ПРАКТИКИ СБОРА И АНАЛИЗА ФОРМАЛИЗОВАННЫХ ДАННЫХ

М.С. Фабрикант
(*Москва*)

МОДЕЛЬ-ОРИЕНТИРОВАННЫЙ ПОДХОД К ОТСУТСТВУЮЩИМ ЗНАЧЕНИЯМ: МНОЖЕСТВЕННАЯ ИМПУТАЦИЯ В МНОГОУРОВНЕВОЙ РЕГРЕССИИ ПОСРЕДСТВОМ *R* (на примере анализа опросных данных)[1]

В статье даются обоснование и описание процедуры множественной импутации отсутствующих значений в массивах данных. Описан способ решения проблемы отсутствующих данных с позиций модель-ориентированного подхода в противоположность дизайн-ориентированному. В теоретической части статьи перечислены и обоснованы преимущества множественной импутации перед более простыми способами оперирования отсутствующими значениями – удалением кейсов по списку и попарно и заменой средним. Указаны ограничения множественной импутации и связанные с ними требования, предъявляемые к данным. В эмпирической части статьи на конкретном примере кросс-культурного исследования, посвященного детерминантам гордости страной, проиллюстрирована процедура множественной импутации и представлен готовый к использованию программный код для диагностики данных и проведения множественной импутации посредством программных пакетов *R VIM* и *mice*.

Маргарита Сауловна Фабрикант – научный сотрудник Лаборатории сравнительных исследований массового сознания Экспертного института НИУ ВШЭ. E-mail: marharyta.fabrykant@gmail.com.

[1] Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

Ключевые слова: пропущенные данные, модель-ориентированный подход, множественная импутация, MCAR, MAR, MNAR, R VIM, R mice.

Введение

В настоящее время существуют два подхода к статистическому анализу опросных данных – дизайн- и модель-ориентированный. Дизайн-ориентированный подход возник раньше и, как следствие, представляется более традиционным и интуитивно понятным. В нем соотношение выборочных данных и характеристик генеральной совокупности определяется дизайном сбора данных – от простой полностью рандомизированной выборки до более сложных стратифицированных, кластеризованных и т.п. Соответственно, статистический вывод из выборочных результатов к параметрам генеральной совокупности рассматривается как проблема *репрезентации*. Модель-ориентированный подход разработан относительно недавно. Его суть заключается в том, что статистический вывод выборочных результатов на генеральную совокупность определяется в первую очередь не дизайном исследования, а параметрами конкретной модели, которая оценивается на основе выборочных данных (разумеется, структура этой модели в свою очередь зависит в том числе и от исследовательского плана, реализованного в ходе сбора первичных данных). Соответственно, статистический вывод в рамках модель-ориентированного подхода рассматривается с позиций не репрезентации, а *предсказания* [1]. Выборочные данные не репрезентируют генеральную совокупность, а позволяют построить модель для предсказания неизвестных данных той части выборки, которая в генеральную совокупность не входит. Таким образом, два вида неизвестных данных – параметры модели и характеристики генеральной совокупности – объединяются в общую категорию неизвестных значений.

Дизайн-ориентированный подход до сих пор остается опцией по умолчанию постольку, поскольку в исследовании используется классическая статистика: статистический смысл уровня значимости как вероятности ошибки первого рода основан на репрезентации и представляет собой оценку репрезентативности. Модель-ориентированный подход в настоящее время используется в основном в связи с байесовской статистикой, которая до сих пор воспринимается многими как экзотическая возможность, к которой имеет смысл прибегать лишь в особых обстоятельствах (например, при слишком маленьком объеме выборки) [2]. Вместе с тем можно указать как минимум на один случай, когда модель-ориентированный подход оказывается не экзотическим, а, напротив, весьма типичным. Речь идет о проблеме пропущенных данных. Если процент пропущенных данных слишком велик (конкретное значение «слишком» зависит от формата и цели каждого конкретного исследования), и им нельзя пренебречь, для решения этой проблемы следует использовать модель-ориентированный метод – множественную импутацию.

Метод множественной импутации возник несколько десятилетий назад и успел доказать свои преимущества по сравнению с другими способами заполнения пропусков (подробнее об этом пойдет речь ниже), однако нельзя сказать, чтобы он стал стандартным способом работы с массивами данных с большим количеством пропущенных значений. В литературе можно встретить следующие возражения против использования множественной импутации.

1. Множественная импутация требует несоизмеримо больших усилий и временных затрат, чем другие способы заполнения пропусков в массивах данных, давая сопоставимые результаты [3].

2. Множественная импутация требует построения моделей, подобно методам проверки гипотез, однако, в отличие от них, не дает никакой новой содержательной информации. Более того, любой новый результат рассматривается как отрицательный и демонстрирует, что импутированные данные существенно отличаются от полученных в результате сбора данных [3].

3. Множественная импутация плохо применима к переменным со сложной формой зависимости от множества других переменных, которую невозможно описать посредством простой модели, т.е. практически ко всему, что изучается в социальных науках [4].

4. Множественная импутация требует малодоступного и трудоемкого в освоении и использовании программного обеспечения [4].

Цель данной статьи – подвергнуть сомнению эти представления о множественной импутации, показав общие основания и конкретные способы и процедуру её использования. В первой части статьи продемонстрированы преимущества множественной импутации по сравнению с другими способами заполнения пропущенных значений, оправдывающие технические трудности и временные затраты на импутирование. Во второй части статьи на примере фрагмента конкретного эмпирического исследования предложен готовый к использованию программный код для проведения множественной импутации с использованием общедоступного инструментария анализа данных – языка программирования R. Выводы сформулированы в форме рекомендаций по использованию множественной импутации в количественных социологических исследованиях не для повышения методологической сложности как самоцели, а для решения проблемы отсутствующих значений наиболее эффективным способом.

Множественная импутация и ее альтернативы

Множественная импутация представляет собой достаточно трудоемкий подход, особенно, как будет показано ниже, в плане временных затрат и требуемого программного кода, поэтому, прежде чем углубиться в её технические детали, имеет смысл кратко проанализировать более ранние альтернативные способы решения проблемы пропущенных данных. Поскольку эти методы намного проще, использовать множественную импутацию имеет смысл

только убедившись, что другие, более ранние, простые и быстрые альтернативы неприемлемы или по крайне мере существенно уступают множественной импутации.

Альтернативных способов решения проблемы множественных данных достаточно много, их подробное русскоязычное описание приведено в статье И.К. Зангиевой [5]. Здесь алгоритмы восстановления пропусков подразделяются на простые (неитеративные) и сложные (итеративные), а вторые в свою очередь – на глобальные (с участием всего массива для заполнения каждого пропуска) и локальные (с участием только части массива, наиболее близкой к пропуску). Согласно этой классификации, множественная импутация относится к сложным глобальным алгоритмам.

Остановимся подробнее только на тех способах, которые, по нашим наблюдениям, чаще всего рассматриваются как более простая альтернатива множественной импутации. К ним относятся четыре простых алгоритма: полное удаление, частичное удаление, замена средним и взвешивание.

В случае полного удаления, или удаления по списку (*listwise*), из процедур обработки данных исключаются все кейсы с отсутствующими значениями хотя бы по одной переменной, которая задействована в каждой отдельной процедуре анализа данных. Этот подход является вариантом по умолчанию во многих пользовательских программах обработки данных и запускается автоматически, когда исследователь просто игнорирует пропуски в матрице данных. Такой способ применим только в том случае, когда доля отсутствующих значений ничтожно мала (до 1%) или условно допустима (1–5%), и можно позволить себе её игнорировать без существенных последствий для результатов статистического анализа. Однако во многих случаях это не так, и процент отсутствующих значений слишком велик (свыше 5%), чтобы не влиять на результат [6].

Иногда в этом случае предлагается использовать смягченный вариант удаления по списку – попарное удаление [7]. Оно при-

менимо тогда, когда процедура обработки данных может быть разделена на этапы, в которых участвуют различные пары переменных, например при построении матрицы интеркорреляций. В этом случае для расчета каждой конкретной корреляции достаточно удалить только те кейсы, чьи значения отсутствуют только по тем двум переменным, между которыми вычисляется корреляция. Остальные переменные в расчете не задействованы, поэтому отсутствующие значения по ним не имеют значения, и удалять соответствующие кейсы не нужно. Этот подход относительно эффективен в том случае, если большинство пропущенных значений по разным переменным относятся к разным кейсам: именно при таком условии большой процент кейсов с пропущенными данными по хотя бы одной из всех переменных может сочетаться с пренебрежимо малым процентом пропущенных значений для отдельно взятой пары переменных. Однако даже при соблюдении такого условия не все методы анализа данных можно разделить на этапы, где были бы задействованы не все переменные. В методах множественного анализа данных, например множественной регрессии, когда все переменные задействованы одновременно, попарное удаление процедурно тождественно удалению по списку и воспроизводит все проблемы, сопутствующие этому варианту.

Прием, который считался стандартным до появления множественной импутации, – метод замены средним значением [8]. Суть его очень проста: все пробелы в матрице данных для всех кейсов заполняются значениями, равными арифметическому среднему по данной переменной. Для номинальных и порядковых шкал возможна аналогичная замена другими мерами центральной тенденции – модой и медианой. Несмотря на простоту, такой подход имеет ряд серьезных недостатков, включая как ограничения по применению, так и заложенное в самой процедуре искажение структуры данных [9; 10; 11]. Смысл процедуры исходит из предположения, что вероятность неизвестного значения на месте неответа тем выше, чем ближе это предполагаемое значение к среднему. Иными словами,

переменная, для которой производится замена средним, не только количественная, но и нормально распределенная (аналогичное ограничение работает для других мер центральной тенденции при других типах шкал). В противном случае, если распределение смещенное, большинство значений будут далеки от среднего.

Если же наблюдается не только выраженная асимметрия, но и отрицательный эксцесс, то массив может содержать относительно мало значений, равных или близких к среднему. В таком случае замена средним приводит к тому, что добавленные значения радикально отличаются от реально собранных в ходе опроса. Наконец, даже для нормально распределенной количественной шкалы замена средним искажает данные, поскольку, необоснованно полагая все отсутствующие значения близкими к среднему, исследователь тем самым недооценивает возможную неоднородность данных, т.е. искусственно уменьшает дисперсию [9]. Помимо неверных описательных статистик, в результате такой операции может ухудшиться любая регрессионная модель, построенная с участием этих переменных, поскольку предикторы с малой дисперсией имеют более низкую предсказательную силу, помимо объективных свойств явления, которое они операционализируют. Таким образом, замена средним относительно приемлема для нормально распределенных количественных переменных с изначально очень малой дисперсией – достаточно редкое сочетание.

Аналогичный недостаток – искажение структуры массива из-за неоправданного предположения, что отсутствующие значения аналогичны имеющимся – характерна и для распространенного метода взвешивания [12], когда кейсам с имеющимися значениями, присваиваются веса больше 1, чтобы дополнить выборку до необходимого объема. При этом, если количество отсутствующих значений для разных переменных не совпадает, то одни и те же кейсы (наблюдения) получают различные веса для разных переменных вне связи с их содержательно обоснованной весомостью, т.е. значимостью для данного массива.

Метод множественной импутации лишен всех перечисленных ограничений и недостатков, однако имеет свои – как организационные, так и содержательные. Основным организационным недостатком этого метода является существенно бóльшая продолжительность процедуры множественной импутации по сравнению с любой из рассмотренных альтернатив. На массиве данных объемом в пару десятков тысяч кейсов процедура множественной импутации может занять, в зависимости от мощности компьютера и точности проверки (о том, что это означает будет сказано ниже), от нескольких часов до нескольких дней. Помимо временных затрат на саму процедуру, использование множественной импутации требует освоения дополнительных программных кодов. SPSS в последних версиях позволяет делать множественную импутацию, однако только для простейших моделей, а оценка чего-то существенно более сложного, чем регрессия методом наименьших квадратов, требует использования других ресурсов для обработки данных. Более сложные виды множественной импутации, в том числе и в многоуровневых регрессиях, осуществимы посредством *R*, что и будет представлено ниже.

Помимо длительности применения, метод множественной импутации имеет ключевое содержательное ограничение, которое следует из самой его сути. В отличие от замены средним, множественная импутация не предназначена для подстановки конкретных значений в базу данных. Иными словами, множественная импутация, несмотря на этимологию слова «импутация», что и означает подстановку, не способна предсказать значение для пропущенного значения по конкретной переменной для конкретного кейса. Вместо дополнения базы данных множественная импутация строит конкретную регрессионную модель исходя из того, как могли бы с наибольшей вероятностью выглядеть параметры этой модели, если бы вместо неответов в базе данных находились те значения, которые предсказывает данная модель. Таким образом, множественная импутация одновременно строит модель исходя

из достраиваемых значений на месте пропусков, и предсказывает значения на месте неответов исходя из самой оцениваемой модели.

Как нетрудно догадаться, этот процесс циклический и осуществляется в несколько итераций – посредством сопоставления тех данных, на которых была построена уточненная модель на каждом новом витке, с теми данными, которые эта модель предсказывает, вплоть до их конвергенции [13; 14]. Чтобы минимизировать элемент случайности при подстановке данных, такую процедуру проводят не один, а несколько раз. Именно поэтому импутация называется множественной. После этого высчитываются средние значения для всех полученных моделей. Количество таких повторов не регламентировано и не зависит от теоретических предпосылок модели. Принятое количество импутаций – 5.

Поскольку, как уже было сказано, каждая процедура импутации занимает достаточно много времени, увеличение количества импутаций существенно удлиняет процедуру обработки данных без существенного улучшения качества результата [15]. Большое количество импутаций имеет смысл задавать, только если при небольшом их количестве для разных импутаций на одном и том же наборе переменных получаются модели с существенно различными параметрами. Впрочем в этом случае лучше проанализировать причины несовпадения, связанные с заданной на входе структурой модели.

Именно структура модели составляет наиболее существенное содержательное ограничение множественной импутации. «Достраивание» массива для оцениваемых параметров модели хорошо лишь настолько, насколько хороша сама модель. Если в модели отсутствуют значимые предикторы, то этот недостаток будет влиять и на качество множественной импутации. Можно сказать, что множественная импутация воспроизводит основное содержательное ограничение – проблему невключенной переменной, – в какой-то мере усиливая ее последствия из-за «достраивания» массива с опорой на искаженные параметры модели. Вместе с тем зависимость от содержательных характеристик модели, качество которых с трудом диагностируется

чисто статистическим, а не теоретическим путем, – недостаток, присущий не только множественной импутации, но и всему модельно-ориентированному подходу.

Еще одно ограничение множественной импутации, специфичное именно для нее, – характер отсутствующих данных. Различают три вида отсутствующих данных в зависимости от причины неответов. Полностью случайно отсутствующими (*missing completely in random*, MCAR) называются пропуски, которые возникают полностью случайно и хаотично. Этот вариант – идеальная абстракция, её не следует ожидать на практике. Не отличим от этого варианта по своей практической значимости второй, более реалистичный вид отсутствующих значений – случайно пропущенные (*missing at random*, MAR). Так называются неответы, которые понемногу связаны с причинами, учтенными внутри модели, используемой при импутировании. В этом случае множественная импутация допустима, в отличие от третьего варианта – не случайно отсутствующих значений (*missing not at random*, MNAR) [9]. Предсказание отсутствующих значений по имеющимся недопустимо, если первые существенно отличаются от вторых по значимым характеристикам вне используемой модели, или, иными словами, если неответы имеют выраженную причинность, не учтенную в модели. Эти причины могут быть неизвестными либо предполагаемыми теоретически, но неучтенными из-за отсутствия эмпирических данных по соответствующим переменным.

Несмотря на эти ограничения, множественная импутация оказалась, как отмечает Зангиева [5], наиболее перспективным методом заполнения пропусков. В качестве основной проблемы при этом указана недоступность необходимого для применения этого метода дорогостоящего коммерческого программного обеспечения. Однако к настоящему времени созданы некоммерческие программные пакеты открытого доступа.

О том, когда и каким образом диагностируется возможность и осуществляется сама процедура множественной импутации на практике, будет рассказано далее на практическом примере.

Множественная импутация в R

Случай использования множественной импутации на практике при помощи языка программирования R взят из сравнительного межстранового исследования гордости страной, проведенного совместно с В.С. Магуном [16; 17]. В этом исследовании были использованы данные Международной программы социальных опросов (*International Social Survey Programme*, далее – ISSP). В отличие от других регулярно проводимых мультистрановых опросов, в ISSP каждая волна тематическая. Мы использовали данные волны 2003 г., последней на момент проведения исследования о гордости страной. Массив включает в себя общий вопрос о гордости принадлежностью к стране (варианты ответа: «очень горжусь», «скорее горжусь», «не очень горжусь», «совсем не горжусь») и 10 вопросов о гордости достижениями страны в конкретных сферах – международной политике, демократии, экономике, социальной защите, науке и технологиях, литературе и искусстве, спорте, вооруженных силах, истории, социальной справедливости (варианты ответа те же, что и для вопроса о гордости в целом).

Основная гипотеза исследования заключалась в том, что общая гордость страной и гордость её конкретными достижениями имеют различные объективные детерминанты. Общей гордости страной в массиве ISSP соответствует всего одна переменная – гордость принадлежностью к стране, однако для измерения гордости конкретными достижениями потребовалось провести факторный анализ на 10 переменных, соответствующих гордости достижениями в каждой из сфер отдельно. Первый фактор в построенной факторной модели (эксплораторный факторный анализ, метод главных осей без вращения) действительно объединил все переменные с положительным знаком. Однако количество наблюдений со значениями для данного фактора оказалось почти на 34% меньшим чем общий объем выборки. На этой стадии и встал вопрос о применении множественной импутации. Однако прежде чем на-

писать программный код для её использования, было необходимо выяснить характер распределения пропущенных данных.

Диагностика проводилась в несколько этапов. Прежде всего посредством простых описательных статистик, а именно – частотного анализа, было выявлено, что 10 переменных, вошедших в фактор, имеют сопоставимые доли пропущенных значений, которые мало отличаются от числа пропущенных значений для переменной общей гордости. Большое количество пропущенных значений для всего фактора оказалось связанным с тем, что для различных переменных отдельные исходные значения отсутствовали для разных кейсов. Факторный анализ одновременно использует все переменные, поэтому попарное удаление на практике тождественно удалению по списку, в результате которого из факторного анализа были исключены все кейсы, для которых отсутствуют значения хотя бы по одной переменной. Иными словами, не было выявлено ни одной переменной, которая давала бы непропорционально большое число отсутствующих значений и тем самым могла бы сделать выборку нерепрезентативной для остальных переменных. После сравнения по всему массиву были проведены аналогичные сравнения для каждой переменной по 33 странам, что также не выявило выраженных диспропорций между странами по количеству пропущенных значений ни по одной из переменных.

Далее была проведена диагностика структуры отсутствующих данных относительно той регрессионной модели, которую предполагалось использовать для проверки гипотезы. Для этого был построен специальный вид графика рассеяния, *margin plot*, отражающий соотношение имеющихся и пропущенных значений по двум переменным. В R такие графики строятся посредством программного пакета VIM. Был использован следующий код.

```
install.packages("VIM")
require(VIM)
marginplot(isspmerged[ ,c("factor1","predictor")], col = c("grey",
"black"), cex =1.8, cex.numbers=0.4, pch=1))
```

Первая строка, как и для других пакетов *R*, инсталлирует VIM, вторая – активизирует его для работы во время данной сессии[1]. Третья строка – собственно код для построения графика. *Margin plot* указывает на вид графика, *isspmerged* – название базы данных, а в квадратных скобках указана та часть массива, для которой требуется построить график. Пропуск перед запятой указывает на необходимость использовать все строки матрицы данных, т.е. все кейсы, а *factor*1 и *predictor* – имена переменных, соответствующих фактору гордости достижениями страны и первому предиктору из регрессионной модели (кавычки вокруг названий переменных здесь необходимы, в отличие от многих других скриптов *R*). Остальная часть кода настраивает оформление графика: col=c(*grey*, *black*) создает вектор цветовых условных обозначений (по умолчанию используются синий и красный цвета; график был сделан ахроматическим специально для данной статьи); значения *cex* и *cex.numbers* задают соответственно размер пространства и осей и шрифта для цифр в подписях делений осей[2], а *pch* – форму точки на графике рассеяния (в данном случае окружности предпочтительнее опции по умолчанию – закрашенных кругов, поскольку на большом массиве данных линии в меньшей степени сливаются в сплошную массу).

Для полной диагностики данных была построена серия таких графиков для каждого предиктора. Один из них – для возраста респондента – представлен на *рис*. 1. Он сочетает в себе диаграмму

[1] Проинсталлировать некий пакет для каждой из установленных версий *R* достаточно единожды, а активизировать необходимо заново для каждой сессии. Активация, в отличие от инсталляции, может осуществляться без подключения к Интернету.

[2] В данном случае относительно мелкий шрифт для цифр специально задан потому, что нам заранее известно, что количество пропущенных значений для фактора гордости страной выражается 5-значным числом, которое могло бы не поместиться в отведенное для него поле на графике, и в таком случае вообще не было бы отображено.

рассеяния, коробчатые диаграммы и одномерные частотные распределения. Черные точки на графике рассеяния обозначают кейсы с пропущенными значениями по одной или обеим переменным, а коробчатые диаграммы на полях – распределения значений: имеющихся – серым цветом, а отсутствующих – черным. Цифры в нижней левой части графика указывают на число отсутствующих значений для каждой переменной и на число кейсов с отсутствующими значениями по обеим переменным. Из графика видно, что, хотя число пропущенных значений для фактора гордости достижениями страны намного больше, чем для возраста (что было нам уже известно и стало причиной использования множественной импутации), формы обеих пар распределений очень похожи. Это сходство свидетельствует о соблюдении критерия случайности пропущенных данных (MAR), которое, как объяснялось выше, выступает условием правомерности множественной импутации посредством выбранной регрессионной модели.

После получения результатов диагностики, подтверждающих условие MAR, была проведена сама множественная импутация. Для этого использовали программный пакет *mice* (аббревиатура *Multiple Imputation by Chained Equations* – множественная импутация цепями уравнений, здесь имеются в виду цепи Маркова), который к настоящему моменту является основным в *R* для решения этой задачи [18]. К числу других пакетов *R* для множественной импутации относятся *Amelia II*, *BaBoon* и *miceadds*. *BaBoon* использует тот же метод импутации, что и mice, но обладает более ограниченными возможностями по количеству функций, и особенно хорошо подходит для дискретных пропущенных значений. *Amelia II* обладает более простым синтаксисом, чем *mice*, и использует другой математический аппарат, который, в отличие от *mice*, ориентирован на работу только с количественными нормально распределенными данными (имеется в виду многомерное нормальное распределение для всей базы данных) и моделями, которые чаще встречаются в эконометрике, чем в социологии

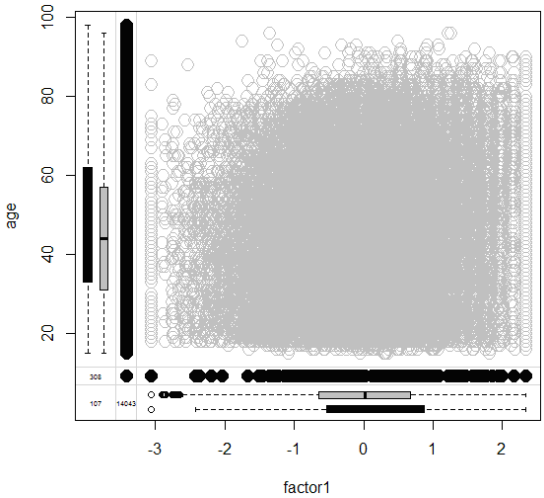

Рис. 1. График margin plot, построенный при помощи программного пакета R VIM (dthcbz R 3.2.0)

(например, временные ряды) [19]. *Miceadds*, как следует из названия, служит дополнением к mice и содержит функции, синтаксис которых нам представляется оправданным только для баз данных с большим количеством переменных, а для остальных случаев – слишком громоздким.

Для множественной импутации был использован следующий синтаксис.

```
Install.packages(“mice”)
require(mice)
ini=mice(isspmerged, maxit=5)
pred=ini$pred
pred["factor1", ]<-c(-2,1,1,1,1,1,0,1,1,2)
imp<-mice(isspmerged, meth=c("", "", "", "", "", "", "", "", "", "", ""),
pred=pred, maxit=5, seed=71152)
```

Первая и вторая строки, как и в представленном выше коде для VIM соответственно инсталлируют и активируют пакет *mice*.

Далее создается пустая матрица данных, куда будет производиться импутация, где *isspmerged* – обозначение рабочей базы данных, которое при самостоятельном использовании кода следует заменить именем собственной базы данных, а *maxit* = 5 задает количество параллельных импутаций, которое в множественной импутации принято задавать равным 5, как отмечалось выше. Далее внутри этой пустой базы данных создается пустой вектор предикторов *pred*, который затем заполняется переменными из базы данных. Место каждого предиктора в общем списке соответствует его месту в рабочей базе данных, а числа задают роль каждого предиктора в модели: 1 обозначает предиктор первого уровня (в данном случае индивидуального), 2 – предиктор второго уровня (в данном случае – странового), а –2 – константу, которую здесь также необходимо включить в модель.

Наконец, последняя строка содержит команду к осуществлению самой множественной импутации. В ней *imp* – название базы данных, дополненной импутированными значениями, на которой потом будет строиться регрессионная модель (её следует отличать от ранее созданной базы данных *ini*, куда помещается отдельно каждая из пяти импутаций), *isspmerged* – исходный массив данных, *meth* отсылает к указанной выше структуре модели, а *pred* – к ранее сформированному вектору предикторов (можно было бы предварительно задать метод отдельной строкой и в этом выражении записать аналогично *meth* = *meth*), *maxit* – количество импутаций, а *set. seed* – случайный параметр для генерирования исходных значений, которые потом корректируются в ходе итераций (точное значении этого параметра нужно для воспроизведения модели с теми же компонентами, включая случайные). Здесь важно еще раз подчеркнуть, что выполнение этих команд может занять несколько часов, особенно для большого объема исходного массива данных [12].

Далее на этих данных для проверки гипотез исследования посредством пакета *lme*4 была построена серия многоуровневых регрессионных моделей, в которых факторная переменная высту-

пала в качестве зависимой переменной индивидуального уровня без агрегирования (подробнее с различными способами работы с данными после множественной импутации можно ознакомиться в трудах Рубина [9] и ван Бурена [19]). Каждая из этих моделей была затем сопоставлена с аналогичной моделью, построенной на исходной базе данных, откуда все кейсы с пропущенными значениями были исключены. Если множественная импутация не привела к искажению данных, то полученные после импутации коэффициенты не должны существенно отличаться от первоначальных, хотя стандартные ошибки могут уменьшиться из-за увеличения количества кейсов за счет импутированных значений. В данном случае так и получилось (гордость страной, основанная на конкретных достижениях, существенно выше у людей более старшего возраста, с более низким уровнем образования, большей религиозностью и более высоким субъективно оцениваемым социальным статусом, проживающих в странах с более низким ВВП на душу населения), хотя эффект для уровня значимости оказался не столь выраженным из-за того, что даже исходный объем данных давал очень высокую значимость.

Однако для меньшего массива данных с большим количеством отсутствующих значений множественная импутация может существенно улучшать модель посредством уменьшения стандартных ошибок коэффициентов и, следовательно, более высокого уровня значимости, с чем и связана необходимость. В приведенном примере множественная импутация выступает как стандартная процедура для массива данных с большим количеством пропущенных значений, результат которой продемонстрировал надежность построенной на этих данных многоуровневой регрессионной модели.

Выводы

На основании приведенного обзора возможностей и ограничений различных способов заполнения пропусков в данных и детально

разобранного примера множественной импутации можно сделать общий вывод о спорности изложенных во введении возражений против её (этой импутации) использования. Вопреки этим распространенным представлениям, множественная импутация не является методологической роскошью, и ее функционал не сводится к тому, что в экономике называется *costly signals* – индикаторы больших затрат и усилий, призванных свидетельствовать об уровне подготовки и мотивации исследователя безотносительно прагматических потребностей, связанных с поставленной им целью. Современные средства обработки данных делают множественную импутацию относительно легкодоступной процедурой. Социальным ученым, которые сталкиваются с данной весьма распространенной проблемой, могут быть предложены следующие рекомендации.

1. Множественную импутацию имеет смысл проводить во всех случаях, когда по включенным в исследование переменным пропущено большое число значений. Как показано в приведенном примере, это относится не только к каждой переменной по отдельности, но и к их сочетаниям. В примере каждая из 10 переменных имеет относительно небольшое число пропусков, но, поскольку пропуски относятся к различным кейсам, факторная переменная из модели, построенной на всех этих переменных, имеет недопустимо большое количество пропусков.

2. Множественная импутация вполне применима к комплексным явлениям со сложной множественной причинностью. В рассмотренном примере использованная при множественной импутации модель, предсказательную силу которой нельзя оценить как высокую, тем не менее дает высокую степень соответствия между аналогичными сериями моделей, построенными на данных до и после импутации. Это связано с тем, что цель множественной импутации – не предсказание значений, а дополнение массива данных с наименьшими искажениями, по сравнению с альтернативными методами.

3. Множественную импутацию следует рассматривать не только как механическое увеличение объема данных, но и как

способ диагностики моделей. В рассмотренном в статье эмпирическом примере соответствие между моделями, построенными на исходных и дополненных массивах, указывает на адекватный подбор предикторов регрессионной модели, построение которой было основной задачей исследования. В тех случаях, когда наблюдается рассогласование, это означает, что на переменную, для которой заполняются пропуски, существенно влияют переменные вне модели (не соблюдается условие MAR), из-за чего результат итераций в значительной степени случайный, и поэтому результаты получаются искаженными. В этом случае следует либо скорректировать модель, включив в нее дополнительные переменные, либо, если сбор данных по необходимым переменным невозможен, скорректировать определение генеральной совокупности, которую репрезентирует массив данных. Таким образом, множественная импутация, в отличие от альтернативных способов восстановления пропущенных значений, представляет собой одновременно чисто техническую вспомогательную процедуру и средство диагностики возможностей и ограничений имеющихся эмпирических данных, результаты которого могут учитываться даже при корректировке цели и задач исследования.

ЛИТЕРАТУРА

1. *Lee E.L., Forthofer R.N.* Analyzing Complex Survey Data. Beverly Hills: Sage, 2006.

2. *Raftery A.E.* Bayesian Model Selection in Social Research // Sociological Methodology. 1995. No. 25. P. 111–164.

3. *Rubin D.B.* Multiple Imputation after 18+ years // Journal of the American Statistical Association.1996. No. 91(434). P. 473–489.

4. *King G. et al.* Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation // American Political Science Association. 2001. No. 95 (1). P. 49–69.

5. *Зангиева И.К.* Проблема пропусков в социологических данных: смысл и подходы к решению // Социология: методология, методы и математическое моделирование. 2011. № 33. С. 28–56.

6. *Acuna E., Rodriguez C.* The Treatment of Missing Values and Its Effect on Classifier Accuracy // Classification, Clustering, and Data Mining Applications. Berlin; Heidelberg: Springer, 2004. P. 639–647.

7. *Graham J.W.* Missing Data Analysis: Making It Work in the Real World // Annual Review of Psychology. 2009. N. 60. P. 549–576.

8. *Raaijmakers Q.A.W.* Effectiveness of Different Missing Data Treatments in Surveys with Likert-type Data: Introducing the Relative Mean Substitution Approach // Educational and Psychological Measurement. 1999. No. 59(5). P. 725–748.

9. *Rubin D.B.* Multiple Imputation for Nonresponse in Surveys. Vol. 81. New York: John Wiley & Sons, 2004.

10. *Schafer J.L., Olsen M.K.* Multiple Imputation for Multivariate Missing-data Problems: A Data Analyst's Perspective // Multivariate Behavioral Research. 1998. No. 33(4). P. 545–571.

11. *Baraldi A.N., Enders C.K.* An Introduction to Modern Missing Data Analyses // Journal of School Psychology. 2010. No. 48(1). P. 5–37.

12. *Koski J.* Defectiveness of Weighting Method in Multicriterion Optimization of Structures // Communications in Applied Numerical Methods. 1985. No. 1(6). P. 333–337.

13. *Van Buuren S., Brand J.P., Groothuis-Oudshoorn C.G., Rubin D.B.* Fully Conditional Specification in Multivariate Imputation // Journal of Statistical Computation and Simulation. 2006. No. 76(12). P. 1049–1064.

14. *Azur M.J., Stuart E.A., Frangakis C., Leaf P.J.* Multiple Imputation by Chained Equations: What Is It and How Does It Work? // International Journal of Methods in Psychiatric Research. 2011. No. 20(1). P. 40–49.

15. *Graham J.W., Olchowski A.E., Gilreath T.D.* How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory // Prevention Science. 2007. No. 8(3). P. 206–213.

16. *Fabrykant M., Magun, V.* Grounded and Normative Dimensions of National Pride in Comparative Perspective. Higher School of Economics Research Paper WP BRP, 62, 2015.

17. *Fabrykant M., Magun V.* Grounded and Normative Dimensions of National Pride in Comparative Perspective // Dynamics of National Identity: Media and Societal Factors of What We Are / Ed. *J. Grimm, L. Huddy, J. Seethaler, P. Schmidt*. London; New York: Routledge, 2016. P. 109–138.

18. *Van Buuren S.* Flexible Imputation of Missing Data. Boca Raton: CRC Press, 2012.

19. *Honaker J, King G, Blackwell M.* Amelia II: A Program for Missing Data // Journal of Statistical Software. 2011. No. 45(7). P. 1–47.

Приложение

В КАКОЙ МЕРЕ ВЫ ГОРДИТЕСЬ РОССИЕЙ ПО КАЖДОЙ ИЗ СЛЕДУЮЩИХ ХАРАКТЕРИСТИК?[1]

(*Дайте <u>один</u> ответ в каждой строке*)	Очень горжусь	В какой-то мере горжусь	Не очень горжусь	Совсем не горжусь	Затрудняюсь ответить
Положение дел с демократией	1	2	3	4	5
Политическое влияние России в мире	1	2	3	4	5
Экономические достижения России	1	2	3	4	5
Система социальной защиты населения (пенсионное обеспечение, помощь многодетным семьям и т.п.)	1	2	3	4	5
Научные и технические достижения России	1	2	3	4	5
Российские достижения в спорте	1	2	3	4	5
Российские достижения в области литературы и искусства	1	2	3	4	5
Российские вооруженные силы	1	2	3	4	5
Российская история	1	2	3	4	5
Положение дел с социальной справедливостью и равноправием всех групп в мире	1	2	3	4	5

[1] Вопросы взяты из русскоязычной версии опросника ISSP-2003 с сайта GESIS [URL: https://dbk.gesis.org/dbksearch/download.asp?db=E&id=6316 (дата обращения: 10.08.2016)].

Fabrykant Marharyta
National Research University Higher School of Economics (NRU HSE), Moscow, marharyta.fabrykant@gmail.com

Model-oriented approach to missing values: Multiple imputation in multilevel regression using R (on the example of analyzing survey data)

The article substantiates and describes the multiple imputation technique and its procedure of dealing with missing values in dataset. It presents the model-oriented approach as opposed to the design-oriented approach to analyzing survey data. The theory section lists the benefits of multiple imputation and states why it should be preferred to other ways of dealing with missing data, such as listwise or pairwise deletion or substitution by the mean. It states the limitations of the multiple imputation technique and the related conditions that the data must satisfy so that multiple imputation could be performed. The empirical section of the article draws on a specific case of the use of multiple imputation in a cross-cultural research on national pride. It shows the procedure in the form of a ready for use program code for diagnosing data and imputing missing values by means of R programming packages *VIM* and *mice*.
Key words: missing values, model-oriented approach, multiple imputation, MCAR, MAR, MNAR, R VIM, R mice

References

1. Lee E..l, Forthofer R.N. *Analyzing complex survey data*. Sage, 2006.
2. Raftery A. E. Bayesian model selection in social research, *Sociological methodology*, 1995, 25, 111–164.
3. Rubin D. B. Multiple imputation after 18+ years, J*ournal of the American Statistical Association*, 1996, 91 (434), 473–489.
4. King G. et al. Analyzing incomplete political science data: An alternative algorithm for multiple imputation, *American Political Science Association*, 2001, 95 (1), 49–69.
5. Zangieva I.K. Problema propuskov v sotsiologicheskih dannyh: smysl I podhody k resheniyu (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2012, 33, 28–56.
6. Acuna E., Rodriguez C. The treatment of missing values and its effect on classifier accuracy, in: *Classification, clustering, and data mining applications*. Berlin, Heidelberg: Springer, 2004. P. 639–647.
7. Graham J. W. Missing data analysis: Making it work in the real world, *Annual review of psychology*, 2009, 60, 549–576.

8. Raaijmakers Q. A. W. Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach, *Educational and Psychological Measurement*, 1999, 59 (5), 725–748.
9. Rubin D. B. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons, 2004.
10. Schafer J.L., Olsen M.K. Multiple imputation for multivariate missing-data problems: A data analyst's perspective, *Multivariate behavioral research*, 1998, 33(4), 545–571.
11. Baraldi A.N., Enders C.K. An introduction to modern missing data analyses, *Journal of School Psychology*, 2010, 48(1), 5–37.
12. Koski J. Defectiveness of weighting method in multicriterion optimization of structures, *Communications in applied numerical methods*, 1985, 1(6), 333–337.
13. Van Buuren S., Brand J.P., Groothuis-Oudshoorn C.G., Rubin D.B. Fully conditional specification in multivariate imputation, *Journal of statistical computation and simulation*, 2006, 76 (12), 1049–1064.
14. Azur M.J., Stuart E.A., Frangakis C., Leaf P.J. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 2011, 20(1), 40-49.
15. Graham J. W., Olchowski A. E., Gilreath T. D. How many imputations are really needed? Some practical clarifications of multiple imputation theory, *Prevention Science*. 2007, 8(3), 206–213.
16. Fabrykant M., Magun, V. *Grounded and Normative Dimensions of National Pride in Comparative Perspective*. Higher School of Economics Research Paper WP BRP, 62, 2015.
17. Fabrykant M., Magun V. Grounded and Normative Dimensions of National Pride in Comparative Perspective, in: Grimm J., Huddy L., Seethaler J., Schmidt P. (eds.) *Dynamics of National Identity: Media and Societal Factors of What We Are*. Routledge, 2016. P. 109-138.
18. Van Buuren S. *Flexible imputation of missing data*. CRC Press, 2012.
19. Honaker J, King G, Blackwell M. Amelia II: A program for missing data, *Journal of Statistical Software*, 2011, 45(7), 1–47.