The consequences of not completing the generational cohort in estimating age-at-menopause

Rui Martins[1,2], Bruno de Sousa[3], Thomas Kneib[4], Maike Hohberg[5], Nadja Klein[6], Elisa Duarte[3], Vítor Rodrigues[7]

[1] Faculdade de Ciências da Universidade de Lisboa (FCUL), Portugal
[2] Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)
[3] Center for Research in Neuropsychology and Cognitive and Behavioral Intervention (CINEICC), University of Coimbra, Portugal
[4] University of Göettingen, Chair of Statistics, Humboldtallee 3, 37073 Goettingen, Germany
[5] Department of Medical Statistics, University Medical Center Göettingen, 37073 Göttingen, Germany
[6] Chair of Uncertainty Quantification and Statistical Learning, Research Center for Trustworthy Data Science and Security; Dep. Stat. Technische Universität Dortmund, Germany
[7] Faculty of Medicine, University of Coimbra, Portugal

E-mail for correspondence: rmmartins@fc.ul.pt

Abstract: When studying age-at-menopause of a particular generation cohort of women the approach where women without an observed menopause are deleted from the study is not advisable because they might convey different informations for the analysis namely about the so called period effect. Generally, the deleted are the youngest who have not yet reached menopause.

The context is a Portuguese breast cancer screening programme in the period 1990–2010 where a late menopause is considered a risk factor. Our aim is to recover missing menopause ages by comparing methods for handling missing (or incomplete) data.

Two imputation approaches are considered: (i) multiple imputation based on a truncated distribution but ignoring the mechanism of missingness; (ii) a bivariate copula-based imputation that simultaneously handles the age-at-menopause and the missing mechanism.

There are contradictory results in current research about whether age-at-menopause is increasing or decreasing in Western countries. We show that both imputation methods unveiled an increasing trend of age at menopause when

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

viewed as a function of the birth year for the youngest generation. This trend is hidden if we model only women with an observed age-at-menopause.

Keywords: menopause; incomplete data; copula regression; GJRM; gamlss.

1 Introduction

Age-at-menopause has an important role in the research about risk factors for breast cancer. However, it is a variable prone to incompleteness, because the time when women participate in a breast cancer screening program overlaps the time when women are most likely to enter menopause. Therefore, the younger women of the generation cohort under analysis tend to have missing information on age-at-menopause. Not recovering those values can lead to wrong conclusions because the parameters estimates for the most recent years will tend to be dominated by these young women.

The question of whether missing values of a variable are related to the underlying value itself allows for classifying the missing data mechanism into three categories (Rubin 1976): missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

We frame the issue of imputing age-at-menopause as a missing data problem since we consider this measure as a covariate in a potential subsequent risk cancer analysis. We therefore ask the same question as in a classical missing value setting: Is the missing mechanism informative or not? Note that recovering the values for age-at-menopause as the dependent variable could also be treated as a censoring or prediction problem but is not the focus of this work.

To test how different strategies to impute missing ages-at-menopause for the youngest women influence the analysis of time- and spatial-trends of that variable, we will analyse the case of a breast cancer screening program in central Portugal. Exploratory analyses show the presence of a geographical pattern of the missing data and a close relation with a woman's year of birth (a.k.a. period effect), implying, at least, a violation of the MAR assumption. Additionally, there is a high percentage of missing values in the variable of interest (23.6%), which precludes an analysis by simply deleting those individuals.

To achieve the goals defined above, we will consider two statistical modelling approaches with the aid of two `R` packages, namely `GJRM` – Generalised Joint Regression Modelling (Marra and Radice, 2017) and `gamlss` – Generalised Additive Models for Location, Scale and Shape (Rigby and Stasinopoulos, 2005). The `GJRM` package allows us to deal simultaneously with two response variables while their specific marginal distributions are conveniently expressed in a joint manner by means of a copula function that binds them together. In this way, we will be able to define a joint distribution for both the process that governs the probability that a woman has not yet reached menopause and for the age-at-menopause itself. The

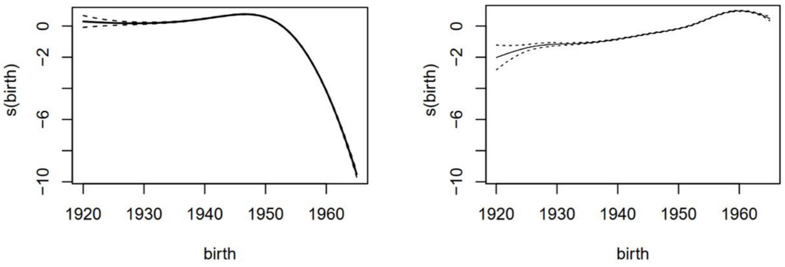

FIGURE 1. Birth year (flexible) effect if only considering women with an observed menopause (left panel). Birth year (flexible) effect after the missing menopause ages have been replaced with the imputations via a truncated Weibull distribution at the screening age (right panel).

gamlss package adopts a method for the imputations which is very flexible and allows imputations from truncated distributions.

2 Data

The dataset that we are working with has of 278 282 women between 1990 and 2010. At the age of 45, all women in each of the 78 municipalities in central Portugal are invited to have a free screening mammogram and every two years thereafter until the age of 69. This region roughly represents 25% of the Portuguese population. At the time of the last screening, 65 765 women (23.6%) stated they had not yet reached menopause (missing information). The variables included in the dataset are: (i) binary characteristics provided by the variables `pregnancy`, `breastfeeding` and the use of `oral contraceptives`; (ii) quantitative information carried by the continuous variables are `age at menopause`, `age at menarche`, `year of birth` and `age at the last screening`; (iii) demographic information given by the `municipality purchasing power index`; and (iv) spatial information embodied in neighbourhood structure of the `municipality of residence`.

2.1 Methodology

The primary goal of this work is to draw inferences about the distribution of the age-at-menopause, Y_i, $i = 1, 2, \dots, n$ given a set of observed covariates, $\mathbf{v}_i$, by considering the primary analysis model $[Y_i \mid \mathbf{v}_i]$. The most popular approach would be to estimate its parameters using only the observed Y_i's, yet estimates from such an analysis would be less efficient than they would be if we had observed Y for every individual. Recovering information via an imputation technique, e.g. multiple imputation (MI), should allow to retrieve some of the information about Y that is not available.

We discuss two different approaches for dealing with missing menopause ages. One considers the data as MNAR and therefore we jointly model the missing data mechanism and the response variable of interest via a bivariate copula. The other considers an MAR data structure and thus only the statistical process of the age-at-menopause is modelled.
The imputations will be obtained by sampling from an approximation to the posterior predictive distribution of the missing data given modelling assumptions and the observed data,

$$f(\mathcal{Y}_{\text{mis}} \mid \mathcal{Y}_{\text{obs}}, \mathbf{v}_i) \approx \int f(\mathcal{Y}_{\text{mis}} \mid \boldsymbol{\Phi}, \mathbf{v}_i)\, \tilde{f}(\boldsymbol{\Phi} \mid \mathcal{Y}_{\text{obs}}, \mathbf{v}_i)\, \mathrm{d}\boldsymbol{\Phi}, \qquad (1)$$

where $\mathcal{Y}_{\text{obs}}$ represents the observed menopause ages and $\mathcal{Y}_{\text{mis}}$ the unobserved ones; $\tilde{f}(\boldsymbol{\Phi} \mid \mathcal{Y}_{\text{obs}}, \mathbf{v}_i)$ is the approximated posterior distribution of all the parameters combined in the vector $\boldsymbol{\Phi}$.

3 Results

An imputation procedure for the missing ages-at-menopause is required if the study aims at analysing the trend of a variable in a setting that includes a cohort of women where the majority has already reached menopause and only a small part has not yet. This is always the case when we have a cohort whose age range includes the more likely age to reach the menopause. In settings, where either all women have already reached menopause, or neither woman is in menopause yet, there is no need to resort to any imputation procedure. From a statistical point of view, the first situation only requires the specification of an analysis model. The second situation cannot be inferred because we do not have information to predict individual menopause, unless we assume that they have the same characteristics as the older cohorts but then we would not be able to study the temporal trends across cohorts.
With a dataset similar to the one that we worked with, not imputing the missing ages-at-menopause means that we will have to wait for all women belonging to the youngest cohorts to reach menopause in order to be able to assess the temporal trends of the menopause for that specific cohort of women. When fulfilling a dataset with imputations made in a proper way, we can model the temporal trends of the age-at-menopause immediately. This means that, in terms of public health, we will be studying the phenomenon of menopause without delays. The naive approach of simply delete the women without an observed menopause leads to biased results (Figure 1 - Left panel).
Finally, we would like to emphasize that age-at-menopause is increasing in the central region of Portugal as a function of the birth year (Figure 1 - Right panel).

Acknowledgments: This work was partially funded by Fundação para a Ciência e a Tecnologia (FCT) through the projects INIC-DAAD – DAAD 441.00, UIDB/ 00006/ 2020 and POCI/ 01/ 0145/ FEDER/ 029443 – SHSADReM – Addressing Social and Health Challenges through new developments in Structured Additive Distributional Regression Models. Nadja Klein acknowledges support through the Emmy Noether grant KL 3037/1-1 of the German research foundation (DFG).

References

Marra G, Radice R. (2017) Bivariate copula additive models for location, scale and shape. *Comput Stat Data An*, **112**, 99 – 113.

Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape (with discussion) *J R Stat Soc Ser C Appl Stat*, **54**(3), 507 – 554.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**(3):581 – 92