Notes on Longitudinal Data Analysis for Epidemiology

Florencia D'Andrea

2024-04-14

Table of contents

Welcome

1 State of the Art - Statistics

All this section is based on Twisk (2013)

1.1 Longitudinal studies

> Longitudinal studies are defined as studies in which the outcome variable is repeateadly measured; i.e. the outcome variable is measured in the same subject on several occasions. – Extracted from Twisk (2013)

Characteristics:

- Observations of one subject over time are not independent of each other.
- Statistics should consider that repeated observations of each subject are correlated.
- This studies bring the illusion than are solving causality but only we can try temporality.

Table 1.1: Statistical notation

	Notation
number of subjects	`i = 1 to N`
number of covariates	`j = 1 to J`
number of times a particular subject is measured	`t = 1 to T`
outcome variable	`Y`
covariates	`X`

1.2 Cohort studies

1.2.1 Observational cohort studies

The question of probable causality remains unanswered.

can be divided into:
* **prospective**. * The only one that can be characterized as longitudinal.

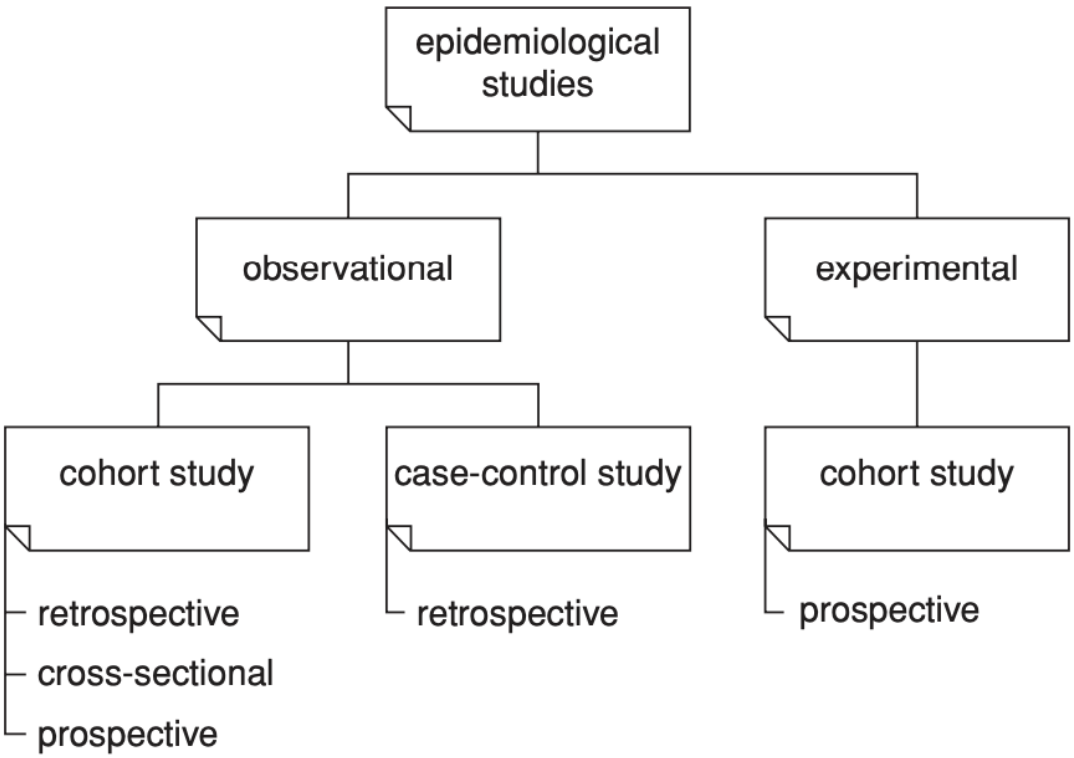


Figure 1.1: Image extracted from Twisk (2013)

* Analyze the longitudinal development of a certain characteristic over time (growth or deterioration).
* **tracking**: "stability" of a certain characteristic over time.

- **retrospective**.
- **cross-sectional**.

Term	Definition
`age`	time from date of birth to date of measurement
`period`	time or moment at which the measurement is taken
`birth cohort`	group of subjects born in the same year

1.2.1.1 Confounding Effects

age, *period* and *cohort* effects could produce variations in the results.

- **period effect**.
  If we measure physical activity during summers, it is likely that we can have more physical activity a hot summer than a rainy one. This can produce a bias in the age trend.
- **cohort effect**. If we want to unify results for the same age for cohorts that start at different ages, we will find that the trend is much flatter than the effects of the cohorts in isolation.

One way to avoid the bias is to use an approach called *multiple longitudinal design* Basically, multiple longitudinal design is to work with more than one cohort at the same time. If all the cohorts show a defined pattern for a particular measure in time, we will be able to detect it with this approach.

- **Test or learning effect**. Individuals start performing better with exposure.
- **Low reproducibility of the measurements**. Inter-period correlation coefficients (IPCs) (van 't Hofand Kowalski, 1979)

1.2.2 Experimental cohort studies (clinical trials)

- There are prospective (ie longitudinal).
- The outcome variable Y is measured at least twice (the classical "pre-test," "post-test" design).
- The issue of causality can be covered

1.3 One continous outcome variable Y is repeatedly measured over time

1.3.1 Two measurements

1.3.1.1 Parametric: paired t-test

Is there a difference in the outcome variable `Y` between `t = 1` and `t = 2`?

The **paired t-test** is used to test the hypothesis that the mean difference between `Yt1` and `Yt2` equals zero.

- Observations within one individual are dependent on each other.
- Use if the number of subjects is quite large (say above 25).

1.3.1.1.1 Assumptions

1 - The observations of different subjects are independent and;
2 - The differences between the two measurements are approximately normally distributed.

1.3.1.2 Non-parametric: (Wilcoxon) signed rank sum

Doesn't assume any distribution.

1.3.2 More than two measurements

Does the outcome variable `Y` change over time?

1.3.3 Multivariate analysis of variance (MANOVA) for repeated measurements

1 - Observations of different subjects. at each of the repeated measurements need to be independent; and. 2 - The observations need to be multivariate normally distributed, which is comparable but slightly more restrictive than the requirement that the differences between subsequent measurements be normally distributed.

1.3.3.1 ANOVA (univariate) vs. MANOVA (multivariate)

To perform an ANOVA there is one extra assumption with 2 parts:

Sphericity assumption (epsilon coefficient)

3 - All correlations in outcome variable Y between repeated measurements are equal, irrespective of the time interval between the measurements.

4 - The variances of outcome variable Y are the same at each of the repeated measurements.

Which approach should be used? If the assumptions are met, ANOVA is more powerful for smaller samples:

> "The restriction of the assumption of sphericity (i.e. equal correlations and equal variances over time) leads to an increase in degrees of freedom, i.e. an increase in power for the "univariate" approach. This increase in power becomes more important when the sample size becomes smaller. Historically, the "multivariate" approach was developed later than the "univariate" approach, especially for situations when the assumption of sphericity does not hold. So, one could argue that **when the assumption of sphericity is violated, the "multivariate" approach should be used**.

– Extracted from Twisk (2013)

MAMOVA in R - https://www.appsilon.com/post/manova-in-r

1.4 One continuous outcome variable Y is compared between different groups.

This design is known as the "one-within, one-between" design. Time is the within-subject factor and the group variable is the between-subjects factor.

> Is there a difference in change over time for outcome variable Y between two or more groups?

- This question can also be answered with MANOVA for repeated measurements if it is assumed that the covariance matrices of the different groups that are compared to each other are homogeneous (independent sample `t-test`).
- Apparently, MANOVA could be biased when you have a lot of drop-offs in the study (Everitt (1998)).

1.5 Continuous outcome variable and several covariates

(which can be either continuous, dichotomous, or categorical)

> **"traditional" methods** tried to reduce the statistical longitudinal problem into a **cross-sectional problem**.

1.5.1 Traditional methods.

- Analysis of the relationships between changes in different parameters between two points in time. (but, you are not using all the data)
- Use individual regression lines with time.

1.5.2 New methods.

With the development of (new) statistical techniques, such as:

1. Generalized estimating equations (GEE) analysis and;
2. Mixed-model analysis,

it has become possible to analyze longitudinal relationships using all available longitudinal data, without summarizing the longitudinal development of each subject into one value.

Crippa (2022) A review of Longitudinal Data Analysis in R: https://rpubs.com/alecri/review_longitudinal

- In this model the coefficients of interest are `1j`, because these regression coefficients show the magnitude of the longitudinal relationship between the outcome variable (`Yit`) and the covariates (`Xijt`).
- Because of the dependency of the repeated observations within one subject, the relationship between X and Y must be adjusted for the subject. There is one by subject and are represented by dummy variables.

1.5.2.1 GEE

- Before carrying out a GEE analysis, the within-subject correlation structure must be chosen.
- In the literature it is assumed that GEE analysis is robust against a wrong choice of the correlation structure. However, this is only the case when there are no missing data and when the model correctly specifies the mean.
- The interpretation of `1` is more complex than in a cross-sectional analysis.

1.5.2.2 Mixed Models

- This type of study design is characterized by a hierarchical structure. In longitudinal studies observations within one subject over time are correlated. The observations over time are nested within the subject.
- The way mixed model analysis adjusts for the subject is different from the way GEE analysis adjusts for the subject.
- when an adjustment is made for the subject (i.e. the id_number), for each subject different intercepts are calculated.
- The first step within a mixed model analysis is therefore to draw a normal distribution around the intercepts and in the second step the variance of that normal distribution is estimated.
- That variance is added to the longitudinal regression model in order to adjust for the subject in an efficient way.
- Because this variance is known as the **random intercept** (`0i`), mixed model analysis is also known as random coefficient analysis.
- The general idea behind a mixed model analysis is that the unexplained variance in outcome variable Y is divided into different components. One of the components is related to the random intercept and another component is related to random slopes.

- A model with a random intercept allows the intercepts to differ between the subjects, but the regression coefficient for the covariate X is the same for all the subjects. In a longitudinal study it is not uncommon that besides the intercepts also the regression coefficients for X differ between the subjects
- When regression coefficients for X differ between subjects, there is an interaction between the covariate X and the subject. As for the adjustment for the subject, also the interaction with the subject has to be added to a cross-sectional regression model with dummy variables.

1.6 GAMM

`gamm` considers that the observations aren't independent (not the same as `gam`)

1.6.1 Random Effects

As we saw in the section about changing the basis, `bs` specifies the type of underlying base function. For random intercepts and linear random slopes we use `bs = "re"`, but for random smooths we use `bs = "fs"`.

3 different types of random effects in GAMMs (`fac` factor coding for the random effect; x0 continuous fixed effect):

- *random intercepts* adjust the height of other model terms with a constant value: `s(fac, bs = "re")`
- *random slopes* adjust the slope of the trend of a numeric predictor: `s(fac, x0, bs = "re")`
- *random smooths* adjust the trend of a numeric predictor in a nonlinear way: `s(x0, fac, bs = "fs", m = 1)`, where the argument `m=1` sets a heavier penalty for the smooth moving away from 0, causing shrinkage to the mean.

https://stats.stackexchange.com/questions/391912/how-to-fit-a-longitudinal-gam-mixed-model-gamm

https://jacolienvanrij.com/Tutorials/GAMM.html#setting-up-a-gamm-model https://r.qcbs.ca/workshop08/pr en/workshop08-pres-en.html#129 https://r.qcbs.ca/workshop08/book-en/introduction-to-gams.html https://bart-larsen.github.io/GAMM-Tutorial/

- We can try to build a more appropriate model by fitting the data with a smoothed (non-linear) term specified by expressions of the form `s(x)`
- `Radj` is the variance explained by the model

2 Scientific articles

stringhini, 2018

Premature mortality reduction from chronic diseases

Biological Risk factors.

- high blood pressure.
- obesity.
- tobacco use.
- excess salt intake.
- diabetes.
- insufficient physical activity.
- alcohol consumption.

Socioeconomic status.

- occupational group.
- educational attainment.
- level of income and wealth.
- place of residence.

2.1 General Additive Mixed Model (GAMM)

semi-parametric model

Let's start with an equation for a Gaussian linear model:

y= 0+x1 1+ , N(0, 2)

What changes in a GAM is the presence of a smoothing term:

y= 0+f(x1)+ , N(0, 2)

This term could be many things.

2.1.1 Walking speed and age

Fixed Effects Predictors.

- age.
- height.

Random Effect.

- study at the intercept and age slope.

2.1.2 Number of years of functioning lost (primary outcome)

It is based on the predictions of the previous model.

Fixed Effects

- age.
- age2.
- height.
- year of birth.
- distances walked.

- risk factor under study (minimally adjusted model).
- all risk factors (mutually adjusted models).

CI - Model based parametric 5000 bootstrap samples

2.1.3 Years of life lost (secundary outcome)

Difference between the areas of the survival curves.

Survival curves Kaplan-Meier adjusted curves, conditional on survival to age 60 years. They run a shared frality Cox model with age as time scale, stratified by the levels of the given risk factor and a year of birth as covariate (for minimally adjusted models) or year of birth and the remaining risk factors as covariates (mutually adjusted models)

schrempft 2022

2.2 Pace of aging

similar to Dunedin Study investigators

1 - Biomarkers were standarized for healthy men and women. Z-scores were reversed for HDL and creatinine clearence.

2 - Mixed-effects models with a random interecept and a random linear slope, were used to calculate participants' personal slopes (change in biomarkers per year) Each year included time, age a t baseline center in the samples mean, and a interaction term between the time and age at the baseline. For biomarkers that show a non linear trajectory, an additionally

3 - The individual slopes for each biomarker (annual change in biomarker Z-score) wee aggregated to create a total Pace of Aging score.

Covariates: - Alcohol $>$ 14 units. - hipertensive or diabetic medication. - pysical inactivity. - smoking status

carmeli, 2019

3 Data Analysis

```
library(haven) # read stata file format
library(readxl) # read excel file format
library(kableExtra)
library(lubridate)
library(tidyr)
library(plotly)
library(ggplot2)
library(ggbeeswarm)
library(PerformanceAnalytics)
library(forcats)
library(gamm4)
library(purrr)
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
```

The data was extracted from the Canadian Longitudinal Study on Aging Webpage.

The baseline study was conducted in 2008, the first follow-up in 2015. It seems to be data for the second follow up (variables ending in `_COF2`), even if there is not information available in the webpage.

3.1 Variables

I will assume that for the second follow-up, the variable names are the same as those described for the first follow-up.

```
# Data
data <- haven::read_dta("data/CLSA_test sample_selected.dta")

# Variable description for the baseline study
var_baseline <- readxl::read_excel("data/baseline_data_dictionaries.xlsx",
                                   col_types = "text") |>
  dplyr::filter(name %in% colnames(data)) |>
  dplyr::select(-SourceTable)
```

```
# Variable description for the first follow-up
var_follow_up1 <-
  readxl::read_excel("data/follow-up1_data_dictionaries_v2.xlsx",
             col_types = "text") |>
  dplyr::filter(name %in% colnames(data)) |>
  dplyr::rename(label = `label:en`,
         question = `question:en`,
         comment = `comment:en`)
```

Some of the variables are labeled. I will extract the labels to can use the information later. For this I create `dlabel()`.

```
# Extract labels and save them in small datasets
dlabel <-  function(variable){
data |>
  dplyr::select({{variable}}) |>
  dplyr::mutate(lab = haven::as_factor({{variable}})) |>
  dplyr::filter (!duplicated({{variable}})) |>
  dplyr::arrange({{variable}})
}

ed_label <- dlabel(ED_UDR04_COM)
sdc_label <- dlabel(SDC_MRTL_COM)
inc_tot_label <- dlabel(INC_TOT_COM)
inc_ptot_label <- dlabel(INC_PTOT_COM)
wea_label <- dlabel(WEA_SVNGSVL_MCQ)
```

This are the variables available in the dataset:

```
table <- rbind(var_baseline, var_follow_up1) |>
  dplyr::select(name, label, everything(), -unit, -table) |>
  dplyr::arrange(name) |>
  kableExtra::kable(caption = "Variables in the dataset") |>
  kableExtra::kable_classic(full_width = F)
table
```

Table 3.1: Variables in the dataset

name	label	valueType	question
AGE_NMBR_COF1	Participant age at beginning of FU1	integer	NA
AGE_NMBR_COM	Age (years)	integer	NA
ED_UDR04_COM	Highest Level of Education - Respondent, 4 Levels	text	NA
INC_PTOT_COM	Total personal income	text	What is your
INC_TOT_COM	Total household income	text	What is your
SDC_MRTL_COM	Marital/partner status	text	What is your
SEX_ASK_COM	Sex	text	Are you male
WEA_SVNGSVL_MCQ	Total value of savings and investments	text	What is the
WEA_SVNGSVL_MCQ	Total value of savings and investments	text	What is the
WLK_TIME_COF1	Total time required to complete 4mWalk (in seconds)	decimal	NA
WLK_TIME_COM	Total time required to complete 4mWalk (in seconds)	decimal	Record time
startdate_COF1	In-Home Questionnaire start date& time	datetime	NA
startdate_COM	Date and time at start of interview	datetime	NA

3.2 Exploratory Data Analysis (EDA)

3.2.1 General observations

- In this dataset there are are reported the baseline values (`*_COM`), as well as the first follow up in 2017/18 (`*_COF1`) and the second follow up (`*_COF2`), that seems to be 2019/20/21 in the data.
- Data were collected between 2012 and 2021.

```
summary(data)
```

```
      id         startdate_COM      startdate_COF1     startdate_COF2
Min.   :    6   Length:1000        Length:1000        Length:1000
1st Qu.:21704   Class :character   Class :character   Class :character
Median :45524   Mode  :character   Mode  :character   Mode  :character
Mean   :44858
3rd Qu.:67622
Max.   :90271

 AGE_NMBR_COM   AGE_NMBR_COF1   AGE_NMBR_COF2   SEX_ASK_COM
Min.   :45.00   Min.   :47.00   Min.   :50.00   Length:1000
1st Qu.:55.00   1st Qu.:58.00   1st Qu.:60.00   Class :character
Median :63.00   Median :65.00   Median :68.00   Mode  :character
```

```
Mean   :63.47   Mean   :66.06   Mean   :68.44
3rd Qu.:72.00   3rd Qu.:74.00   3rd Qu.:76.00
Max.   :86.00   Max.   :89.00   Max.   :91.00
                NA's   :81      NA's   :148
 ED_UDR04_COM    SDC_MRTL_COM    INC_TOT_COM     INC_PTOT_COM
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:4.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
Median :4.000   Median :2.000   Median :3.000   Median :2.000
Mean   :3.577   Mean   :2.308   Mean   :3.494   Mean   :2.792
3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:3.000
Max.   :9.000   Max.   :5.000   Max.   :9.000   Max.   :9.000

WEA_SVNGSVL_MCQ  WLK_TIME_COM     WLK_TIME_COF1       WLK_TIME_COF2
Min.   :1.000   Min.   :-88.000   Min.   :-99999.00   Min.   :-99999.00
1st Qu.:2.000   1st Qu.:  3.600   1st Qu.:     3.60   1st Qu.:-99993.00
Median :3.000   Median :  4.090   Median :     4.09   Median :     3.22
Mean   :3.116   Mean   :  4.172   Mean   : -6439.67   Mean   :-46395.00
3rd Qu.:3.000   3rd Qu.:  4.683   3rd Qu.:     4.72   3rd Qu.:     4.18
Max.   :9.000   Max.   : 14.410   Max.   :    10.63   Max.   :    20.32
NA's   :66      NA's   :4         NA's   :81          NA's   :148
```

- The cases were `NA` values were reported for `COF1` and `COF2` were removed. The `NAs` for `WEA_SVNGSVL_MCQ` were also eliminated from the analysis.
- Variable `height` is missing.

```
data$startdate_COF1[data$startdate_COF1 == ""] <- NA
data$startdate_COF2[data$startdate_COF2 == ""] <- NA


data <- data |>
  dplyr::mutate(
    startdate_COM = as.POSIXct(startdate_COM,
                               format = "%Y-%m-%dT%H:%M:%OS"), # convert dates
    startdate_COF1 = as.POSIXct(startdate_COF1,
                                format = "%Y-%m-%dT%H:%M:%OS"),
    startdate_COF2 = as.POSIXct(startdate_COF2,
                                format = "%Y-%m-%dT%H:%M:%OS")
  ) |>
  dplyr::filter(
                is.na(data$startdate_COF1) == FALSE,
                is.na(data$startdate_COF2) == FALSE,
                is.na(data$WEA_SVNGSVL_MCQ) == FALSE) # remove NAs
```

3.2.2 Data wrangling

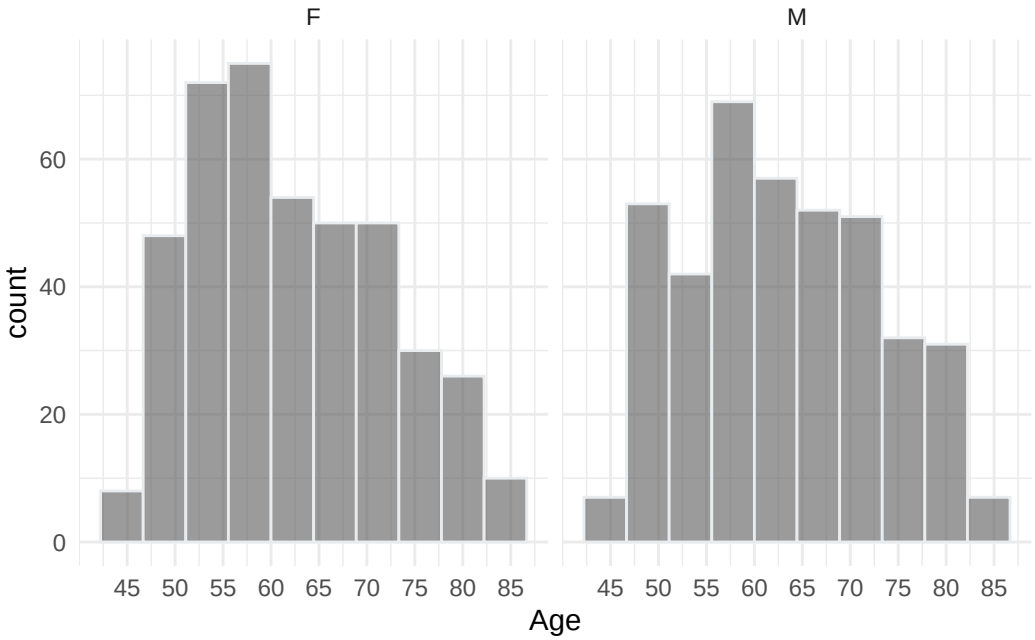
- The dataset is divided in subtables to make it easier to use.
- The variable `Walking time` is converted to `Walking speed`.
- Following Stringhini et al. (2018), the ages have been reduced from 45 to 90.

```
# Socieconomic variables sub-table
soc_vars <- data |>
  haven::zap_labels() |> # I have already saved the labels in tables
  tidyr::pivot_longer(col = c(9, 11, 12, 13),
                      values_to = "SOC_VALS",
                      names_to = "SOC_VARS") |>
  dplyr::filter(!SOC_VALS < 0) |>
  dplyr::select(id, SOC_VALS, SOC_VARS)

# Gender sub-table
gender <- data |>
  haven::zap_labels() |>
  dplyr::select(id, SEX_ASK_COM)

# Walking speed sub-table
wlk_time <- data |>
  haven::zap_labels() |>
  tidyr::pivot_longer(14:16, values_to = "WLK_TIME", names_to = "COHORT") |>
  dplyr::filter(!WLK_TIME < 0) |>
  tidyr::separate(COHORT, into = c("A", "B", "COHORT"), sep = "_") |>
  dplyr::select(id, COHORT, WLK_TIME)

# Age sub-table
age <- data |>
  haven::zap_labels() |>
  pivot_longer(5:7,
               values_to = "AGE_VALS", names_to = "AGE_VARS")  |>
  dplyr::select(id, AGE_VALS, AGE_VARS) |>
  dplyr::filter(!is.na(AGE_VALS)) |>
  tidyr::separate(AGE_VARS, into = c("A", "B", "COHORT"), sep = "_") |>
  dplyr::select(id, COHORT, AGE_VALS)

# Final table
data_by_cohort <-  dplyr::left_join(age, wlk_time)
data_plot <- dplyr::left_join(data_by_cohort, soc_vars)
```

```r
# Convert Walking time to Walking speed
data_plot <- data_plot |>
  dplyr::mutate(WLK_TIME = (4 / WLK_TIME)) |>
  dplyr::filter(AGE_VALS >= 45,
                AGE_VALS <= 90)
```

3.2.3 Demographic variables

```r
  data_plot |>
    dplyr::left_join(gender) |>
    select(id, AGE_VALS, SEX_ASK_COM) |>
    filter (!duplicated(id)) |>
    ggplot() +
    scale_fill_brewer() +
    geom_histogram(
      aes(AGE_VALS),
      bins = 10,
      color = "#e9ecef",
      alpha = 0.6,
      position = 'identity'
    ) +
    scale_x_continuous(breaks = round(seq(45,90, by = 5),1)) +
    xlab("Age") +
    facet_wrap(vars(SEX_ASK_COM)) +
    theme_minimal()
```



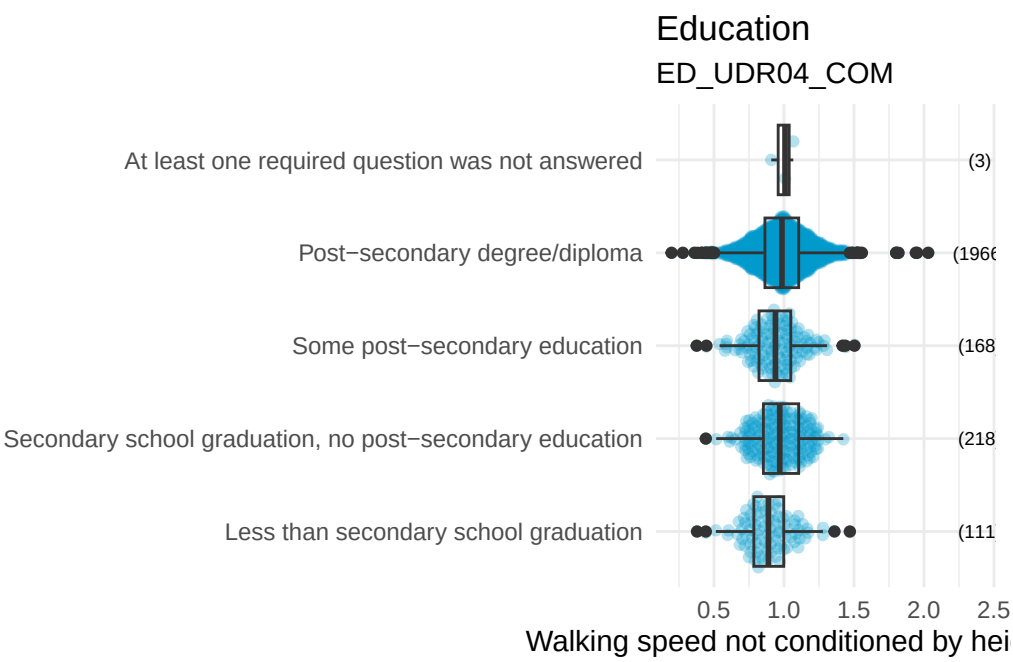
Comments:

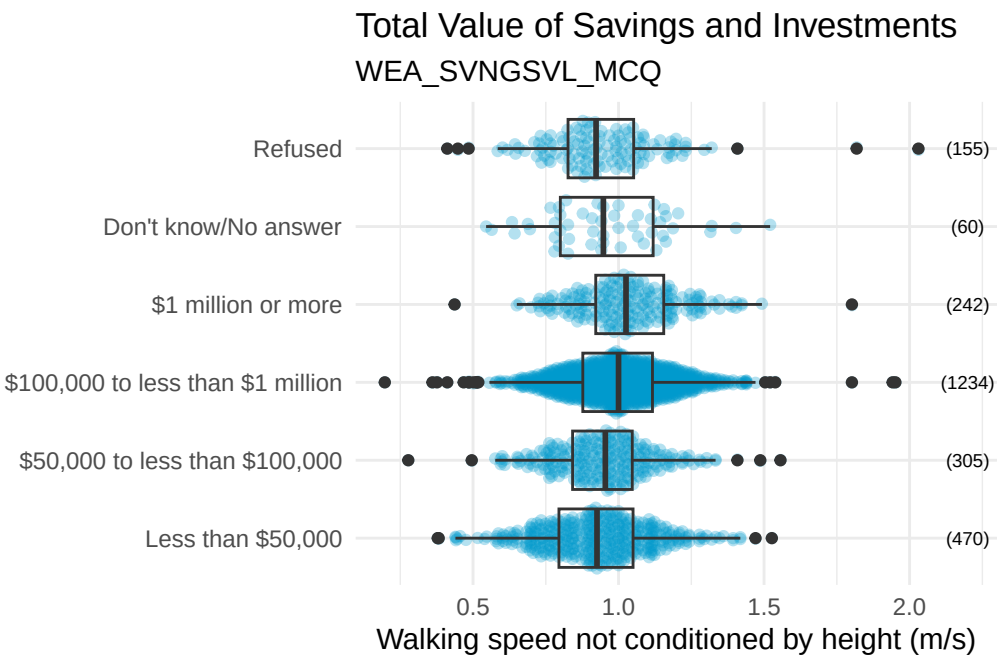
- There are fewer observations for older ages.
- There are 824 observations in the dataset. 48.7 % are men and 51.3 % are women.

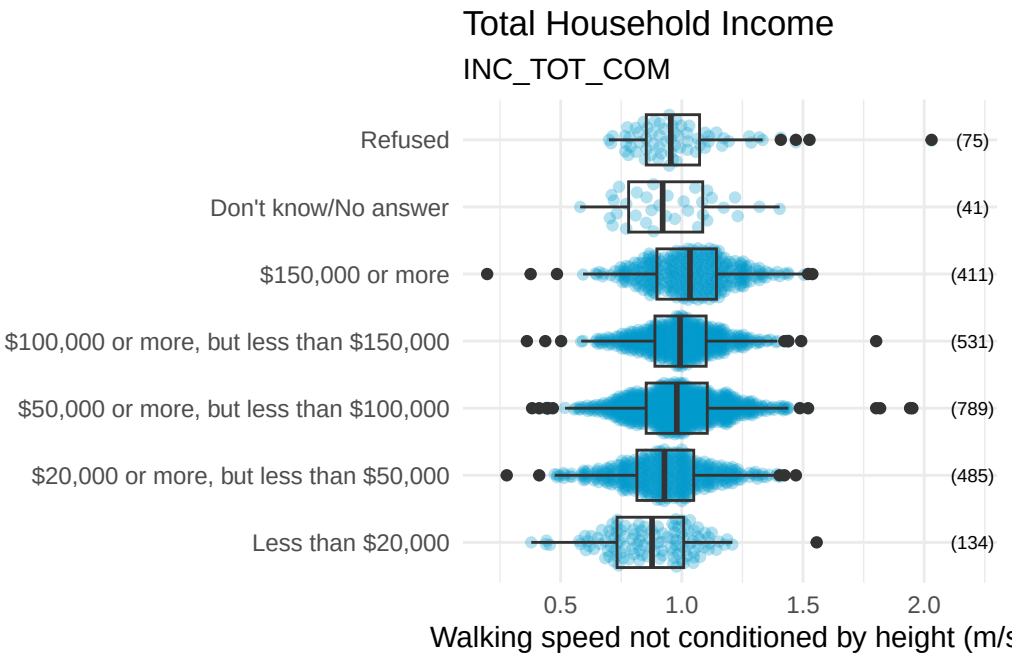
3.2.4 Socioeconomic Variables

```
data_boxplot <-  data_plot |>
  dplyr::filter(SOC_VARS == 'ED_UDR04_COM') |>
  dplyr::left_join(ed_label, by = dplyr::join_by(SOC_VALS == ED_UDR04_COM))
# Number of points by category
obs1 <- data_boxplot |> dplyr::count(lab)

data_boxplot |>
  ggplot(aes(x = as.factor(lab), y = WLK_TIME)) +
  geom_quasirandom(alpha = 0.3, color = "deepskyblue3") +
  geom_boxplot(fill = 'transparent') +
  coord_flip() +
  labs(title = "Education", subtitle = "ED_UDR04_COM") +
  xlab("") +
  ylab("Walking speed not conditioned by height (m/s)") +
```

```
geom_text(data = obs1, aes(y = 2.4,
                           label = paste0("(", n, ")")),
          size = 2.5) +
theme_minimal()
```







Considering that we don't have `height` values to associate with walking speed, we will solely make comments about the amount of observations in each category

Comments:

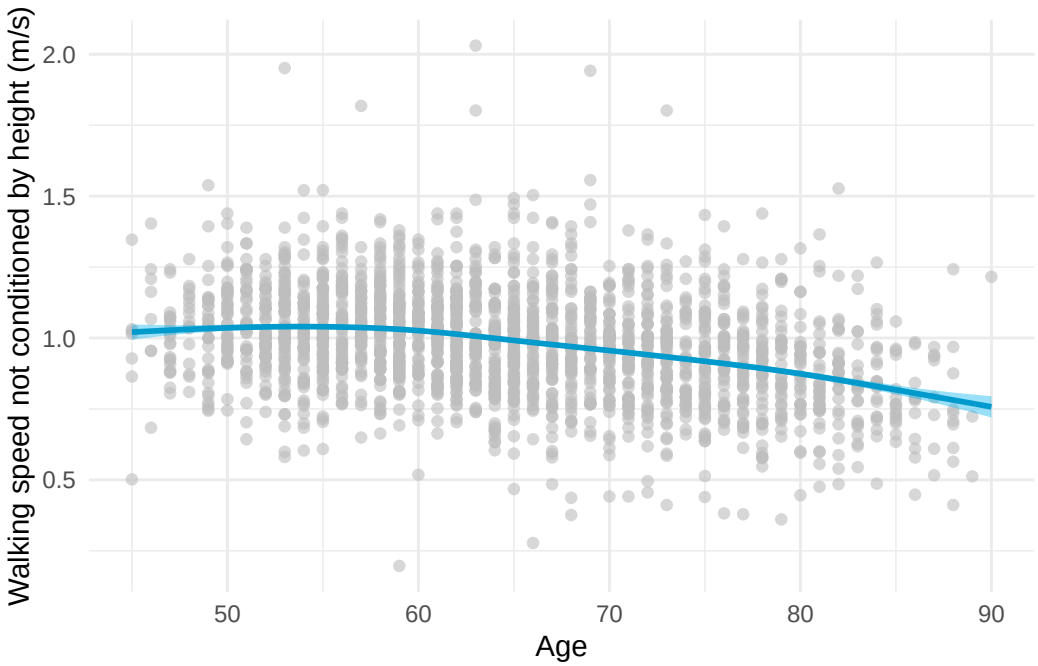
- 79.7% of the participants hold a post-secondary degree/diploma.
- 50% of the participants have `Total Value of Saving and Investments` between `$100,000 to less than $1 million dollars`.
- However, the `Total Household Income` is more homogeneous, being reported `$50,000 or more, but less than $100,000` by 32% of the participants. This could indicate that some study participants with a post-secondary degree and $100,000 - $1 million dollars of savings and investments belongs to more than one `Total Household Income` category. More analysis should be conducted to confirm this.
- Data reported as `refused` or `No answer` should be removed of the analysis.

3.3 Walking Speed as function of age.

- Obervations are not independent
- Outcome variable: walking speed

First, let's visualize the data. `ggplot2` function `geom_smooth()` adjusts the data to `walking-speed ~ s(age, bs = "cs")`.

```
data_plot |>
  ggplot(aes(y = WLK_TIME,
             x = AGE_VALS)) +
  geom_point(alpha = 0.2, color = 'gray') +
  geom_smooth(color = "deepskyblue3", fill = "deepskyblue2") +
  xlab("Age") +
  ylab("Walking speed not conditioned by height (m/s)") +
  theme_minimal()
```



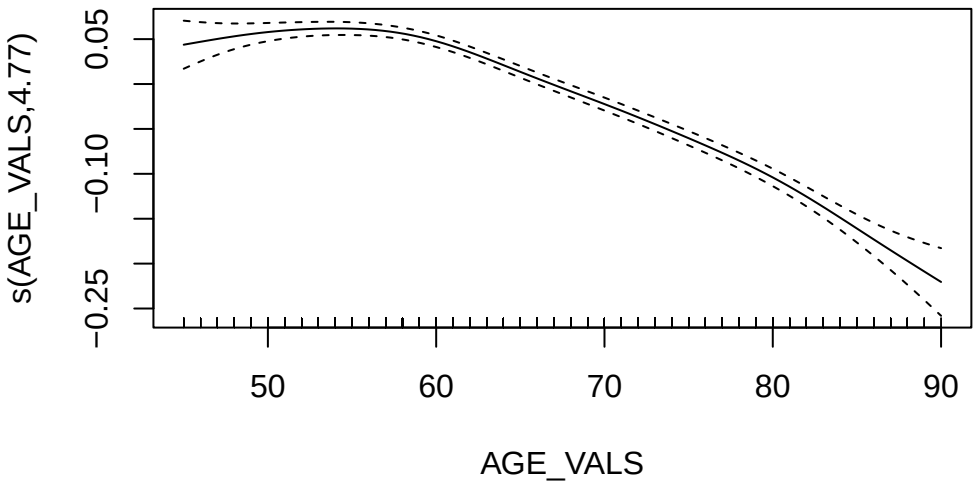
If we fit this basic first model:

```
model <- gam(data = data_plot, WLK_TIME ~ s(AGE_VALS))
summary(model)
```

```
Family: gaussian
Link function: identity

Formula:
WLK_TIME ~ s(AGE_VALS)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.978068   0.001996   490.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df     F p-value
s(AGE_VALS) 4.773  5.843 172.1  <2e-16 ***
---
```

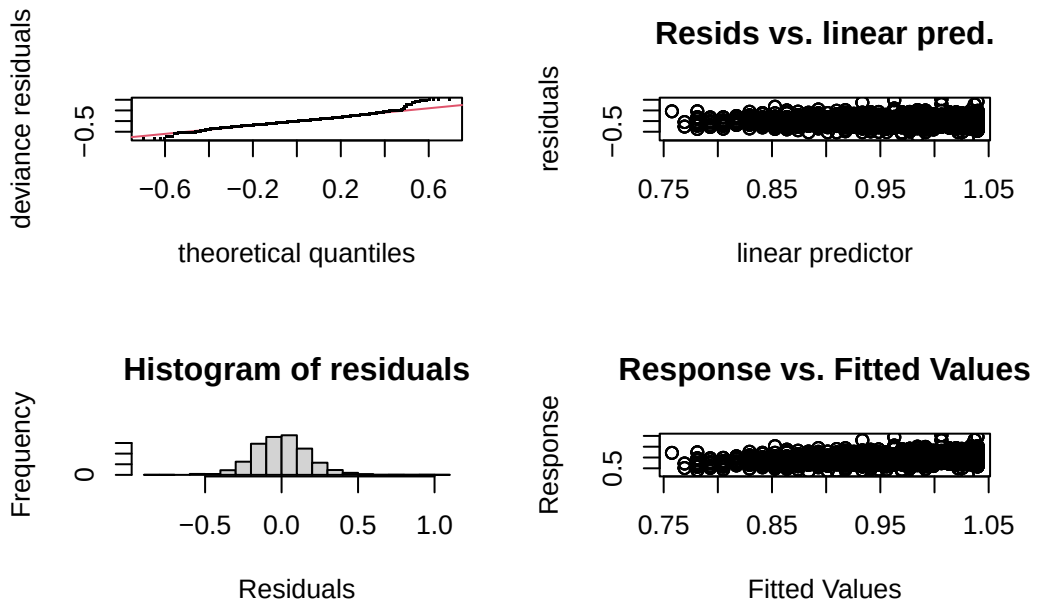
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.109   Deviance explained = 10.9%
GCV = 0.032821  Scale est. = 0.032798  n = 8236
```

```r
plot(model)
```



Despite we don't have `height` values, it is possible to capture a decreasing trend of the walking speed with the age.

```r
gam.check(model)
```



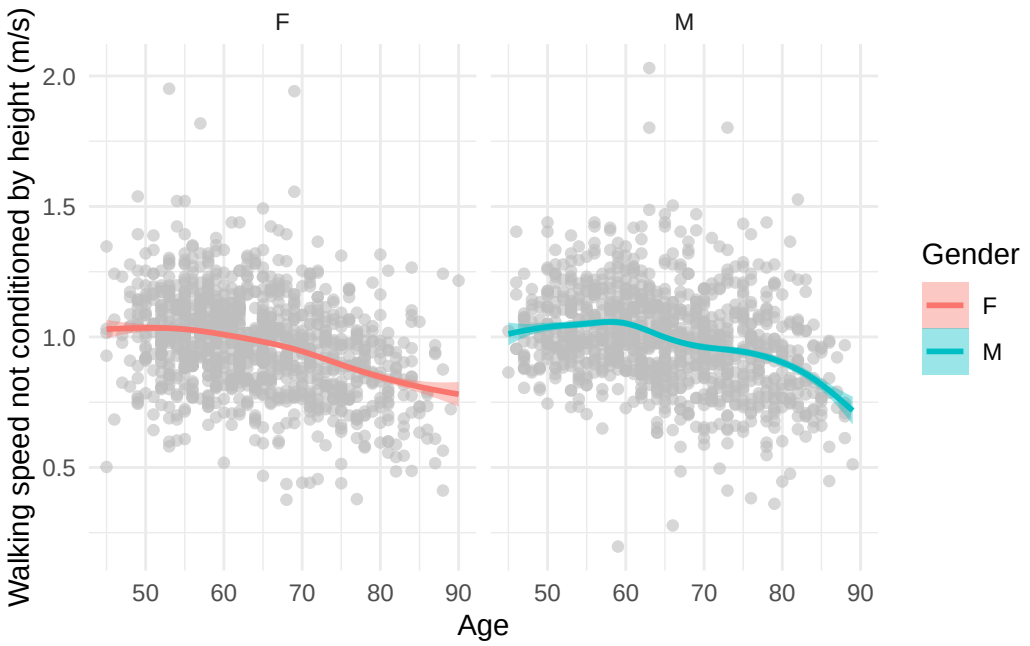

```
Method: GCV   Optimizer: magic
Smoothing parameter selection converged after 4 iterations.
The RMS GCV score gradient at convergence was 4.021007e-07 .
The Hessian was positive definite.
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

              k'  edf k-index p-value
s(AGE_VALS) 9.00 4.77       1    0.53
```

To be continued...

If we split the data by gender:

```
data_plot |>
  dplyr::left_join(gender) |>
  filter(!WLK_TIME < 0) |> # outlier: walking speed around -70 removed
  ggplot(aes(y = WLK_TIME,
             x = AGE_VALS)) +
  geom_point(alpha = 0.2, color = 'gray') +
```

```
  geom_smooth(aes(color = SEX_ASK_COM, fill = SEX_ASK_COM)) +
  facet_wrap(vars(SEX_ASK_COM)) +
  xlab("Age") +
  ylab("Walking speed not conditioned by height (m/s)") +
  labs(fill = "Gender", color = "Gender") +
  theme_minimal()
```

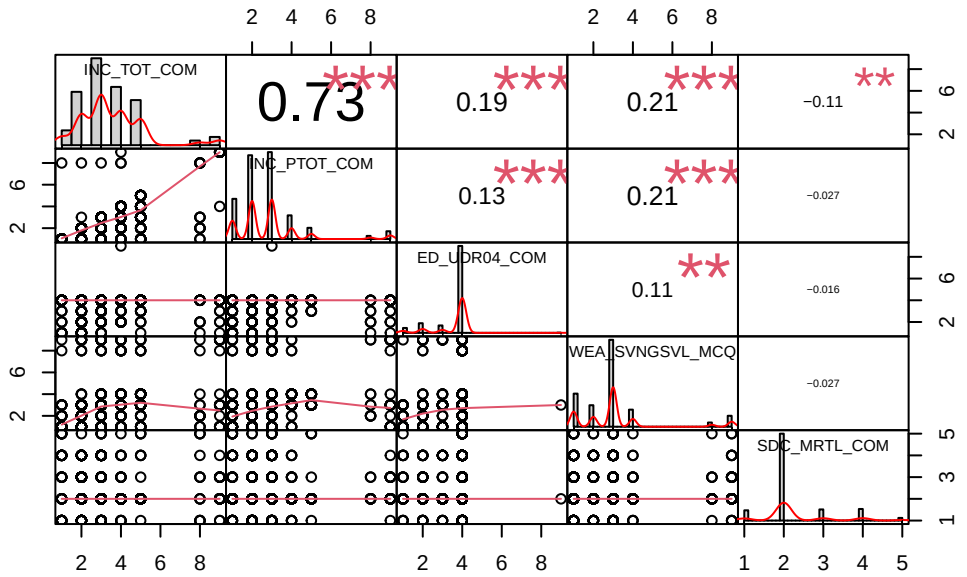


Next steps:

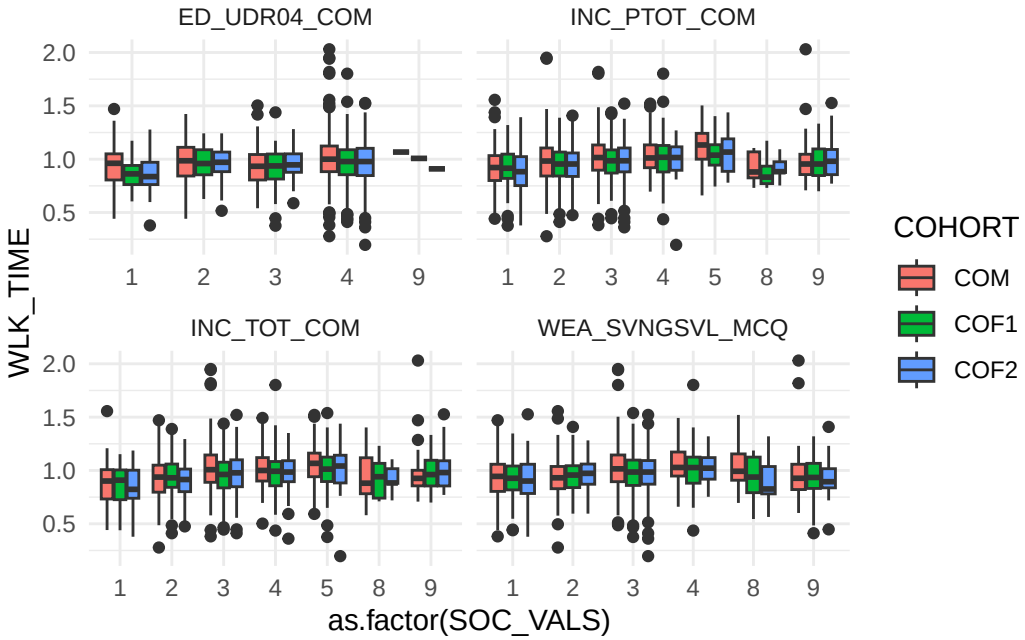
- Add `age`, `height`, and the socioeconomic variables as fixed effects using a generalized additive mixed model (GAMM).
- This could be performed for one socioeconomic variable at a time or to all the variables at the same time.
- In Stringhini et al. (2018), `study` is used as a random effect at the intercept and age slope, but in that article, multiple studies from different countries were used.
  In this case, we could try using `id` as a random effect.

3.4 Removed

```
socvars <- data |>
  select(starts_with("INC"), ED_UDR04_COM, WEA_SVNGSVL_MCQ, SDC_MRTL_COM)

chart.Correlation(socvars, histogram=TRUE, pch=19)
```




```
data_plot |>
ggplot( aes(x = as.factor(SOC_VALS), y = WLK_TIME)) +
  geom_boxplot(aes(fill = fct_relevel(COHORT, "COM", "COF1", "COF2"))) +
  facet_wrap(vars(SOC_VARS), scales = "free_x" ) +
  labs(fill = "COHORT") +
  theme_minimal()
```




```
# library(gamm4)
#
# model1 <- gamm(WLK_TIME ~ s(AGE_VALS) + INC_TOT_COM,
#                data = data_plot)
#
# model2 <- gamm(WLK_TIME ~ s(AGE_VALS) + INC_TOT_COM +
#                ED_UDR04_COM + WEA_SVNGSVL_MCQ,
#                random = ~ (1 | id),
#                data = data_plot)
#
# model3 <- gamm(WLK_TIME ~ s(AGE_VALS, k = 4) +
#                 s(INC_TOT_COM, k=4) +
#                 s(ED_UDR04_COM, k=4) +
#                 s(WEA_SVNGSVL_MCQ, k=4),
#                 random = ~ (1 | id)
#                 data = data_soc)
#
#
#
# summary(model1)
# summary(model2)
```

```
# summary(model3)
```

References

Everitt, BS. 1998. "Analysis of Longitudinal Data: Beyond MANOVA." *The British Journal of Psychiatry* 172 (1): 7–10.

Stringhini, Silvia, Cristian Carmeli, Markus Jokela, Mauricio Avendaño, Cathal McCrory, Angelo d'Errico, Murielle Bochud, et al. 2018. "Socioeconomic Status, Non-Communicable Disease Risk Factors, and Walking Speed in Older Adults: Multi-Cohort Population Based Study." *Bmj* 360.

Twisk, Jos W. R. 2013. *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. 2nd ed. Cambridge University Press.