        as.integer(f2_s9_900) == 2 | as.integer(f2_s9_900) == 3 ~ "cohabiting",
                        TRUE ~ "non_partnered")), levels = c("non_partnered", "cohabiting", "married"))

# e) 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("no", "yes")))

# f) 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")

##############################################################################
#####                    3.3.2 Daughter-related variables                #####
##############################################################################

# a) not being enrolled in school
daughters$not_enrolled <- factor(ifelse(as.integer(daughters$f1_s2_17) == 1, "no", "yes"),
                                 levels = c("no", "yes"))

# b) lacked knoledge about the period
# Girls were asked whether they knew what was happening to them when they had their first period
daughters$lack_period_knowledge <-
```

```r
  factor(ifelse(as.integer(daughters$f2_s8_841) == 1, "no", "yes"), levels = c("no", "yes"))

# c) sexuality knowledge
# Girls were asked if they had ever received info about sexual intercourse, if they
# answered "yes" they were asked from who they had received the most info (family, school, other)
daughters$sexuality_knowledge <-
  factor(with(daughters,
              case_when(is.na(f2_s8_800f) ~ NA_character_,
                        as.integer(f2_s8_800f) == 2 ~ "no_info",
                        as.integer(f2_s8_801f) == 1 ~ "family",
                        as.integer(f2_s8_801f) == 2 ~ "school",
                        TRUE ~ "other")), levels = c("no_info", "family", "school", "other"))

##############################################################################
#####                    3.3.3 Household-related variables                #####
##############################################################################

# a) 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")))

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

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

# d) 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)) # we also get the number of household members

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

```r
daughters_tidy <- daughters %>%  select(household_id, subject_id, mother_id,
   early_sexual_activity, coercion_1st_intercourse, contraception_1st_intercourse,
   teenage_pregnancy, rural, h_internet, minority, not_enrolled, lack_period_knowledge,
   sexuality_knowledge) %>% left_join(income, by = c("household_id" = "household_id"))

mothers_tidy <- mothers %>% select(subject_id, m_age, m_marital_status, m_job, m_education,
   m_teenage_birth, m_lack_empowerment, m_marital_status)

data <- daughters_tidy %>% left_join(mothers_tidy, by = c("mother_id" = "subject_id"))

nrow(data) # the initial sample contains 1636 subjects
```

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

```r
# Removing the NAs ------------------------------
sapply(data, function(x){
  na <- is.na(x)
  na_total <- sum(na)
  return(na_total)
}) # we create this function to know how many more NAs there are
```

```
##                  household_id                    subject_id
##                             0                             0
##                     mother_id         early_sexual_activity
##                           289                           138
##      coercion_1st_intercourse contraception_1st_intercourse
##                          1296                          1301
##             teenage_pregnancy                         rural
##                           118                             0
##                    h_internet                      minority
##                             0                             0
##                  not_enrolled         lack_period_knowledge
##                             0                           119
##           sexuality_knowledge                      h_income
##                           118                             0
##                         m_age              m_marital_status
##                           289                           573
##                         m_job                   m_education
##                           289                           289
##               m_teenage_birth            m_lack_empowerment
##                           579                           573
```

```r
# we eliminate those who didn't live with their mothers (289 individuals)
data <- data %>% filter(!is.na(mother_id))

# we eliminate those who didn't provide information about their sexual activity (120 individuals)
data <- data %>% filter(!is.na(early_sexual_activity)) %>%
  filter(!(early_sexual_activity == "yes" & is.na(contraception_1st_intercourse)))

# we eliminate those whose mother did not report their sexual activity (249 individuals)
data <- data %>% filter(!is.na(m_lack_empowerment) & !is.na(m_teenage_birth))
```

```r
nrow(data) # the final sample has 978 people
```

```
## [1] 978
```

```r
summary(data)
```

```
##  household_id        subject_id         mother_id         early_sexual_activity
##  Length:978         Length:978         Length:978         no :819
##  Class :character   Class :character   Class :character   yes:159
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  coercion_1st_intercourse contraception_1st_intercourse teenage_pregnancy
##  no  :102                 no  : 86                      no :909
##  yes : 57                 yes : 73                      yes: 69
##  NA's:819                 NA's:819
##
##
##
##  rural     h_internet minority  not_enrolled lack_period_knowledge
##  no :585   no :557    no :768   no :899      no :777
##  yes:393   yes:421    yes:210   yes: 79      yes:201
##
##
##
##
##  sexuality_knowledge    h_income             m_age            m_marital_status
##  no_info:110         Min.   :-143850.0   Min.   :29.00   non_partnered:201
##  family : 65         1st Qu.:    250.0   1st Qu.:36.00   cohabiting   :283
##  school :731         Median :    539.0   Median :39.00   married      :494
##  other  : 72         Mean   :    601.6   Mean   :39.61
##                      3rd Qu.:    900.0   3rd Qu.:43.00
##                      Max.   :  18100.0   Max.   :49.00
##  m_job        m_education  m_teenage_birth m_lack_empowerment
##  no :387   none     : 31   no :475         no :882
##  yes:591   primary  :427   yes:503         yes: 96
##            secondary:370
##            tertiary :150
##
##
```

```r
################################################################################
#####                          4 Summary statistics                        #####
################################################################################

# We first separate the factor variables into different variables
data_copy <- data
data_copy$value <- TRUE
data_copy <- spread(data_copy, sexuality_knowledge, value, fill = FALSE, sep = "_")

data_copy$value <- TRUE
```

```r
data_copy <- spread(data_copy, m_education, value, fill = FALSE, sep = "_")

data_copy$value <- TRUE
data_copy <- spread(data_copy, m_marital_status, value, fill = FALSE, sep = "_")

# We create a variable to binarize the data
binarize <- function(x){
  x = ifelse(as.integer(x) == 2, TRUE, FALSE) # and apply it to all the columns
  return(x)
}

binary_var <- c("early_sexual_activity", "coercion_1st_intercourse",
                "teenage_pregnancy", "contraception_1st_intercourse", "minority", "rural",
                "h_internet", "not_enrolled", "lack_period_knowledge", "m_job",
                "m_teenage_birth", "m_lack_empowerment")

data_copy[, binary_var] <- sapply(data_copy[, binary_var], FUN = binarize)

# if(!require(haven)) install.packages("haven", repos = "http://cran.us.r-project.org")
# write_dta(data_copy, "early_sexual_activity_data.dta")

# We will use the chi square test and t test to compare variables within groups (early
# sexual initiators and no early sexual initiators)

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

indep_var <- c("early_sexual_activity", "teenage_pregnancy", "contraception_1st_intercourse")

cat_var <- c("m_lack_empowerment", "m_teenage_birth", "minority", "rural", "h_internet",
             "not_enrolled", "lack_period_knowledge", "sexuality_knowledge_no_info",
             "sexuality_knowledge_family", "sexuality_knowledge_school", "sexuality_knowledge_other",
             "m_job", "m_marital_status_non_partnered", "m_marital_status_cohabiting",
             "m_marital_status_married", "m_education_none", "m_education_primary",
             "m_education_secondary", "m_education_tertiary")

# We calculate the chi-square of each combination of categorical variables
# We create an empty matrix where we will store the p-value of the chi-square tests
chi_sq_test <- matrix(NA, nrow = length(cat_var), ncol = length(indep_var))
rownames(chi_sq_test) <- cat_var
colnames(chi_sq_test) <- indep_var

for(x in cat_var) {
  for(y in indep_var) {
    chi_sq_test[x,y] <- chisq.test(data_copy[,x], data_copy[,y])$p.value
  }
}

colnames(chi_sq_test) <- paste("p_value", colnames(chi_sq_test), sep = "_")

# Now, we calculate the percentage levels
cat_var_mean_T <- matrix(NA, nrow = length(cat_var), ncol = length(indep_var))
rownames(cat_var_mean_T) <- cat_var
colnames(cat_var_mean_T) <- indep_var
```

```r
for(x in cat_var) {
  for(y in indep_var) {
    cat_var_mean_T[x,y] <- mean(data_copy[data_copy[,y] == TRUE, x], na.rm = TRUE)
  }
}

cat_var_mean_F <- matrix(NA, nrow = length(cat_var), ncol = length(indep_var))
rownames(cat_var_mean_F) <- cat_var
colnames(cat_var_mean_F) <- indep_var

for(x in cat_var) {
  for(y in indep_var) {
    cat_var_mean_F[x,y] <- mean(data_copy[data_copy[,y] == FALSE, x], na.rm = TRUE)
  }
}

colnames(cat_var_mean_F) <- paste("no", colnames(cat_var_mean_F), sep = "_")

# means and t-test (continuous variables) ------------------------------
# We calculate the t-test of each combination of variables
# We create an empty matrix where we will store the p-value of the t-tests

cont_var <- c("m_age", "h_income")

t_test <- matrix(NA, nrow = length(cont_var), ncol = length(indep_var))
rownames(t_test) <- cont_var
colnames(t_test) <- indep_var

for(x in cont_var) {
  for(y in indep_var) {
    t_test[x,y] <- t.test(data_copy[,x] ~ data_copy[,y], var.equal = TRUE)$p.value
  }
}

colnames(t_test) <- paste("p_value", colnames(t_test), sep = "_")

# Now, we take the mean levels
cont_var_mean_T <- matrix(NA, nrow = length(cont_var), ncol = length(indep_var))
rownames(cont_var_mean_T) <- cont_var
colnames(cont_var_mean_T) <- indep_var

for(x in cont_var) {
  for(y in indep_var) {
    cont_var_mean_T[x,y] <- mean(data_copy[data_copy[,y] == TRUE, x], na.rm = TRUE)
  }
}

cont_var_mean_F <- matrix(NA, nrow = length(cont_var), ncol = length(indep_var))
rownames(cont_var_mean_F) <- cont_var
colnames(cont_var_mean_F) <- indep_var

for(x in cont_var) {
  for(y in indep_var) {
```

```
    cont_var_mean_F[x,y] <- mean(data_copy[data_copy[,y] == FALSE, x], na.rm = TRUE)
  }
}

colnames(cont_var_mean_F) <- paste("no", colnames(cont_var_mean_F), sep = "_")

mean_total  <- sapply(c(cat_var, cont_var), function(x){
  mean <- mean(data_copy[, x], na.rm = TRUE)
  return(mean)
})

mean_total_indep_var  <- sapply(indep_var, function(x){
  mean <- mean(data_copy[, x], na.rm = TRUE)
  return(mean)
})

mean_total_indep_var <- round(mean_total_indep_var, digits = 2)

mean_total_indep_var <-
  as.data.frame(matrix(c(mean_total_indep_var, rep("", 27)), nrow = 3, ncol = 10))

# Now, we combine all the matrixes into one matrix
table_sum_stat <- rbind(cbind(chi_sq_test, cat_var_mean_T, cat_var_mean_F),
                        cbind(t_test, cont_var_mean_T, cont_var_mean_F))

table_sum_stat <- cbind(mean_total, table_sum_stat)

table_sum_stat <-
  table_sum_stat[,c("mean_total", "early_sexual_activity", "no_early_sexual_activity",
                    "p_value_early_sexual_activity", "teenage_pregnancy", "no_teenage_pregnancy",
                    "p_value_teenage_pregnancy", "contraception_1st_intercourse",
                    "no_contraception_1st_intercourse", "p_value_contraception_1st_intercourse")]

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

table_sum_stat[, c(1:3, 5, 6, 8, 9)] <-
  round(table_sum_stat[, c(1:3, 5, 6, 8, 9)], digits = 2)
table_sum_stat[, c(4, 7, 10)] <- round(table_sum_stat[, c(4, 7, 10)], digits = 3)

table_sum_stat_copy <- as.data.frame(table_sum_stat)

p_value_stars <- function(x){
  x <- case_when(x <= 0.01 ~ paste(format(x, digits = 3), "***"),
                 x <= 0.05 ~ paste(format(x, digits = 3), "**"),
                 x <= 0.1 ~ paste(format(x, digits = 3), " *"),
                 TRUE ~ format(x, digits = 3))
  return(x)
}

table_sum_stat_copy[,c(4,7,10)] <- sapply(table_sum_stat_copy[,c(4,7,10)], p_value_stars)

table_sum_stat_copy[, c(1:3, 5, 6, 8, 9)] <-
  format(table_sum_stat_copy[, c(1:3, 5, 6, 8, 9)], digits = 2)
```

```r
# We reorder the rows in the table
table_sum_stat_copy <- table_sum_stat_copy[
  c("m_lack_empowerment", "m_teenage_birth", "m_age",
    "m_job", "m_marital_status_non_partnered", "m_marital_status_cohabiting",
    "m_marital_status_married", "m_education_none", "m_education_primary",
    "m_education_secondary", "m_education_tertiary", "not_enrolled",
    "lack_period_knowledge", "sexuality_knowledge_no_info", "sexuality_knowledge_family",
    "sexuality_knowledge_school", "sexuality_knowledge_other", "minority", "rural",
    "h_internet", "h_income"),]

colnames(mean_total_indep_var) <- colnames(table_sum_stat_copy)
rownames(mean_total_indep_var) <- indep_var
table_sum_stat_copy <- rbind(mean_total_indep_var, table_sum_stat_copy)

variables <- c("Early sexual initiation", "Teenage pregnancy",
               "Contraception use", "Mother lacks sexual empowerment",
               "Mother had a teenage birth",
               "Age", "Employed", "Non-partnered", "Cohabiting", "Married", "No education",
               "Primary education", "Secondary education", "Tertirary education",
               "Not enrolled in school", "No knowledge about period", "No knowledge",
               "Knows from family", "Knows from school", "Knows from other sources",
               "Ethnic minority", "Rural area", "Internet access", "Household income")

rownames(table_sum_stat_copy) <- variables
```

```r
# This code creates a latex table with the summary statistics
kable(table_sum_stat_copy, format = "latex", booktabs = T, linesep = "", escape = F,
      col.names = linebreak(c("Total","Yes","No", "p value","Yes", "No", "p value", "Yes", "No",
                              "p value"), align = "c"),
      caption = "Percentage and mean levels of explanatory variables by group") %>%
  add_header_above(c(" " = 2, "Early sexual initiation" = 3,
                     "Teenage pregnancy" = 3, "Contraception use" = 3)) %>%
  add_header_above(c(" " = 2, "Daughters' sexual outcomes" = 9)) %>%
  kable_styling(latex_options = c("hold_position", "scale_down")) %>%
  pack_rows("Daughters' sexual outcomes", 1, 3, latex_gap_space = "0.8em", italic = T, bold = F) %>%
  pack_rows("Main explanatory variables", 4, 5, latex_gap_space = "0.8em", italic = T, bold = F) %>%
  pack_rows("Mother-related variables", 6, 14, latex_gap_space = "0.8em", italic = T, bold = F) %>%
  pack_rows("Daughter-related variables", 15, 16, latex_gap_space = "0.8em", italic = T, bold = F) %>%
  pack_rows("Daughters' knowledge about contraception", 17, 20, latex_gap_space = "0.8em", italic = T, b
  pack_rows("Household-related variables", 21, 24, latex_gap_space = "0.8em", italic = T, bold = F) %>%
  footnote(general = "p values for comparison of percentages using chi-square. p values for comparison o
           threeparttable = T, footnote_as_chunk = T, fixed_small_size = F) %>% kable_styling(font_size
```

Table 1: Percentage and mean levels of explanatory variables by group

		Daughters' sexual outcomes								
		Early sexual initiation			Teenage pregnancy			Contraception use		
	Total	Yes	No	p value	Yes	No	p value	Yes	No	p value
Daughters' sexual outcomes										
Early sexual initiation	0.16									
Teenage pregnancy	0.07									
Contraception use	0.46									
Main explanatory variables										
Mother lacks sexual empowerment	0.10	0.14	0.09	0.086 *	0.13	0.10	0.469	0.12	0.15	0.782
Mother had a teenage birth	0.51	0.67	0.48	0.000 ***	0.70	0.50	0.003 ***	0.64	0.70	0.581
Mother-related variables										
Age	39.61	39.01	39.73	0.065 *	38.68	39.68	0.075 *	38.89	39.10	0.761
Employed	0.60	0.68	0.59	0.043 **	0.65	0.60	0.474	0.67	0.69	0.977
Non-partnered	0.21	0.25	0.20	0.143	0.20	0.21	1.000	0.29	0.22	0.434
Cohabiting	0.29	0.35	0.28	0.105	0.41	0.28	0.038 **	0.34	0.35	1.000
Married	0.51	0.40	0.53	0.006 ***	0.39	0.51	0.066 *	0.37	0.43	0.541
No education	0.03	0.07	0.02	0.007 ***	0.09	0.03	0.018 **	0.01	0.12	0.026 **
Primary education	0.44	0.43	0.44	1.000	0.48	0.43	0.550	0.37	0.49	0.180
Secondary education	0.38	0.39	0.38	0.810	0.33	0.38	0.502	0.41	0.37	0.736
Tertirary education	0.15	0.11	0.16	0.098 *	0.10	0.16	0.285	0.21	0.02	0.001 ***
Daughter-related variables										
Not enrolled in school	0.08	0.26	0.05	0.000 ***	0.36	0.06	0.000 ***	0.15	0.36	0.005 ***
No knowledge about period	0.21	0.29	0.19	0.006 ***	0.32	0.20	0.024 **	0.19	0.37	0.020 **
Daughters' knowledge about contraception										
No knowledge	0.11	0.08	0.12	0.229	0.10	0.11	0.918	0.03	0.13	0.044 **
Knows from family	0.07	0.09	0.06	0.308	0.09	0.06	0.647	0.12	0.06	0.244
Knows from school	0.75	0.65	0.77	0.004 ***	0.57	0.76	0.001 ***	0.68	0.63	0.558
Knows from other sources	0.07	0.18	0.05	0.000 ***	0.25	0.06	0.000 ***	0.16	0.19	0.882
Household-related variables										
Ethnic minority	0.21	0.30	0.20	0.009 ***	0.26	0.21	0.414	0.15	0.42	0.000 ***
Rural area	0.40	0.42	0.40	0.776	0.38	0.40	0.755	0.26	0.55	0.000 ***
Internet access	0.43	0.30	0.46	0.000 ***	0.26	0.44	0.005 ***	0.45	0.17	0.000 ***
Household income	601.64	724.23	577.84	0.722	532.07	606.92	0.900	996.21	493.36	0.054 *

Note: p values for comparison of percentagges using chi-square. p values for comparison of means using t-test. N=978 for early sexual initiation and teenage pregnancy. N=159 for contraception use. *p < .1; **p < .05; ***p < .01

```r
################################################################################
#####                          5 The logit models                          #####
################################################################################

# Model 1: -------------------------------
# early sexual activity

logit_m1 <-
  glm(early_sexual_activity ~ m_lack_empowerment + m_teenage_birth + # explanatory variables
        m_age + m_job + m_marital_status + m_education + # mother-related variables
        not_enrolled + lack_period_knowledge  + sexuality_knowledge+ # daughter-related variables
        minority + rural + h_internet + h_income, # household-related variables
      data = data, family = "binomial")

# we create a function to transform log odds to odds ratios and add some confidence intervals
logit_table <- function(x){
  odds_logit <- cbind(exp(cbind(OR = coef(x), confint(x))),summary(x)$coef[,4])
  colnames(odds_logit) <- c("OR", "2.5%", "97.5%","p.value")
  odds_logit <- as.data.frame(odds_logit)
  odds_logit[,1:3] <- sapply(odds_logit[,1:3], round, digits = 2)
  odds_logit$CI <- paste(format(odds_logit[,2], digits = 2),
                         format(odds_logit[,3], digits = 2), sep = " -")
  odds_logit$OR <-
    case_when(odds_logit$p.value <= 0.01 ~ paste(format(odds_logit$OR, digits = 2), "***"),
              odds_logit$p.value <= 0.05 ~ paste(format(odds_logit$OR, digits = 2), "**"),
              odds_logit$p.value <= 0.1 ~ paste(format(odds_logit$OR, digits = 2), "*"),
              TRUE ~ format(odds_logit$OR, digits = 2))
  odds_logit <- odds_logit[,c(1,5)]
  return(odds_logit)
}

logit_m1 <- logit_table(logit_m1)

# Model 2: -------------------------------
# teenage pregnancy

logit_m2 <-
  glm(teenage_pregnancy ~ m_lack_empowerment + m_teenage_birth + # explanatory variables
        m_age + m_job + m_marital_status + m_education + # mother-related variables
        not_enrolled + lack_period_knowledge + sexuality_knowledge + # daughter-related variables
        minority + rural + h_internet + h_income, # household-related variables
      data = data, family = "binomial")

logit_m2 <- logit_table(logit_m2)

# Model 3: -------------------------------
# contraception 1st intercourse

logit_m3 <-
  glm(contraception_1st_intercourse ~ m_lack_empowerment + m_teenage_birth + # explanatory variables
        m_age + m_job + m_marital_status + m_education + # mother-related variables
        not_enrolled + lack_period_knowledge + sexuality_knowledge + # daughter-related variables
        minority + rural + h_internet + h_income, # household-related variables
```

```
      data = data, family = "binomial")

logit_m3 <- logit_table(logit_m3)

# we write some code to create a nice latex table
table_logit_results <-  cbind(logit_m1, logit_m2, logit_m3)

variables <- c("Mother lacks sexual empowerment", "Mother had a teenage birth",
               "Age", "Employed", "Cohabiting", "Married", "Primary education",
               "Secondary education", "Tertirary education", "Not enrolled in school",
               "No knowledge about period", "Knows from family",
               "Knows from school", "Knows from other sources", "Ethnic minority",
               "Rural area", "Internet access", "Household income")

table_logit_results <- table_logit_results[-1,]
rownames(table_logit_results) <- variables
colnames(table_logit_results) <- c("OR","95\\% CI", "OR", "95\\% CI", "OR", "95\\% CI")

kable(table_logit_results, format = "latex", booktabs = T, linesep = "", escape = F,
      caption = "Odds ratio from logistic regression models predicting daughters' sexual outcomes") %>%
  add_header_above(c(" " = 1, "Early sexual initiation" = 2,
                     "Teenage pregnancy" = 2, "Contraception use" = 2)) %>%
  add_header_above(c(" " = 1, "Daughters' sexual outcomes" = 6)) %>%
  kable_styling(latex_options = c("hold_position", "scale_down")) %>%
  pack_rows("Main explanatory variables", 1, 2, latex_gap_space = "0.8em", italic = T, bold = F) %>%
  pack_rows("Mother-related variables", 3, 9, latex_gap_space = "0.8em", italic = T, bold = F) %>%
  pack_rows("Daughter-related variables", 10, 11, latex_gap_space = "0.8em", italic = T, bold = F) %>%
  pack_rows("Daughters' knowledge about contraception", 12, 14, latex_gap_space = "0.8em", italic = T,
  pack_rows("Household-related variables", 15, 18, latex_gap_space = "0.8em", italic = T, bold = F) %>%
  footnote(general = "N=981 for early sexual initiation and teenage pregnancy. N=159 for contraception
           threeparttable = T, footnote_as_chunk = T, fixed_small_size = F)
```

Table 2: Odds ratio from logistic regression models predicting daughters' sexual outcomes

	Daughters' sexual outcomes					
	Early sexual initiation		Teenage pregnancy		Contraception use	
	OR	95% CI	OR	95% CI	OR	95% CI
Main explanatory variables						
Mother lacks sexual empowerment	1.72 *	0.96 – 2.97	1.57	0.67 – 3.4	0.74	0.19 – 2.7
Mother had a teenage birth	1.98 ***	1.29 – 3.08	1.74 *	0.92 – 3.3	0.57	0.22 – 1.4
Mother-related variables						
Age	0.99	0.95 – 1.04	0.97	0.91 – 1.0	1.03	0.93 – 1.1
Employed	1.56 **	1.03 – 2.39	1.47	0.82 – 2.7	0.70	0.27 – 1.8
Cohabiting	0.81	0.47 – 1.38	1.25	0.58 – 2.8	0.74	0.21 – 2.5
Married	0.72	0.44 – 1.19	1.16	0.55 – 2.6	0.43	0.13 – 1.3
Primary education	0.41 *	0.16 – 1.13	0.49	0.15 – 1.8	8.09 *	0.95 – 186.7
Secondary education	0.60	0.23 – 1.70	0.51	0.15 – 1.9	8.40 *	0.99 – 190.5
Tertirary education	0.51	0.17 – 1.62	0.62	0.14 – 2.9	34.82 **	2.57 –1102.0
Daughter-related variables						
Not enrolled in school	8.58 ***	4.85 –15.45	8.05 ***	4.16 –15.6	0.51	0.19 – 1.3
No knowledge about period	1.59 **	1.01 – 2.49	1.70 *	0.90 – 3.1	0.77	0.31 – 1.9
Daughters' knowledge about contraception						
Knows from family	4.88 ***	1.83 –13.54	3.36 *	0.89 –13.0	2.78	0.32 – 31.2
Knows from school	2.74 **	1.32 – 6.27	1.66	0.64 – 5.0	1.36	0.23 – 11.0
Knows from other sources	9.66 ***	4.02 –24.89	8.06 ***	2.78 –26.5	1.25	0.19 – 11.0
Household-related variables						
Ethnic minority	1.57 *	0.99 – 2.49	1.04	0.52 – 2.0	0.36 **	0.14 – 0.9
Rural area	0.81	0.52 – 1.25	0.56 *	0.29 – 1.0	0.58	0.25 – 1.4
Internet access	0.64 *	0.40 – 1.02	0.55 *	0.27 – 1.1	2.09	0.80 – 5.6
Household income	1.00	1.00 – 1.00	1.00	1.00 – NA	1.00	1.00 – 1.0

Note: N=981 for early sexual initiation and teenage pregnancy. N=159 for contraception use. *p < .1; **p < .05; ***p < .01