NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Svetlana Zhuchkova
Aleksei Rotmistrov

A COMPARISON OF THE MISSING-INDICATOR METHOD AND COMPLETE CASE ANALYSIS IN CASE OF CATEGORICAL DATA

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: SOCIOLOGY
WP BRP 87/SOC/2019

Svetlana Zhuchkova[1], *Aleksei Rotmistrov*[2]

A COMPARISON OF THE MISSING-INDICATOR METHOD AND COMPLETE CASE ANALYSIS IN CASE OF CATEGORICAL DATA[3]

The research aims to provide a complex analysis of missing-indicator method's performance in case of a categorical independent variable in regression in comparison with complete case analysis. While the latter seems to be the most popular way to handle missing data, the former appears to be a simple and effective alternative that allows making a full sample available for analysis. By means of a statistical experiment and simulated data, we examined how these methods perform in conditions that differ in a mechanism of missingness, proportion of missing data, and model specification. The final results show that, overall, both methods produce unbiased estimates of regression coefficients, but crucially biased estimates of their standard errors and additional statistics such as R^2, adjusted R^2, and F-statistic, especially in case of a missing-indicator method. We explain these results by contribution of a missing-indicator variable, coefficient of which always turns out to be significant and far away from zero.

JEL Classification: C18, C35, C51.

Keywords: categorical data, missing data, missing indicator method, regression analysis.

[1] National Research University Higher School of Economics. Faculty of Computer Science. Master Student; E-mail: szhuchkova@hse.ru

[2] National Research University Higher School of Economics. Faculty of Social Science. Associate Professor; E-mail: arotmistrov@hse.ru

[3] The publication was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2018 (grant №18-05-0031) and by the Russian Academic Excellence Project "5-100".

Introduction

Presence of missing data is a standard problem in quantitative social research: it arises in any survey research and becomes even more common in contemporary studies based on so-called big data. Traditionally, there are three ways to handle missing data: complete case analysis (also known as a listwise deletion), imputation of missing values, and a missing-indicator method. The latter corresponds to a situation when all missing data of a variable are coded as a single value and a new dummy variable is created that indicates presence or absence of missing data – a missing-indicator variable. Originally appeared in medical studies devoted to etiology [Miettinen 1985], now this approach is becoming popular in social science as well. The missing-indicator method deserves attention due to several reasons. First, it does not reduce the statistical power of methods in comparison with the complete case analysis. Second, it is much easier to implement and interpret in comparison with imputation of missing values (especially multiple imputation) [Groenwold et al. 2012]. Third, in social research, a missing-indicator variable may reflect a particular substantive reason why participants hide some of their characteristics. For example, in the recent study of freelance online marketplace, the authors use such variable and assume that freelancers may not indicate some of their socio-demographic characteristics (such as sex, age, a country of residence) intentionally [Strebkov et al. 2019] – for example, in order to increase their chances to find a job and not to be discriminated against in a freelance competition.

Every approach to handling missing data aims to make a full sample available for analysis and reduce parameter estimates' bias [van Kuijk et al. 2016]. However, the last property in context of the missing indicator method is still under discussion. Currently, several studies dedicated to this method's properties are published, and the results obtained in these studies are quite contradictory. Hence, the results of some studies show that the missing-indicator method produces the most biased estimates in comparison with the other methods of handling missing data [Donders et al. 2006; van der Heijden et al. 2006; Henry et al. 2013; Jones 1996; Knol et al. 2010], while in other studies, this method does not produce bias at all [Groenwold et al. 2012; White and Thompson 2005]. This difference may be explained by the fact that there is no standardized methodology for comparison of the methods. Besides, all current studies are medical ones, and we highlight two main reasons why the results of these studies should not be directly transferred to social research. Firstly, most[4] of the mentioned studies are based on real data, and, therefore, the results may be influenced by these data's peculiarities. Secondly, the authors of these studies consider mostly continuous variables that are quite rare in social research. Indeed, categorical data are the most common in sociological practice, and implementation of the missing-indicator method to categorical variables has not yet

[4] The only exceptions here are the articles [Choi et al. 2018] and [Donders et al. 2006] where the authors use simulated data but limit their analysis to the cases of continuous variables only.

been thoroughly investigated[5]. Besides, it seems more appropriate to apply the missing-indicator method to categorical data since it is more 'natural' to add a discrete category (referred to missing data) to a discrete variable rather than to a continuous one [Donders et al. 2006].

In our study, we use simulated data and examine how the missing-indicator method performs in case of a categorical independent variable. The conducted statistical experiment differ in three criteria: a mechanism of missingness, proportion of missing data, and model specification. All the results are compared with the situation when the complete case analysis is used. We do not consider imputation of missing values since the results may depend on the chosen method of imputation [Akande et al. 2017] and aggregation of data [Zangieva and Suleymanova 2016]. The final results show that, overall, both methods produce unbiased estimates of regression coefficients, but crucially biased estimates of their standard errors and additional statistics such as R^2, adjusted R^2, and F-statistic, especially in case of the missing-indicator method. We explain these results by contribution of a missing-indicator variable, coefficient of which always turns out to be significant and far away from zero.

Approaches to handle missingness of categorical data

In this chapter, we examine the mentioned alternatives for handling missingness (complete case analysis, imputation of data, and the missing-indicator method) and describe whether they solve the problem of making a full sample available for analysis, how do they handle continuous and categorical variables, and how they are affected by the type of a missingness mechanism. We adopt the three types of a missingness' mechanism defined in [Rubin 1976]: i) data are *missing completely at random* when missingness does not depend on either observed or unobserved data, ii) data are *missing at random* when missingness depends on observed data, and iii) data are *missing not at random* when missingness depends on unobserved data, i.e., on missing values themselves.

Obviously, among the mentioned three alternatives, the complete case analysis neither solves the problem of making a full sample available for analysis nor differs in how it handles variables of different types. There may be a misconception that this approach is not affected by the type of the missingness generating mechanism. Indeed, it is appropriate only under the condition of the missingness completely at random since in this case 'the reduced sample of individuals with only complete data can be regarded as a simple random subsample from the original data' [Ratner 2008: 270]. Nevertheless, an exploration of a missingness mechanism is conducted and mentioned in published studies rarely, and the complete case analysis still seems to be the most widespread way to handle missing data in real research.

[5] The exception is paper [Henry et al. 2013] where the variable of race contains missing values, but in this study, the authors use the real data and do not control any factors that may affect the results of comparison.

Among the methods of data imputation, most were directed to handle continuous variables originally; then, lots of them were adapted to handle missingness of categorical data, in particular, multiple imputation approaches. One of the first proposed approaches for multiple imputation of categorical data was log-linear analysis [Schafer 1997] that allows testing even high-level interactions of categorical data but may be applied only to few variables at once since this method requires to build a multivariate contingency table [Vermunt et al. 2008]. An alternative popular approach is to use a multivariate normal model, which was actually developed for continuous variables, and round the outcome to obtain discrete values [Alisson 2000]. Some studies, however, confirm that this approach leads to a crucial bias of estimates, and the authors conclude that this approach should never be applied, especially to categorical variables [Alisson 2005]. Besides, this approach does not allow to test possible interactions of categorical variables automatically as well as other methods based on linear models [Akande et al. 2017]. Another possible alternative called hot-deck imputation uses the principle of a near neighbor and also results in biased estimates [Schafer and Graham 2002]. Two modern methods of multiple imputation – latent class analysis [Vermunt et al. 2008] or correspondence analysis [Greenacre and Pardo 2006; Hendry et al. 2017; Stavseth et al. 2019] – potentially overcome all the mentioned restrictions and do not produce bias, but their performance has not yet been widely examined under different research circumstances and compared to other methods. In regard of the missingness mechanism, all methods of multiple imputation require missing data to be generated at random [Rubin 1976], and this requirement is easily violated in real research [Ratner 2008: 270].

Under such circumstances, the missing-indicator method appears to be a noteworthy alternative. This method has earned popularity, especially in computer science, where some methods of analysis (primarily decision trees) use this approach by default [Gentle et al. 2012; Rokach and Maimon 2010]. In studies based on big data, in which a proportion of missing values of some variables often exceeds 50% and may even reach 90% or more, the missing-indicator method becomes the only possible solution to handle missingness and remain an available sample for analysis. Besides, in many social studies (including those based on surveys), this method is used as well. While it is impossible to collect and present precise statistics on this issue, here are some recent examples of social studies in which the missing-indicator method is applied: [Chen and Hossler 2017; Gesser-Edelsburg et al. 2018; Rickles et al. 2018; Strebkov et al. 2019; Trevizo and Lopez 2016; Weiss et al. 2017; Zhelyazkova and Ritschard 2018].

Nonetheless, there is still no consensus on how the missing-indicator method performs even in case of continuous data. Most of current studies based on real data reveal that the missing-indicator method results in biased estimates [Donders et al. 2006; van der Heijden et al. 2006; Henry et al. 2013; Jones 1996; Knol et la. 2010], but in such studies, authors are not able to control

all factors associated with obtained results. There is a limited number of simulation studies (for example, [Choi et al. 2018; Zhuchkova and Rotmistrov 2018]), but their authors either do not pay attention to categorical data or do not consider different mechanisms of missingness. Our research aims to fill this gap and provide complex analysis of how the missing-indicator method performs in different analytical situations.

Methodology

The main method of this research is a statistical experiment. In a statistical experiment, it is assumed that all true parameters of a model are known, and a researcher may estimate how these parameters are affected by any changes he or she makes. In relation to our experiment, the idea behind it was the following: to simulate data that correspond to a regression model with specific parameters and examine how the use of the missing-indicator method affects these parameters. Here, we consider only regression modeling among other methods of data analysis since the missing-indicator method is widely used specifically in regression. In contrast to previous studies, we use simulated data in order to avoid a hypothetical influence of real data's peculiarities and, thus, be able to control all the experiment's factors, make the experiment as 'pure' as possible.

In our experiment, we consider three factors, or criteria, that may determine the missing-indicator method's effect: a mechanism of missingness, a proportion of missing data, and a model's specification. Below, we justify the necessity of these criteria and describe how they were implemented in our study.

1. ***The mechanism of missingness (considered in the preceding chapter):*** *completely at random (MCAR), at random (MAR), and not at random (MNAR).* According to the previous studies, the mechanism of missingness is regarded as one of the most important factors related to the missing-indicator method's performance. But current results are rather counterintuitive: one of the last relevant study's authors conclude that the missing-indicator method produces unbiased results in randomized trials, in which missingness is not associated with any studied characteristics, and biased results in nonrandomized trials, in which missingness is likely associated with other variables and external factors [Groenwold et al. 2012]. The former corresponds to MCAR, and the latter corresponds to MAR and MNAR. At the same time, if a missing-indicator variable indicates the presence of some hidden reason why data are not presented, then its missing values should be understood as MNAR. As it was mentioned above, the practice of providing a new variable to indicate the substantive reason for missing data is becoming popular in social science. Thus, when missingness is associated with some hidden reason and regarded as just additional substantive value of a variable, it is expected to produce unbiased results – that is why we find the current findings counterintuitive. This argumentation is partly confirmed by the findings of [Choi et al. 2018].

In our research, MCAR data are those that were chosen randomly by means of a random number generator, MAR data depend on values of another observed variable, and MNAR data are those that depend on the same variable, values of which are replaced with missing values.

2. ***A proportion of missing data:*** *10%, 25%, 50%.* It is obvious that the results of modeling should become worse with an increasing number of missing data, but the question is to what extent this effect occurs with a different proportion of missing values. Thus, in [Zhuchkova and Rotmistrov 2018], a statistical experiment on missing data was conducted with CHAID – a method that treats missing values as a single value by default. It is necessary to point out that only the data missing completely at random were considered in that study. It was revealed that low (10%) and medium (25%) proportions of missing data lead to similar results that do not significantly differ from the benchmark. In contrast, a high proportion (50%) causes misleading and incorrect results. Practically, this criterion may be helpful to get a threshold of a missingness proportion when a researcher should refuse to use the missing-indicator method as a strategy of handling missing data.

3. ***A model specification****: a model with one categorical regressor, or with one categorical and one continuous regressor, or with one categorical regressor, one continuous regressor, and their interactions.* This criterion aims to overcome the previous studies' limitations with respect to the simplicity of investigated models and their relevance to social science. While authors of previous studies operated with simple linear regression containing a single continuous predictor, here we also consider multiple regression and pay special attention to categorical predictors that are widespread in social research. The need to add another predictor into a model is explained by the fact that a coefficient of one variable may be influenced by coefficients of other variables, so it is not enough to limit hypothetical situations to a model with a single explanatory variable. In addition, we also extended our models with interactions of both predictors since interactions may play a significant role in modeling complex social phenomena [Morgan and Sonquist 1961].

In order to simulate regression, it is necessary to start with generating independent variables, and a dependent variable is then modeled on the basis of these variables (including an intercept and a random error). In our case, we simulated data based on a model containing one categorical, one continuous regressor, and their interactions. A categorical regressor (variable 'A') with three valid categories was randomly simulated from a discrete uniform distribution (and then dichotomized) while a continuous regressor (variable 'B') and a random error were randomly simulated from a

normal distribution with different parameters[6]. Then, a dependent variable (variable 'C') was received by means of these variables.

One additional binary variable (variable 'D') was created to use it when simulating the MAR data. The sample size was 2000, since samples in sociological surveys usually are of this size or somewhat less. All the true parameters of the models are presented in Appendix 1.

To conduct future analysis of bias provided by different methods of handling missing data, we should have got not only point estimates of initial parameters but also their interval estimates. To get it, we used bootstrapping – a statistical technique of generating new samples with replacement out of an original sample [Efron and Tibshirani 1993]. Generally, the experiment's design included nine steps:

1. generate a new sample (N = 2000) with replacement from the initial simulated data,
2. dichotomize the categorical variable 'A' and estimate parameters of all three models' specifications – *estimation of true parameters*,
3. choose 10%, 25%, or 50% of observations (depending on the necessary proportion of missing data) from that sample using the following rule: for MCAR, choose randomly; for MAR, choose randomly those observations that have a certain value for the additional variable 'D'; for MNAR, chose randomly those observations that do not have the certain value for the variable 'A',
4. for the chosen observations, replace values of the original non-dichotomized variable 'A' with missing values,
5. recode the missing values into a new single value ('4'), as it is required by the missing indicator method,
6. dichotomize the updated variable 'A' and estimate the parameters of all three models' specifications – *the missing indicator method*,
7. ignore observations with the value '4' for the variable 'A' and estimate parameters of all three models' specifications using the rest observations – *the complete case analysis*,
8. repeat the steps 1-7 2000 times,
9. calculate point and interval estimates of all required parameters using the data obtained from 2000 repetitions. The point estimate of the parameters was calculated as a mean of all the estimates, and bounds of the confidence intervals were obtained as 2.5 and 97.5 percentiles of the estimates. The gained point estimates expectedly appeared statistically similar to the respective initial coefficients.

Since we fixed the same value of random seed before every set of the experiment's iterations, every set was carried out on the same sample. It means that 2000 estimates with 25% of

[6] All the simulations were carried out using 'numpy' [Oliphant 2006; van der Walt et al. 2011] and 'pandas' [McKinney 2010] python modules, and all the regression models were built using 'statsmodels' python module [Seabold and Perktold 2010]. All the technical files are available upon request.

missing data simulated as *MNAR* using the missing-indicator method were obtained from the same data as 2000 estimates with the same criteria but using *MAR* or *MCAR*. Analogously, 2000 estimates with 25% of missing data simulated as MNAR using *the missing-indicator method* were obtained from the same data as 2000 estimates with the same criteria but using *the complete case analysis*, and so on. This approach allows us to be sure that there is no influence of uncontrolled factors, compare the estimates directly, and replicate the experiment in the future.

Along with regression coefficients and their standard errors, we also estimated some additional parameters such as R^2, adjusted R^2, and F-statistic. The point estimates of parameters were compared directly, and the respective interval estimates were compared using the following metric:

$$\Delta = \frac{|U_t - U_e| + |L_t - L_e|}{U_t - L_t} * 100\%, \quad (1)$$

where U_t and L_t are upper and lower bounds of the *true* 95% confidence interval respectively, and U_e and L_e are upper and lower bounds of the *estimated* 95% confidence interval respectively. The metric was introduced in the studies of a similar topic [Zangieva and Suleymanova 2016; Zangieva and Timonina 2014], and it indicates a relative degree of deviation of confidence intervals. This metric is easy to interpret: the closer the value to zero, the more similar the confidence intervals are. Thus, we are interested in smaller values of this metric. When comparing values of this metric, we consider the difference equal to or less than 10 as non-significant difference.

Based on the literature and our experience, we stated the following hypotheses according to the criteria of the experiment:

Hypothesis 1 (the mechanism of missingness). *The complete case analysis performs better in case of the missingness completely at random (MCAR), and the missing-indicator method performs better in case of the missingness not at random (MNAR).* As we explained above, when data are missing completely at random, the subsample remained for the complete case analysis becomes a random subsample of the initial data. When data are missing not at random, an additional category created for the missing-indicator method may be regarded as a 'natural' category of the initial variable.

Hypothesis 2 (proportion of missing data). *In general, with an increasing proportion of missing data, both methods produce more biased estimates, but the missing-indicator method results in more biased estimates in comparison with the complete case analysis.* Here, we base on the results of previous studies, according to which the missing-indicator method, overall, performs worse than the complete case analysis [Jones 1996].

Hypothesis 3 (model specification). *For the missing-indicator method, the more complicated a model specification is, the more biased parameter estimates become.* We suppose

that the missing-indicator binary variable added to the model may affect all the rest coefficients of the model, and the more predictors the model has, the greater changes the missing-indicator variable makes.

Results

For assessing how the complete case analysis and the missing indicator method (hereinafter – CCA and MIM respectively) perform, we aggregated Δ-metrics (presented in Appendix 2) through all the relevant specifications. For instance, aggregated Δ-metric for Intercept is an average value of Δ-metrics for Intercepts from Specifications 1-3, since all the specifications contain their Intercepts. Analogously, since only Specifications 2 and 3 contain continuous variable 'B', aggregated Δ-metric for 'B' is an average value of Δ-metrics for variable 'B' from Specifications 2 and 3. The point estimates and original, non-aggregated Δ-metrics are presented in Appendices 1 and 2.

Testing **Hypothesis 1** requires to compare values from Table 1, from the columns related to CCA and MIM. The closer a value of Δ-metric to zero, the more similar the confidence intervals are, and the better a method of handling missingness performs. Within column MCAR, MAR, and MNAR, CCA and MIM perform almost identically. Thereof, the bias of regression coefficients is roughly similar if handling missingness by both CCA and MIM.

Unexpectedly, **additional pattern** appears. From Table 1, it is well seen that Δ-metrics strongly differ within columns MCAR/MAR and MNAR for both CCA and MIM. In other words, both methods of handling missingness perform worse if missingness is not at random than if it is completely at random or at random.

Tab. 1. Regression coefficients in context of the different mechanisms of missingness: a degree of deviation (Δ) of the confidence intervals

Parameter	Complete Case Analysis			Missing-Indicator Method		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
Intercept	22	24	44	21	23	45
A_2	27	26	54	27	26	54
A_3	20	20	30	20	20	29
B	22	27	41	23	24	39
A_2*B	21	25	52	21	25	52
A_3*B	19	21	24	19	21	24

In regard to the first part of **Hypothesis 2**, indeed, both methods produce more biased estimates with an increasing proportion of missing data. Thus, if shifting from the left to the right through columns within CCA in Table 2, the values of Δ-metric rise. They do the same within the columns of MIM. But the second part of Hypothesis 2, postulating that MIM performs worse than CCA, appears to be disproved by the experiment, because for each level of the missingness' proportion, the values within CCA and MIM are quite similar.

For example, taking point estimates from Appendix 1, Intercept in Specification 1 under conditions of MCAR and 10% of missingness is, on average, 248,7 for both CCA and MIM. If a missingness' proportion grows up to 25%, Intercept is, on average, 248,8 for both methods. And if a missingness' proportion grows up to 50%, Intercept is, on average, 248,9 for both methods. Meanwhile, the true value of Intercept equals 248,5.

If, then, analogously taking the respective interval estimates' Δ-metric from Appendix 1, they equal 8, 21, and 49 for both CCA and MIM. The pattern of **the similarity** of both point and interval estimates of regressions coefficients for CCA and MIM seems to be more or less general for all the specifications.

Tab. 2. Regression coefficients in context of different proportions of missing data: a degree of deviation (Δ) of the confidence intervals

Parameter	Complete Case Analysis			Missing-Indicator Method		
	10%	25%	50%	10%	25%	50%
Intercept	7	20	63	6	20	62
A_2	11	25	71	11	25	71
A_3	5	15	50	5	15	49
B	10	23	58	10	23	53
A_2*B	10	22	66	10	22	66
A_3*B	7	14	44	7	14	44

Hypothesis 3 was dedicated to a model specification. Examined Specifications 1-3 are presented in columns within CCA (on the left) and MIM (on the right) in Table 3. Moving through columns 1-3 within MIM, we might not distinguish a clear pattern of changing the values. Thereof, Hypothesis 3 seems to be refused. Additionally, the same pattern is seen in columns 1-3 within CCA; and comparing the left and the right sides of the table, it is clearly seen that the values are almost identical. Hence, it corroborates the regularity that CCA and MIM perform roughly similarly regarding regression coefficients irrespectively the mechanism of missingness, a proportion of missing data, and a model specification's complexity.

Tab. 3. Regression coefficients in context of model specifications: a degree of deviation (Δ) of the confidence intervals

Parameter	Complete Case Analysis			Missing-Indicator Method		
	1	2	3	1	2	3
Intercept	34	28	28	34	26	28
A_2	38	36	33	38	36	33
A_3	26	24	20	26	24	20
B		25	35		23	35
A_2*B			33			33
A_3*B			22			22

Notes: Specification 1 is a model with one categorical predictor,
Specification 2 is a model with one categorical and one continuous predictors,
Specification 3 is a model with one categorical predictor, one continuous predictor,

Parameter	Complete Case Analysis			Missing-Indicator Method		
	1	2	3	1	2	3
and their interactions.						

Besides the proposed hypotheses, some additional important results were gained. We found out that in spite of similar good performance of CCA and MIM regarding regression coefficients, both methods seem to give worse results within one specific combination of conditions: MNAR and 50% of missingness. In Appendix 2, Δ-metrics within this combination are three-four times bigger than within other combinations. Originally, it happens because the estimated confidence intervals of the regression coefficients within this combination are two times longer than the true interval and the intervals within other combinations. However, the true interval is nested in all the estimated intervals, which means that the estimates of regression coefficients here are not biased, but imprecise. Perhaps, we may consider this proportion as a threshold when a researcher should refuse to handle data missing not at random: if data are missing not at random and comprise half of the sample or more, then using both CCA and MIM lead to the quite imprecise estimates of regression coefficients highly likely. These findings to some extent replicate the results of the similar study devoted to use of the missing indicator method in decision trees [Zhuchkova and Rotmistrov 2018]. For the situations of MCAR and MAR, the threshold shoulb be smaller but our research does not allow to identify it.

Other surprising results were gained in regard to regression coefficients' standard errors and additional statistics: R^2, adjusted R^2, and F-statistic. We found out that here CCA and MIM perform rather poorly, and MIM performs even worse than CCA. Considering Tables 4-6, it is well seen that they contain values of Δ-metric several times bigger than the respective values in Tables 1-3; almost all of them exceed 100, which in this case means that confidence intervals for estimated parameters do not overlap confidence intervals for respective true parameters at all.

Tab. 4. Regression coefficients' standard errors and R^2 in context of the different mechanisms of missingness: a degree of deviation (Δ) of the confidence intervals

Parameter	Complete Case Analysis			Missing-Indicator Method		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
Intercept (st. error)	442	453	824	1175	1180	2275
A_2 (st. error)	559	567	1050	1509	1510	2911
A_3 (st. error)	503	507	496	1369	1370	1902
B (st. error)	446	446	533	853	848	1459
A_2*B (st. error)	394	404	897	1492	1486	2786
A_3*B (st. error)	392	401	503	1481	1483	2071
R^2	22	23	257	994	985	1473
Adjusted R^2	22	23	258	996	987	1476
F-statistic	289	287	399	671	669	759

Tab. 5. Regression coefficients' standard errors and R^2 in context of the of different proportions of missing data: a degree of deviation (Δ) of the confidence intervals

Parameter	Complete Case Analysis			Missing-Indicator Method		
	10%	25%	50%	10%	25%	50%
Intercept (st. error)	132	390	1197	418	1131	3082
A_2 (st. error)	166	491	1519	532	1445	3954
A_3 (st. error)	115	343	1048	443	1174	3024
B (st. error)	118	341	966	334	848	1979
A_2*B (st. error)	125	373	1197	544	1440	3780
A_3*B (st. error)	95	288	912	513	1307	3216
R^2	17	56	228	402	1015	2035
Adjusted R^2	17	57	229	403	1017	2038
F-statistic	113	285	577	513	705	880

Tab. 6. Regression coefficients' standard errors and R^2 in context of model specifications: a degree of deviation (Δ) of the confidence intervals

Parameter	Complete Case Analysis			Missing-Indicator Method		
	1	2	3	1	2	3
Intercept (st. error)	629	581	510	1508	1379	1743
A_2 (st. error)	750	770	656	1799	1888	2244
A_3 (st. error)	518	500	488	1376	1358	1906
B (st. error)		528	422		675	1432
A_2*B (st. error)			565			1921
A_3*B (st. error)			432			1678
R^2	82	74	145	845	852	1756
Adjusted R^2	83	75	146	846	853	1759
F-statistic	311	309	356	668	629	802

Notes: Specification 1 is a model with one categorical predictor,
Specification 2 is a model with one categorical and one continuous predictors,
Specification 3 is a model with one categorical predictor, one continuous predictor, and their interactions.

Regarding standard errors, it means that respective coefficients' significance is estimated poorly. Appendices 1 and 2 show that the standard errors are estimated much higher than their true values. Consequently, p-values of respective regression coefficients are overestimated as well, which means that the regression coefficients are estimated less significant than they are in truth. It happens both for CCA and MIM, but the latter performs remarkably worse.

The difference in the methods' performance regarding R^2 and adjusted R^2 is quite bigger since Δ-metrics within the combination of CCA and MCAR or MAR (Table 4) and within the combination of CCA and 10% or 25% of missingness (Table 5) are acceptably small meanwhile the respective combinations with MIM exceed the threshold of 100 almost ten times.

Why are the standard errors, R^2, and adjusted R^2 estimated poorly and why does MIM perform dramatically worse than CCA? In case of the latter, increase of standard errors is explained

by reduction of the available sample: the lower a sample is, the higher standard errors are. For example, from Appendix 2, Δ-metric for Intercept's standard error in Specification 1 under conditions of *CCA*, MCAR, and 10% of missingness is 130. If missingness' proportion grows up to 25%, Δ-metrics grows up to 373. And if missingness' proportion grows up to 50%, Δ-metrics grows up to 1008.

In contrast, in case of *MIM*, increase of standard errors is not related to sample size because it remains the same. It is rather explained by increase of residual sum of squares (RSS). RSS rises due to a contribution of a missing-indicator variable, which, according to MIM, should be added to a set of original variables to indicate presence or absence of missing data. While coefficients of other variables appear to be unbiased, their performance is deteriorated by the missing-indicator variable's coefficient. Thus, residuals rise, and a model's overall prediction worsens.

To illustrate this idea, look at Appendix 2. Δ-metrics for R^2 in Specification 1 under conditions of *CCA*, MCAR, and 10% of missingness is 4. If missingness' proportion grows up to 25%, Δ-metrics grows up to 14. And if missingness' proportion grows up to 50%, Δ-metrics grows up to 47. Thus, the bias of R^2 estimates grows but not crucially. Indeed, in Appendix 1, the true R^2 and the estimated ones when using *CCA* equal 0,62 irrespective the missingness' proportion.

In contrast, in Appendix 2, Δ-metrics for R^2 in Specification 1 under conditions of *MIM*, MCAR, and 10% of missingness is 254. If missingness' proportion grows up to 25%, Δ-metrics grows up to 633. And if missingness' proportion grows up to 50%, Δ-metrics grows up to 1278. I.e., the bias of R^2 estimates and its grows are crucial. Indeed, in Appendix 1, the estimated ones when using *MIM* equal 0,56, 0,46 and 0,31, respectively while the true R^2 still equals 0,62.

To sum up, in case of CCA, the estimates of R^2 and adjusted R^2 are unbiased but imprecise, and in case of MIM, ones are significantly lower than the respective true values. Considering that all the regression coefficients are identical among the two methods, we conclude that the bias appears because of the missing-indicator variable's coefficient; besides, in our experiment, the missing-indicator variable was always significant.

The last rows in Tables 4-6 contain F-statistic. Judging by its Δ-metrics, CCA and MIM perform poorly, and, anew, the latter performs worse than the former. The explanation relies on F-statistic's formula, which includes respective R^2 and available sample size in its numerator. In case of *CCA*, the mentioned reduction of the available sample leads to smooth reduction of estimated F-statistic comparing to its true value. In case of *MIM*, the also mentioned reduction of R^2 leads to steep reduction of estimated F-statistic comparing to its true value.

Conclusion

Our research aims to provide a complex analysis of the missing-indicator method performance in case of a categorical variable in comparison with the complete case analysis. While

the latter seems to be the most popular way to handle missing data in real research, the former appears to be a simple and effective alternative that allows making a full sample available for analysis. By means of the statistical experiment and simulated data, we examined how these methods perform in conditions that differ in the mechanism of missingness, a proportion of missing data, and a model specification.

Although the results might seem ambiguous in the presented aggregated form, we clearly identified one common pattern among all possible combinations of the criteria, which we have tried to describe above. Regression coefficients themselves may be unbiased for both complete case analysis and missing-indicator method, while standard errors of these coefficients may at the same time be completely incorrect. This fact makes it difficult to estimate coefficients significance accurately. Besides, for the missing-indicator method, an additional variable of the new category for missing data increases residuals of the model and thus worsens the prediction, which manifests itself in crucially underestimated values of R^2. It brings us to the main conclusion of the conducted research. In social science, where the key objective of analysis is often to detect and estimate a relationship between some variables, both methods are appropriate tools for handling missing data since the estimates are unbiased in most cases. But when it comes to forecasting, missing-indicator method should not be used. Regardless all the factors, the missing-indicator variable always turns out to be significant, and its coefficient, which is far away from zero, greatly deteriorates a model's predictive quality and does not allow to adequately understand real predictive power of the chosen variables. Surprisingly, this conclusion completely contradicts the results of the similar study, where instead of regression, authors examined decision trees [Zhuchkova and Rotmistrov 2018].

In our research, we have tried to overcome some limitations of the previous studies by using simulated data and providing an extended set of criteria. However, it has own limitations as well. Firstly, we only inserted missing data in one variable. In real research, it is more probable that several variables have missing values. But such an approach would only complicate design of our experiment and would not allow us to control the rest factors. Secondly, when simulating missingness at random (MAR), we used a variable that was not included in our regression model. Probably, if we used one that was included in the model (variable 'B') to simulate missingness at random, the results would be different, and regression coefficients would be more biased. However, we address this issue to further studies.

References

1. Akande, O., Li, F., & Reiter, J. (2017). An Empirical Comparison of Multiple Imputation Methods for Categorical Data. The American Statistician, 71(2), 162–170. doi:10.1080/00031305.2016.1277158

2. Allison, P. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods and Research, 28, 301–9.

3. Allison, P. (2005). Imputation of Categorical Variables with PROC MI. Proceedings of the SAS Users Group International Conference (SUGI) 30, 113-130. URL: https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/113-30.pdf

4. Chen, J., & Hossler, D. (2016). The Effects of Financial Aid on College Success of Two-Year Beginning Nontraditional Students. Research in Higher Education, 58(1), 40–76. doi:10.1007/s11162-016-9416-0

5. Choi, J., Dekkers, O. M., & le Cessie, S. (2018). A comparison of different methods to handle missing data in the context of propensity score analysis. European Journal of Epidemiology, 34(1), 23-36. doi:10.1007/s10654-018-0447-z

6. Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59(10), 1087–1091. doi:10.1016/j.jclinepi.2006.01.014

7. Efron, B., & Tibshirani, R. (1993) An Introduction to the Bootstrap. Chapman and Hall, New York, London.

8. Gentle, J.E., Härdle, W. K., & Mori, Y. (2012) Handbook of Computational Statistics: Concepts and Methods. Berlin: Springer.

9. Gesser-Edelsburg, A., Zemach, M., Lotan, T., Elias, W., & Grimberg, E. (2018). Perceptions, intentions and behavioral norms that affect pre-license driving among Arab youth in Israel, Accident Analysis & Prevention, 111, 1–11.

10. Greenacre, M., & Pardo, R. (2006). Subset Correspondence Analysis. Sociological Methods & Research, 35(2), 193–218. doi:10.1177/0049124106290316

11. Groenwold, R. H. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., & Moons, K. G. M. (2012). Missing covariate data in clinical research: when and when not to use missing-indicator method for analysis. Canadian Medical Association Journal, 184(11), 1265–1269. doi:10.1503/cmaj.110977

12. Hendry, G.M., Zewotir, T., Naidoo, R.N., & North, D. (2017) A comparative study of multiple imputation and subset correspondence analysis in dealing with missing data. South African Statistical Journal, 51(1), 183-198.

13. Henry, A. J., Hevelone, N. D., Lipsitz, S., & Nguyen, L. L. (2013). Comparative methods for handling missing data in large databases. Journal of Vascular Surgery, 58(5), 1353–1359.e6. doi:10.1016/j.jvs.2013.05.008

14. Jones, M. P. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. Journal of the American Statistical Association, 91(433), 222–230. doi:10.1080/01621459.1996.10476680

15. Knol, M. J., Janssen, K. J. M., Donders, A. R. T., Egberts, A. C. G., Heerdink, E. R., Grobbee, D. E., Moons, K. G. M., & Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. Journal of Clinical Epidemiology, 63(7), 728–736. doi:10.1016/j.jclinepi.2009.08.028

16. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.

17. Miettinen, O.S. (1985). Theoretical epidemiology: principles of occurrence research. New York (NY): John Wiley & Sons.

18. Morgan, J., & Sonquist, J. (1963) Problems in the Analysis of Survey Data, and a Proposal. Journal of the American Statistical Association, 58(302), 415—434. doi:10.1080/01621459.1963.10500855.

19. Oliphant, T.E. (2006). A guide to NumPy. USA: Trelgol Publishing.

20. Ratner, B. (2012). Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. Boca Raton: CRC Press.

21. Rickles, J., Heppen, J. B., Allensworth, E., Sorensen, N., & Walters, K. (2018). Online Credit Recovery and the Path to On-Time High School Graduation. Educational Researcher, 47(8), 481–491. doi:10.3102/0013189X18788054

22. Rokach, L., Maimon, O. (2010). Decision Trees. Data Mining and Knowledge Discovery Handbook. Boston: Springer, 165–192.

23. Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581–592.

24. Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

25. Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. Psychological Methods, 7, 147–77.

26. Seabold, S., & Perktold, J. (2010) Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference.

27. Stavseth, M. R., Clausen, T., & Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. SAGE Open Medicine, 7, 1-12. doi:10.1177/2050312118822912

28. Strebkov, D., Shevchuk, A., Lukina, A., Melianova, E., & Tyulyupo, A. (2019). Social Factors of Contractor Selection on Freelance Online Marketplace: Study of Contests Using “Big Data”. Journal of Economic Sociology, 20(3), 25-66.

29. Trevizo, D., & Lopez, M. J. (2016). Neighborhood Segregation and Business Outcomes. Sociological Perspectives, 59(3), 668–693. doi:10.1177/0731121416629992

30. van der Heijden, G. J. M. G., T. Donders, A. R., Stijnen, T., & Moons, K. G. M. (2006). Imputation of missing values is superior to complete case analysis and missing-indicator method in multivariable diagnostic research: A clinical example. Journal of Clinical Epidemiology, 59(10), 1102–1109. doi:10.1016/j.jclinepi.2006.01.015

31. van der Walt, S. S., Colbert, C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering, 13, 22-30, doi:10.1109/MCSE.2011.37

32. van Kuijk, S.M.J., Viechtbauer, W., Peeters, L.L., & Smits, L. (2016) Bias in regression coefficient estimates when assumptions for handling missing data are violated: A simulation study. Epidemiology Biostatistics and Public Health, 13(1), e11598.

33. Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2008). Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. Sociological Methodology, 38(1), 369–397. doi:10.1111/j.1467-9531.2008.00202.x

34. Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence From Past Multisite Randomized Trials. Journal of Research on Educational Effectiveness, 10(4), 843–876. doi:10.1080/19345747.2017.1300719

35. White, I. R., & Thompson, S. G. (2005) Adjusting for partially missing baseline measurements in randomized trials. Stat Med, 24, 993-1007.

36. Zangieva, I., & Suleymanova, A. (2016). Comparative analysis of approaches to aggregation of multiple imputation results. Sociology: methodology, methods, mathematical modeling, 42, 7-60.

37. Zangieva, I., & Timonina, E. (2014) Comparing imputation algorithms efficiency respective to the data analysis methods. The monitoring of public opinion: economic and social changes journal, 1(119), 41-55.

38. Zhelyazkova, N., & Ritschard, G. (2018). Parental Leave Take-Up of Fathers in Luxembourg. Population Research and Policy Review, 37(5), 769-793. doi:10.1007/s11113-018-9470-8

39. Zhuchkova, S., & Rotmistrov, A. (2018) Handling missing data with CHAID: results of a statistical experiment. Sociology: methodology, methods, mathematical modeling, 0(46), 85-122.

Appendix 1. Point estimates for linear regression in terms of all the possible combinations

	True Value	Complete Case Analysis			Missing Indicator Method			Complete Case Analysis			Missing Indicator Method			Complete Case Analysis			Missing Indicator Method		
		10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%	50%
		MCAR						MAR						MNAR					
Specification 1																			
Intercept	248,7	248,7	248,8	248,9	248,7	248,8	248,9	248,3	247,6	246,1	248,3	247,6	246,1	248,5	248,6	248,6	248,5	248,6	248,6
A_2	-428,6	-428,7	-428,9	-429,0	-428,7	-428,9	-429,0	-428,1	-427,2	-425,0	-428,1	-427,2	-425,0	-428,6	-428,8	-428,5	-428,6	-428,8	-428,5
A_3	-211,4	-211,4	-211,5	-211,5	-211,4	-211,5	-211,5	-211,2	-210,9	-210,8	-211,2	-210,9	-210,8	-211,3	-211,4	-211,3	-211,3	-211,4	-211,3
Intercept (st. error)	5,3	5,6	6,2	7,6	6,1	7,3	10,2	5,6	6,2	7,6	6,1	7,3	10,2	5,7	6,5	9,5	6,5	8,5	15,7
A_2 (st. error)	7,6	8,0	8,7	10,7	8,6	10,3	14,4	8,0	8,7	10,7	8,6	10,4	14,4	8,1	9,2	13,5	9,1	12,1	22,3
A_3 (st. error)	7,6	8,0	8,8	10,7	8,6	10,4	14,4	8,0	8,8	10,8	8,6	10,4	14,5	7,8	8,4	10,7	8,8	10,9	17,7
R	0,62	0,62	0,62	0,62	0,56	0,46	0,31	0,62	0,62	0,61	0,56	0,46	0,30	0,61	0,59	0,50	0,53	0,39	0,16
Adjusted R	0,62	0,62	0,62	0,62	0,56	0,46	0,31	0,62	0,62	0,61	0,55	0,46	0,30	0,61	0,59	0,50	0,53	0,39	0,16
F-statistic	1611,1	1451,4	1211,6	808,7	834,3	576,6	298,9	1448,4	1201,2	793,1	831,2	569,6	291,8	1402,2	1083,5	509,4	739,1	422,9	125,0
A_4					-213,4	-213,9	-214,1				-211,7	-210,4	-209,0				-213,8	-214,6	-214,7
A_4 (st. error)					12,2	10,3	11,7				12,2	10,4	11,8				12,7	11,6	17,0
Specification 2																			
Intercept	232,3	232,3	232,4	232,4	232,0	231,6	230,8	232,1	231,7	230,9	231,5	230,3	228,2	231,9	231,3	229,0	231,7	231,0	229,7
A_2	-426,8	-426,9	-427,1	-427,2	-426,8	-427,0	-427,0	-426,4	-425,7	-424,1	-426,4	-425,5	-423,9	-426,7	-426,8	-426,2	-426,7	-426,8	-426,3
A_3	-211,5	-211,5	-211,6	-211,6	-211,5	-211,6	-211,6	-211,3	-211,2	-211,4	-211,3	-211,3	-211,5	-211,4	-211,5	-211,4	-211,4	-211,5	-211,4
B	2,3	2,3	2,3	2,3	2,3	2,4	2,5	2,3	2,2	2,2	2,3	2,4	2,6	2,3	2,4	2,7	2,3	2,4	2,6
Intercept (st. error)	5,6	5,9	6,5	7,9	6,4	7,6	10,4	5,9	6,5	7,9	6,4	7,6	10,4	6,0	6,8	9,6	6,7	8,8	15,9
A_2 (st. error)	7,4	7,8	8,6	10,5	8,5	10,2	14,2	7,8	8,6	10,5	8,5	10,2	14,2	8,0	9,0	13,1	9,0	11,9	22,0
A_3 (st. error)	7,5	7,9	8,6	10,6	8,5	10,3	14,3	7,9	8,6	10,6	8,5	10,3	14,3	7,7	8,2	10,4	8,7	10,8	17,5
B (st. error)	0,3	0,3	0,3	0,4	0,3	0,3	0,4	0,3	0,3	0,4	0,3	0,3	0,4	0,3	0,3	0,3	0,3	0,3	0,4
R	0,63	0,63	0,63	0,63	0,57	0,48	0,33	0,63	0,63	0,63	0,57	0,48	0,32	0,62	0,61	0,53	0,54	0,40	0,18
Adjusted R	0,63	0,63	0,63	0,63	0,57	0,48	0,32	0,63	0,63	0,63	0,57	0,47	0,32	0,62	0,61	0,53	0,54	0,40	0,17
F-statistic	1134,8	1022,4	853,7	570,6	660,9	458,0	241,2	1019,9	845,6	558,1	659,1	453,7	237,1	991,3	773,5	381,4	585,6	337,6	106,1

	True Value	Complete Case Analysis			Missing Indicator Method			Complete Case Analysis			Missing Indicator Method			Complete Case Analysis			Missing Indicator Method		
		10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%	50%
		MCAR						MAR						MNAR					
A_4					-212,8	-213,3	-213,4				-211,7	-210,6	-209,9				-212,8	-213,5	-213,7
A_4 (st. error)					12,0	10,2	11,6				12,0	10,2	11,7				12,5	11,4	16,9
								Specification 3											
Intercept	163,0	163,0	163,0	163,0	163,0	163,0	163,0	163,0	163,2	164,2	163,0	163,2	164,2	162,8	162,9	162,8	162,8	162,9	162,8
A_2	-288,9	-289,0	-289,0	-289,0	-289,0	-289,0	-289,0	-288,7	-288,3	-288,0	-288,7	-288,3	-288,0	-288,8	-289,0	-288,9	-288,8	-289,0	-288,9
A_3	-148,8	-148,8	-148,7	-148,8	-148,8	-148,7	-148,8	-148,8	-149,3	-150,8	-148,8	-149,3	-150,8	-148,6	-148,7	-148,6	-148,6	-148,7	-148,6
B	11,9	11,9	11,9	11,9	11,9	11,9	11,9	11,9	11,9	12,0	11,9	11,9	12,0	11,9	11,8	11,9	11,9	11,8	11,9
A_2*B	-20,3	-20,3	-20,3	-20,3	-20,3	-20,3	-20,3	-20,3	-20,4	-20,6	-20,3	-20,4	-20,6	-20,3	-20,3	-20,3	-20,3	-20,3	-20,3
A_3*B	-8,7	-8,7	-8,7	-8,7	-8,7	-8,7	-8,7	-8,7	-8,8	-8,9	-8,7	-8,8	-8,9	-8,7	-8,7	-8,7	-8,7	-8,7	-8,7
Intercept (st. error)	4,6	4,8	5,3	6,5	5,7	7,5	11,2	4,8	5,2	6,3	5,7	7,4	11,0	5,0	5,8	9,2	6,2	9,1	18,1
A_2 (st. error)	6,4	6,7	7,4	9,0	7,9	10,4	15,6	6,7	7,3	8,9	7,9	10,4	15,5	6,9	8,1	12,9	8,7	12,7	25,3
A_3 (st. error)	6,5	6,9	7,5	9,2	8,1	10,6	16,0	6,9	7,5	9,1	8,1	10,6	15,8	6,8	7,5	10,4	8,6	11,7	20,4
B (st. error)	0,3	0,4	0,4	0,5	0,4	0,6	0,8	0,4	0,4	0,5	0,4	0,6	0,8	0,4	0,4	0,7	0,5	0,7	1,4
A_2*B (st. error)	0,5	0,5	0,6	0,7	0,6	0,8	1,2	0,5	0,6	0,7	0,6	0,8	1,2	0,5	0,6	1,0	0,7	1,0	1,9
A_3*B (st. error)	0,5	0,5	0,6	0,7	0,6	0,8	1,2	0,5	0,6	0,7	0,6	0,8	1,2	0,5	0,6	0,8	0,6	0,9	1,5
R	0,80	0,80	0,80	0,80	0,73	0,61	0,41	0,80	0,80	0,80	0,73	0,61	0,41	0,79	0,77	0,67	0,69	0,51	0,22
Adjusted R	0,80	0,80	0,80	0,80	0,72	0,61	0,41	0,80	0,80	0,80	0,72	0,61	0,41	0,79	0,77	0,67	0,69	0,51	0,22
F-statistic	1624,4	1463,7	1221,0	815,7	753,7	442,8	200,0	1464,3	1223,3	821,8	753,0	441,4	199,1	1379,3	1015,5	414,5	626,7	299,2	80,3
A_4					-146,7	-147,1	-147,1				-148,1	-147,7	-148,7				-145,6	-146,1	-146,1
A_4 (st. error)					11,4	10,5	12,9				11,4	10,5	12,8				12,1	12,3	19,6
A_4*B					-9,1	-9,1	-9,1				-8,9	-8,9	-9,0				-9,3	-9,3	-9,3
A_4*B (st. error)					0,9	0,8	1,0				0,9	0,8	1,0				0,9	0,9	1,5

Appendix 2. Δ-metrics of interval estimates for linear regression in terms of all the possible combinations

	Complete Case Analysis			Missing Indicator Method			Complete Case Analysis			Missing Indicator Method			Complete Case Analysis			Missing Indicator Method		
	10%	**25%**	**50%**	**10%**	**25%**	**50%**	**10%**	**25%**	**50%**	**10%**	**25%**	**50%**	**10%**	**25%**	**50%**	**10%**	**25%**	**50%**
	MCAR						**MAR**						**MNAR**					
Specification 1																		
Intercept	8	21	49	8	21	49	7	17	54	7	17	54	12	28	108	12	28	108
A_2	10	22	49	10	22	49	12	23	59	12	23	59	17	33	116	17	33	116
A_3	4	15	43	4	15	43	5	17	42	5	17	42	6	20	77	6	20	77
Intercept (st. error)	130	373	1008	329	892	2185	128	372	1014	329	900	2211	171	540	1921	501	1456	4768
A_2 (st. error)	155	451	1205	396	1068	2613	153	449	1208	396	1074	2630	209	640	2284	600	1733	5679
A_3 (st. error)	138	401	1074	351	954	2344	137	408	1089	352	964	2370	79	260	1074	419	1149	3483
R	4	14	47	254	633	1278	4	16	45	257	649	1301	30	108	472	376	951	1907
Adjusted R	4	14	47	255	634	1279	4	16	45	258	650	1302	31	109	474	377	952	1909
F-statistic	99	245	490	476	633	804	95	248	500	477	638	809	125	321	673	533	728	910
Specification 2																		
Intercept	7	19	48	6	15	39	2	14	42	3	15	37	9	24	84	6	26	88
A_2	9	19	48	10	20	49	12	19	54	11	19	55	15	32	116	15	31	117
A_3	5	14	41	5	14	39	5	17	42	5	16	38	5	20	70	5	19	70
B	10	20	47	7	17	33	4	17	43	6	20	44	8	21	57	9	23	44
Intercept (st. error)	126	365	974	313	828	1999	122	351	963	304	826	2005	161	490	1678	470	1339	4329
A_2 (st. error)	163	465	1252	416	1122	2747	161	466	1255	413	1127	2758	211	653	2306	633	1824	5955
A_3 (st. error)	137	392	1045	348	940	2314	137	401	1064	350	950	2334	75	244	1005	416	1136	3435
B (st. error)	150	436	1191	228	540	1002	155	440	1190	228	545	1016	118	327	747	331	774	1414
R	7	16	46	255	641	1287	4	14	48	257	651	1306	27	93	412	382	959	1931
Adjusted R	7	16	47	256	642	1288	4	14	48	258	652	1308	27	94	415	383	960	1933
F-statistic	96	245	492	414	591	781	98	252	502	415	595	786	123	313	657	480	697	901
Specification 3																		

	Complete Case Analysis			Missing Indicator Method			Complete Case Analysis			Missing Indicator Method			Complete Case Analysis			Missing Indicator Method		
	10%	**25%**	**50%**	**10%**	**25%**	**50%**	**10%**	**25%**	**50%**	**10%**	**25%**	**50%**	**10%**	**25%**	**50%**	**10%**	**25%**	**50%**
	MCAR						**MAR**						**MNAR**					
Intercept	6	16	43	6	16	43	2	12	42	2	12	42	8	26	100	8	26	100
A_2	7	24	47	7	24	47	5	17	45	5	17	45	15	34	105	15	34	105
A_3	3	13	42	3	13	42	5	13	36	5	13	36	4	8	57	4	8	57
B	14	24	48	14	24	48	8	19	43	8	19	43	17	34	109	17	34	109
A_2*B	9	19	48	9	19	48	8	14	40	8	14	40	13	33	109	13	33	109
A_3*B	8	15	41	8	15	41	7	12	38	7	12	38	4	14	53	4	14	53
Intercept (st. error)	98	277	728	438	1105	2534	95	260	675	428	1091	2483	159	482	1814	646	1740	5227
A_2 (st. error)	123	351	941	553	1416	3262	120	334	882	548	1410	3227	196	613	2339	828	2233	6717
A_3 (st. error)	119	339	915	536	1380	3159	114	324	855	529	1363	3111	98	317	1311	683	1729	4665
B (st. error)	81	224	597	349	901	2066	80	225	584	351	904	2076	124	396	1488	519	1421	4298
A_2*B (st. error)	105	300	807	465	1207	2787	102	294	786	468	1213	2794	168	525	1999	699	1901	5758
A_3*B (st. error)	98	297	808	467	1205	2778	103	295	778	472	1204	2768	84	273	1151	600	1512	4102
R	7	16	47	526	1331	2661	6	17	44	524	1332	2670	67	213	888	785	1991	3979
Adjusted R	7	16	48	528	1334	2665	6	17	45	526	1335	2674	67	215	894	787	1995	3984
F-statistic	108	271	539	580	789	951	106	268	533	580	790	951	164	407	806	665	885	1031

Svetlana Zhuchkova
National Research University Higher School of Economics. Faculty of Computer Science. Master Student; E-mail: szhuchkova@hse.ru

Aleksei Rotmistrov
National Research University Higher School of Economics. Faculty of Social Science. Associate Professor; E-mail: arotmistrov@hse.ru