DOI: 10.14515/monitoring.2021.4.940



С. В. Жучкова, А. Н. Ротмистров, Е. А. Шабанова

ИМЕЕТ ЛИ МЕТОД ИНДИКАТОРНОЙ ПЕРЕМЕННОЙ ПРЕИМУЩЕСТВА ПЕРЕД АНАЛИЗОМ ПОЛНЫХ НАБЛЮДЕНИЙ ПРИ ОБРАБОТКЕ ПРОПУСКОВ В КАТЕГОРИАЛЬНЫХ РЕГРЕССОРАХ?

Правильная ссылка на статью:
Жучкова С. В., Ротмистров А. Н., Шабанова Е. А. Имеет ли метод индикаторной переменной преимущества перед анализом полных наблюдений при обработке пропусков в категориальных регрессорах? // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 4. С. 23—52. https://doi.org/10.14515/monitoring.2021.4.940.
For citation:
Zhuchkova S. V., Rotmistrov A. N., Shabanova E. A. (2021) Should the Missing-Indicator Method be Preferred to Complete Case Analysis When Handling Missingness in a Categorical Regressor? *Monitoring of Public Opinion: Economic and Social Changes.* No. 4. P. 23–52. https://doi.org/10.14515/monitoring.2021.4.940. (In Russ.)

DOI: 10.14515/monitoring.2021.4.940

ИМЕЕТ ЛИ МЕТОД ИНДИКАТОРНОЙ ПЕРЕМЕННОЙ ПРЕИМУЩЕСТВА ПЕРЕД АНАЛИЗОМ ПОЛНЫХ НАБЛЮДЕНИЙ ПРИ ОБРАБОТКЕ ПРОПУСКОВ В КАТЕГОРИАЛЬНЫХ РЕГРЕССОРАХ?

ЖУЧКОВА Светлана Васильевна — младший научный сотрудник, Центр социологии высшего образования, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-MAIL: szhuchkova@hse.ru
https://orcid.org/0000-0002-4425-725X

РОТМИСТРОВ Алексей Николаевич — кандидат социологических наук, доцент кафедры методов сбора и анализа социологической информации, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-MAIL: alexey.n.rotmistrov@gmail.com
https://orcid.org/0000-0003-2386-8710

ШАБАНОВА Екатерина Алексеевна — стажер-исследователь, Международная лаборатория прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-MAIL: ekaterina.u.shabanova@gmail.com
https://orcid.org/0000-0002-6430-1297

Аннотация. Если в категориальном регрессоре есть пропущенные значения, то что лучше применить: анализ полных наблюдений или метод индикаторной переменной? Суть первого подхода состоит в исключении из анализа (в нашем случае — линейного регрессионного) наблюдений, содержащих пропуски хотя бы по одной из изучаемых переменных. Этот подход применяется по умолчанию во многих популярных приложениях, и, вопреки сложившимся

SHOULD THE MISSING-INDICATOR METHOD BE PREFERRED TO COMPLETE CASE ANALYSIS WHEN HANDLING MISSINGNESS IN A CATEGORICAL REGRESSOR?

Svetlana V. ZHUCHKOVA[1] *— Junior Research Fellow, Centre of Sociology of Higher Education*
E-MAIL: szhuchkova@hse.ru
https://orcid.org/0000-0002-4425-725X

Alexey N. ROTMISTROV[1] *— Cand. Sci. (Soc.), Associate Professor, Department of Sociological Research Methods*
E-MAIL: alexey.n.rotmistrov@gmail.com
https://orcid.org/0000-0003-2386-8710

Ekaterina A. SHABANOVA[1] *— Research Assistant, International Laboratory for Applied Network Research*
E-MAIL: ekaterina.u.shabanova@gmail.com
https://orcid.org/0000-0002-6430-1297

[1] HSE University, Moscow, Russia

Abstract. If missingness is encountered in a categorical regressor, which approach is preferable: complete case analysis or the missing-indicator method? The former approach implies including in analysis (linear regression in our research) only the cases without missingness across analyzed variables. This approach is embedded in many statistical applications by default, and despite the opinion that its applicability

DOI: 10.14515/monitoring.2021.4.940

представлениям о его ограниченности, все больше исследований подтверждают его универсальность — даже в случае неслучайных пропусков. Метод индикаторной переменной, при котором пропущенные значения заменяются на валидные, а в пару исходной переменной создается дополнительная индикаторная, выступает более новой альтернативой, которая, в отличие от первого подхода, позволяет использовать информацию из всех наблюдений и при этом, гипотетически, также не приводит к искажению изучаемых статистических параметров. Посредством статистического эксперимента на симулированных данных, контролируя механизм порождения пропусков, их долю и спецификацию регрессионной модели, мы сравниваем полученные на основе каждого из подходов статистические оценки регрессионных коэффициентов на предмет их искажений: смещения и неэффективности. Согласно результатам, оба подхода не приводят к заметному смещению, однако метод индикаторной переменной приводит к менее эффективной оценке.

is rather restricted, up-to-date studies provide evidence for its wide applicability — even to missingness not at random. The missing-indicator method, according to which missing data are replaced with a single valid value and a new missing-indicator variable is created, pretends to be an alternative that keeps a full sample available for analysis and, hypothetically, does not lead to the deterioration of parameter estimates. By means of simulated data and a statistical experiment, controlling the factors of missingness mechanism, missingness proportion, and a regression model's specification, we compare parameter estimates produced by each approach to handling missingness — how biased and inefficient they are. According to the results, no approach leads to crucially biased estimates, but the missing-indicator method produces ineffective estimates.

Ключевые слова: категориальные данные, пропуски в данных, случайные пропуски, неслучайные пропуски, анализ полных наблюдений, метод индикаторной переменной, регрессионный анализ, статистический эксперимент, метод Монте-Карло, симуляция данных, смещение, coverage

Keywords: categorical data, missing data, missingness at random, missingness not at random, complete case analysis, missing indicator method, regression analysis, simulated data, statistical experiment, Monte Carlo technique, bias, coverage

Благодарность. Публикация подготовлена в ходе проведения исследования «Комплексное сравнение методов обработки пропущенных данных в социологических исследованиях» (№ 20-04-016) в рамках Программы «Научный фонд Национального исследователь-

Acknowledgments. The publication was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2020 (grant No. 20-04-016) and by the Russian Academic Excellence Project «5-100».

DOI: 10.14515/monitoring.2021.4.940

ского университета „Высшая школа экономики“ (НИУ ВШЭ)» в 2020 г. и в рамках государственной поддержки ведущих университетов Российской Федерации «5-100».

Введение

Проводя количественные исследования, социологи постоянно сталкиваются с проблемой пропусков в данных. Она может быть сформулирована так: оценки статистических параметров, рассчитанных на основе наблюдений, содержащих пропуски, с высокой вероятностью искажены относительно этих же статистик, рассчитанных на основе этих же наблюдений, если бы они не содержали пропусков. Вслед за К. Доугерти [Dougherty, 2016] под искажением статистической оценки параметра мы подразумеваем ее смещение и неэффективность. Смещением оценки называется ее отклонение от истинного значения параметра. В большинстве исследований проблемы пропусков речь идет именно о смещении [Morris, White, Crowther, 2019]. Неэффективность оценки в контексте проблемы пропусков — это расширение доверительного интервала статистики, построенного на основе наблюдений с пропусками, относительно доверительного интервала той же статистики, построенного на основе тех же данных, в которых вместо пропусков стоят истинные значения [Rubin, 1996].

Переход от традиционных опросных данных к большим данным не избавляет от проблемы пропусков, как это порой кажется, а усугубляет ее, поскольку, собирая опросные данные, социологи действуют превентивно — составляют качественную анкету, подбирают наиболее подходящий формат опроса, мотивируют и стимулируют респондентов, а собирая большие данные, они не могут (и не должны в парадигме нереактивных данных[1]) влиять на наполненность источников информации[2].

Обнаружив в своих данных пропуски, прежде чем перейти к запланированным методам анализа данных, социологи прибегают[3] к тем или иным подходам к обработке этих пропусков. Подходов к обработке пропусков много; все они в какой-то мере нацелены на предотвращение смещения оценок и на сохранение исходного размера совокупности [van Kuijk, Viechtbauer, Peeters, Smits, 2016]. Разнообразие подходов к обработке пропусков обусловлено множеством факторов, опосредующих их влияние на искажение оценок изучаемых параметров. Среди них упоминаемыми во многих, если не большинстве методологических работ, посвященных проблеме пропусков, оказываются механизм порождения пропусков и их доля.

Мы придерживаемся традиции различать три основных механизма пропусков, предложенных Л. Рубиным [Rubin, 1976][4]:

[1] Нереактивные данные — данные, собираемые о человеке помимо его участия.

[2] Больше о проблеме пропусков в контексте больших данных см. [Giest, Samuels, 2020].

[3] Порой даже не осознавая этого, так как популярные статистические приложения зачастую обрабатывают пропуски по умолчанию.

[4] В оригинале эти механизмы называются соответственно: MCAR — missing completely at random, MAR — missing at random, MNAR — missing not at random.

DOI: 10.14515/monitoring.2021.4.940

1) полностью случайные пропуски (далее ПСП) — их возникновение не связано ни с наблюдаемыми, ни с ненаблюдаемыми данными (переменными и наблюдениями);

2) случайные пропуски (далее СП), возникновение которых в одной переменной не связано с ее собственными значениями, но связано со значениями некоторой другой переменной;

3) неслучайные пропуски (далее НП), возникновение которых обязательно связано с ее собственными значениями и опционально — со значениями некоторой другой переменной.

Для лучшего понимания этих механизмов можно представить себе онлайн-опрос, содержащий переменную о материальном положении респондентов. Пусть заметная доля малообеспеченных респондентов пропускает ее. Если причины пропусков многообразны и разнородны (например, невнимательность, нестабильное интернет-соединение и т. д.), то такую ситуацию мы отнесли бы к полностью случайному механизму возникновения пропусков. Если же доминирует одна причина — нестабильное интернет-соединение, и при этом она более характерна именно для малообеспеченных респондентов, то такую ситуацию мы отнесли бы к случайному механизму. Наконец, если малообеспеченные респонденты склонны не отвечать на вопрос о материальном положении, поскольку стесняются его, то такую ситуацию мы отнесли бы к неслучайному механизму.

При выборе подхода к обработке пропусков для предотвращения искажения оценок того или иного параметра важным представляется учет его природы. Другими словами, предполагая разные методы последующего анализа данных, содержащих пропуски, следует выбирать и разные подходы к обработке пропусков. В этом контексте водораздел проходит между одномерными методами анализа (описательными статистиками) и многомерными, поскольку во втором случае выбор подхода к обработке пропусков может повлиять на связи между анализируемыми переменными. В этой статье мы изучаем подходы к обработке пропусков в контексте популярного метода анализа данных — линейной регрессии (методом наименьших квадратов) — посредством статистического эксперимента на симулированных данных для стандартизации сравнения изучаемых подходов. В эксперименте контролируются следующие факторы: спецификация регрессионной модели, механизм порождения пропусков и их доля.

Статья построена следующим образом: сначала мы на уровне теории сравниваем наиболее популярные подходы к обработке пропусков и обосновываем выбор методов полных наблюдений и индикаторной переменной. Затем мы описываем методологию проводимого эксперимента — его дизайн, контролируемые факторы и используемые для анализа результатов метрики, а также формулируем гипотезы. После этого представляем результаты проведенного эксперимента — как в разрезе сформулированных гипотез, так и по вопросам, эксплицитно не ставившимся до его проведения.

Теоретическое сравнение подходов к обработке пропусков

Более или менее широко известны четыре подхода к обработке пропусков: 1) анализ полных наблюдений (complete case analysis, или listwise deletion), 2) анализ доступных наблюдений (available case analysis, или pairwise deletion), 3) импу-

DOI: 10.14515/monitoring.2021.4.940

тация значений, или заполнение пропусков (imputation) и 4) метод индикаторной переменной (the missing indicator method).

Анализом полных наблюдений (далее — АПН) называется исключение из дальнейшего анализа тех наблюдений, которые имеют пропуск хотя бы по одной анализируемой переменной. Из этого определения следует, в частности, что применение АПН в контексте регрессионного анализа может приводить к широкой вариации объема анализируемой совокупности в зависимости от числа анализируемых переменных, если последние содержат пропуски. Анализ доступных наблюдений похож на АПН, но предполагает более гибкое удаление наблюдений с пропусками: если применяемый метод анализа многоэтапен, то на разных его этапах удаляться из анализа могут разные наблюдения. Поясним эту гибкость на примере регрессионного анализа: поскольку он включает этап расчета корреляций между регрессорами и зависимой переменной, то можно для каждой такой корреляции брать во внимание все наблюдения, не имеющие пропусков только по двум рассматриваемым переменным, игнорируя, что по остальным регрессорам эти наблюдения могут иметь пропуски. Напротив, при использовании АПН в регрессионном анализе в расчете каждой корреляции участвуют только те наблюдения, которые не имеют пропусков и по остальным регрессорам.

Импутация — большая группа методов, к которой, в частности, относятся импутация на основе меры центральной тенденции (скажем, среднего арифметического значения) [Little, Rubin, 2002], множественной линейной регрессии [там же], деревьев решений [Ratner, 2011], так называемых hot-deck и cold-deck алгоритмов [Little, Rubin, 2002], логлинейного анализа [Schafer, 1997], кластеризации и латентно-классового анализа [Vermunt et al., 2008], анализа соответствий [Greenacre, Pardo, 2006] и т. д.

Метод индикаторной переменной (далее — МИП) предполагает добавление в переменную, содержащую пропуски, нового значения, отличного от всех исходных и маркирующего пропуски, а также создание в пару к ней новой бинарной или индикаторной переменной, показывающей наличие или отсутствие пропуска в соответствующих ей исходной переменной (столбец) и наблюдении (строка).

Среди перечисленных четырех подходов к обработке пропусков мы фокусируемся на сравнении АПН и МИП и **не** рассматриваем анализ доступных наблюдений и импутацию, поскольку применимость последних несколько ограничена. Дело в том, что анализ доступных наблюдений по самой своей сути релевантен только в контексте методов анализа данных, вычисления в которых можно разбить на этапы с участием двух переменных (скажем, меры парной связи, регрессия и т. п.) [Фабрикант, 2015]. В свою очередь, согласно многочисленным исследованиям, методы импутации релевантны только для данных, пропуски в которых относятся к механизму СП [Bartlett et al., 2014; Hughes et al., 2019]. Кроме того, методы импутации крайне многообразны и выбор из них следует основывать в том числе на содержательной специфике анализируемых данных [Akande, Li, Reiter, 2017], а симулированные данные для планируемого нами эксперимента очищены от содержательной специфики.

Для упорядочения теоретического сравнения мы вводим три критерия:

1) сохраняется ли в рамках каждого подхода исходный размер выборки;

DOI: 10.14515/monitoring.2021.4.940

2) чувствительны ли результаты применения подходов к типу шкалы переменной с пропусками (категориальная и континуальная);

3) чувствительны ли они к механизму порождения пропусков (ПСП, СП, НП).

Исходя из своего названия, АПН не сохраняет исходный размер выборки. Этот вид анализа не чувствителен к типу шкалы, поскольку при его использовании происходит удаление наблюдений с пропусками независимо от того, в переменной с каким типом шкалы пропуски находятся. В отношении критерия чувствительности к механизму порождения пропусков многие авторы считают, что анализ полных наблюдений допустим исключительно при условии ПСП, поскольку «только в этом случае подвыборка полных наблюдений может репрезентировать исходную выборку» [Ratner, 2011: 270]. Однако, как было замечено еще П. Эллисоном [Allison, 2005] и подтверждено недавними исследованиями [Bartlett et al., 2014; Bartlett, Harel, Carpenter, 2015; Hughes et al., 2019], эта точка зрения неверна: как минимум в контексте регрессионного моделирования применение АПН обычно не приводит к смещениям оценок параметров модели. Смещения могут возникнуть, лишь если пропуски в регрессоре обусловлены значениями зависимой переменной. Рассмотрим варианты этой ситуации. Пропуски в некотором регрессоре могут обусловливаться, во-первых, значениями только зависимой переменной (механизм СП) или же, во-вторых, и значениями зависимой переменной, и значениями самого этого регрессора (механизм НП), плюс те же два варианта, но с добавлением обусловленности пропусков в регрессоре значениями других регрессоров (тоже механизмы СП и НП соответственно). Было установлено, что первые два варианта могут привести к смещению оценки только константы, а вторые два — еще и к смещению оценок регрессионных коэффициентов при регрессорах [Bartlett et al., 2015: 732]. Тем не менее генерализовать эти противоречивые свидетельства contra и pro касательно почти универсальной применимости АПН следует с осторожностью, так как все они основаны на эмпирических, а не на симулированных данных.

МИП новее АПН. Возникнув в медицинских исследованиях из области этиологии [Miettinen, 1985], этот метод снискал популярность и в других науках, в том числе компьютерных и социальных, поскольку имеет ряд преимуществ по сравнению с АПН и другими конкурирующими подходами. Так, он сохраняет исходный размер выборки и, следовательно, применяемые после него методы анализа данных сохраняют свою статистическую мощность. При этом он более прост в применении и интерпретации по сравнению с импутацией (особенно множественной), которая тоже сохраняет исходный размер выборки [Groenwold et al., 2012]. Особенно это преимущество ценится в компьютерных науках, где распространены исследования на больших данных, в которых доля пропусков часто превышает 50 %, а иногда достигает и 90 % [Anagnostopoulos, Triantafillou, 2014]. Тогда ради сохранения исходного размера выборки прибегают именно к методу индикаторной переменной, причем часто по умолчанию, скажем, при применении деревьев решений [Gentle, Hardle, Mori, 2012; Rokach, Maimon, 2010]. В социальных науках этот подход тоже прижился — можно составить обширный список исследований, в которых он используется (например, [Chen, Hossler, 2017; Gesser-Edelsburg et al., 2018; Rickles

DOI: 10.14515/monitoring.2021.4.940

et al., 2018; Trevizo, Lopez, 2016; Weiss et al., 2017; Zhelyazkova, Ritschard, 2018; Стребков и др., 2019] и т. д.).

В контексте регрессионного моделирования сохранение исходного размера выборки выглядит заманчиво, поскольку при большом числе регрессоров даже умеренная доля пропусков в каждом из них может привести к сокращению анализируемой выборки в разы, если использовать АПН. Что же касается МИП, то он гипотетически позволяет достичь баланса между сохранением исходного размера выборки и избеганием смещения оценок статистик [van Kuijk et al., 2016].

По второму критерию нашего анализа МИП претендует на нечувствительность к типу шкалы переменной с пропусками. Его применение выглядит особенно гармоничным по отношению к распространенным в социальных науках категориальным признакам, а интерпретация индикаторной переменной в социологическом контексте приобретает дополнительное содержание. Дело в том, что математически, с точки зрения частотного распределения значений (валидных и пропусков), именно категориальные признаки выглядят предпочтительными для использования метода индикаторной переменной, по крайней мере в рамках регрессионного моделирования [Donders et al., 2006]. Действительно, в случае категориального регрессора каждая его категория должна быть превращена в бинарную (фиктивную) переменную, и новая бинарная переменная (индикаторная) оказывается вполне однородна с ними. Напротив, в рамках континуального регрессора при использовании МИП все пропущенные значения концентрируются в один сильно выбивающийся из общего распределения столбец, что приводит к изменению параметров распределения регрессора относительно истинных. Однако поиск работ, в которых предпринято сравнение МИП и АПН в их применении к категориальным данным, почти не дал результатов [5].

Дискуссионными вопросами остаются нечувствительность МИП к механизму порождения пропусков и в целом его способность давать неискаженные оценки. Согласно одним методологическим исследованиям, метод дает наиболее смещенные оценки параметров по сравнению со всеми конкурирующими подходами [Donders et al., 2006; van der Heijden et al., 2006; Henry et al., 2013; Jones, 1996; Knol et al., 2010]. Согласно другим, он, напротив, обеспечивает несмещенные оценки [Groenwold et al., 2012; White, Thompson, 2005]. Р. Грёнволд и соавторы [Groenwold et al., 2012] показывают, что МИП не приводит к смещениям, если применяется в рамках рандомизированных испытаний (в случае этой статьи — медицинских), в которых пропуски обычно не связаны с другими изучаемыми переменными и внешними факторами, и, напротив, приводит к смещениям, если применяется в рамках нерандомизированных испытаний, где велика вероятность иметь неслучайные пропуски. В терминах типологии Л. Рубина [Rubin, 1976] первая ситуация относится к механизму ПСП, а вторая к СП и НП. Если эти результаты верны и не обусловлены только рандомизацией или ее отсутствием, то они представляются противоречащими набирающей популярность в социальных науках логике наделять социальным (или социально-психологическим) смыслом пропущенные значения. Согласно этой логике, пропуски в социологиче-

[5] Исключение составляет статья [Henry et al., 2013], в которой пропущенные значения содержатся в категориальной переменной «раса»; но эта работа основана на реальных данных, и в ней не контролируются никакие факторы, влияющие на результат сравнения.

DOI: 10.14515/monitoring.2021.4.940

ских данных часто обусловлены желанием информантов скрыть сенситивную или нежелательную информацию. Если исследователь помечает такое значение специальным кодом и интерпретирует не как отсутствие знания об информанте, а как намек на наличие у него скрываемой характеристики, то пропущенное значение перестает быть пропущенным, и математическая модель на этих данных должна иметь те же параметры, как если бы пропусков в данных не было вовсе. Например, в недавнем исследовании одной из онлайн-платформ для фрилансеров [Стребков и др., 2019] авторы допустили, что фрилансеры намеренно не заполняют некоторые социально-демографические пункты своего профиля, такие как пол, возраст, страна проживания, чтобы не столкнуться с дискриминацией и повысить шансы получения заказов. Описанная логика нашла подтверждение и в результатах статистического эксперимента, проведенного С. Жучковой и А. Ротмистровым [Жучкова, Ротмистров, 2018]. Так, когда внесенные вероятностным образом в категориальный предиктор дерева классификации пропуски концентрировались вместе с каким-либо из исходных валидных значений этого предиктора, в дереве наблюдалось наименьшее искажение его первоначальной структуры, правда, при условии, что доля пропусков невелика, а предиктор оказывался далеко от корня дерева.

Таким образом, относительно применимости МИП к пропускам, обусловленным разными механизмами их порождения, складывается своего рода научный пазл, требующий новых исследований. Не следует забывать, что и АПН долгое время позиционировался как метод, чувствительный к механизму порождения пропусков [Little, Rubin, 2002]. Затем его чувствительность была поставлена под сомнение [Allison, 2005] и, наконец, несколько раз опровергнута [Bartlett et al., 2014, 2015; Hughes et al., 2019].

Противоречивость научных представлений о нечувствительности изучаемых подходов к механизму порождения пропусков и в целом их способность давать неискаженные оценки может быть объяснена рядом причин. Во-первых, отсутствует стандартизированная методология сравнения методов обработки пропусков. Во-вторых, большинство[6] проведенных сравнений базировались на эмпирических данных, причем на континуальных медицинских, поэтому в результатах могли отразиться и свойства методов, и особенности самих данных. Методологических исследований на симулированных данных крайне мало (например, [Choi, Dekkers, le Cessie, 2018; Жучкова, Ротмистров, 2018]), в них рассмотрены либо континуальные переменные, либо один механизм порождения пропусков. Цель нашей статьи — применить изучаемые подходы к категориальному регрессору, в том числе при наличии в модели континуального регрессора в разрезе всех механизмов порождения пропусков.

Методология

Исследование проводится в формате статистического эксперимента, идея которого заключается в том, чтобы задать исходные (истинные) параметры изучаемых моделей и переменных и выявлять вариацию оценок этих параметров под воздействием какого-либо фактора, контролируя все остальные факторы.

[6] Исключения составляют статьи [Choi et al., 2018] и [Donders et al., 2006], в которых симулируются данные, но в этих статьях рассмотрены только континуальные переменные.

DOI: 10.14515/monitoring.2021.4.940

Мы используем линейную регрессионную модель[7], метод наименьших квадратов. Зависимая переменная является континуальной, регрессоры — категориальной и континуальной переменными. В качестве исходно задаваемых параметров выступают параметры распределения переменных и регрессионные коэффициенты. В ходе эксперимента мы выявляем вариации их оценок, а также стандартных ошибок под воздействием трех факторов: механизма порождения пропусков, доли пропусков и спецификации регрессионной модели. Далее мы раскроем эти факторы и обоснуем важность их контроля.

1) Исходя из **механизма порождения пропусков** выделяют *полностью случайные пропуски (ПСП), случайные пропуски (СП)* и *неслучайные пропуски (НП).* Раскрытая в предыдущем разделе методологическая традиция учитывать механизм порождения пропусков и противоречивость представлений о нечувствительности изучаемых подходов к этому механизму делают данный фактор ключевым в нашем эксперименте.

2) **Доля пропусков** рассматривается нами в трех значениях: *10%, 25%* и *50%.* Интуитивно ясно, что чем выше доля, тем труднее изучаемым подходам обработать пропуски корректно. Однако научная позиция требует измерить степень «ухудшения» работы каждого из них. Так, авторы данной статьи уже задавались этим вопросом, но относительно дерева решений CHAID, а не множественной линейной регрессии, и установили, что при доле пропусков 10% и 25% использование МИП позволяет получить примерно ту же структуру дерева, что и в случае отсутствия пропусков [Жучкова, Ротмистров, 2018]. Если проводить аналогию с регрессией, это соответствовало бы ситуации отсутствия серьезных смещений оценок параметров регрессионной модели. Однако при 50% пропусков дерево имеет крайне искаженную структуру, что ведет к ошибочным содержательным результатам [там же]. Наше предыдущее исследование тоже основывалось на статистическом эксперименте и симулированных данных, но рассматривался только один механизм пропусков — ПСП [там же]. В настоящей работе мы проверяем релевантность выводов, сделанных в 2018 г., в контексте множественной линейной регрессии и дополняем их рассмотрением остальных механизмов пропусков, а также применением АПН.

3) В рамках фактора **спецификация регрессионной модели** выделяют *модель с категориальным регрессором; с категориальным и континуальным регрессорами; с категориальным и континуальным регрессорами и эффектами их взаимодействия*. В этой статье мы делаем акцент на категориальном регрессоре как более релевантном для социальных наук и как менее изученном в контексте проблемы пропусков по сравнению с континуальными регрессорами. При этом совсем отказаться от рассмотрения последних было бы серьезным упрощением, далеким от реальной практики. Наконец, в наш эксперимент включены и эффекты взаимодействия категориального и континуального регрессоров, поскольку в объяснении социальных явлений, характеризующихся комплексностью, эффекты взаимодействия зачастую играют бóльшую роль, чем главные эффекты [Morgan, Sonquist, 1963].

Статистический эксперимент возможен на реальных и на симулированных данных. Мы выбрали второе, чтобы не зависеть от возможной специфики реальных

[7] В ее рамках все чаще применяется МИП, хотя АПН сохраняет свой приоритет. Конкурируют с регрессией в этом отношении разве что деревья решений.

DOI: 10.14515/monitoring.2021.4.940

эмпирических данных, которая могла бы привнести вариацию оценок параметров, не связанную с экспериментальными факторами, а также чтобы не сужать генерализацию выводов из эксперимента до некоторой содержательной области. Этим наше исследование отличается от абсолютного большинства описанных выше работ, в частности тех, в которых демонстрируется универсальность АПН [Bartlett et al., 2014; Bartlett, Harel, Carpenter, 2015; Hughes et al., 2019].

Процесс симуляции данных[8]

Для регрессионного моделирования процесс симуляции данных начинается с симуляции регрессоров и переменной, отвечающей за случайную ошибку, которые затем подставляются в регрессионное уравнение нужного семейства функций, включающее константу. Для каждого из регрессоров и константы задаются коэффициенты, и на основании них, а также переменной, отвечающей за случайную ошибку, рассчитываются значения зависимой переменной. Для нашего эксперимента регрессоры и случайная ошибка симулируются, исходя из следующих теоретических распределений:

$$X \sim U(1, 3),$$

$$Z \sim N(0, 1),$$

$$\varepsilon \sim N(0, 25).$$

Поясним нотацию: категориальный регрессор «*X*» симулируется из теоретического дискретного равномерного распределения с тремя категориями: «1», «2», «3». Почему из равномерного? Дело в том, что для включения в регрессию категориальный регрессор обычно превращается в набор бинарных фиктивных переменных, один из недостатков которых — частотное преобладание категории «Другое» (обычно кодируемой как «0») над значащей категорией (обычно кодируемой как «1»). Наименее выражена эта диспропорция в случае, если категориальный регрессор распределен равномерно. Кроме того, равномерное распределение вполне характерно для исходных номинальных переменных. Континуальный регрессор «*Z*» симулируется из теоретического нормального распределения с математическим ожиданием 0 и стандартным отклонением 1. Переменная, отвечающая за случайную ошибку, симулируется из теоретического нормального распределения с математическим ожиданием 0 и стандартным отклонением 5.

Чтобы симулировать зависимую переменную «*Y*» для каждой спецификации (с категориальным регрессором; с категориальным и континуальным регрессорами; с категориальным и континуальным регрессорами и их эффектами взаимодействия), для трех созданных переменных (регрессор «*X*» дихотомизируется с отнесением в референтную группу категории «1») задаются истинные значения регрессионных коэффициентов, отличия от которых после многократного внесения в данные пропусков и их обработки и выступают предметом анализа.

[8] Все симуляции осуществлялись на языке программирования Python средствами пакетов «numpy» [Oliphant, 2006; van der Walt, Colbert, Varoquaux, 2011] и «pandas» [McKinney, 2010], а регрессионное моделирование — средствами пакета «statsmodels» [Seabold, Perktold, 2010].

DOI: 10.14515/monitoring.2021.4.940

В нашем эксперименте истинные значения регрессионных коэффициентов были следующие:

$$Y \sim 100 + 15 \cdot X_2 - 15 \cdot X_3 + \varepsilon,$$

$$Y \sim 100 + 15 \cdot X_2 - 15 \cdot X_3 + 12 \cdot Z + \varepsilon,$$

$$Y \sim 100 + 15 \cdot X_2 - 15 \cdot X_3 + 12 \cdot Z - 20 \cdot Z \cdot X_2 - 10 \cdot Z \cdot X_3 + \varepsilon.$$

В течение эксперимента регрессоры и переменная, отвечающая за случайную ошибку, а также зависимая переменная для каждой из трех спецификаций симулируются многократно и включают 2000 наблюдений, поскольку в социологических исследованиях обычно используются выборки такого или меньшего размера[9]. Назовем итерацией симуляцию одного набора переменных и создание из него наборов данных с пропусками разных механизмов и долей для АПН и МИП. Получается, что «физически» итерация охватывает 54 набора данных: $3 \times 3 \times 3$ уровней экспериментальных факторов $\times 2$ изучаемых подхода.

Пропуски трех разных механизмов вносятся только в категориальный регрессор «*X*» и не вносятся в континуальный регрессор «*Z*», поскольку последнее усложнило бы дизайн эксперимента и затруднило бы интерпретацию его результатов. Это будет разумно после того, как станут ясны особенности работы изучаемых подходов в категориальном регрессоре в разрезе экспериментальных факторов.

Пропуски механизма НП вносятся в «*X*» путем вероятностной[10] замены двух его значений («1» и «2») из трех на пропущенные, причем при таком внесении пропусков не делается различие между этими значениями. Пропуски механизма СП вносятся, ориентируясь на значение «1» вспомогательной бинарной переменной с биномиальным распределением «*D*» ~ *Bin*(2000, 0,8), пропуски механизма ПСП — путем вероятностной[11] замены любых его значений на пропущенные.

Пропуски вносятся для всех трех спецификаций, то есть три раза по три механизма — НП, СП и ПСП, причем для каждого механизма пропуски вносятся в 10 % наблюдений, в 25 % и в 50 % — итого, как было отмечено, на каждой итерации симулируются 27 наборов данных, к каждому из которых применяются параллельно АПН и МИП. На основе полученных 54 массивов рассчитываются точечные оценки регрессионных коэффициентов, их стандартные ошибки и доверительные интервалы — итого на каждой итерации получается 54 набора указанных показателей.

Важная ремарка — в рамках одной итерации каждая симулируемая спецификация имеет свой истинный образец, то есть симулируемая спецификация с категориальным регрессором симулируется на основе истинной спецификации с категориальным регрессором, симулируемая спецификация с категориальным и континуальным регрессорами симулируется на основе истинной спецификации с категориальным и континуальным регрессорами и т. д. Альтернативным мог бы быть

[9] В текст статьи не вошло дополнение к эксперименту, в котором размер совокупности был снижен до 75 наблюдений (фактически это дополнительный экспериментальный фактор). Результаты основного эксперимента подтвердились при таком снижении.

[10] Вероятности соответствуют уровням фактора «Доля пропусков».

[11] То же.

DOI: 10.14515/monitoring.2021.4.940

дизайн, предполагающий одну истинную спецификацию, скажем, с категориальным и континуальным регрессорами и эффектами их взаимодействия, на основе которой симулировались бы три изучаемые спецификации: с категориальным регрессором, с категориальным и континуальным регрессорами, с категориальным и континуальным регрессорами и их эффектами взаимодействия. Вероятно, наш дизайн несколько более далек от реальной практики, чем альтернативный, так как в реальной практике истинная спецификация априорно неизвестна. Однако в альтернативном дизайне фактически вводился бы дополнительный экспериментальный фактор — совпадение истинной и симулируемой спецификаций, что усложнило бы его. Мы же предпочитаем двигаться от простого к сложному, отдавая себе отчет в ограничениях своей работы, которые можно преодолеть в будущем.

Наш эксперимент включает 2 000 итераций[12], в результате чего получается 2 000 наборов точечных оценок, к которым применяются специальные метрики (о них ниже в этом разделе), позволяющие судить об искажении оценок. Выбранный дизайн эксперимента, таким образом, представляет собой симуляции Монте-Карло. Пошагово эксперимент выглядит следующим образом:

1) Выбрать одну из трех спецификаций (далее — «текущая спецификация»).

2) Вероятностным образом симулировать переменные «*X*», «*Z*», ε и «*D*» в объеме 2 000 наблюдений из заданных теоретических распределений.

3) Найти взвешенную сумму переменных «*X*», «*Z*» и ε с учетом истинных для текущей спецификации регрессионных коэффициентов, чтобы получить значения зависимой переменной «*Y*». Регрессор «*X*» для этого дихотомизируется, а дихотомизированная переменная «X_1» (отнесенная к референтной группе) не участвует в расчете значений зависимой переменной.

4) Выбрать долю пропусков (далее — «текущая доля пропусков») и механизм их порождения (далее — «текущий механизм»).

5) Внести пропуски в регрессор «*X*» согласно текущим механизму и доле пропусков. Если текущий механизм НП, то для внесения пропусков вероятностно (согласно текущей доле пропусков) выбрать те наблюдения, у которых по этой переменной имеются значения «1» и «2». Если текущий механизм СП, то ориентироваться на вспомогательную бинарную переменную «*D*»: для внесения пропусков в регрессор «*X*» вероятностно (согласно текущей доле пропусков) выбрать те наблюдения, у которых вспомогательная переменная имеет значение «1». Если текущий механизм ПСП, то для внесения пропусков в регрессор «*X*» вероятностно (согласно текущей доле пропусков) заменить любые его значения на пропущенные. Важная ремарка: поскольку между категориальным регрессором и зависимой переменной задана умеренная корреляция, а между ним и вторым регрессором — незначимая корреляция, то при текущем механизме НП пропуски в категориальном регрессоре связаны как с его собственными значениями, так и со значениями зависимой переменной. Следовательно, реализуется вариант, описанный в предыдущем разделе, который должен приводить к смещениям оценок регрессионных коэффициентов для АПН. При текущем механизме СП пропуски в категориальном регрессоре **не** связаны со значениями зависимой переменной, следовательно,

[12] В русле традиции компьютерных наук симулировать число итераций, примерно равное объему совокупности.

DOI: 10.14515/monitoring.2021.4.940

при таком варианте смещения оценок регрессионных коэффициентов для АПН **не** должны возникать. Таким образом, дизайн нашего эксперимента покрывает сразу две альтернативы, в которых для АПН следует ожидать противоположных исходов: присутствия смещения при механизме НП и отсутствия смещения при механизме СП. В этом отношении наш дизайн выглядит близким к реальной практике — он не покрывает только ситуацию выраженной связи между двумя регрессорами, но и в реальной практике исследователи стараются избегать таких ситуаций как коллинеарных.

6) Из массива данных с пропусками сделать две копии: одна для применения анализа полных наблюдений (далее — массив АПН), другая — для метода индикаторной переменной (далее — массив МИП). Из массива АПН наблюдения с пропусками удалить, а в массиве МИП каждое пропущенное значение заменить на категорию «4».

7) На массивах АПН и МИП построить регрессионную модель, соответствующую текущей спецификации; «записать» оценки коэффициентов, их стандартные ошибки и доверительные интервалы в базу с результатами.

8) Повторить шаги 1—7 для каждой комбинации уровней экспериментальных факторов. При этом проконтролировать, что начальное значение генератора случайных чисел (random seed) — одно и то же для всех комбинаций уровней экспериментальных факторов.

9) Выполнить 2 000 итераций для шагов 1—8, фиксируя для каждой новой итерации новое начальное значение генератора случайных чисел. Правило такое: для каждой итерации свое начальное значение. Это позволяет воспроизводить эксперимент повторно и непосредственно сравнивать оценки регрессионных параметров внутри итерации.

Результаты экспериментов анализируются с помощью традиционных для симуляционных исследований метрик [Morris et al., 2019]: среднего (mean или estimate), смещения (bias), среднеквадратичной ошибки (MSE) и покрытия доверительного интервала (coverage или coverage probability; далее — «покрытие ДИ»). Среднее — среднее арифметическое оценок анализируемого параметра (скажем, регрессионного коэффициента) по всем симуляциям, относящимся к рассматриваемому уровню фактора. Смещение — разница между рассчитанным средним и истинным значением параметра, которое задавалось при симуляции данных. Среднеквадратичная ошибка — средний квадрат отклонений всех оценок от истинного значения параметра. Смещение отражает отклонение оценок от истинного значения, а среднеквадратичная ошибка — еще и вариацию этого отклонения. Покрытие ДИ — доля доверительных интервалов, которые включают в себя истинное значение параметра, задаваемое при генерации данных. Эта метрика может говорить как о несмещенности оценки, так и о ее эффективности (возвращаясь к изложенному в разделе «Введение»). Как показано в работе К. Доугерти [Dougherty, 2016], эти два свойства оценки могут приходить в противоречие, которое проявляется в интерпретации покрытия ДИ: малая ее величина вызывается тем, что истинное значение параметра попадает в малое число доверительных интервалов, следовательно, свидетельствует о смещенности; слишком большая ее величина обусловливается тем, что доверительные интервалы слишком широки из-за высокой стандартной ошибки

DOI: 10.14515/monitoring.2021.4.940

оценки, следовательно, свидетельствует о неэффективности. Условной границей между этими двумя ситуациями принято считать значение покрытия 0,95 [Morris et al., 2019]. Для оценок регрессионных коэффициентов рассчитываем все четыре метрики, а для стандартных ошибок коэффициентов — только среднее.

Ориентирами для интерпретации результатов эксперимента выступают гипотезы, опирающиеся на проанализированные выше статьи и оперирующие уровнями экспериментальных факторов:

Гипотеза 1 (механизм пропусков). *Анализ полных наблюдений приводит к меньшим смещениям оценок регрессионных коэффициентов в случае ПСП и СП, а метод индикаторной переменной — только в случае НП.* Часть этой гипотезы касательно АПН опирается на упомянутые в предыдущем разделе исследования этого подхода, показавшие, что при его использовании в рамках регрессии решающий фактор смещения оценок — это обусловленность пропуска в регрессоре значением зависимой переменной [Bartlett et al., 2014, 2015; Hughes et al., 2019]. Часть гипотезы относительно МИП объясняется описанной в том же разделе логикой, что именно неслучайный характер пропусков может оправдывать создание для них специальной категории.

Гипотеза 2 (доля пропусков). *С увеличением доли пропусков в данных оба подхода (АПН и МИП) приводят к более смещенным оценкам, но для метода индикаторной переменной смещение растет быстрее, чем для анализа полных наблюдений.* Эта гипотеза опирается на исследование М. Джонса [Jones, 1996], где в целом показывается, что МИП дает более смещенные оценки, чем АПН.

Гипотеза 3 (спецификация модели). *Чем сложнее спецификация регрессионной модели, тем более смещенные оценки ее параметров дает метод индикаторной переменной.* Эта гипотеза сформулирована, исходя из следующей логики: наличие индикаторной переменной может влиять на значения коэффициентов всех исходных переменных, поэтому чем больше в модели исходных переменных, тем больше потенциал этого влияния. Для АПН мы не ожидаем увидеть вариацию смещения при разных спецификациях, поскольку при симуляции данных исключили связь между наличием/отсутствием пропуска в категориальном регрессоре с одной стороны и континуальным регрессором и эффектами взаимодействия — с другой. Как мы уже подчеркивали, такая связь может приводить к смещению константы и коэффициентов регрессоров при условии связи наличия/отсутствия пропуска и зависимой переменной [Bartlett et al., 2015].

Во всех трех гипотезах речь идет именно о метрике смещения. В этом отношении мы придерживаемся традиции, подмеченной Т. Моррисом и соавторами: «зачастую смещение [bias] — центральный показатель результатов применения подхода. Частотная интерпретация вероятности опирается на несмещенность [unbiasedness] как на ключевое свойство статистической оценки» [Morris et al., 2019: 2086].

Результаты

На каждой из 2000 итераций эксперимента возникает по 27 наборов оценок параметров для АПН (анализа полных наблюдений) и МИП (метода индикаторной переменной), соответствующих комбинациям уровней экспериментальных факторов. Таким образом, результат статистического эксперимента выглядит как

DOI: 10.14515/monitoring.2021.4.940

массив из (27 + 27) × 2000 строк, а столбцами этого массива выступают оценки изучаемых параметров (регрессионных коэффициентов и их стандартных ошибок). Для удобства рассмотрения гипотез исследования, отталкивающихся от факторов, полученный массив агрегируется в разрезе каждого фактора.

В таблице 1 представлены три метрики (смещение, среднеквадратичная ошибка и покрытие ДИ) для оценок всех регрессионных коэффициентов в разрезе *механизмов* пропусков и изучаемых подходов, то есть информация для проверки **гипотезы 1**, согласно которой для механизмов пропусков ПСП и СП ожидаются более низкие абсолютные величины смещения при использовании АПН, чем МИП, а для механизма НП ожидаются, наоборот, более высокие абсолютные величины смещения при использовании АПН, чем МИП. В реальности при использовании обоих подходов смещение пренебрежимо мало, а получаемые оценки максимально близки к истинным. Например, смещение оценки коэффициента при регрессоре «X_2» для обоих подходов и всех механизмов пропусков варьирует от 0,002 до 0,007, что составляет 0,01 %—0,05 % от истинного значения параметра. Нет разницы между подходами и в степени вариации оценок коэффициентов. Следовательно, **гипотеза 1 не принимается**.

Таблица 1. ***Смещение, среднеквадратичная ошибка и покрытие ДИ оценки регрессионных коэффициентов в разрезе механизмов пропусков***

Параметр	Механизм	ПСП		СП		НП	
		АПН	МИП	АПН	МИП	АПН	МИП
Константа	Смещение	–,001	–,001	–,001	–,001	–,007	–,007
Среднеквадратичная ошибка		,056	,056	,056	,056	,088	,088
Покрытие ДИ (%)		,953	,996	,947	,996	,948	,987
X_2	Смещение	,007	,007	,003	,002	,007	,007
Среднеквадратичная ошибка		,112	,112	,110	,110	,177	,177
Покрытие ДИ (%)		,950	,996	,954	,996	,946	,985
X_3	Смещение	,003	,003	–,004	–,004	,005	,005
Среднеквадратичная ошибка		,108	,108	,112	,112	,125	,125
Покрытие ДИ (%)		,955	,996	,954	,997	,950	,988
Z	Смещение	,003	,003	,001	,002	,001	,003
Среднеквадратичная ошибка		,037	,044	,037	,045	,052	,053
Покрытие ДИ (%)		,952	,972	,952	,976	,952	,973
$Z \cdot X_2$	Смещение	,000	,000	,001	,001	–,007	–,007
Среднеквадратичная ошибка		,105	,105	,108	,108	,168	,168
Покрытие ДИ (%)		,958	,997	,955	,997	,953	,995
$Z \cdot X_3$	Смещение	–,005	–,005	–,002	–,002	–,002	–,002
Среднеквадратичная ошибка		,115	,115	,115	,115	,125	,125
Покрытие ДИ (%)		,944	,997	,942	,998	,946	,996

DOI: 10.14515/monitoring.2021.4.940

Различие между подходами обнаружилось в величине покрытия ДИ[13]: для АПН оно в пределах нормы, то есть около 0,95, а для МИП превышает 0,95 и приближается к 1. Как мы упоминали, слишком высокая величина покрытия ДИ свидетельствует о слишком широких доверительных интервалах и, при отсутствии смещения (а в его отсутствии мы только что убедились), — о завышении стандартной ошибки. При завышенной стандартной ошибке оценки коэффициента последняя становится менее эффективной, чем была бы при отсутствии пропусков, из-за чего регрессор выглядит как менее статистически значимый. Завышение стандартных ошибок при использовании МИП характерно для всех трех механизмов пропусков.

МИП предполагает создание индикаторной переменной, которая при его применении включена в регрессию как главный эффект и, в соответствующей спецификации, — как часть эффекта взаимодействия. Коэффициенты при этих эффектах не имеют истинного значения, так как их нет в модели при отсутствии пропусков. Однако математически можно считать, что при отсутствии пропусков истинные значения коэффициентов равны нулю. Как показал наш эксперимент, даже если пропуски есть, но относятся к механизмам ПСП и СП, то средние оценок коэффициентов главных эффектов индикаторной переменной примерно равны нулю, а эффектов взаимодействия — существенно отклоняются от нуля. Для пропусков механизма НП оценки коэффициентов и главных эффектов, а также эффектов взаимодействия существенно отклоняются от нуля. Мы интерпретируем это так, что оценкам коэффициентов при индикаторной переменной удается отразить неслучайный характер пропусков.

В таблице 2 представлены те же метрики в разрезе изучаемых подходов и *долей* пропусков, то есть информация для проверки **гипотезы 2**, согласно которой с увеличением доли пропусков в данных ожидается рост абсолютных величин смещения при использовании АПН и МИП, причем при использовании второго — быстрее. Наблюдается ли этот рост? В общем случае — нет, монотонного увеличения не наблюдается, смещение увеличивается только для континуального регрессора «B» (причем интенсивнее именно в случае МИП). Однако сама величина смещения, вероятно, слишком мала, чтобы принимать ее во внимание, поэтому мы заключаем, что доля пропусков не сказывается на смещении регрессионных коэффициентов, то есть **отвергаем гипотезу 2**. Как видно, в разрезе долей пропусков повторяются закономерности, выявленные ранее в контексте механизмов пропусков: смещения не наблюдаются, вариация оценок между подходами не отличается, при этом доверительные интервалы оценок обладают гораздо более высоким покрытием в случае МИП, чем в случае АПН, снова свидетельствуя о проблеме завышения стандартных ошибок при применении первого подхода.

В таблице 3 представлены те же метрики в разрезе изучаемых подходов и *спецификаций моделей*, то есть информация для проверки **гипотезы 3**, согласно которой с усложнением спецификации (слева направо) ожидается рост абсолютных величин смещения при использовании МИП. Наблюдается ли рост этого смещения? Нет, поэтому **гипотеза 3 отвергается**.

[13] В гипотезе речи о покрытии ДИ не было.

DOI: 10.14515/monitoring.2021.4.940

Таблица 2. ***Смещение, среднеквадратичная ошибка и покрытие ДИ оценок регрессионных коэффициентов в разрезе долей пропусков***

Параметр	Доля	10%		25%		50%	
		АПН	МИП	АПН	МИП	АПН	МИП
Константа	Смещение	–,003	–,003	–,005	–,005	–,002	–,002
Среднеквадратичная ошибка		,041	,041	,054	,054	,104	,104
Покрытие ДИ (%)		,952	,986	,949	,994	,947	,999
X_2	Смещение	,007	,007	,009	,009	,000	,000
Среднеквадратичная ошибка		,085	,085	,107	,107	,206	,206
Покрытие ДИ (%)		,952	,985	,950	,995	,948	,998
X_3	Смещение	,001	,001	,003	,003	,000	,000
Среднеквадратичная ошибка		,081	,081	,100	,100	,164	,164
Покрытие ДИ (%)		,956	,986	,950	,996	,953	,999
Z	Смещение	,001	,002	,002	,003	,002	,004
Среднеквадратичная ошибка		,028	,030	,035	,040	,063	,071
Покрытие ДИ (%)		,953	,972	,952	,975	,950	,974
$Z \cdot X_2$	Смещение	,000	,000	,000	,000	–,007	–,007
Среднеквадратичная ошибка		,083	,083	,102	,102	,196	,196
Покрытие ДИ (%)		,957	,991	,958	,999	,952	1,000
$Z \cdot X_3$	Смещение	,000	,000	–,002	–,002	–,006	–,006
Среднеквадратичная ошибка		,085	,085	,102	,102	,167	,167
Покрытие ДИ (%)		,944	,991	,945	1,000	,943	1,000

Таблица 3. ***Смещение, среднеквадратичная ошибка и покрытие ДИ оценок регрессионных коэффициентов в разрезе спецификаций модели***

Параметр	Спецификация*	1		2		3	
		АПН	МИП	АПН	МИП	АПН	МИП
Константа	Смещение	–,003	–,003	–,003	–,003	–,003	–,003
Среднеквадратичная ошибка		,066	,066	,066	,066	,066	,066
Покрытие ДИ (%)		,949	,991	,949	,991	,950	,997
X_2	Смещение	,006	,006	,006	,006	,006	,006
Среднеквадратичная ошибка		,133	,133	,133	,133	,133	,133
Покрытие ДИ (%)		,950	,991	,950	,991	,951	,996
X_3	Смещение	,001	,001	,001	,001	,001	,001
Среднеквадратичная ошибка		,115	,115	,115	,115	,115	,115
Покрытие ДИ (%)		,952	,992	,953	,992	,953	,997

DOI: 10.14515/monitoring.2021.4.940

Параметр	Спецификация*	1		2		3	
		АПН	МИП	АПН	МИП	АПН	МИП
Z	Смещение	—	—	,001	,003	,002	,002
Среднеквадратичная ошибка		—	—	,019	,029	,065	,065
Покрытие ДИ (%)		—	—	,950	,950	,953	,998
$Z \cdot X_2$	Смещение	—	—	—	—	–,002	–,002
Среднеквадратичная ошибка		—	—	—	—	,127	,127
Покрытие ДИ (%)		—	—	—	—	,956	,997
$Z \cdot X_3$	Смещение	—	—	—	—	–,003	–,003
Среднеквадратичная ошибка		—	—	—	—	,118	,118
Покрытие ДИ (%)		—	—	—	—	,944	,997

* *Примечание*: спецификация 1 — модель только с категориальным регрессором, спецификация 2 — модель с категориальным и континуальным регрессорами, спецификация 3 — модель с категориальным и континуальным регрессорами и их эффектами взаимодействия.

Та же картина наблюдается и для АПН. Таким образом, находит дополнительное подтверждение эмпирически выявленная закономерность о пренебрежимо малом смещении оценок регрессионных коэффициентов при обработке пропусков АПН и МИП — независимо от механизма пропусков в данных, их доли и сложности спецификации регрессионной модели. Как и ранее, видно, что подходы различаются только величинами покрытия ДИ: независимо от спецификации модели для МИП они и, следовательно, стандартные ошибки оценок коэффициентов существенно выше, чем для АПН.

Итак, ни одна из гипотез не нашла подтверждения. Помимо этого, эксперимент дал ряд важных результатов, которые касаются стандартных ошибок оценок регрессионных коэффициентов. Выше на основе величин покрытия ДИ мы сделали предварительный вывод, что применение МИП приводит к существенному увеличению стандартных ошибок оценок регрессионных коэффициентов. Средние значения стандартных ошибок в разрезе экспериментальных факторов приведены в таблице 4: они всегда в 1,5—2 раза выше для МИП, чем для АПН, что наглядно подтверждает упомянутый предварительный вывод.

Таблица 4. ***Средние стандартные ошибки оценок регрессионных коэффициентов в разрезе экспериментальных факторов***

В разрезе ***механизмов пропусков***						
Механизм	**ПСП**		**СП**		**НП**	
Параметр	**АПН**	**МИП**	**АПН**	**МИП**	**АПН**	**МИП**
Константа	,234	,404	,234	,403	,281	,406
X_2	,331	,571	,331	,571	,398	,575
X_3	,331	,571	,331	,571	,344	,492
Z	,185	,309	,185	,309	,209	,314

DOI: 10.14515/monitoring.2021.4.940

$Z \cdot X_2$	,332	,620	,332	,619	,400	,686
$Z \cdot X_3$	,332	,620	,332	,619	,346	,584
MIV	—	,572	—	,572	—	,541
$Z \cdot$ **MIV**	—	,619	—	,619	—	,638
В разрезе ***долей пропусков***						
Доля	**10%**		**25%**		**50%**	
Параметр	**АПН**	**МИП**	**АПН**	**МИП**	**АПН**	**МИП**
Константа	,206	,258	,231	,357	,312	,598
X_2	,292	,365	,326	,505	,441	,846
X_3	,288	,361	,315	,490	,403	,783
Z	,162	,205	,180	,281	,236	,446
$Z \cdot X_2$	,292	,389	,327	,562	,444	,974
$Z \cdot X_3$	,289	,385	,316	,543	,405	,896
MIV	—	,513	—	,499	—	,675
$Z \cdot$ **MIV**	—	,548	—	,554	—	,775
В разрезе ***спецификаций модели***						
Спецификация*	**1**		**2**		**3**	
Параметр	**АПН**	**МИП**	**АПН**	**МИП**	**АПН**	**МИП**
Константа	,250	,380	,250	,381	,250	,452
X_2	,353	,538	,353	,538	,353	,640
X_3	,335	,514	,335	,514	,336	,606
Z	—	—	,135	,168	,251	,454
$Z \cdot X_2$	—	—	—	—	,354	,642
$Z \cdot X_3$	—	—	—	—	,337	,608
MIV	—	,531	—	,531	—	,564
$Z \cdot$ **MIV**	—	—	—	—	—	,625

* *Примечание*: спецификация 1 — модель только с категориальным регрессором; спецификация 2 — модель с категориальным и континуальным регрессорами; спецификация 3 — модель с категориальным и континуальным регрессорами и их эффектами взаимодействия.

В таблице 4 заметны дополнительные закономерности. Во-первых, для обоих подходов характерно увеличение стандартных ошибок при росте доли пропусков. Для АПН это ожидаемо, ведь при росте доли пропусков уменьшается размер анализируемой выборки, для МИП такое объяснение не применимо. Во-вторых, при использовании обоих подходов стандартные ошибки выше, когда пропуски внесены согласно механизму НП, нежели ПСП и СП. Исключение составляют стандартные ошибки оценок коэффициентов регрессоров X_3 и $Z \cdot X_3$, которые ниже, когда пропуски внесены согласно механизму НП, нежели ПСП и СП. Особенность этих регрессоров в том, что они включают в себя переменную X_3 — именно эта категория осталась «нетронутой» при внесении пропусков согласно механизму

DOI: 10.14515/monitoring.2021.4.940

НП, при внесении же пропусков согласно механизмам ПСП и СП пропуски вносились и в нее. В-третьих, при использовании МИП стандартные ошибки оценок коэффициентов регрессоров, которые участвуют во всех спецификациях (X_2 и X_3), выше в рамках самой сложной спецификации.

Что означает завышение стандартных ошибок на практике? Если два гипотетических исследователя строят одинаковые регрессионные модели на одних и тех же данных с пропусками, но один применяет АПН, а второй — МИП, то оба получат одинаковые оценки регрессионных коэффициентов, однако у первого они будут выглядеть более статистически значимыми (с точки зрения p-value), чем у второго. Возможно даже, что второй исключит какие-то регрессоры из уравнения как незначимые, а первый эти же регрессоры оставит в уравнении. В таком случае исследователи могут сделать разные содержательные выводы о наличии или отсутствии связи между рассматриваемыми регрессорами и зависимой переменной. Поскольку в социальных науках принято учитывать значимость связи при ее интерпретации, использование МИП не выглядит предпочтительным подходом.

Проверка трех гипотез и дополняющий ее анализ покрытия ДИ создают общее впечатление, что смещение оценок регрессионных коэффициентов пренебрежимо мало и почти не варьирует между уровнями экспериментальных факторов, тогда как стандартные ошибки этих оценок, во-первых, выше для МИП, чем для АПН, во-вторых, заметно варьируют между уровнями экспериментальных факторов. Чтобы подтвердить это впечатление строго статистически, мы прибегаем к дисперсионному анализу, который попутно позволяет сравнить вклады факторов (по величине суммы квадратов отклонений) в вариацию смещений и стандартных ошибок. В дисперсионном анализе в качестве зависимых переменных на первом этапе используются смещения оценок регрессионных коэффициентов, а на втором — их стандартные ошибки. В роли главных эффектов выступают три экспериментальных фактора, а также подход к обработке пропусков.

Поскольку в этой статье мы сфокусированы на категориальном регрессоре и во всех изучаемых спецификациях участвуют только созданные из него фиктивные переменные и константа (к которой относится одна из его категорий), далее мы рассматриваем только их. В следующих работах мы предполагаем постепенно преодолевать эти ограничения.

В таблице 5 мы видим похожую закономерность для смещений оценок константы и коэффициентов категориального регрессора: заметная сумма квадратов отклонений объясняется главными эффектами доли и механизма пропусков (их строки подкрашены серым). Хотя для каждого из коэффициентов только один из этих факторов преодолевает порог значимости 0,05, их влияние гораздо более заметно по сравнению с факторами спецификации регрессионной модели и подхода к обработке пропусков. Казалось бы, этот статистический результат противоречит сделанному ранее выводу, что смещение оценок регрессионных коэффициентов пренебрежимо мало и почти не варьирует между уровнями экспериментальных факторов. Однако учитывая, что статистическая мощность совокупности данных, на которых проводился эксперимент, велика, и что это может приводить к статистической значимости даже минимальной разницы между групповыми средними, мы предлагаем следующую дополнительную проверку:

DOI: 10.14515/monitoring.2021.4.940

рассмотреть величину смещения для самой неблагоприятной комбинации уровней факторов, то есть той, которая приводит к наибольшему абсолютному смещению. Чтобы найти такую комбинацию, мы применяем метод ChAID к смещениям оценок константы и коэффициентов категориального регрессора в качестве его зависимых переменных, и к экспериментальным факторам и подходам к обработке пропусков — в качестве его предикторов.

Таблица 5. ***Дисперсионный анализ, моделирующий смещения оценок константы и регрессионных коэффициентов категориального регрессора факторами: спецификация регрессионной модели, доля и механизм пропусков, подход к их обработке***

Зависимые переменные	**Смещение константы (%)**		**Смещение X_2 (%)**		**Смещение X_3 (%)**	
Главные эффекты	**Сумма квадратов**	**p-value**	**Сумма квадратов**	**p-value**	**Сумма квадратов**	**p-value**
Спецификация	,00	1,00	,00	1,00	,00	,99
Доля пропусков	,12	,39	1,50	,00	,20	,42
Подход	,00	,99	,00	,99	,00	1,00
Механизм	,70	,01	,50	,15	1,58	,00

ChAID[14] — это метод, позволяющий автоматически отбирать предикторы, которые оказывают значимый эффект на зависимую переменную, упорядочивать их по этой значимости, объединять категории предикторов, статистически значимо не различающиеся по своему эффекту на зависимую переменную, и находить комбинации значений предикторов, предсказывающие искомые значения зависимой переменной. Мы подаем в ChAID континуальные зависимые переменные, к которым в нем применяется F-статистика, что роднит его с дисперсионным анализом. ChAID относят к парадигме «data-driven». В сочетании с методами, относимыми к парадигме «theory-driven», он помогает развить и сделать более удобными для обзора и интерпретации результаты применения изучаемых подходов. Следуя этой логике, мы дополняем результаты дисперсионного анализа результатами ChAID. Конечно, можно было бы в рамках дисперсионного анализа рассмотреть групповые средние зависимой переменной для каждого уровня факторов и эффекты взаимодействия последних, но это гораздо более трудоемкая процедура, нежели применение ChAID (в этом и проявляется близость ChAID к парадигме «data-driven» по сравнению с дисперсионным анализом).

ChAID подтвердил выводы, сделанные нами на основе дисперсионного анализа: в трех построенных деревьях для константы и регрессионных коэффициентов первой расщепляющей переменной выступили механизм пропусков, их доля и снова механизм соответственно. К наибольшему абсолютному смещению для константы приводит неслучайный механизм пропусков (комбинации уровней факторов не сформировались). Для коэффициента X_2 к наибольшему абсолютному

[14] ChAID — Chi-square automatic interaction detection.

DOI: 10.14515/monitoring.2021.4.940

смещению приводит, как ни странно, доля пропусков 10 %—25 %. Для коэффициента X_3 к наибольшему абсолютному смещению приводит комбинация случайного механизма пропусков и их доля 50 %. На наш взгляд, эти статистические выводы подтверждают сформировавшееся ранее впечатление, что смещение оценок регрессионных коэффициентов пренебрежимо мало и почти не варьирует между уровнями экспериментальных факторов.

Для статистической проверки вывода, что стандартные ошибки оценок константы и регрессионных коэффициентов категориального регрессора, во-первых, выше для МИП, чем для АПН, а во-вторых, они заметно варьируют между уровнями экспериментальных факторов, мы также прибегаем к ChAID. Построенные им деревья приведены на рисунке 1 в Приложении. Они довольно разветвленные, поскольку стандартные ошибки действительно значимо варьируют между уровнями экспериментальных факторов, подходами к обработке пропусков, а также их комбинациями. Важно, что все три дерева демонстрируют похожую закономерность: самый значимый предиктор — спецификация регрессионной модели (стандартные ошибки ниже в спецификациях без эффектов взаимодействия и выше в спецификациях с эффектами); за ним в порядке убывания значимости следуют доля пропусков (стандартные ошибки тем ниже, чем ниже доля пропусков) и подход к обработке пропусков (стандартные ошибки ниже при применении АПН и выше при применении МИП). Последним по значимости расположился механизм пропусков, причем во всех узлах его уровня объединились полностью случайный и случайный механизмы. Что еще более интересно и в какой-то мере релевантно нашей первой гипотезе (хотя в ней мы предполагали преимущество МИП для неслучайных пропусков с точки зрения несмещенности, а не эффективности), — при применении АПН стандартные ошибки ниже для полностью случайного и случайного механизмов и выше для неслучайного механизма, а при применении МИП — наоборот. В целом же мы считаем статистически подтвержденным впечатление о большей статистической эффективности оценок при применении АПН, чем при применении МИП.

Чтобы этот вывод выглядел ближе к исследовательской практике, мы применили ChAID к длине доверительных интервалов оценок константы и коэффициентов категориального регрессора, выраженным в процентах от истинных значений этих коэффициентов (для сравнимости). Построенные деревья[15] ожидаемо подтвердили описанную только что закономерность. Однако обратимся к комбинациям предикторов, приводящим к наиболее широким доверительным интервалам и различающимся только подходом к обработке пропусков, — ее составляют спецификация с эффектами взаимодействия и доля пропусков 50 %[16]. Такая комбинация приводит к доверительному интервалу, ширина которого при использовании АПН равна 0,01 % от истинного значения константы, а при использовании МИП — 0,03 %. Она же приводит к доверительному интервалу с шириной 0,12 % от истинного значения коэффициента при X_2, если использовать АПН, и 0,25 %, если использовать МИП. Наконец, она приводит к доверительному

[15] По структуре уровней и узлов они дублируют деревья с рисунка 1, поэтому мы не вставили их в статью.

[16] Это те же самые узлы деревьев и рисунка 1, которые там приводят к максимальным стандартным ошибкам, что логично.

DOI: 10.14515/monitoring.2021.4.940

интервалу с шириной 0,11 % от истинного значения коэффициента при X_2 в случае АПН и 0,23 % — в случае МИП. Таким образом, АПН действительно обеспечивает более эффективную (до нескольких раз) оценку константы и коэффициентов категориального регрессора, чем МИП.

Заключение

Наше исследование было нацелено на комплексное изучение работы МИП (метода индикаторной переменной) с пропусками в категориальном регрессоре линейной регрессии методом наименьших квадратов. Своего рода эталоном выступил АПН (анализ полных наблюдений) — наиболее популярный подход для работы с пропусками, неожиданная универсальность которого становится предметом обсуждения в последние годы. Мы хотели внести свою лепту как в проверку универсальности АПН, так и в поиск ниши для МИП. Первое получилось даже сверх ожиданий: обнаружена неизвестная прежде нечувствительность АПН к пропускам в регрессоре, наличие/отсутствие которых связано с зависимой переменной (тем самым поставлены под сомнение некоторые результаты работы Дж. Бартлетта и соавторов [Bartlett et al., 2015]).

Априорным преимуществом МИП перед АПН считается сохранение исходного размера выборки для дальнейшего анализа без математических и интерпретативных сложностей. Мы показали, что если методом дальнейшего анализа выступает линейная регрессия, то сохранение исходного размера выборки не улучшает оценку регрессионных коэффициентов. Этому неподтвержденному нами преимуществу сопутствует неожиданно обнаруженный недостаток в виде завышенных стандартных ошибок оценок и расширенных доверительных интервалов. Кроме того, применение МИП оказалось не проще, чем АПН. Специфически социологическая логика применения МИП к неслучайным пропускам тоже не нашла своего подтверждения: выяснилось, что (а) МИП приводит не к меньшим смещениям, чем АПН, при неслучайных пропусках и (б) в части стандартных ошибок оценок регрессионных коэффициентов результаты МИП не всегда лучше при неслучайных пропусках, чем при случайных. Таким образом, мы не только не подтвердили статистически наличие ниши для МИП, но и поставили под сомнение предполагаемые его преимущества.

Исследование было проведено посредством статистического эксперимента на симулированных данных с контролем трех экспериментальных факторов: механизма пропусков, их доли и спецификации регрессионной модели. Такими строгостью и стандартизацией мы старались преодолеть ограничения предшественников. Однако и наша работа имеет ограничения. Во-первых, для внесения пропусков механизма СП мы использовали вспомогательную переменную, не участвующую непосредственно в регрессионной модели. Гипотетически, если вместо вспомогательной переменной использовать другой регрессор или зависимую переменную, результат мог бы отличаться от нашего. Во-вторых, мы вносили пропуски только в один регрессор; в реальной практике более вероятны пропуски сразу в нескольких регрессорах. В-третьих, важными экспериментальными факторами могут выступить размер выборки и совпадение/несовпадение истинной и симулируемой спецификаций. Включение в дизайн эксперимента этих факторов усложнит его как для проведения, так и для восприятия полученных результатов.

DOI: 10.14515/monitoring.2021.4.940

Тем не менее, основываясь на уже полученных в данном исследовании результатах, на наш взгляд, весьма весомых и неожиданных, можно двигаться дальше, постепенно усложняя эксперимент и приближая его условия к реальной практике.

Список литературы (References)

Жучкова С. В., Ротмистров А. Н. Возможность работы с пропущенными данными при использовании CHAID: результаты статистического эксперимента // Социология: методология, методы, математическое моделирование. 2018. № 46. С. 85—122.
Zhuchkova S., Rotmistrov A. (2018) Handling Missing Data with CHAID: Results of a Statistical Experiment. *Sociology: Methodology, Methods, Mathematical modeling*. No. 46. P. 85—122.

Стребков Д. О., Шевчук А. В., Лукина А. А., Мелианова Е. Г., Тюлюпо А. В. Социальные факторы выбора контрагентов на бирже удаленной работы: исследование конкурсов с помощью «больших данных» // Экономическая социология. 2019. Т. 20. № 3. С. 25—65. https://doi.org/10.17323/1726-3247-2019-3-25-65.
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*. Vol. 20. No. 3. P. 25—65. https://doi.org/10.17323/1726-3247-2019-3-25-65.

Фабрикант М. С. Модель-ориентированный подход к отсутствующим значениям: множественная импутация в многоуровневой регрессии посредством R (на примере анализа опросных данных по гордости страной) // Социология: методология, методы, математическое моделирование. 2015. № 41. С. 7—29.
Fabrykant M. (2015) Model-Oriented Approach to Missing Values: Multiple Imputation in Multilevel Regression Using R (On the Example of Analyzing Survey Data). *Sociology: Methodology, Methods, Mathematical Modeling*. No. 41. P. 7—29.

Akande O., Li F., Reiter J. (2017) An Empirical Comparison of Multiple Imputation Methods for Categorical Data. *The American Statistician.* Vol. 71. No. 2. P. 162—170. https://doi.org/10.1080/00031305.2016.1277158.

Allison P. (2005) Imputation of Categorical Variables with PROC MI. In: *Proceedings of the SAS Users Group International Conference (SUGI).* No. 30. P. 113—130. URL: https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/113-30.pdf (accessed: 21.04.2021).

Anagnostopoulos C., Triantafillou P. (2014) Scaling Out Big Data Missing Value Imputations: Pythia vs. Godzilla. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. P. 651—660. https://doi.org/10.1145/2623330.2623615.

Bartlett J. W., Carpenter J. R., Tilling K., Vansteelandt S. (2014) Improving upon the Efficiency of Complete Case Analysis When Covariates Are MNAR. *Biostatistics*. Vol. 15. No. 4. P. 719—730. https://doi.org/10.1093/biostatistics/kxu023.

DOI: 10.14515/monitoring.2021.4.940

Bartlett J. W., Harel O., Carpenter J. R. (2015) Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. *American Journal of Epidemiology*. Vol. 182. No. 8. P. 730—736. https://doi.org/10.1093/aje/kwv114.

Chen J., Hossler D. (2017) The Effects of Financial Aid on College Success of Two-Year Beginning Nontraditional Students. *Research in Higher Education*. Vol. 58. No. 1. P. 40—76. https://doi.org/10.1007/s11162-016-9416-0.

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*. Vol. 34. No. 1. P. 23—36. https://doi.org/10.1007/s10654-018-0447-z.

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*. Vol. 59. No. 10. P. 1087—1091. https://doi.org/10.1016/j.jclinepi.2006.01.014.

Dougherty C. (2016) Introduction to Econometrics. Oxford: Oxford University Press.

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

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 and Prevention*. No. 111. P. 1—11. https://doi.org/10.1016/j.aap.2017.11.005.

Giest S., Samuels A. (2020) 'For Good Measure': Data Gaps in a Big Data World. *Policy Sciences*. Vol. 53. No. 3. P. 559—569. https://doi.org/10.1007/s11077-020-09384-1.

Greenacre M., Pardo R. (2006) Subset Correspondence Analysis: Visualizing Relationships Among a Selected Set of Response Categories From a Questionnaire Survey. *Sociological Methods & Research*. Vol. 35. No. 2. P. 193—218. https://doi.org/10.1177/0049124106290316.

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*. Vol. 184. No. 11. P. 1265—1269. https://doi.org/10.1503/cmaj.110977.

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*. Vol. 58. No. 5. P. 1353—1359. https://doi.org/10.1016/j.jvs.2013.05.008.

Hughes R. A., Heron J., Sterne J. A. C., Tilling K. (2019) Accounting for Missing Data in Statistical Analyses: Multiple Imputation Is Not Always the Answer. *International Journal of Epidemiology*. Vol. 48. No. 4. P. 1294—1304. https://doi.org/10.1093/ije/dyz032.

Jones M. P. (1996) Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. *Journal of the American Statistical Association*. Vol. 91. No. 433. P. 222—230. https://doi.org/10.1080/01621459.1996.10476680.

DOI: 10.14515/monitoring.2021.4.940

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*. Vol. 63. No. 7. P. 728—736. https://doi.org/10.1016/j.jclinepi.2009.08.028.

Little R. J. A., Rubin D. B. (2002) Statistical Analysis with Missing Data (2nd ed.). Hoboken, NJ: Wiley.

McKinney W. (2010) Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*. P. 51—56. URL: https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf (accessed: 01.08.2021).

Miettinen O. S. (1985) Theoretical Epidemiology: Principles of Occurrence Research. New York, NY: John Wiley & Sons.

Morgan J., Sonquist J. (1963) Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*. Vol. 58. No. 302. P. 415—434. https://doi.org/10.1080/01621459.1963.10500855.

Morris T. P., White I. R., Crowther M. J. (2019) Using Simulation Studies to Evaluate Statistical Methods. *Statistics in Medicine*. Vol. 38. No. 11. P. 2074—2102. https://doi.org/10.1002/sim.8086.

Oliphant T. E. (2006) A Guide to NumPy. USA: Trelgol Publishing.

Ratner B. (2011) Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. Boca Raton: CRC Press.

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*. Vol. 47. No. 8. P. 481—491. https://doi.org/10.3102/0013189X18788054.

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

Rubin D. B. (1976) Inference and Missing Data. *Biometrika*. Vol. 63. No. 3. P. 581—592. https://doi.org/10.2307/2335739.

Rubin D. B. (1996) Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*. Vol. 91. No. 434. P. 473—489. https://doi.org/10.1080/01621459.1996.10476908.

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

Seabold S., Perktold J. (2010) Statsmodels: Econometric and Statistical Modeling with Python. In: *Proceedings of the 9th Python in Science Conference*. URL: https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf (accessed: 01.08.2021).

Trevizo D., Lopez M. J. (2016) Neighborhood Segregation and Business Outcomes: Mexican Immigrant Entrepreneurs in Los Angeles County. *Sociological Perspectives*. Vol. 59. No. 3. P. 668—693. https://doi.org/10.1177/0731121416629992.

DOI: 10.14515/monitoring.2021.4.940

van der Heijden G. J. M. G., Donders A. R. T., 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*. Vol. 59. No. 10. P. 1102—1109. https://doi.org/10.1016/j.jclinepi.2006.01.015.

van der Walt S. S., Colbert C., Varoquaux G. (2011) The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*. No. 13. P. 22—30. https://doi.org/10.1109/MCSE.2011.37.

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*. Vol. 13. No. 1. P. e11598-e11598—8. URL: https://cris.maastrichtuniversity.nl/ws/portalfiles/portal/64128744/Viechtbauer_2016_Bias_in_regression_coefficient_estimates_when_assu.pdf (accessed: 01.08.2021).

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*. Vol. 38. No. 1. P. 369—397. https://doi.org/10.1111/j.1467-9531.2008.00202.x.

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*. Vol. 10. No. 4. P. 843—876. https://doi.org/10.1080/19345747.2017.1300719.

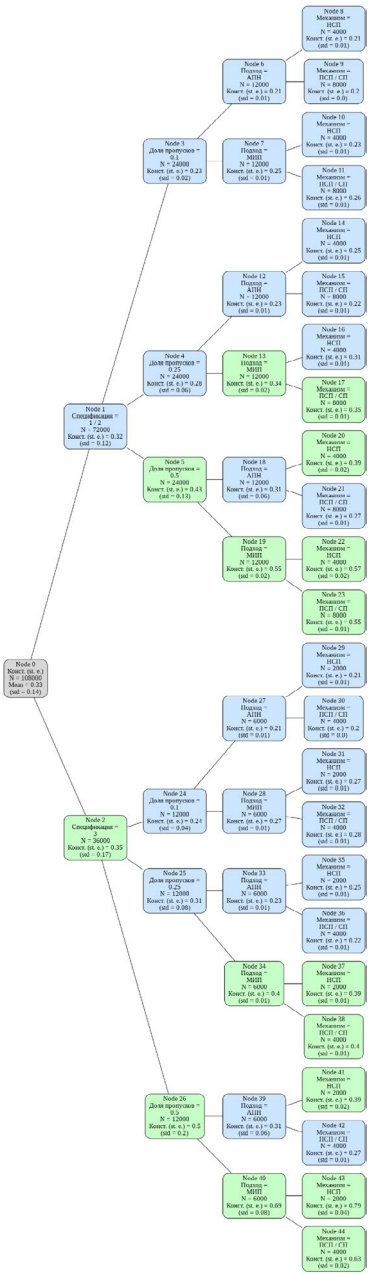
White I. R., Thompson S. G. (2005) Adjusting for Partially Missing Baseline Measurements in Randomized Trials. *Statistics in Medicine*. Vol. 24. No. 7. P. 993—1007. https://doi.org/10.1002/sim.1981.

Zhelyazkova N., Ritschard G. (2018) Parental Leave Take-Up of Fathers in Luxembourg. *Population Research and Policy Review*. Vol. 37. No. 5. P. 769—793. https://doi.org/10.1007/s11113-018-9470-8.

DOI: 10.14515/monitoring.2021.4.940

Приложение

Рис. 1. ***Деревья ChAID, моделирующие стандартные ошибки оценок константы и регрессионных коэффициентов категориального регрессора предикторами: спецификация регрессионной модели, доля и механизм пропусков, подход к их обработке***



DOI: 10.14515/monitoring.2021.4.940