Final exam

Mohammad Hasan

February 8, 2023

```
knitr::opts_chunk$set(echo = TRUE)
```

Question 2 - Import, extract and save data

Question 2 (a) : Download the SPSS data file KiGGS03_06.sav from moodle and import it into R.

```
library(foreign)

data <- read.spss("Data/KiGGS03_06.sav", to.data.frame=TRUE)
```

```
## re-encoding from CP1252
```

Question 2 (b) : Create a new dataframe in R named kiggs, which contains all variables (and only these) for the analysis (E070M, E070V, E072, E074, arztZ01, sex, age2, schichtz, e065z)

```
kiggs <- data[,c('E070M','E070V','E072','E074','arztZ01','sex','age2','schichtz','e065z')]
```

Question 2 (c) : Run the formatting steps} in the provided Rmd file data_formatting.Rmd. Save this formatted dataframe on your computer, e.g. on your desktop.

```
# Formatting the data
kiggs$E070M    <- factor(kiggs$E070M,    labels = c("yes, daily", "yes, occasionally", "no"))
kiggs$E070V    <- factor(kiggs$E070V,    labels = c("yes, daily", "yes, occasionally", "no"))
kiggs$E072     <- factor(kiggs$E072,     labels = c("yes, regularly", "yes, from time to time", "no, ne
kiggs$E074     <- factor(kiggs$E074,     labels = c("yes, regularly", "yes, from time to time", "no, ne
kiggs$sex      <- factor(kiggs$sex,      labels = c("boys", "girls"))
kiggs$age2     <- factor(kiggs$age2,     labels = c("0-1y", "2-3y", "4-5y", "6-7y", "8-9y", "10-11y", "
kiggs$schichtz <- factor(kiggs$schichtz, labels = c("low social status", "medium social status", "high

#Saving the data in my computer
write.csv(kiggs, "Data/kiggs.csv", row.names=FALSE)
```

Question 3 - Data transformations and data checks

Question 3 (a) : Check that the variables E070M, E070V, E072, E074 are all factors. If they are not, transform them into factors

```
#Checking whether the variables E070M, E070V, E072, E074 are all factors or not
is.factor(kiggs$E070M)
```

```
## [1] TRUE
```

```
is.factor(kiggs$E070V)
```

```
## [1] TRUE
```

```
is.factor(kiggs$E072)
```

```
## [1] TRUE
```

```
is.factor(kiggs$E074)
```

```
## [1] TRUE
```

Question 3 (a) : Set the value "has not breastfed" of variable E074 to NA for all children.

```
#Setting the value "has not breastfed" of variable E074 to NA for all children

kiggs["E074"][kiggs["E074"] == "has not breastfed"] <- "NA"
```

Question 3 (a) : Delete this now empty factor level from the variable

```
# Dropping the empty factor level or NA value from the variable
kiggs = kiggs[complete.cases(kiggs[,c("E074")]),]
```

Question 3 (a) : Check whether these two steps worked as intended.

```
#Checking whether "has not breastfed" value is available or not
library(stringr)

check1 = kiggs[str_detect(kiggs$E074, "has not breastfed"), ]
print(paste(nrow(check1), "row found with has not breastfed value"))
```

```
## [1] "0 row found with has not breastfed value"
```

```r
# Checking whether empty factor has been removed or not
check2 = kiggs[str_detect(kiggs$E074, "NA"), ]
print(paste(nrow(check2), "row found with NA value"))
```

```
## [1] "0 row found with NA value"
```

Question 3 (a) : Now calculate the new variable burdenS as the sum of the ranks of the four variables E070M, E070V, E072, E074 for each person (i.e. sum of the numerical factor levels).

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# The rank is "yes, regularly" = 1, "yes, from time to time" = 2, "no, never" = 3. Keeping NA value sam

RankForE070M = dense_rank(kiggs$E070M)

RankForE070V = dense_rank(kiggs$E070V)

RankForE072 = dense_rank(kiggs$E072)

RankForE074 = dense_rank(kiggs$E074)

burdenS = RankForE070M + RankForE070V + RankForE072 + RankForE074
```

Question 3 (a) : What is the meaning of this new variable, does a high value mean that the children were exposed to high levels of smoking, or that they were exposed to low levels of smoking?

As no smoking has the highest value in ranking, a high value of burdenS mean that the children were exposed to low levels of smoking.

Question 3 (b) : Add this variabes burdenS to the dataset kiggs, and save it in its updated form as an RData file (overwrite the previous file).

```r
# Adding the burdenS variable to the kiggs dataset
kiggs["burdenS"] = burdenS

#updating the kiggs file that is saved in computer
write.csv(kiggs, "Data/kiggs.csv", row.names=FALSE)
```

Question 4 - Descriptive statistics

Calculate absolute frequency tables for age2, sex, burdenS, and mean + standard deviation for arztZ01. Display them in a table or describe them in continuous text.

```
# Absolute frequency table

AFage2 <- table(kiggs$age2)
AFsex <-  table(kiggs$sex)
AFburdenS <- table(kiggs$burdenS)

AFage2
```

```
##
##   0-1y   2-3y   4-5y   6-7y   8-9y 10-11y 12-13y 14-15y 16-17y
##   1517   1493   1502   1587   1659   1556   1460   1434   1292
```

```
AFsex
```

```
##
##  boys girls
##  6816  6684
```

```
AFburdenS
```

```
##
##    4    5    6    7    8    9   10   11   12
##  134   76  459  319  961  561 2457 1083 6599
```

```
library(qwraps2)
options(qwraps2_markup = "markdown")

# Define format for the table:
Data_summary1 <-
  list("age2" =
       list("0-1y" = ~ qwraps2::n_perc0(age2 == "0-1y", show_symbol = TRUE, na_rm = TRUE),
            "2-3y"  = ~ qwraps2::n_perc0(age2 == "2-3y", show_symbol = TRUE, na_rm = TRUE),
            "4-5y"  = ~ qwraps2::n_perc0(age2 == "4-5y", show_symbol = TRUE, na_rm = TRUE),
            "6-7y"  = ~ qwraps2::n_perc0(age2 == "6-7y", show_symbol = TRUE, na_rm = TRUE),
            "8-9y"  = ~ qwraps2::n_perc0(age2 == "8-9y", show_symbol = TRUE, na_rm = TRUE),
            "10-11y"  = ~ qwraps2::n_perc0(age2 == "10-11y", show_symbol = TRUE, na_rm = TRUE),
            "12-13y"  = ~ qwraps2::n_perc0(age2 == "12-13y", show_symbol = TRUE, na_rm = TRUE),
            "14-15y"  = ~ qwraps2::n_perc0(age2 == "14-15y", show_symbol = TRUE, na_rm = TRUE),
            "16-17y"  = ~ qwraps2::n_perc0(age2 == "16-17y", show_symbol = TRUE, na_rm = TRUE)),
      "sex" =
       list("boys" = ~ qwraps2::n_perc0(sex == "boys", show_symbol = TRUE, na_rm = TRUE),
            "girls"  = ~ qwraps2::n_perc0(sex == "girls", show_symbol = TRUE, na_rm = TRUE)),
       "burdenS" =
       list("4" = ~ qwraps2::n_perc0(burdenS == "4", show_symbol = TRUE, na_rm = TRUE),
            "5"  = ~ qwraps2::n_perc0(burdenS == "5", show_symbol = TRUE, na_rm = TRUE),
            "6"  = ~ qwraps2::n_perc0(burdenS == "6", show_symbol = TRUE, na_rm = TRUE),
            "7"  = ~ qwraps2::n_perc0(burdenS == "7", show_symbol = TRUE, na_rm = TRUE),
            "8"  = ~ qwraps2::n_perc0(burdenS == "8", show_symbol = TRUE, na_rm = TRUE),
            "9"  = ~ qwraps2::n_perc0(burdenS == "9", show_symbol = TRUE, na_rm = TRUE),
            "10"  = ~ qwraps2::n_perc0(burdenS == "10", show_symbol = TRUE, na_rm = TRUE),
            "11"  = ~ qwraps2::n_perc0(burdenS == "11", show_symbol = TRUE, na_rm = TRUE),
            "12"  = ~ qwraps2::n_perc0(burdenS == "12", show_symbol = TRUE, na_rm = TRUE)),
      "arztZ01" =
      list("Mean (SD)" = ~ qwraps2::mean_sd(arztZ01, denote_sd = "paren", na_rm = TRUE, show_n = "never
       )

summary_table(kiggs, Data_summary1)
```

	kiggs (N = 13,500)
age2	
0-1y	1,517 (11%)
2-3y	1,493 (11%)
4-5y	1,502 (11%)
6-7y	1,587 (12%)
8-9y	1,659 (12%)
10-11y	1,556 (12%)
12-13y	1,460 (11%)
14-15y	1,434 (11%)
16-17y	1,292 (10%)
sex	
boys	6,816 (50%)
girls	6,684 (50%)
burdenS	
4	134 (1%)
5	76 (1%)
6	459 (4%)
7	319 (3%)
8	961 (8%)
9	561 (4%)
10	2,457 (19%)
11	1,083 (9%)
12	6,599 (52%)
arztZ01	
Mean (SD)	2.46 (3.40)

Also indicate how many missing values each of these 4 variables has, and how many observations have complete data for these 4 variables.

```
library(qwraps2)
options(qwraps2_markup = "markdown")

Data_summary2 <-
  list("age2" =
       list("Observations with complete data: "    = ~ sum(complete.cases(age2)),
           "Missing values: "  = ~ sum(!complete.cases(kiggs$age2))),
      "sex" =
       list("Observations with complete data: "    = ~ sum(complete.cases(sex)),
           "Missing values: " = ~ sum(!complete.cases(kiggs$sex))),
       "burdenS" =
       list("Observations with complete data: "    = ~ sum(complete.cases(burdenS)),
           "Missing values: " = ~ sum(!complete.cases(kiggs$burdenS))),
    "arztZ01" =
        list("Observations with complete data: "    = ~ sum(complete.cases(arztZ01)),
           "Missing values: " = ~ sum(!complete.cases(arztZ01)))
       )
summary_table(kiggs, Data_summary2)
```

	kiggs (N = 13,500)
age2	
Observations with complete data:	13500
Missing values:	0
sex	
Observations with complete data:	13500
Missing values:	0
burdenS	
Observations with complete data:	12649
Missing values:	851
arztZ01	
Observations with complete data:	13141
Missing values:	359

Question 5 - Linear Regression

5. (a) Calculate a linear regression, with arztZ01 as outcome and the predictors burdenS, sex, age2, schichtz and e065z.

```
#Transforming variables
predictorarztZ01<-as.numeric(as.character(kiggs$arztZ01))

#Converting factor frequency and amount factor variables to numeric
socialClass<-as.numeric(kiggs$schichtz)
totalSleep<-as.numeric(kiggs$e065z)
age2_num<-as.numeric(kiggs$age2)

#Fitting the linear model
regression_model<-lm(predictorarztZ01 ~ kiggs$burdenS + kiggs$sex + age2_num + socialClass + totalSleep)
regression_model
```

```
##
## Call:
## lm(formula = predictorarztZ01 ~ kiggs$burdenS + kiggs$sex + age2_num +
##     socialClass + totalSleep)
##
## Coefficients:
##    (Intercept)   kiggs$burdenS  kiggs$sexgirls        age2_num     socialClass
##        6.18923        -0.03136        -0.11773        -0.84083        -0.09282
##     totalSleep
##        0.06972
```

In the regression model, the variable to be predicted (arztZ01) is initially a factor of 34 levels. This variable has been converted into a metric variable.

Furthermore, variables that have more than 2 factor levels have been converted to numeric variables. Variables social class, total sleep and age are ordinal. These variables have been taken into the model as numeric variables. The sex variable is a factor of two levels and has been kept as it is (nominal). Lastly, the burdenS variable is numeric and have been kept as it is (Metric).

5. (b) To answer the question of whether the smoking behavior of parents has an influence on the health of children, adjusting for possible influencing factors, consider the significance test of the regression coefficient of burdenS in this regression. Report the regression coefficient of burdenS, interpret the coefficient, report its 95% confidence interval, and report its p value of the significance est.

What is your conclusion: Is there an association or not? In which direction?

```
summary(regression_model)$coefficients[,1]
```

```
##    (Intercept)  kiggs$burdenS kiggs$sexgirls       age2_num    socialClass
##     6.18923498    -0.03135777    -0.11772605    -0.84083464    -0.09281736
##     totalSleep
##     0.06971976
```

Interpretation of regression coefficients

burdenS: The estimated -0.03135777 regression coefficient indicates that for an increase in the burdenS level by a unit, the expected number of pediatrician visits level decreases by the value of the regression coefficient, given that the other variables are held constant.

```
# Estimation of 95% confidence intervals of the regression coefficient
confint(regression_model, level = 0.95)
```

```
##                        2.5 %      97.5 %
## (Intercept)      5.432342552  6.94612741
## kiggs$burdenS  -0.072824357  0.01010881
## kiggs$sexgirls -0.268024597  0.03257249
## age2_num       -0.907458492 -0.77421078
## socialClass    -0.200514353  0.01487962
## totalSleep      0.008868533  0.13057100
```

```
#Reporting p values of the predictors
summary(regression_model)$coefficients[,4]
```

```
##    (Intercept)  kiggs$burdenS kiggs$sexgirls        age2_num    socialClass
##   6.493593e-57   1.382769e-01   1.247158e-01  4.006689e-130   9.117625e-02
##     totalSleep
##   2.473448e-02
```

Based on the p-values it can be stated that burdenS does not have a significant effect on the level of pediatrician visits. The P value of burdenS is 0.1383, which is greater than 0.05.

5. (c) Since there is evidence that individuals drawn from the same area are correlated with each other, but we are not interested in the effect of the area on the health ... what would be a suitable strategy for accounting for this correlation?

One possible strategy is to include area as a random effect in our statistical model.

Question 6 – Sample size calculation

6. (a) Look at the literature or think for yourself based on expert knowledge what effect size you would expect. State the effect size that you are assuming and explain why.

There are two research examples given below which show the effect of smoking during pregnancy on weight of the baby. Both suggest that there are substantial effect of smoking during pregnancy on child weight. We also found in our research that burdenS has -0.03135777 regression coefficient and it indicates that when there are less smoking within the family (mother and father), the pediatrician visits decreases, which could mean child is more healthy when there is less smoking in household.

Based on the literature review from this two research and finding of this research, I expect the effect size would be MEDIUM (0.5 for Cohen's d and [.3 to .5 or -.3 to -.5] for Pearson's r).

The reason behind this assumption is, both the mentioned research and this experiment do not show enough strength and proof that the effect size has the power to be large. Also, the mentioned research did not mention anything about the amount of cigarettes rather than smoking generally. So, I expect the effect size would be medium.

1. https://www.sciencedirect.com/science/article/pii/0895435688900509
2. https://www.bmj.com/content/2/5806/127

6. (b) Choose an appropriate statistical model for the sample size calculation and explain why.

A two-sample t-test is appropriate in this study we can build the model by comparing the mean birth weight of babies born to two different groups of mothers, those who smoke 10 cigarettes per day and those who do not smoke at all. The two groups are independent of each other, which mean that the birth weight of one baby is not influenced by the birth weight of another baby. Also, the outcome variable (birth weight) is continuous, which makes the two-sample t-test an appropriate statistical test for comparing the means of the two groups.

The sample size calculation for this test would depend on several factors such as the desired level of statistical power, the significance level, the expected difference in mean birth weight, and the standard deviation of birth weight in both groups.

6. (c) Now compute the minimum necessary sample size for a power of 80% and a significance threshold of alpha = 0.05, for example by using a function in the R package pwr. What is the sample size?

```r
library(pwr)

d <- 0.5
power <- 0.8
sig.level <- 0.05

pwr.t.test(d = d, power = power, sig.level = sig.level, alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
```

```
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The function returns a result of 63.76, which we round up to the nearest whole number of 64. Therefore, the minimum necessary sample size is 64 participants per group (128 participants in total).

6. (d) Do you think this is a good study, or do you see any major weaknesses in the study design?

There are several potential weaknesses in the study design:

In the study, mothers who smoke during pregnancy may differ systematically from mothers who do not smoke in terms of other factors that could affect birth weight, such as maternal age, socioeconomic status, and diet. Which is not covered in the study. There is no information about the timing of smoking during pregnancy and the study does not investigate the effects of smoking more or less than this amount. There is no information about factors that could affect birth weight, such as maternal weight gain during pregnancy, gestational age, and the presence of medical conditions such as hypertension and diabetes.