Bulletin of Taras Shevchenko National University of Kyiv. Economics, 2022; 3(220): 41-45
УДК 330.4
JEL classification: C
DOI: https://doi.org/10.17721/1728-2667.2022/220-3/5

A.-N. Fataliieva, PhD Student
ORCID ID: 0000-0001-5541-8509,
D. Shamaida, PhD Student
ORCID ID: 0000-0001-6080-2425
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

THE APPLICATION OF PATTERN MIXTURE MODELS AND TIPPING POINT ANALYSIS IN SOCIAL RESEARCH

Nowadays, social economics focuses on many critical issues; among them, public health and morbidity are among the highest priorities since they directly impact human capital formation, which is an important component in the development of the economy. Within public health issues, one of the crucial directions is the analysis of the effectiveness of drugs, which is typically performed on micro-level involving patients in hospitals. The data collected usually is not complete, and it causes problems during the analysis as if a significant part of the critical data is missed, that invalidates finding. The multiple imputation method is one of the most common approaches in dealing with this problem.

Both primary and sensitivity analysis were performed involving multiple imputation approaches. While the preliminary analysis was performed assuming that the missing-data values are overlooked at random, the sensitivity analysis was conducted on the two approaches of missing not-at-random algorithm – the pattern mixture models and the tipping point method. In the paper, the methodological aspects of the usage of these methods were highlighted. Also, the practical implementation of these methods was given in the example of imputing the missing values of the laboratory parameter at different time points with subsequent calculation of AUC and testing the hypothesis of drug efficacy using the analysis of the covariance model. The primary analysis showed the effectiveness of the new drug compared to the placebo. A sensitivity analysis proved the results of the primary analysis. The tipping point method showed that if the assumption that the mean value of dropout is more significant than observed values for more than 196, then the result of the primary analysis is questionable.

Keywords: public health; morbidity; human capital; missing data; multiple imputations.

Introduction. The problems of public health and morbidity are important social issues nowadays. Human health is an important factor in human capital formation. It impacts demographic processes, quality of life, and productivity, which directly affect the level of development of the national economy.

Considering the specifics of data collection and processing methods, the methodology of analysis of phenomena and processes in the field of health care requires a separate thorough study.

After all, the collection of clinical and morbidity data is usually carried out at the micro-level by medical staff, where information comes directly from the study subjects (patients) in most cases. Therefore, this process is influenced by human factors, which often leads to incomplete data collection for qualitative analysis.

One of the most effective data imputation methods is the multiple imputation method. It creates several different plausible imputed data sets that allow considering the uncertainty about the missing data. The multiple imputation method offers many approaches depending on data specifics. Two approaches have gained increasing popularity in recent years. It is a pattern mixture model and tipping point analysis.

Research analysis and problem definition. The subject of this paper is the methodological principles and features of the usage of multiple imputation methods in the sensitivity analysis of social research results. The object of this research is to the missing data of the medical indicators results in the framework of the liver incidence study.

The research interest of pattern mixture model and tipping point analysis methods is to show the approaches that confirm the correctness and validity of the results received by classic and well-researched forms of data imputation based on MAR (missing at random). Additionally, this problem has not been deeply studied by Ukrainian scientists, so the research on this issue becomes of great interest for the development of domestic statistical science.

Ukrainian researchers such as N. Kovtun, A.-N Fatalieva, O. Mishchuk, and R. Tkachenko studied the statistical methods of data imputation. Foreign scientists such as Ratitch and O'kelly, Yuan, Smuk, Rubin D. B, and Little have made significant contributions to developing these methods.

N. Kovtun, A.-N. Fatalieva investigated different approaches to implementing automated methods for recovering missed data [1]. O. Mishchuk and R. Tkachenko considered the methods of imputing the missing data in ecological monitoring [2]. Ratitch and O'kelly developed a strategy for implementing a pattern mixture model and tipping point analysis using standard SAS/STAT procedures [3]. Yuan reviewed the concepts of multiple imputations and explained how to apply the pattern-mixture model in the analysis [4]. Smuk compared the results of filling in missing data using different methods of data imputation and confirmed the feasibility of using the tipping point method as part of the sensitivity analysis [5].

Methodology. The research used methods of analysis and synthesis, system and logical analysis and also specific statistical methods such as multiple imputation method, and analysis of covariance (ANCOVA).

One of the areas of research in the field of public health is the analysis of the effectiveness of drugs. Usually, it includes comparing new treatments with control treatments (mostly placebo). Typically, this process involves conducting two types of analysis: primary and sensitivity analysis.

The most common method to handle missing data in the primary analysis is the method of multiple imputations under MAR (missing at random) assumption. MAR assumption means the missing-data values do not contain any additional information given observed data about the missing-data mechanism. Thus, the process that causes missing data can be ignored. However, the MAR assumption may not always be clinically probable. That is why it is good practice to perform the sensitivity analysis that allows evaluating the robustness of the results to the deviations from the MAR assumption. The sensitivity analysis takes into account the

possibility of the data being missing, not at random (MNAR). Missing not at random is defined when a variable's missing values are related to that variable's values.

Typically analysis of the data with multiple imputation methods includes three phases:

1) The missing data are imputed with estimated values, and a complete data set is created. This process of fill-in is repeated *m* times. As a result, there are *m* complete datasets identical to the observed data entries but differ in the imputed values.

2) The parameters of interest from each imputed dataset are analyzed using a statistical method of interest. This includes, for example, the FREQ, MEANS, MIXED, and GENMOD SAS procedures.

3) The parameter estimates (e.g., coefficients and standard errors) obtained from each analyzed data set are then combined for inference [4].

In this paper, the focus is conducted on the two approaches of sensitivity analysis – the pattern mixture models and the tipping point method.

The pattern mixture model (PMM) is a general approach that contains several distinct methods and allows several approaches to their implementation. Still, all of them have a common idea: assume there are 2 components – observed data (Y_{obs}) and drop-outs (Y_{mis}). Let R represents a matrix of indicators of missingness. PMM methods allow to decompose the joint probability of data and missingness as follows:

$$p(X) = p(X)\, p(R,X) = \; p(X)\, p(R,X), p(Y_{mis}|Y_{obs},R,X)$$

In the PMM methods, subjects are supposed to be grouped into cohorts so that subjects in the same cohort share the same pattern of missing data. $p(X)$ can be considered as a probability distribution of various missingness patterns. A complete data analysis model $p(R,X)$ is then estimated within each pattern (cohort).

$p(R,X)$ represents a model for observed data within each pattern, and $p(Y_{mis}|Y_{obs},R,X)$ represents a model for missing data conditioned on observed data within each pattern.

The partial case of this method is PMM with control-based imputation. The idea of the method is that subjects that take new treatment after drop-out follow the path of control (placebo) group subjects with the same future evolution of the disease. Subjects that withdraw from the placebo group are assumed to evolve in the same way as to control subjects that remain in the study [6,7].

Tipping point analysis (TPA) allows to identify and discuss clinical plausibility of assumptions (the "tipping points") under which there is no longer evidence of efficacy – usually performed as sensitivity analysis and doubt the results obtained in the primary analysis.

The assumption in this method is that the true mean value for study completers is greater or smaller (depending on the nature of the parameter) by the delta value (also called shift) than the true mean in dropouts. And the idea is to test different shifts to determine the one that could doubt the results of the primary analysis.

Typically, delta-adjustment is performed only for a new drug, while allowing the control (placebo) drug to follow the MAR assumption.

Implementing TPA for continuous endpoints includes the following steps:

1. Determine the adjustment group (typically, it is a new drug) and adjustment direction.
2. Impute values and generate *m* complete data sets.
3. The *m* complete data sets are analyzed by using standard procedures (ANCOVA, etc.).
4. The results from the *m* complete data sets are combined.
5. Then step 2 should be repeated to generate multiple imputed data sets with a specified shift parameter (with adjustment the experimental or both experimental and control groups).
6. Steps 3 and 4 should be repeated for various shift values.
7. See if the p-value obtained from the analysis is still less than the given level of significance.
8. Find the 'tipping point' value of the parameter when p-value greater than the given level of significance [8].

The implementation of these methods was done in SAS.

Conducting research and results. The statistical study of the drug's effectiveness was conducted involving patients who have liver disease. There was a laboratory parameter (Bilirubin) collected weekly from Week 1 to Week 5. All patients were treated by either placebo or a new treatment. They were categorized by age group (6-11 years, 12-17 years and >=18 Years) and by type of disease (where 1 – mild, 2 – severe). Table 1 represents the structure of collected data where x – observed data, . – missing data.

***Table* 1. Input dataset structure**

Subject ID	Treatment	Age group	Disease type	Week 1	Week 2	Week 3	Week 4	Week 5
1	Placebo	12-17 Years	2	x	X	x	.	.
2	New	>= 18 Years	1	x	X	.	x	x
3	New	6-11 Years	1	x	.	x	x	.
4	New	>= 18 Years	2	x	X	x	x	.
5	Placebo	>= 18 Years	1	x	.	x	x	x
6	Placebo	>= 18 Years	1	x	X	.	x	x
7	Placebo	12-17 Years	1	x	X	.	x	x

Source: compiled by the authors.

The primary efficacy endpoint is the area under the curve (AUC) from Week 1 to Week 5, based on the percent change from baseline (the latest assessment before treatment) in collected laboratory parameters. Multiple imputation (MI) approach under MAR assumption was used to handle the missing data. Then, the AUC-related efficacy endpoint was calculated involving the following steps:

1) A wholly conditional specification (FCS) regression-based multiple imputation model was used in SAS PROC MI procedure. The outcome at that visit is the dependent variable and observed data (other visits, treatment, age

category, type of disease, and baseline value) are independent variables. The missing data were filled in 30 times to generate 30 complete data sets.

2) The corresponding AUC endpoint was calculated based on the imputed data for each patient and for each of the 30 completed analysis datasets.

3) 30 complete data sets were analyzed by using the ANCOVA analyses. Primary efficacy parameter was compared between treatment groups using the treatment group as the main effect, age category, type of disease, and baseline values of lab parameter as covariates.

The results were combined for the inference using the SAS MIANALYZE procedure and presented in Table 2.

***Table* 2. Multiple imputations under MAR assumption**

	New treatment	Placebo
LS Mean (SE)	–744.6 (177.11)	390.4 (262.82)
95 % CI	(–1091.9, –397.2)	(–125.2, 906.0)
LS Mean difference New vs Placebo (SE)	–1134.9 (243.23)	x
95 % CI	(–1611.8, –658.1)	x
P-value	<0.0001	

Source: compiled by the authors.

The least square (LS) means obtained from ANCOVA model calculated for $AUC_{week\ 1\text{-}week\ 5}$ was equal to –744.6 for the new treatment group and 390.4 – for the placebo, and the LS mean difference was -1134.9. P-value <0.0001 allowed rejecting the null hypothesis of no significant difference in for $AUC_{week\ 1\text{-}week\ 5}$ between new and placebo treatment.

The first approach in the sensitivity analysis was the pattern mixture model. Control-based pattern imputation explored the assumption that patients revert to the control group (placebo-treated group) after drop-out. Thus, it was assumed that after discontinuation, the unobserved values in the new treatment group followed the path of observed values in the placebo group. This approach considered a rather pessimistic scenario of imputation, assuming the effect of the study drug stopped and did not remain after drop-out.

After imputing missing values under MNAR assumption, the same subsequent steps, as were done for the primary analysis, were performed for this sensitivity analysis: the $AUC_{week1\text{-}week5}$ was derived for each of the imputed datasets. The ANCOVA analysis was performed separately for each of the 30 completed analysis data sets, and the results were combined into one multiple imputation inference using the SAS MIANALYZE procedure. The following results were obtained and presented in Table 3.

***Table* 3. The pattern-mixture model with control-based pattern imputation**

	New treatment	Placebo
LS Mean (SE)	–590.42 (189.13)	323.85 (270.65)
95 % CI	(–961.35, –219.49)	(–206.75, 854.44)
LS Mean difference New vs Placebo (SE)	–914.26 (258.73)	x
95 % CI	(-1421.44, –407.09)	x
P-value	0.0004	

Source: compiled by the authors.

Table 3 shows that the point estimate of LS means obtained from ANCOVA model calculated for $AUC_{week1\text{-}week5}$ using the pattern-mixture model with control-based pattern imputation were the following: -590.42 for the new treatment group and 323.85 – for placebo, and the LS mean difference was equal to -914.26. P-value that was equal to 0.0004 allowed rejecting the null hypothesis of no significant difference between the new and placebo group for this sensitivity analysis; hence there was a significant difference in AUC between these two treatment groups.

So, following the assumption of this sensitivity analysis (that drop-outs in the new treatment group follow the path of the placebo group), the conclusion is that the results obtained in the primary analysis under the MAR assumption are true, and the drug is effective.

The second approach of the sensitivity analysis was the tipping point analysis. In this analysis there was an assumption that the assessments were missed due to the high value of the evaluated parameter (missing under MNAR). Hence the supposition was that the true mean in completers and dropouts differs by the particular shift value. The true shift value is unknown, so the examination of several possible shift values between the particular range was performed. Taking into account the nature of the parameter and range that it follows, testing was performed with the assumption that the shift is between 100 and 500. It was assumed that missing values for the new treatment group had greater values compared to the observed values, which meant the pessimistic scenario.

For the purpose of conducting the analysis, the modification of the SAS macro, described by Yang Yuan in his paper [4] was applied. The following results were obtained and presented in Table 4.

***Table* 4. P-values for Shift Parameters between 100 and 520 with step=30**

	P-value	Shift value
1	0.0050	100
2	0.0118	130
3	0.0244	160
4	0.0469	190
5	0.0784	220
6	0.1241	250
7	0.1807	280
8	0.2272	310
9	0.2949	340
10	0.3642	370
11	0.4325	400
12	0.4975	430
13	0.5579	460
14	0.6145	490
15	0.6639	520

Source: compiled by the authors.

Table 4 shows that for a two-sided Type I error level of 0.05, the tipping point for the shift parameter is between 190 and 220. Graphically the changing of p-value depending on shift value is presented in Figure 1.

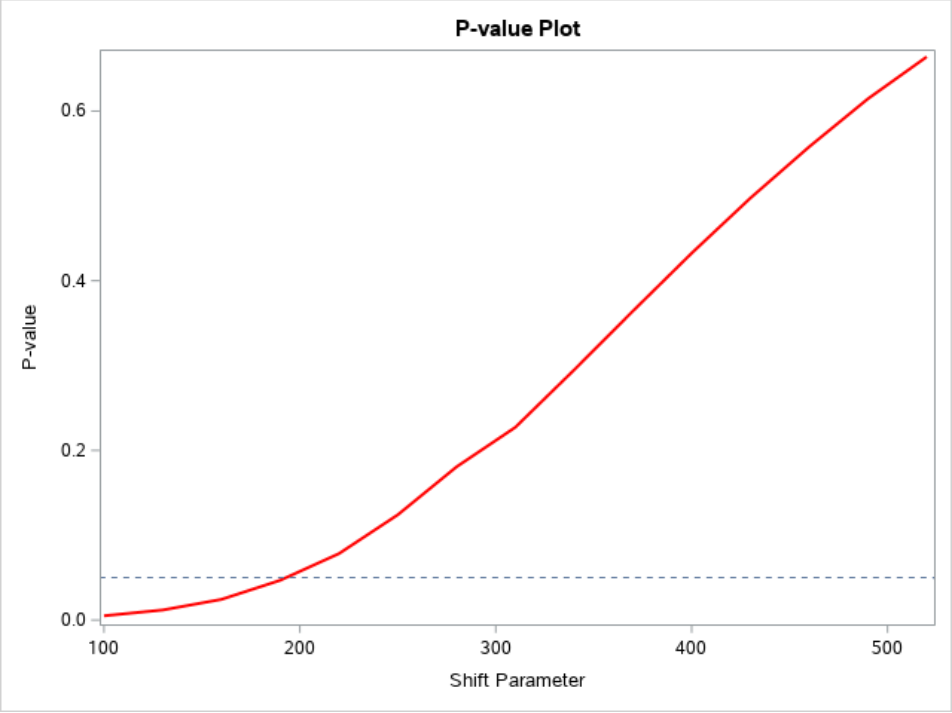

Fig. 1. P-values for Shift Parameters between 100 and 520

Source: compiled by the authors.

The next step was to define the accurate value of shift in the range between 190 and 220 which calls into question the results of the analysis assuming MAR. Table 5 shows the subset of the results.

***Table* 5. P-values for Shift Parameters between 190 and 220 with step=1**

	P-value	Shift value
1	0.0469	190
2	0.0478	191
3	0.0488	192
4	0.0477	193
5	0.0486	194
6	0.0496	195
7	0.0506	196
8	0.0516	197
9	0.0526	198
10	0.0536	199
…		

Source: compiled by the authors

Table 5 shows that the study conclusion under MAR is reversed when the shift parameter is 196. So, following the assumption of this sensitivity analysis, the deduction is that the results obtained in the primary analysis under the MAR assumption could be true if the difference between observed and dropouts is less than 196; otherwise, the conclusions under the MAR assumption are questionable.

Prospects for further studies and conclusions. Solving the problem of missing data allows for conducting high-quality public health research.

Multiple imputation as a common method of dealing with missing data, usually assumes that data is missed at random, but this assumption could not be verified for sure. For the purpose of testing the validity of results obtained under MAR, the sensitivity analysis is performed assuming data is missed not at random (MNAR), and such approaches as the pattern mixture models and the tipping point method could be used for this.

This paper presents a brief overview of these methods and implements them in SAS on the example of imputing the missing values of the laboratory parameter at different time points (week 1-week 5) with subsequent calculation of AUC and testing the hypothesis of drug efficacy using the ANCOVA model.

A pattern-mixture model with control-based pattern imputation assumes that unobserved values in one treatment group (new) follow the path of observed values in another treatment group (placebo); considering the pessimistic scenario, that effect of the drug is leveled out after the drop-out. This sensitivity analysis proved the results of the primary analysis and showed the effectiveness of the new drug compared to the placebo since there was a statistically significant difference in AUC between these two treatment groups.

The tipping point method as another approach used in sensitivity analysis proceeds from the assumption that the true mean of observed data and drop-outs differ by a given value (called shift value). The idea of the analysis is to test cases with different shifts and find the one that questions the validity of the preliminary result. In case if shift value is plausible, taking into account the nature of the parameter that was analyzed, then there is a high probability that results under MAR are not valid. The implementation of this method showed that if the assumption that the mean value of dropout is more significant than for observed values for more than 196, then the result of the primary analysis is questionable.

The approaches highlighted in the article allow covering the problems of missing data for continuous variables, but there are cases that involve the analysis of binary or categorical variables. This problem requires further research and the development of statistical tools.

References

1. *Kovtun N. V., Fataliieva A.-N. Y.* New trends in evidence-based statistics: data imputation problems. *Statistics of Ukraine.* 2019. *№ 87(4).* P. 4–13.
2. *Mishchuk O. S., Tkachenko R. O.* Methods of processing and filling of missing parameters in ecological monitoring data. *Scientific Bulletin of UNFU.* 2019. № *29(6)*. P. 119–122.
3. *Ratitch B., O'Kelly M.*. Implementation of Pattern-Mixture Models Using Standard SAS/STAT Procedures. Proceedings of PharmaSUG. 2011. URL: https://www.pharmasug.org/proceedings/2011/SP/PharmaSUG-2011-SP04.pdf
4. *Yuan Y.* Sensitivity Analysis in Multiple Imputation for Missing Data. *Paper SAS Institute Inc.* 2014.
5. *Smuk M.* Missing Data Methodology: Sensitivity analysis after multiple imputation. *PhD thesis, London School of Hygiene & Tropical Medicine.* 2015.
6. *Little R. J. A.* Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association.* 1993. № 88. P. 125–134.
7. *Ratitch B., O'Kelly M., Tosiello R.* Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics.* 2013. Vol. 12, Is. 6. P. 337–347.
8. Tipping point analysis – multiple imputation for stress test under missing not at random (MNAR). URL: https://onbiostatistics.blogspot.com/2015/08/tipping-point-analysis-multiple.html
9. *Rubin D. B.* Multiple imputation for nonresponse in surveys. New York : John Wiley & Sons, Inc., 1987.
10. *Little R., Yau L.* Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics.* 1996. Vol. 52. P.1324–1333.
11. *Brand J. P. L.* Development, implementation, and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. Ph.D. thesis, Erasmus University, 1999. URL: https://core.ac.uk/download/pdf/18508128.pdf
12. *Berglund P. and Heeringa S.* Multiple imputation of missing data using SAS, Cary. NC : SAS Institute Inc., 2014. URL: https://support.sas.com/content/dam/SAS/support/en/books/multiple-imputation-of-missing-data-using-sas/65370_excerpt.pdf
13. *Kenward M. G.* The handling of missing data in clinical trials. *Clin. Investig. (Lond.).* 2013. № 3. P. 241–250. URL: https://www.openaccessjournals.com/articles/the-handling-of-missing-data-in-clinical-trials.pdf
14. *Molenberghs G., Kenward M. G.* Missing data in clinical studies. New York : John Wiley & Sons, 2007. URL: https://download.e-bookshelf.de/download/0000/5740/97/L-G-0000574097-0002359047.pdf
15. *Molenberghs G.* Incomplete data in clinical studies: analysis, sensitivity, and sensitivity analysis. *Drug information journal.* 2009. № 43(4). P. 409–429.
16. *Carpenter J. R., Kenward M. G.* Multiple Imputation and Its Application. New York : John Wiley & Sons, 2013.
17. *Van Buuren S.* Flexible Imputation of Missing Data. Boca Raton, FL : Chapman & Hall/CRC, 2012.

Received: 05/09/2022
1st Revision: 13/09/2022
Accepted: 04/10/2022

Author's declaration on the sources of funding of research presented in the scientific article or of the preparation of the scientific article: budget of university's scientific project

А.-Н. Фаталієва, асп.,
Д. Шамайда, асп.
Київський національний університет імені Тараса Шевченка, Київ, Україна

ЗАСТОСУВАННЯ МОДЕЛЕЙ, ЩО ВРАХОВУЮТЬ МЕХАНІЗМ ВИБУВАННЯ ТА МЕТОД ПЕРЕЛОМНОЇ ТОЧКИ В СОЦІАЛЬНИХ ДОСЛІДЖЕННЯХ

У наш час соціальна економіка зосереджена на багатьох критичних питаннях, серед яких охорона здоров'я та захворюваність є одними з найпріоритетніших, оскільки вони безпосередньо впливають на формування людського капіталу, який є важливим складником розвитку економіки.

Пропущені дані є актуальною проблемою багатьох соціальних досліджень, що вносить елементи упередженості та робить висновки сумнівними. Існують різні підходи до опрацювання пропущених даних, найпотужнішим серед яких є метод множинної імпутації. У цьому документі представлено огляд двох підходів до аналізу пропущених даних, які стають популярними в останні роки, – моделі, що враховують механізм вибування, і метод переломної точки. Усі розрахунки проводились у пакеті статистичних програм SAS 9.4.

Ключові слова: громадське здоров'я; захворюваність; людський капітал; пропущені дані; множинна імпутація.