Stochastic Process Model and its Applications to Imputation of Censored Longitudinal Data

Ilya Y. Zhbannikov, Konstantin G. Arbeev, Anatoliy I. Yashin

Abstract

Longitudinal data are widely used in medicine, demography, sociology and other areas related to population studies. Incomplete observations in such data often confound the results of analysis. A plethora of data imputation methods have already been proposed to alleviate this problem. The Stochastic Process Model (SPM) represents a general framework for modeling joint evolution of repeatedly measured variables and time-to-event outcome typically observed in longitudinal studies of aging, health and longevity. It is perfectly suitable for imputing missing observations in longitudinal data. We applied SPM to the problem of imputation of missing longitudinal data. This model was applied both to the Framingham Heart Study and Cardiovascular Health Study data as well as to simulated datasets. Comparing to the other best available tools, we show that our proposed methodology in many cases outperforms the current best available solutions both on simulated and real-world censored longitudinal data. R package *stpm* is available under the following link: https://cran.r-project.org/package=stpm

Introduction

Missing values are considered to be one of the major problems in analysis of longitudinal data. The presence of missing values or not handling them properly may lead to inaccurate inference about the data. There are various sources of incomplete observations: accidental data loss or data entry errors, study design (e.g., due to budget constraints some variables can be measured for a subsample of study participants) and specific situations when non-response occurs. For example, a participant may not respond because some questions are sensitive or due to stress, fatigue or lack of knowledge. These absent answers would be considered as missing values. Missing data also appear due to attrition (subject's death, relocation, or drop-out because of different other reasons).

Currently, a set of different data imputation methodologies has been proposed in order to alleviate this problem. These methodologies are implemented in statistical computing platforms such as R and SAS. Methodologies such as MICE (Multivariate Imputation through Chained Equations) [1], EM (Expectation-Maximization) [2] and EMB (Expectation-Maximization with Bootstrapping) [3], and Random Forest [4] are widely accepted and commonly used for data imputation. The method MICE utilizes an approach known as "Chained equations". This is an iterative approach where the values on the next iteration depend on values obtained at the previous iteration. MICE assumes that the missing data satisfies the Missing at Random (MAR), meaning that the probability that an observation is missing depends exclusively on observed values and does not depend on the unobserved ones. Missing values then can be predicted using regression (e.g., linear or logistic depending on a variable type). An R package *mice* is one of the widely used and well-established R packages for multiple imputation implementing the MICE approach. The algorithm establishes several imputed datasets in parallel and later uses statistical inference [12] to come with final imputed dataset. This operation is also known as ‘pooling’. EMB (Expectation-Maximization with Bootstrapping) is another method used for data imputation. This approach assumes that all variables (covariates) in data satisfy the multivariate normal distribution (MVN) and also satisfy the MAR assumption. The R package *Amelia* implements this methodology and together with multiple imputation scenario produces acceptable results even for a high rate of missing data. Bayesian regression approach to data imputation is realized in the R package `mi` (Multiple Imputation with diagnostics) [5]. The methodology implemented in `mi` makes two assumptions: (i) Bayesian version of regression models in order to handle the issue of separation which arises in binary data when the maximum likelihood estimation method fails to find a complete separation of data points; (ii) addition of the background noise to the imputation process in order to avoid the problem of additive constraints. Using a random forest (RF) for data imputation produces results comparable to other methods and it is implemented in R package *missForest* [7]. Its general idea is to employ the RF method to predict missing values. Thus, for each variable the method fits a random forest on the observed data and then predicts the missing part using fitted model from the previous step. The algorithm continues repeating these two steps until a pre-defined stopping criterion is satisfied or maximum number of iteration is passed. The general problem with random forest is slow convergence therefore its runtime highly depends on a size of the dataset. Sets of different imputation methods are implemented in the R package *Hmisc* [6]. This package allows also combinations as different imputation scenarios for each variable, from using user defined statistical method (mean, median, random) to imputation using regression models, bootstrapping, and predictive mean matching.

These approaches and implemented programming solutions based on them imply the MAR assumption (except some special cases in the MICE method), however the reality can be more complicated. For example, because of the attrition due to mortality, where mortality may depend on the trajectories of the variables being imputed, one can face the “missing not at random” (MNAR) situation. Another deficiency of general imputation methods is their inability to cope with the structure of the data, and a longitudinal dataset with consecutive observations represents a good example of such specific situations. To cope with these limitations, joint-models can be applied. Joint models [20, 21] are specific type of statistical models which contain a longitudinal part, which usually can be in a form of a linear mixed effect model (LME) and a survival part, represented by the Cox proportional hazards model. Joint models can be used in situations where MNAR patter is presented. Joint models offer clear and simple framework and interpretations, still however have some limitations, for example, when the hazards model, commonly presented as Cox proportional hazards, may not describe the reality in some situations. Since we were not able to find any software employing JMs in data imputation, we had to implement our own, using an R package *JM* [20].

We propose a methodology of multivariate data imputation with the Stochastic Process Model (SPM) wrapped to the newly developed R package *stpm* [13], which itself implements the general SPM methodology. The SPM was developed several decades ago and since that time different versions have been applied for analyses of clinical, demographic, epidemiologic data as well as in many other studies that relate stochastic dynamics of repeated measures to the probabilities of end-points. The idea of SPM was first described in [8]. Later it was extended in several publications [8-10]. SPM links the dynamic of stochastic variables represented by multivariable autoregressive or stochastic differential equations to hazard rates presented as quadratic functions of the state variables (covariates). The minimum of a quadratic function is a point (or domain) in the variable state space, which corresponds to the optimal system status (e.g., the “normal” health status) at a specific time (e.g. age). The SPM models the stochastic dynamics of covariates assuming that the respective process satisfies a certain stochastic model which better describes the reality in many situations. This is an advantage of this methodology. In this article we show how the SPM can be applied to imputation of censored longitudinal data. Our package is freely available from CRAN under the following link: https://cran.r-project.org/package=stpm (stable release) or directly from its website: https://github.com/izhbannikov/spm (developing version).

Methods

Description of the Stochastic Process Model

In the specification of the SPM described in the 2007 paper by Yashin and colleagues [10] the stochastic differential equation describing the age dynamics of a covariate is:

$$d\boldsymbol{Y}(t) = \boldsymbol{a}(t)\big(\boldsymbol{Y}(t) - \boldsymbol{f}_1(t)\big)dt + \boldsymbol{b}(t)d\boldsymbol{W}(t), \boldsymbol{Y}(t = t_0) \qquad (1)$$

In this equation, $\boldsymbol{Y}(t)$ (a k – length vector) is the value of particular covariates at a time (age) t; $\boldsymbol{f}_1(t)$ (a k – length vector) which corresponds to the long-term mean value of the stochastic process $\boldsymbol{Y}(t)$, and $\boldsymbol{Y}(t)$ describes a trajectory of covariates affected by various factors that can be described by a random Wiener process $\boldsymbol{W}(t)$. Coefficient $\boldsymbol{a}(t)$ (a k x k matrix) characterizes the rate at which the stochastic process $\boldsymbol{Y}(t)$ reverts to its mean. In the area of research on biodemography of aging, $\boldsymbol{f}_1(t)$ is the mean allostatic trajectory and $\boldsymbol{a}(t)$ is the adaptive capacity of the individual organism or a technical system. Coefficient $\boldsymbol{b}(t)$ (a k – length vector) characterizes the strength of the random perturbations received from the Wiener process $\boldsymbol{W}(t)$.

The following function $\mu\big(t; \boldsymbol{Y}(t)\big)$ represents a hazards rate:

$$\mu\big(t; \boldsymbol{Y}(t)\big) = \mu_0(t) + \big(\boldsymbol{Y}(t) - \boldsymbol{f}_0(t)\big)^T \times \boldsymbol{Q}(t) \times (\boldsymbol{Y}(t) - \boldsymbol{f}_0(t)) \qquad (2)$$

where μ_0 is the baseline hazard, which represents the risk when $\mathbf{Y}(t)$ follows its optimal trajectory; $\boldsymbol{f}_0(t)$ (a k – length vector) represents the optimal trajectory that minimizes the risk and $\boldsymbol{Q}(t)$ (a k x k matrix) represents a sensitivity of risk function to deviations from the norm. The model can accept any reasonable number of covariates therefore it can be considered for multivariate data analysis. The model coefficients $\boldsymbol{a}$, $\boldsymbol{f}_1$, $\boldsymbol{Q}$, $\boldsymbol{f}_0$, $\boldsymbol{b}$, $\mu_0(t)$ can be estimated by maximum likelihood optimization method from the dataset containing only complete cases (no missing observations). In this work for simplicity we assumed that all model coefficients are time-independent. Having these coefficient estimates, one can predict missing values of $\boldsymbol{Y}$ based on the known available observations. By default the baseline hazard follows the Gompertz model: $\mu_0(t) = ce^{\theta t}$ and parameters c and θ are estimated as well.

Multiple imputations

The package *stpm* (function `spm.impute(…)`) uses multiple imputation strategy [12] that is currently widely used since it produces unbiased results. On the first stage of the imputation the algorithm starts from imputing missing values using equations (1) and (2) thereby fully utilizing the Stochastic Process Model methodology. Following classical multiple imputation theory, the algorithm produces m several independent imputed datasets in parallel. On this stage algorithm uses parameter estimates $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{f}}_1$, $\hat{\boldsymbol{Q}}$, $\hat{\boldsymbol{f}}_0$, $\hat{\boldsymbol{b}}$, $\hat{\mu}_0$ found from complete observations.

The second stage is parameter estimation from these intermediate datasets: $\widehat{\boldsymbol{a}}^{(i)}$, $\widehat{\boldsymbol{f}}_{\boldsymbol{1}}^{(i)}$, $\widehat{\boldsymbol{Q}}^{(i)}$, $\widehat{\boldsymbol{f}}_{\boldsymbol{0}}^{(i)}$, $\widehat{\boldsymbol{b}}^{(i)}$, $\hat{\mu}_0^{(i)}$ (i = 1…m). On the third and final stage, the algorithm performs final parameter estimation from those obtained on the previous stage using statistical inference assuming that parameters follow normal distribution and produces a final imputed dataset.

Experimental setup

In order to evaluate the imputation ability of the Stochastic Process Model and compare it to the best available methodologies and tools we developed a set of tests based on simulation and real-world data. We computed the mean absolute difference (AD) and root mean square error (RMSE) between true and predicted (imputed) values with the following equations:

$$AD = \frac{1}{Nobs}\sum_{i=1}^{Nobs}|x_i - \widehat{x_\imath}| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{Nobs}\sum_{i=1}^{Nobs}(x_i - \widehat{x_\imath})^2} \tag{4}$$

Where x_i and $\widehat{x_\imath}$ are true and predicted values; N_{obs} is a number of observations.

Simulation study

Simulation study was performed according to the following strategy for one, two and five covariates in the Stochastic Process Model:

1. *Nd* = 100 datasets were simulated (see example in Table 1), each dataset contained *N* = 1,000 hypothetical individuals with starting age of 30 (years); for simulation we used a single hypothetical covariate; 2 years between each observations; a single dichotomous outcome variable (0 - alive, 1 – deceased) was used. We simulated these datasets in a way to be close to the Framingham Heart Study data [11].
2. From each of 100 simulated datasets (obtained at the previous step) we then randomly (using a uniform distribution) excluded observations in order to come up with approximately 25% of incomplete cases. In the end, there were 100 datasets containing missing observations to be imputed. These datasets with incomplete cases were used as test datasets.
3. For each of 100 datasets containing missing observations: missing values were imputed using the Stochastic Process Model implemented in the corresponding R package *stpm*.
4. In parallel, for each of 100 datasets containing missing observations: missing values were imputed using tools (methods) *mice, mi, Amelia, missForest, Hmisc* and *JM.*
5. Then the AD and RMSE were computed from the results in steps 3 and 4, along with the statistical characteristics: mean, median and standard deviation computed across the 100 experiments.
6. These steps above were repeated for two and five covariates.

In Table 1 we show an example of simulated data sets. The data has the following structure: the first column: id, represents a subject ID; case – indicator showing whether an event happened or not (0 - event not happened or 1 - event happened); t1 – current observation time; t2 – next observation time; y1, y2, y3 – current values of three covariates; y1.next, y2.next, y3.next – next value of corresponding covariates.

.

Table 1. Example of longitudinal data set with three covariates: y1, y2, y3.

id	case	t1	t2	y1	y1.next	y2	y2.next	y3	y3.next
1	0	30	32.00	100.00	101.84	200.00	205.03	80.00	78.47
1	0	32	34.00	101.84	101.48	205.03	200.52	78.47	79.47
1	0	34	36.00	101.48	100.40	200.52	207.58	79.47	78.58
1	0	36	38.00	100.40	100.07	207.58	207.83	78.58	75.65
…	…	…	…	…	…	…	…	…	…
1	1	76	77.88	96.54	NA	200.82	NA	77.56	NA
2	0	30	32.00	100.00	101.01	200.00	200.75	80.00	81.08
2	0	32	34.00	101.01	94.58	200.75	204.19	81.08	81.71
2	0	34	36.00	94.58	92.14	204.19	206.35	81.71	82.63
2	0	36	38.00	92.14	93.29	206.35	209.13	82.63	81.16
…	…	…	…	…	…	…	…	…	…

Simulation results are presented in Table 2.

Application to the Framingham Heart Study data

We applied the SPM to imputation of longitudinal data from the Framingham Heart Study (FHS). This analysis was performed in order to show the ability of the SPM methodology to deal with the typical real-world problems. The Framingham data on the original cohort contain repeated measurements of physiological and other variables and as well as time-to-event observations of 5,079 (both females and males) individuals. From the FHS dataset, the following covariates were used: DBP (diastolic blood pressure), BMI (body mass index), BG (blood glucose), PP (pulse pressure), HC (hematocrit) and we compared results to other tools (*mice*, *mi*, *Amelia*, *missForest*, *Hmisc, JM*). To mimic missing data, we randomly excluded some observations in order to have 25% of incomplete cases. Results of data imputation for the SPM (R package *stpm*) and other tools are shown in Table 2.

Application to the Cardiovascular Health Study data

We also applied the model to imputation of longitudinal data from the Cardiovascular Heart Study (CHS) data with the same methodology that was applied to FHS data. We used the same five covariates: DBP, BMI, BG, PP, and HC. The total number of individuals in CHS dataset was 5,184 (both females and males).

In this work, we used the FHS and CHS data provided by the National Heart, Lung, and Blood Institute's (NHLBI) Biologic Specimen and Data Repositories Information Coordinating Center (BioLINCC) resource (https://biolincc.nhlbi.nih.gov/home/).

Parameters used in software packages across all experiments

Amelia

```
m=5, parallel = "multicore"
```

mice

```
m=5, maxit = 5
```

mi

All parameters were left at their default

hmisc

```
impute(<covariate_name>, median)
```

missForest

```
maxiter = 1*
```

* Because runtime of the imputation with *missForest* was quite large for our sample size (~1 hour for sample size ~5,000), we decided to use a single iteration. In addition, no significant differences were seen for 2 and three iterations.

JM

```
lme(fixed=log(y1)~years*t1, random= ~ 1 | id, data = x, na.action = na.omit)
coxph(Surv(years, case) ~ 1, data = y, x = TRUE, na.action = na.omit)
jointModel(meFit, survFit, timeVar = timevar, method=”weibull-AFT-GH”)
```

Here `y1` is a covariate; `years` is a total observation time for a particular subject, `id` – subject identification number. We used “weibull-AFT-GH” because the R package *JM* failed on the available proportional hazards methods (which are more appropriate in our case) such as “weibull-PH-aGH”, “weibull-PH-GH”, “piecewise-PH-aGH”, “piecewise-PH-GH”, “Cox-PH-aGH”, “Cox-PH-GH”, “spline-PH-aGH”, “spline-PH-GH”.

All programming scripts used in simulation study are available under the following link: https://drive.google.com/drive/folders/0B042UhX99EetTFBlWFJOMER0UTg?usp=sharing

Results

Simulation

According to Table 2 our approach produces minimal mean absolute difference and minimum sum of squared errors thereby outperforms other approaches across all simulation tests. *missForest* and *Hmisc* produced the best results among other tools (not including *stpm*).

Table 2. Simulation results: statistical characteristics of the difference between observed and predicted values. The *AD* is the absolute mean differences between true x_i and predicted $\widehat{x_i}$ values: $AD = \frac{1}{Nobs}\sum_{i=1}^{Nobs}|x_i - \widehat{x_i}|$. *RMSE* is the root mean square error between true and predicted values, $RMSE = \sqrt{\frac{1}{Nobs}\sum_{i=1}^{Nobs}(x_i - \widehat{x_i})^2}$. N_{obs} is the number of observations in a simulated dataset. In this table *AD* and *RMSE* are the mean AD/*RMSE values* across 100 experiments. Here we also present lower (0.05) and upper (0.95) boundaries of empirical confidence interval (in brackets).

Application (method)	**Number of covariates**					
	1		**2**		**5**	
	AD	**RMSE**	**AD**	**RMSE**	**AD**	**RMSE**
stpm	**1.339*** [1.303, 1.377]	**3.444** [3.341,3.533]	**1.104** [1.075,1.134]	**2.909** [2.826,2.997]	1.259 [0.858,1.808]	3.508 [2.396,5.472]
Amelia	2.644 [2.575,2.711]	7.205 [7.035, 7.398]	2.041 [1.998, 2.093]	5.642 [5.494, 5.797]	1.201 [1.175, 1.228]	3.606 [3.536, 3.683]
mice	2.632 [2.571, 2.697]	7.209 [7.043, 7.391]	2.023 [1.97, 2.076]	5.634 [5.477, 5.779]	1.189 [1.163, 1.217]	3.6 [3.527, 3.676]
mi	2.647 [2.574, 2.714]	7.211 [7.02, 7.4]	2.044 [1.993, 2.095]	5.652 [5.503, 5.79]	1.204 [1.174, 1.228]	3.611 [3.527, 3.683]
hmisc	2.547 [2.473, 2.61]	6.606 [6.441, 6.748]	1.939 [1.886, 1.988]	5.126 [4.973, 5.257]	1.057 [1.034, 1.081]	3.167 [3.101, 3.226]
missForest	1.759 [1.712, 1.806]	4.943 [4.794, 5.071]	1.366 [1.328, 1.403]	3.971 [3.845, 4.082]	**0.793** [0.77, 0.814]	**2.532** [2.464, 2.608]
JM	2.612 [2.502, 2.723]	6.711 [6.476, 6.941]	2.075 [1.872, 2.017]	5.477 [4.988, 5.381]	1.058 [0.83, 1.109]	3.152 [2.849, 3.266]

*Best values shown in bold.

Table 3. Results of imputation of the Framingham Heart Study data with $N = 5,079$ individuals (original cohort).

Application (package)	**Number of covariates**							
	1 (DBP)		**2 (DBP, BMI)**		**3 (DBP, BMI, BG)**		**5 (DBP, BMI, BG, PP, HC)**	
	AD	**RMSE**	**AD**	**RMSE**	**AD**	**RMSE**	**AD**	**RMSE**
stpm	**0.985****	**2.738**	**0.684**	**2.098**	**1.122**	**3.232**	2.102	6.482
Amelia	2.037	5.763	1.432	4.493	1.898	5.816	2.223	8.051
*mice**	2.187	5.689	1.538	4.325	2.163	6.048	2.266	7.985
mi	2.034	5.767	1.440	4.541	1.901	5.820	2.224	8.010
hmisc	2.185	5.709	1.543	4.361	2.139	6.116	2.202	8.116
missForest	1.379	3.963	0.999	3.163	1.325	4.164	**1.409**	**5.814**
JM	2.159	5.439	1.504	4.175	1.901	5.318	2.045	7.501

* "*mice.pmm*" method failed due to a software error, "*mice.mean*" method was used instead

** Best results shown in bold

Table 4. Results of imputation of the Cardiovascular Health Study data with $N = 5,184$ individuals.

Application (package)	**Number of covariates**							
	1 (DBP)		**2 (DBP, BMI)**		**3 (DBP, BMI, BG)**		**5 (DBP, BMI, BG, PP, HC)**	
	AD	**RMSE**	**AD**	**RMSE**	**AD**	**RMSE**	**AD**	**RMSE**
stpm	**0.770****	**2.344**	**0.857**	**2.430**	**0.932**	**4.160**	**1.406**	**4.591**
Amelia	1.672	5.053	1.580	4.686	3.097	11.436	2.793	10.367
*mice**	1.644	5.014	3.036	6.685	2.761	11.151	2.516	10.263
mi	1.673	4.993	1.564	4.644	3.133	11.519	2.804	10.442
hmisc	2.053	5.232	1.666	4.322	2.694	10.952	2.605	9.879
missForest	1.179	3.565	1.115	3.381	1.948	8.079	1.858	7.626
JM	2.349	5.612	1.673	4.359	2.830	11.338	1.449	5.064

* *mice.pmm* method was used

** Best results shown in bold

Discussion

The Stochastic Process Model describes dynamic mechanisms of changes of physiological variables (covariates) with time (age) and, thereby, allows researchers utilizing the full potential of longitudinal data. It also allows the study of differences, for example, in genotype-specific hazards. In application to living organisms, e.g. humans, the presence of so-called hidden components (that cannot be measured directly with common statistical methods) such as adaptive capacity, allostatic load, resistance to stresses, physiological norm plays an important role in aging-related processes. Stochastic Process Model can uncover influences of such hidden components providing researchers with a new way of analyzing longitudinal data.

Our package *stpm* allows for multiple imputations of censored longitudinal data. However it should be used with cautions using appropriate assumptions on the nature of the data. In opposite to general-purpose multiple imputation tools such as *mice*, *Amelia*, *mi*, *missForest*, *Hmisc*, *JM*. the imputation methodology inside the R package *stpm* is specifically designed for dealing with censored longitudinal data with relationships between the outcome and covariates. Therefore applying SPM to imputation of longitudinal and time-to-event data can also be considered as a special case of imputation. We showed that this is working and plausible approach for such kind of data. In spite that SPM and its software implementation, *stpm*, outperforms other methods described above for smaller number of covariates, it still performs average in cases of more than three covariates. Average results in comparison to other methods (e.g. Miss Forest) for higher number of covariates (more than 3) may suggest that for higher number of covariates the matrix Q (see Methods section) becomes not positive-definite. Suggestions are to use lower number of covariates or impute missing values independently for each covariate.

Stochastic Process Model assumes the specific form of the hazard of risk function which should be taken into account while analyzing of longitudinal data. By default we assume that the hazard rate function, which is related to changing physiological variable with age, is a *U*- or *J*-shaped curve. This assumption is biologically justified with empirical observations [14-19]. In some cases the real (true) shape of hazard rate function is unclear and, since it is impossible to estimate the true form from the given data, an incorrectly assumed hazard may introduce additional bias. To alleviate this issue, additional sensitivity analyses may be needed in order to estimate the impacts of various hazard forms on sustainability of results.

Conclusion

We presented a methodology of imputation of censored longitudinal data based on the Stochastic Process Model. For appropriate specified parameters of the model, this methodology is more precise than other data imputation methods considered in this article. The package *stpm* and corresponding R-scripts used in this article are freely available from the following link: https://github.com/izhbannikov/spm and https://CRAN.R-project.org/package=stpm.

Acknowledgements

This work was supported by the National Institute on Aging of the National Institutes of Health (NIA/NIH) under Award Numbers P01AG043352, R01AG046860, and P30AG034424. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIA/NIH. This manuscript was prepared using FRAMCOHORT and CHS Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the CHS, FRAMCOHORT or the NHLBI. The Cardiovascular Health Study (CHS) was conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the CHS Investigators. This manuscript was prepared using a limited access data set obtained by the NHLBI and does not necessarily reflect the opinions or views of the CHS Study or the NHLBI.

References

1. Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.
2. A.P. Dempster, N.M. Laird, and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39, 1-38.
3. James Honaker, Gary King, Matthew Blackwell (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47.
4. Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.
5. Gelman A and Hill J (2011). Opening Windows to the Black Box. Journal of Statistical Software, 40.
6. https://cran.r-project.org/web/packages/Hmisc/index.html
7. Stekhoven D. J., & Buehlmann, P. (2012). MissForest - non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
8. Woodbury, M.A., Manton, K.G.: A random-walk model of human mortality and aging. Theoretical Population Biology 11(1), 37{48} (1977).
9. Yashin, A.I., Manton, K.G., Vaupel, J.W.: Mortality and aging in a heterogeneous population: A stochastic process model with observed and unobserved variables. Theoretical Population Biology 27(2), 154{175} (1985).
10. Yashin, A.I., Arbeev, K.G., Akushevich, I., Kulminski, A., Akushevich, L., Ukraintseva, S.V.: Stochastic model for analysis of longitudinal data on aging and mortality. Mathematical Biosciences 208(2), 538{551} (2007).
11. S. S. Mahmood, L. Daniel, R. S. Vasan, and T. J.Wang. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. The Lancet, 383(9921): 999 – 1008, (2013).
12. Donald B. Rubin (1978). Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics).
13. Ilya Y. Zhbannikov, Konstantin G. Arbeev, Igor Akushevich, , Eric Stallard, Anatoliy I. Yashin, stpm: an R package for stochastic process model (2017) BMC Bioinformatics, 18:125, DOI: 10.1186/s12859-017-1538-7.
14. Witteman JCM, Grobbee DE, Valkenburg HA, Stijnen T, Burger H, Hofman A, et al. J-shaped relation between change in diastolic blood pressure and progression of aortic atherosclerosis. The Lancet. 1994;343(8896):504 – 507. Originally published as Volume 1, Issue 8896.
15. Allison DB, Faith MS, Heo M, Kotler DP. Hypothesis Concerning the U-shaped Relation between Body Mass Index and Mortality. American Journal of Epidemiology. 1997;146(4):339–349.
16. Boutitie F, Gueyffier F, Pocock S, Fagard R, Boissel JP. J-Shaped Relationship between Blood Pressure and Mortality in Hypertensive Patients: New Insights from a Meta-Analysis of Individual-Patient Data. Annals of Internal Medicine. 2002;136(6):438–448.

17. Kuzuya M, Enoki H, Iwata M, Hasegawa J, Hirakawa Y. J-shaped relationship between resting pulse rate and all-cause mortality in community-dwelling older people with disabilities. Journal of the American Geriatrics Society. 2008;56(2):367–368.
18. Mazza A, Zamboni S, Rizzato E, Pessina AC, Tikhonoff V, Schiavon L, et al. Serum uric acid shows a J-shaped trend with coronary mortality in non-insulin-dependent diabetic elderly people. The CArdiovascular STudy in the ELderly (CASTEL). Acta Diabetologica. 2007;44(3):99–105.
19. Okumiya K, Matsubayashi K, Wada T, Fujisawa M, Osaki Y, Doi Y, et al. A U-Shaped Association Between Home Systolic Blood Pressure and Four-Year Mortality in Community-Dwelling Older Men. Journal of the American Geriatrics Society. 1999;47(12):1415–1421.
20. Dimitris Rizopoulos (2010). JM: An R Package for the Joint Modeling of Longitudinal and Time-to-Event Data. Journal of Statistical Software, 35(9), 1-33.
21. Dimitris Rizopoulos (2016). The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. Journal of Statistical Software, 72(7), 1-45, doi:10.18637/jss.v072.i07