ПРАКТИКИ СБОРА И АНАЛИЗА ФОРМАЛИЗОВАННЫХ ДАННЫХ

И.К. Зангиева
(*Москва*)

ПРОБЛЕМА ПРОПУСКОВ В СОЦИОЛОГИЧЕСКИХ ДАННЫХ: СМЫСЛ И ПОДХОДЫ К РЕШЕНИЮ

Содержание настоящей статьи может послужить первым шагом к выработке рекомендаций по выбору способа работы с отдельными пропусками в данных сегодня. В статье предлагаются методы учета конкретной исследовательской ситуации при выборе способов работы с пропусками после сбора данных; выделяются условия, при которых имеет смысл искусственно заполнять пропуски; предлагается способ сравнения алгоритмов заполнения пропусков на основе специально организованного эксперимента; рассматриваются возникающие при этом трудности; строится типология используемых для заполнения пропусков алгоритмов, дающая возможность сократить количество последних при сравнении.

Ключевые слова: пропуски в данных, неответы, заполнение пропусков, исследовательская ситуация.

1. Постановка исследовательской задачи

В настоящее время многие стратегические решения в сфере экономического и социального регулирования, маркетинге и консалтинге базируются на результатах массовых социологических исследований. Типичным недостатком большинства социологиче-

Ирина Казбековна Зангиева – аспирант и преподаватель кафедры методов сбора и анализа социологической информации НИУ «ВШЭ». E-mail: zangieva.irina@gmail.com.

ских опросов является большое количество частичных пропусков в данных (неответов респондентов на отдельные вопросы).

В различных руководствах и учебниках при изложении основ проектирования выборки нередко принимаются идеальные предпосылки. Предполагается, что отобранный респондент либо ответит на абсолютно все вопросы, либо совсем откажется участвовать. Иначе говоря, мы получим либо абсолютно полное наблюдение (*full response*), либо не получим наблюдения вообще (*unit – nonresponse*), что приведёт к возникновению проблемы недостижимости объекта.

На самом деле, чаще всего имеет место третья, промежуточная ситуация – респондент отвечает только на некоторые вопросы, другие же по тем или иным причинам остаются без ответа. В этом случае мы имеем дело с неполным наблюдением или отдельными пропусками (*item/partial nonresponse*). В данной статье мы не будем затрагивать вопросы недостижимости, а обратимся к проблеме наличия в данных отдельных неполных наблюдений (отдельных пропусков), потому что они присутствуют практически в каждом исследовании.

Это касается даже крупных исследований, которые в других отношениях могут служить примером методической организации массового анкетного опроса, например, НОБУС (Национальное обследование благосостояния домохозяйств и участия в социальных программах, 2003 г.), РМЭЗ (Российский мониторинг экономического положения и здоровья населения, ведущийся с 1996 г. по настоящее время), ОНПЗ (Опрос население по проблемам занятости, осуществляющийся ежеквартально начиная с 1995 г.).

Наличие в данных пропусков чревато следующими негативными последствиями для качества всего исследования:

1) сокращением общего размера доступной для анализа выборки;

2) потерей в качестве результатов из-за снижения валидности статистических выводов;

3) в некоторых случаях − возникновением в данных систематической ошибки измерения;

4) ограничениями в применении некоторых видов анализа. Так, например, при построении регрессионных моделей растет ошибка измерения, следовательно, страдает точность получаемого с их помощью прогноза [1, p. 492; 2, p. 84–88].

Поэтому проблему наличия пропусков в данных массовых социологических исследований нельзя недооценивать. Она требует подробного рассмотрения. Необходимо понять, почему возникают пропуски, какие бывают виды пропусков и какие существуют способы работы с ними.

Исходя из этого цель данной статьи – комплексное рассмотрение проблемы наличия в социологических данных пропусков. Комплексность достигается путем решения следующих задач (каждой из них посвящен отдельный параграф статьи).

1. Проанализировать основные причины возникновения отдельных пропусков.

2. Предложить классификации пропусков в зависимости от степени их случайности и описания конструктивных способов ее (случайности) оценки.

3. Описать общие подходы к работе с отдельными пропусками. Рассматриваться будут *post hoc* подходы, предполагающие работу с пропусками по факту их наличия, т.е. уже после сбора данных. За пределами внимания останутся меры по профилактике возникновения пропусков.

4. Подробно рассмотреть активно развивающийся подход к заполнению пропусков («множественное заполнение пропусков») и описать несколько алгоритмов такого заполнения.

5. Предложить методику экспериментального сравнения эффективности отдельных алгоритмов заполнения пропусков для дальнейшей разработки практических рекомендаций по их выбору.

2. Причины, виды пропусков и способы борьбы с ними

Отдельные пропуски могут быть двух видов: реальные и артефактные. *Реальные* пропуски возникают, если, несмотря на все усилия (методики стимулирования ответа, различные формы вопроса), от респондента не удалось получить ответа на вопрос. *Артефактные* (искусственные) пропуски возникают после удаления из массива нереалистичных, заведомо ложных значений, нарушающих логику последовательности ответов, которые приходится удалять самому исследователю на этапе чистки массива [3, p. 145]. Можно привести следующие примеры. Подросток в качестве уровня своего образования указывает «кандидат наук», а человек без определенного места жительства – площадь квартиры, в которой он якобы проживает на данный момент. Чтобы не вносить в данные заведомо ложную информацию, эти ответы из массива будут удалены, в результате чего и возникнут искусственные, или, как их еще называют, «артефактные» пропуски.

Мы будем говорить только о реальных пропусках, возникших из-за того, что респондент не ответил на некоторый вопрос. И начать необходимо с описания причин, по которым отдельные пропуски возникают (из-за чего респонденты не отвечают на вопрос).

Причины пропусков могут быть весьма различны. Респондент может не ответить на вопрос как по собственной «вине» (по причине своей некомпетентности в теме исследования, скрытности, психологической неконтактности и т.д.), так и по вине исследователя (неграмотно составленная анкета, обилие сенситивных вопросов, неудовлетворительная работа с интервьюерами), интервьюера (некорректно проводящего опрос) или кодировщика (допустившего ошибку при вводе информации в компьютер).

Ошибки исследователя и интервьюера можно отнести к методологическим причинам неответов, кодировщика – к техническим, а скрытность и неконтактность респондента – к психологическим.

Таким образом, методологические причины неответов охватывают все ошибки и недочеты, допущенные социологом при разработке инструментария и организации полевого этапа исследования, а также интервьюерами и анкетерами – непосредственно при сборе данных, тогда как психологические причины связаны с ситуацией опроса, рассматриваемой на уровне межличностного взаимодействия между интервьюером (анкетером) и респондентом. Следовательно, с психологической точки зрения отказ от ответа на вопрос обусловлен коммуникативными и психологическими особенностями различных социально-демографических групп.

Устранение методологических причин может потребовать доработки инструментария исследования и более тщательного инструктажа интервьюеров, технических – дополнительного обучения кодировщиков, психологических – применения в процессе интервью (анкетирования) психологических приемов преодоления недоверия, неконтактности респондентов.

Я не буду останавливаться здесь на ошибках исследователя и кодировщика, ограничившись рассмотрением некоторых аспектов «вины» интервьюера и респондента как основных участников коммуникации в рамках опроса. Далее основное внимание будет уделено описанию и устранению психологических причин неответов на вопросы (появления пропусков в данных).

Влияние личных особенностей интервьюера и респондента на результат их коммуникации друг с другом (получение или неполучение ответа от респондента)

Рассматривая роль интервьюера, естественно задаться вопросом: не будет ли способствовать преодолению нежелания респондента отвечать на предлагаемые ему вопросы и, соответственно, снижению вероятности возникновения пропусков, обеспечение близости интервьюера и респондента по каким-то их характеристикам?

В литературе показано, что высокая вероятность неответа респондента обусловлена прежде всего значительной величиной социальной дистанции между ним и интервьюером, степенью эмоционального отчуждения между ними. Чем сильнее выражены в ситуации интервью эти характеристики, тем выше вероятность того, что между участниками интервью не возникнет успешной коммуникации, и человек либо откажется участвовать в опросе в принципе, либо не ответит на многие вопросы [4, p. 236–238].

В данной статье мы не ставим целью вникать в глубины процесса коммуникации, а хотим затронуть лишь взаимосвязи простейших социально-демографических характеристик интервьюеров и респондентов. Так, например, эмпирически было установлено, что совпадение возраста респондента и интервьюера повышает вероятность получения ответа на вопрос, а совпадение уровня образования значимого влияния на нее (такую вероятность) не оказывает [1, p. 492; 2, p. 86].

Поэтому подбирая интервьюера из примерно той же возрастной группы, что и респондент, можно существенно повысить эффективность коммуникации между ними, снизить количество неответов. Однако на практике добиться такого соответствия можно далеко не всегда. Так, в случае использования при опросе маршрутной выборки запланировать совпадение возраста интервьюера и респондента невозможно.

Если же говорить о «вине» респондента в возникновении пропусков, то нужно понять, почему люди зачастую не отвечают на вопрос. Здесь можно указать три основные причины:

– некомпетентность респондента в теме вопроса [5, p. 50–54];

– неконтактность [6, p. 68–70];

– нежелание отвечать из-за недоверия к самой процедуре опроса и к личности интервьюера в частности [7, p. 422–425].

Если некомпетентность и неконтактность на практике преодолеть довольно трудно, то уровень доверия иногда можно повысить. Так, например, У. Ханнефельд – участник исследовательской группы, проводившей германскую социоэкономическую панель в

80-х годах XX в., отмечал, что в рамках данного проекта удалось добиться повышения доверия между респондентами и интервьюерами за счет гарантированного вознаграждения за участие в опросе (причем не всегда материального) и проведения до начала интервью установочной беседы, направленной на установление доверительных отношений между участниками [8, S. 178–180].

Перейдем к рассмотрению типов пропусков, которые могут возникнуть в данных.

Степень случайности пропусков как основание для их типологии

Любой способ работы с пропусками, как и всякий другой социологический метод работы с данными, опирается на определенные модельные представления исследователя об изучаемой ситуации. Об этом требуется сказать несколько слов.

На практике исследователю бывает трудно однозначно определить причину, по которой респонденты не ответили на отдельные вопросы (мы не рассматриваем ситуацию, когда неответы провоцируются искусственно в целях методического эксперимента). Кроме того, связь между причинами пропусков и способами «борьбы» с ними изучена слабо. В определенной мере разработана лишь одна линия: влияние причины пропуска на степень его случайности (в частности, на отсутствие таковой) и опосредованно – на выбор способа «борьбы» с пропусками.

В литературе выделяется три типа пропусков, в зависимости от их случайного или систематического характера: полностью случайные пропуски, случайные пропуски и неслучайные пропуски. Д. Рубин и Р. Литтл описывают эти пропуски следующим образом.

1. *Полностью случайными неответами (полностью случайными пропусками; missing completely at random* – MCAR) на какой-либо вопрос анкеты называются такие неответы, вероятность возникновения которых не зависит ни от истинного ответа,

который был бы получен, ни от ответов на другие вопросы. Например, неназывание возраста будет полностью случайным, если его вероятность одинакова среди респондентов вех возрастных групп (детей, подростков, взрослого населения, пожилых людей), т.е. она не зависит от истинного возраста респондента. При этом вероятность неуказания возраста должна быть одинакова для мужчин и женщин, людей с разным образованием и т.д., т.е. не должна зависеть от других характеристик респондента. Если отобрать из массива респондентов, полностью случайно не ответивших на вопрос (т.е. отобрать анкеты с полностью случайными пропусками), то они составят простую случайную выборку из всех опрошенных, не зависящую от наблюдаемых или ненаблюдаемых ответов. Такая независимость исключает систематические различия между людьми, ответившими и не ответившими на вопрос. Иначе говоря, наличие в данных полностью случайных пропусков равноценно непопаданию объекта в полностью случайную выборку объема, достаточного для достижения заданного уровня точности оценок. Поэтому имеющиеся в массиве данных полностью случайные пропуски не вызывают смещения в результатах анализа данных, а только снижают статистическую мощность при проверке гипотез [9; цит. по: 10, p. 155].

2. *Случайные неответы (случайные пропуски; missing at random* – MAR) отличаются от полностью неслучайных тем, что обусловлены известными значениями других переменных, но не связаны с переменной, значение которой пропущено. Так, неуказание своего возраста будет случайным, если его вероятность зависит от пола респондента, но не зависит от его истинного возраста. Например, если женщины не отвечают на вопрос о возрасте чаще чем мужчины, но при этом женщины разного возраста не отвечают на этот вопрос одинаково часто. Если проводить аналогию с выборкой, то можно сказать, что совокупность людей, случайно не ответивших на вопросы, представляет собой простую случайную выборку из страты, определяемой известным значением другого

признака. Например, в рассматриваемом случае с неназыванием возраста в роли стратифицирующего признака выступает пол [9; цит. по: 10, p. 155–156].

3. *Неслучайные, систематические неответы (неслучайные, систематические пропуски; not missing at random* – NMAR) возникают, если вероятность ответа на вопрос зависит от смысла самого вопроса. Иначе говоря, неслучайным будет отказ указать не просто возраст, а, например, возраст начала курения, потому что в обществе курение является социально неодобряемым поведением, и респондент предпочтет не ответить на вопрос, чем признаться, что начал курить в юности. Причем отказ будет обусловлен именно боязнью общественного порицания за курение в юном возрасте, а не просто тем фактом, что человек курит с юности. Вместе с тем неслучайные пропуски могут быть методически обусловленными. Например, ошибками в формулировке вопроса или перечне альтернатив [9; цит. по: 10, p. 156]. В последнем случае необходимо определить саму причину возникновения пропусков и устранить ее.

Поскольку предположение о характере случайности пропуска является существенным модельным предположением, заложенным в большинстве методов работы с пропусками, то задаче оценки степени случайности пропусков мы уделим особое внимание.

Прежде чем переходить к устранению пропусков в данных, нужно быть уверенными в том, что мы точно знаем, к какому из описанных типов эти пропуски относятся, поскольку, как мы отметили выше, от этого напрямую зависит выбор способа работы с пропусками уже после окончания сбора данных. Уверенность эта должна быть основана на применении конструктивных критериев оценивания степени случайности пропусков. Под конструктивными мы подразумеваем объективные формальные статистические критерии, основанные на проверке статистических гипотез. Рассмотрим критерии, часто упоминаемые в методической литературе, прежде всего зарубежной. Мы рассмотрим только те способы,

которые предполагают работу с уже собранными данными (*post hoc*), и не предполагают исследования инструментария или процедур сбора данных, т.е. изучения возможных причин появления пропусков.

Возможные способы конструктивного определения степени случайности пропусков

Не существует критериев, которые бы помогали однозначно определить, полностью случайны пропуски или нет. Однако есть критерии, позволяющие подтвердить или опровергнуть гипотезу о неполной случайности пропусков (NMCAR). Неполная случайность пропусков предполагает либо их просто случайность (MAR), либо систематичность (NMAR).

Чаще всего в литературе, посвященной работе с пропущенными данными, упоминаются три способа установления степени случайности пропусков: показатели DRSS и DXX, процедура Дж. и П. Коэнов (J. Cohen, P. Cohen).

Все названные способы в качестве основного средства используют регрессионные уравнения и анализ изменений их параметров в зависимости от того, каким способом были заполнены пропуски в данных, на которых эти уравнения строились. Причем в этих моделях обязательно должны быть известны все значения зависимой переменной.

Показатель DRSS оценивает изменение суммы квадратов остатков в регрессионном уравнении, построенном на нескольких массивах данных, пропуски в которых заполнены несколькими способами. Если суммы квадратов значимо различаются, когда пропуски заполнены разными методами, значит, пропуски не полностью случайны. Если же суммы квадратов совпадают, пропуски можно считать полностью случайными. Показатель DXX построен на оценке изменения матрицы $X'X$, где X – ковариационная матрица для независимых переменных. В данном случае считается, что пропуски не полностью случайны, если при заполнении их раз-

личными методами значимо изменяется ковариационная матрица признаков [11, p. 15–18; 12, p. 206–207; 13, p. 17–22].

Дж. Коэн и П. Коэн предложили следующий способ выявления неполной случайности пропусков. Для каждой изучаемой независимой переменной в регрессионной модели создается индикаторная переменная со значениями 0, если информация у объекта по данной переменной присутствует, и 1 – если она отсутствует. Затем оценивается разница в средних значениях зависимой переменной *y* (для которой известны все значения) в группах с известными и неизвестными значениями каждой независимой переменной.

Проверяется статистическая гипотеза $H_0 : \overline{y}_o = \overline{y}_m$, где y – значения зависимой переменной в рассматриваемой регрессионной модели, индекс ***o*** (от англ. *observed)* соответствует объектам, для которых известно значение рассматриваемой независимой переменой в количестве ($n - m$), индекс m – объектам, для которых значение рассматриваемой независимой переменной пропущено в количестве ($n - o$), где n – общий объем выборки.

Этот способ диагностики основан на предположении о том, что если пропуски в данных полностью случайны, то разница в значениях *y* при известном и неизвестном *x* будет статистически незначимой, а если расхождение между $\overline{y}_0$ и $\overline{y}_m$ значимо, то можно сделать вывод о том, что неответы на отдельные вопросы имеют неслучайный характер [14, p. 281–289].

Перейдем к рассмотрению подходов к работе с пропущенными данными.

3. Подходы к работе с пропусками в данных

Существует три основных подхода к работе с пропусками: исключение из анализа неполных наблюдений (анкет), взвешивание данных, заполнение пропусков (импутирование). Дадим характеристику каждого подхода, а затем более подробно рассмотрим подход, состоящий в заполнении пропусков.

Исключение неполных наблюдений

Этот подход заключается в том, что наблюдения (анкеты, случаи), для которых отсутствуют значения хотя бы одного интересующего нас признака, исключаются из анализа.

Основными преимуществами подхода являются легкость и быстрота его реализации. Если пропусков в данных немного, то удаление неполных наблюдений, по мнению многих авторов, лучшее решение проблемы, так как влечет за собой только незначительное снижение статистической мощности выводов. Однако это справедливо только для полностью случайных и случайных пропусков в данных. О том, как соотносятся типы пропусков по степени случайности и допустимые подходы к работе с ними, будет сказано далее.

Однако при этом теряется информация по признакам (переменным), не содержавшим пропущенных значений, так как наблюдение удаляется целиком (списочное удаление) (*listwise deletion*) [15, p. 106–107]. Такие потери информации можно уменьшить, используя попарное удаление пропусков (*pairwise deletion*), когда, например, при расчёте коэффициентов корреляции между двумя признаками из анализа исключаются только объекты, для которых неизвестно значение хотя бы одной из двух переменных, а не все объекты с пропусками, как при списочном удалении.

Проблема стоит тем более остро, что специфика социологических данных диктует постоянную необходимость «чистки» массива не только в отношении пропущенных данных, но и для удаления артефактов, откровенно ложных ответов и т.д.

Взвешивание данных

Взвешивание данных заключается в том, что исследователь сначала удаляет все неполные наблюдения, а затем использует оставшиеся данные во взвешенном виде.

Взвешивание выборки осуществляется таким образом, чтобы сохранить ее структуру (квоту) по некоторым заданным исследо-

вателем признакам. Пропуски в данном случае по-прежнему не заполняются, а неполные анкеты заменяются другими, взятыми из имеющихся полностью заполненных [3, p. 83–86]. Поясним сказанное на примере.

Предположим, что исследователь заинтересован в сохранении исходной структуры выборки по полу и что в первичной выборке была 1000 женщин. Пусть на вопрос о возрасте не ответили 500 женщин. Соответствующие 500 анкет исключаются из рассмотрения. Естественно, структура совокупности по полу в результате меняется. Чтобы ее восстановить, необходимо искусственно добавить 500 женщин, чтобы их снова стало 1000 и тем самым соотношение между количеством мужчин и количеством женщин сохранилось. Это делается с помощью дублирования всех анкет женщин, оставшихся в выборке (т.е. ответивших на вопрос о возрасте). В таком случае говорят, что мы использовали целочисленный весовой коэффициент 2 для полных наблюдений. Если же нам необходимо восполнить, например, 40% женщин (на вопрос не ответили 400 женщин), тогда мы из ответивших 600 женщин случайным образом отберем 400, и продублируем в массиве уже только их анкеты. Таким образом, ответивших женщин окажется в массиве в 1,6 раза больше, чем было изначально. В данном случае мы имеем дело с дробным весовым коэффициентом для полных наблюдений, равным 1,6.

Однако и взвешивание данных как решение проблемы пропусков не лишено недостатков. Перечислим их, опираясь на [16, с. 112–114].

1. Структура выборочной совокупности по выбранной переменной (в примере выше это был пол) смещается в сторону той структуры, которая имеет место для полных (относительно возраста) наблюдений. В ней начинают преобладать значения целевого признака (переменной, значения которой необходимо восстановить; в нашем примере – возраста), характерные для взвешенных полных наблюдений, тем самым уменьшается дисперсия целевой переменной.

2. Появляются новые смещения или увеличиваются уже имеющиеся. Новые смещения появляются за счет того, что при взвешивании данных по одной переменной (в нашем примере – по полу) при анализе других переменных-признаков (значения которых могут и не быть пропущенными) веса сохраняются, в результате чего выводы по ним получатся искаженными. Сохранив квоту по выбранному признаку (в приведенном выше примере это пол), мы может резко нарушить первоначальные пропорции по каким-то другим признакам (скажем, по образованию), в результате чего общие выводы могут стать весьма искаженными.

Принимая решение об использовании весов, необходимо учитывать все перечисленные факторы. Иногда, когда масштабы пропусков в данных незначительны относительно общего объема выборочной совокупности, лучше отказаться от взвешивания и анализировать только полные наблюдения.

Исключение неполных наблюдений и взвешивание данных, описанные выше, это традиционные подходы к работе с пропусками. Ниже будет описан более прогрессивный и активно развивающийся подход – заполнение пропусков, или импутирование (от англ. *imputation*).

Заполнение пропусков (импутирование)

Основной задачей исследователя, решившего заполнить пропуски, является восстановление исходной структуры данных, которая имела бы место, если бы отсутствующие значения не были пропущены.

Сначала при помощи различных способов пропуски в данных заполняются, а затем полученный массив полных данных обрабатывают при помощи заранее предусмотренных методов анализа как массив без пропусков. При этом не делается никаких различий между наблюдениями, полными изначально, и наблюдениями, значения характеристик которых были приписаны [17, S. 416–420; 18, p. 230–232].

Естественно, что такой подход не лишен и недостатков. При заполнении пропусков мы будем иметь дело с частичным отходом от репрезентативности за счет возможного возникновения в результате его применения смещений в структуре массива относительно реальной ситуации [19, p. 17–36]. Но названный недостаток ни в коей мере не перекрывает основного и наиболее важного достоинства заполнения пропусков. Заполнение пропусков позволяет на выходе получить массив полных данных рассчитанного на этапе дизайна выборки размера и реализовывать на нем методы анализа данных, требующие их абсолютной полноты.

После отдельного рассмотрения типов пропусков по степени случайности, с одной стороны, и основных способов работы с ними – с другой, необходимо соотнести их между собой, для того чтобы понять, каким образом можно работать с пропусками, характеризующимися разной степенью случайности.

Связь типов пропусков и способов работы с ними

Если возникновение пропусков в данных полностью случайно или случайно, то их называют *игнорируемыми*, поскольку их можно удалить (когда их немного, о чем мы уже упоминали выше) или заполнить. Под игнорируемостью в данном случае понимается возможность смириться с наличием пропусков и скорректировать их (удалить или заполнить с помощью существующих алогритмов заполнения пропусков, некоторые из которых будут описаны ниже).

Если же пропуски не случайны, а носят систематический характер, тогда они считаются *неигнорируемыми*. И просто так ограничиться заполнением или удалением пропусков нельзя, так как оно внесет в данные систематическую ошибку

При наличии неслучайных пропусков, как говорилось выше, необходима доработка самого способа сбора данных и инструментария для устранения ошибок, повлекших возникновение систематических пропусков, либо, если доля неслучайных про-

пусков незначительна – использование при заполнении поправок, статистически нивелирующих систематичность пропусков [20, с. 286].

Неслучайные пропуски, как уже было отмечено выше, скорректировать нельзя, их можно только предупредить еще до сбора данных. Поэтому далее мы сосредоточимся на работе со случайными (полностью случайными) пропусками, которые можно «вылечить» уже после сбора данных, при этом рассмотрим только наиболее сложный и наиболее перспективный способ «лечения» – заполнение пропусков.

На данный момент разработано значительное количество (несколько десятков) отдельных алгоритмов заполнения пропусков в данных. Здесь мы рассмотрим только некоторые из них, наиболее часто упоминаемые в современной методической литературе как традиционные (подстановка мер центральной тенденции), так и более современные (*HotDeck*, EM, регрессионное моделирование пропусков и множественное заполнение пропусков).

4. Краткое описание некоторых алгоритмов заполнения пропусков

Заполнение пропусков мерами центральной тенденции

Подстановка среднего арифметического − самый простой алгоритм, с которого обычно в литературе начинается описание имеющихся способов заполнения пропусков в данных.

Но среднее арифметическое имеет смысл только для признаков, измеренных по шкалам не ниже интервальной. Поэтому в случае категориальных (номинальных и порядковых) шкал нужно говорить о заполнении пропусков не средним арифметическими, а подходящими для них мерами центральной тенденции: модой для номинальных шкал и модой или медианой – для порядковых.

Данный метод не требует обязательного применения специального программного обеспечения, так как он реализуем даже в пакете *Microsoft Excel*. Однако заполнение пропусков мерами центральной тенденции «усредняет» данные, уменьшая степень разброса значений признака, и приближает структуру итогового массива к структуре только полных наблюдений, не учитывая возможные особенности объектов с пропусками [21, p. 55–56].

HotDeck

HotDeck[1], или как он еще называется метод «ближайшего соседа», представляет собой подстановку на место пропуска известного значения признака, принадлежащего объекту, наиболее близкому к рассматриваемому объекту с пропуском.

При подборе ближайшего объекта (наблюдения) от рассматриваемого наблюдения с пропуском рассчитывается расстояние *d* до каждого из полных наблюдений. Значения признаков, по которым рассчитывается расстояние, должны быть известны как для объекта с пропуском, так и для всех объектов, из которых выбирается ближайших к нему. Меру расстояния выбирает сам исследователь исходя из типа используемых данных, своих представлений о характере связи между переменными и стоящих перед ним конкретных задач. Чаще всего используются евклидово расстояние или его квадрат, расстояние Манхеттен, расстояние Миньковского и так далее [22, p. 104–105].

[1] В данном случае под «колодой» понимается именно набор полных наблюдений, наиболее близких к наблюдению с пропусками. Сначала задача исследователя эту колоду сформировать, а затем уже из нее выбирать объект, у которого будет взято для подстановки известное значение признака. Так, например, в Американском бюро переписей при заполнении пропусков в ответах одного респондента, в качестве такой «колоды» использовались респонденты из того же района и уже из их числа случайным образом отбирался респондент, известный ответ которого и приписывался не ответившему на данный вопрос респонденту.

EM-алгоритм

EM-алгоритм (от англ. *expectation maximization* − максимизация ожидания) в самом общем смысле представляет собой итерационную процедуру по оптимизации некоторого функционала, отражающего степень соответствия между ожидаемыми и реально подставляемыми на место пропуска значениями.

В случае использования EM-алгоритма для заполнения пропусков, его можно описать следующим образом.

Выделяются два семейства законов плотности условных распределений двух взаимосвязаных переменных x и y, зависящих от параметра φ: $f(x|\varphi)$ и $g(y|\varphi)$. Функция $f(x|\varphi)$ описывает полные наблюдения переменной x, $g(y|\varphi)$ – наблюдения, содержащие пропуски по переменной y. Причем, функция для неполных наблюдений определяется интегрированием функции полных наблюдений f:

$$g(y|\varphi) = \int_{X(y)} f(x|\varphi)\, dx.$$

Задача EM-алгоритма найти значение параметра φ, максимизирующее функцию $g(y|\varphi)$ при наблюдаемом значении переменной y, через использование соответствующего распределения переменной x. Необходимо учитывать, что несколько различных значений функции $f(x|\varphi)$ могут давать в результате одно и то же значение $g(y|\varphi)$ [22, p. 98–100].

Регрессионное моделирование пропусков

В большинстве случаев, заполнение пропусков при помощи регрессионных моделей осуществляется в два этапа:

- на первом этапе на совокупности полных наблюдений строится регрессионная модель и оцениваются коэффициенты в уравнении, где в качестве зависимой переменной выступает целевая переменная, пропущенные значения которой необходимо восстановить;
- затем, по полученному на предыдущем этапе уравнению, куда подставляются известные значения независимых переменных-

предикторов, для каждого неполного наблюдения рассчитывается отсутствующее значение по зависимой целевой переменной. В случае интервальных и абсолютных переменных рассчитывается конкретное значение, а для порядковых и номинальных переменных с некоторой вероятностью предсказывается категория, к которой должен быть отнесен объект.

Существует несколько модификаций классической регрессионной модели, которые могут быть использованы для заполнения пропусков по переменным, измеренным по шкалам ниже интервального уровня [23].

Множественное заполнение пропусков

На сегодня это наиболее активно развиваемый и описываемый в литературе алгоритм.

Техника множественного заполнения пропусков, разработанная Д. Рубиным и Р. Литтлом, предусматривает подстановку сразу нескольких значений на место каждого из пропусков. Причем существенный разброс подобранных для заполнения одного пропуска значений свидетельствует о неопределенности типа пропусков и причин их возникновения [24, с. 473–489].

Данные с каждым набором заполненных пропусков сохраняются в общем массиве, который теперь имеет размер $k \times n$, где k – количество итераций алгоритма, а n – объем исходного массива данных. Чтобы затем достичь исходного размера выборки при последующих расчётах, полученный массив перевзвешивается и каждому его объекту присваивается весовой коэффициент $1/k$. Также возможно получение одного агрегированного массива, который затем анализируется как обычный полный массив [25, p. 246–248].

Множественное заполнение в практике современного анализа данных массовых опросов становится все более и более распространенным и реализуется во многих специализированных

пакетах, предназначенных для анализа данных. Например, в SPSS, SOLAS, MICE [26, p. 246–248].

Каждый из наборов получается с использованием методов, основанных на одной из трех моделей множественного заполнения пропусков: 1) предиктивной; 2) степени предрасположенности; 3) дискриминантной. Рассмотрим каждую из них. Следует отметить, что первые две модели предназначены для работы только с непрерывными переменными, а третья модель – с дискретными.

Множественное заполнение пропусков на основе предиктивной модели (predictive model based method)

В предиктивной (прогнозной) модели предполагается, что вместо пропущенного значения подставляется ближайшее к спрогнозированному реальное значение переменной, принадлежащее полному наблюдению. Для расчета прогнозируемого значения пропуска в модели на совокупности полных данных строится линейное регрессионное уравнение вида: $E\,[Z_j|\varphi] = \phi_0 + \phi_1 Z_1 + \phi_2 Z_2 + \ldots + \phi_{j-1} Z_{j-1}$, где Z_i – значения рассматриваемой переменной для полных наблюдений [26, p. 245].

Затем, после оценки регрессионных коэффициентов на совокупности полных наблюдений, которые обозначаются как φ^*, рассчитывается прогнозное значение для каждого пропуска, что выглядит следующим образом: $Z_j^{(i)}\phi_0^* + \phi_1^* Z_1 + \phi_2^* Z_2 + \ldots + \phi_{j-1}^* Z_{j-1} + \sigma^* \varepsilon$, где σ^* – рассчитанная исходя из модели дисперсия значений переменной, а ε – нормально распределенная случайная величина [27, p. 9].

Множественное заполнение пропусков на основе модели степени предрасположенности (propensity scores model based method)

Данная модель основана на оценках предрасположенности (*propensity scores*) респондента к ответу, т.е. на вероятности получить его ответ на вопрос, и реализована в виде нескольких ал-

горитмов множественного заполнения пропусков [28, p. 586–587]. Далее мы лишь кратко опишем приложение оценок предрасположенности респондента для ответа только в множественном заполнении пропусков, без их детального описания.

При использовании модели степени предрасположенности для подстановки выбирается значение целевой характеристики у объекта, ближайшего к целевому объекту. В каждом случае используется разная функция расстояния. В качестве меры близости служит логистическая модель индикаторов неполноты наблюдений, которые принимают значение 1 – если наблюдение полное, и 0 – если оно неполное.

Однако многие авторы отмечают, что множественное заполнение пропусков на основе степени *предрасположен*ности в некоторых случаях вызывает существенное смещение в структуре данных и результатов [29, с. 306].

Множественное заполнение пропусков на основе дискриминантной модели (discriminant model based method)

Дискриминантная модель используется в случае любых дискретных переменных (порядковых или номинальных).

Предсказание отсутствующего значения здесь опирается на вероятность $P(Z_j^{mis} = k \left| Z^{obs} \right.)$ того, что отсутствующее значение в реальности принимает конкретное значение, где Z – значение целевой переменной, отсутствующее или имеющееся соответственно. Это вероятность может быть рассчитана по теореме Байеса из оценок параметров совместного распределения значений целевой переменной (Z) [30, p. 768].

Несколько авторов отмечают, что при небольших объемах игнорируемых пропусков (полностью случайных или случайных) множественное заполнение пропусков даже с небольшим количеством итераций позволяет получить состоятельные оценки и валидные выводы [24, p. 475; 27, p. 308].

Однако максимальную точность данный метод дает в случае количественных, непрерывных переменных, что несколько сокращает область его применения [31, p. 20–21].

Алгоритмов заполнения пропусков существует много, в том числе и рассчитанных на схожие методы анализа данных. Выше были описаны только пять из них, наиболее распространённые в современной исследовательской практике. Поэтому очевидно встает проблема выбора подходящего алгоритма заполнения пропусков.

5. Проблема выбора подходящего алгоритма заполнения пропусков

Эксперимент как основа для разработки рекомендаций по выбору алгоритма заполнения пропусков

Осмысление того или иного алгоритма заполнения пропусков в методической литературе сводится обычно к простому описанию теоретической сути этого алгоритма и демонстрации практического применения процедуры. Привлекательность алгоритма обычно объясняют его математическими достоинствами или недостатками, без намека на то, что полученный результат может быть обусловлен не только объективными свойствами алгоритма, но и его эффективностью в зависимости от других параметров, например метода анализа данных, который предполагается использовать после заполнения пропусков, и их количества. Заполнение пропусков будет адекватным и точным, только если выбирать конкретный алгоритм с учетом количества имеющихся пропусков и метода, с помощью которого предполагается затем анализировать данные. Скорее всего одни алгоритмы лучше работают с одними методами анализа данных, а другие с другими, некоторые неэффективны, когда в данных большое количество пропусков, а другие эффективны даже в этом случае.

При современном состоянии науки, когда разработано уже большое количество разнообразных алгоритмов заполнения пропусков, эффективность каждого из них в конкретных исследовательских ситуациях можно оценить только экспериментально. Для этого необходимо сравнить результаты, полученные после заполнения определенным способом искусственно внесённых в массив пропусков, с идеальным случаем – результатами, полученными на массиве полных данных до внесения в него искусственных пропусков.

Только при учете упомянутых параметров в качестве критериев для экспериментального сравнения алгоритмов заполнения пропусков возможно получение валидных оценок эффективности каждого из них. Получение этих оценок позволит построить пошаговую схему для выбора подходящего способа заполнения пропусков. Подобная схема может лечь в основу практических рекомендаций по выбору метода заполнения пропусков в каждом конкретном случае.

Соответствующий методический эксперимент, в котором количество пропусков и метод последующего анализа данных выступают в качестве параметров для сравнения эффективности отдельных алгоритмов, можно организовать следующим образом:

1) сформировать массив, состоящий только из полных наблюдений;

2) оценить на нем параметры распределений переменных и осуществить интересующие виды анализа. Полученные только на полных наблюдениях результаты станут эталоном − базой для дальнейших сравнений;

3) внести в массив полных наблюдений различное количество случайных пропусков. В итоге получится совокупность массивов данных разной степени полноты;

4) на каждом из них повторить шаг 2 и оценить смещения в оценках, возникшие в результате наличия в данных разного количества пропусков;

5) заполнить в каждом массиве пропуски с использованием нескольких сравниваемых методов заполнения пропусков. Повторить на каждом полном массиве шаг 2.

6) сравнить полученные при использовании каждого из методов заполнения пропусков результаты с эталонными результатами, полученными на шаге 2.

7) выбрать для каждой ситуации наиболее эффективный метод заполнения пропусков и отразить их схематически.

Теоретически при реализации эксперимента нельзя исключать гипотетическую возможность того, что на полном массиве могут быть получены смещенные оценки, а на массиве с восстановленными пропусками – нет. Это возможно, если допустить, что на отдельные вопросы не отвечают респонденты, чьи истинные ответы, которые могли бы быть даны, подвержены некоторым систематическим смещениям. Заполнение пропусков с помощью подходящего алгоритма могло бы эти смещения устранить. Однако такая ситуация маловероятна и поэтому ею можно пренебречь и при реализации эксперимента не рассматривать.

Однако при планировании эксперимента возникает серьезная проблема – невозможность сравнения в рамках одного эксперимента всех существующих алгоритмов заполнения пропусков, которых, как уже было отмечено выше, насчитывается несколько десятков. Встает еще одна задача – выбор способа обоснованного сокращения количества рассматриваемых алгоритмов. Наши предложения по этому поводу представлены в следующем параграфе.

Типология алгоритмов заполнения пропусков как обоснование сокращения количества сравниваемых алгоритмов

Основные алгоритмы заполнения случайных пропусков можно разделить на две основные группы: простые и сложные алгоритмы. Сложные алгоритмы в свою очередь подразделяются на локальные и глобальные. Идею выделять простые и сложные,

глобальные и локальные алгоритмы предложил Р. Литтл [31, p. 9]. Мы же постарались наполнить эту классификацию конкретными примерами, и прежде всего отразить в ней не только «старые», описанные самим Литтлом алгоритмы, такие как заполнение пропусков средним арифметическим и регрессионное моделирование пропусков, но и относительно новые и распространенные в современной практике алгоритмы, которые не могли быть известны Литтлу на момент публикации им соответствующей статьи. Например, *HotDeck*[1], EM-алгоритм и множественное заполнение пропусков. Тем самым нами фактически показано, что его классификация актуальна до сих пор.

Итак, покажем, что новые алгоритмы вписываются в схему Литтла, представленную на *рис.* 1. Простые алгоритмы – неитеративные алгоритмы, основанные на простых арифметических операциях, мерах расстояния между объектами: например, заполнение пропусков мерами центральной тенденции, регрессионное моделирование пропусков. Нами в эту группу был отнесен алгоритм *HotDeck*.

Сложные алгоритмы – итеративные алгоритмы, предполагающие оптимизацию некоторого функционала, отражающего степень соответствия подставляемых на место пропусков значений структуре полных наблюдений или оценкам структуры данных в генеральной совокупности. Они в свою очередь подразделяются на глобальные и локальные алгоритмы.

Глобальные алгоритмы – в оценивании (предсказании) каждого пропущенного значения участвуют все объекты рассматриваемой совокупности. В данную группу мы отнесли *EM*-алгоритм.

[1] В данном случае в качестве относительно новых упомянута обновленная версия *HotDeck*, в котором «колода» полных наблюдений формируется через отбор наблюдений, наиболее близких (с наименьшим расстоянием по нескольким признакам) к неполному. В традиционной версии алгоритма, зародившейся, как уже было отмечено, в Бюро переписей США, подбор наблюдения для заполнения пропусков в более ранний период осуществлялся на основе рандомизации.

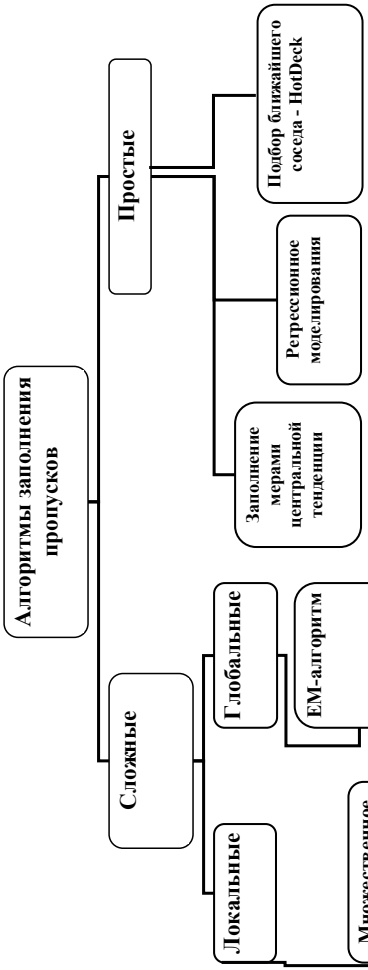

Рис. 1. Классификация методов заполнения пропусков [31, p. 9–11]

Локальные алгоритмы – в оценивании (предсказании) каждого пропущенного значения участвуют полные наблюдения, находящиеся в некоторой окрестности предсказываемого объекта: например множественное заполнение пропусков.

Таким образом, имея данную классификацию алгоритмов заполнения пропусков, основания которой заложены Литтлом, и еще дополнив группы, для проведения эксперимента из каждой группы нам необходимо отобрать один способ, и уже только для них проводить экспериментальное сравнение по описанной выше методике.

Это позволит существенно сократить временные затраты на проведение эксперимента, но при этом сравнить между собой все имеющиеся типы алгоритмов заполнения пропусков.

Заключение

Итак, мы обсудили проблему наличия в данных социологических опросов отдельных пропусков, возникающих, если респонденты не отвечают на те или иные вопросы. Были описаны причины возникновения, типы пропусков в данных и подходы к работе с ними.

Особое внимание было уделено одному из наиболее перспективных и активно развивающихся подходов – заполнению пропусков (импутированию). Предложена методика экспериментальной оценки эффективности отдельных алгоритмов заполнения пропусков с целью формулирования практических рекомендаций по выбору подходящего алгоритма в зависимости от конкретной ситуации. В дальнейшем результаты предлагаемого методического эксперимента и сформулированные на их основе рекомендации по выбору алгоритма заполнения пропусков будут представлены в виде отдельной публикации.

ЛИТЕРАТУРА

1. *Lillard L., Smith J.P., Welch F.* What Do We Really Know About Wages? The Importance of Nonresponce and Census Imputation // Journal of Political Economy. 1986. No. 5. P. 489–506.

2. *Sousa-Poza A., Hanneberger F.* Wage Data Collected by Telephone Interviews: an Empirical Analysis of the Item Nonresponce Problem and Its Implications for the Estimation of Wage Functions // Schweizerishee Zeitschift fur Volkswirtschaft und Statistik. 2000. No. 136. S. 79–98.

3. *Sande I.* Imputation in Surveys: Coping with Reality // The American Statistician. 1982. No. 36(3). P. 145–153.

4. *Dilman D.A.* Mail and Telephone Surveys: The Total Design Method. N.Y.: John Willey and Sons, 1978. P. 230–245.

5. *Pickery J., Billiet J.* Item Non-response as a Predictor of Unit Non-response in a Panel Survey // International Conference on Survey Non-response. Portland Oregon (USA), 1999. P. 48–67.

6. *Shrapler J-P*. Respondent Behavior in Panel Studies: A Case Study of the German Socio – Economic Panel (GSOEP) // DIW Discussion papers. DIW – Berlin. 2001. No. 224. P. 65–74.

7. *Hill D., Willis R.J.* Reducing Panel Attrition: A Search for Effective Policy Instruments // Journal of Human Recourses. 2001. No. 36 (3). P. 416–438.

8. *Hanefeldt U.* Das sozio-ekonomische Panel: Grundlagen und Konzeption. Frankfurt/Main: Campus Verlag, 1987. S. 171–186.

9. *Little R.J., Rubin D.B.* Statistical Analysis with Missing Data. N.Y.: John Willey and Sons, 1978.

10. *De Leeuw E. D., Joop H., Huisman M.* Prevention and Treatment of Item Nonresponse // Journal of Official Statistics. 2003. No. 19(2). P. 145–160.

11. *Simonoff J.S.* Regression Diagnostics to Detect Nonrandom Missingness in Linear Regression // Technometrics. 1988. No. 30(2).

12. *Hoaglin D.C., Welsch R. E.* The Hat Matrix in Regression and ANOVA // The American Statistician. 1978. No. 32. P. 17–22.

13. *Cohen J., Cohen P.* Applied Multiple Regression // Correlation Analysis for the Behavioral Sciences. 2nd ed. N.J.: Lawrence Erlbaum, 1983. P. 281–289.

14. *Zweimuller J.* Survey Non-response and Biases in Wage Regressions // Economic Letters. 1992. No. 39. P. 105–109.

15. *Biewen M.* Item Non-response and Inequality Measurement: Evidence from the German Earnings Distribution // Allgemeines Statistiches Archive. 2001. No. 85. P. 409–425.

16. *Чуриков А.В.* Основы формирования выборки: лекции для студентов направления 521200 (Социология). М.: ГУ–ВШЭ, 2005.

17. *Daubler T.* Nonresponseanalysen der Stichprobe F des SOEP // DIW Berlin. 2002. No. 15. S. 219–249.

18. *Barnard. J.,Meng. X.L.* Applications of Multiple Imputation in Medical Studies: from AIDS to NHANES // Statistical Methods in Medical Research. 1999. No. 8. P. 227–244.

19. *Little R.J.* Survey Nonresponse Adjustments for Estimates of Means // International Statistical Review / Revue Internationale de Statistique. 1976. No. 54(8). P. 139–157.

20. *Злоба Е., Яцкив И.* Статистические методы восстановления пропущенных данных // Computer Modelling &New Technologies. 2002. № 6. С. 55–56.

21. *Delyon B., Lavielle M., Moulines E.* Convergence of a Stochastic Approximation Version of the EM Algorithm // The Annals of Statistics. 1999. No. 27. P. 94–128.

22. *Wu C.F.* On the Convergence Properties of the EM-algorithm // The Annals of Statistics. 1983. No. 11(1). P. 95–103.

23. SPSS Missing Value Analysis 12.0 [on-line]. URL: http://www.spss.ru/products/missing_value/mva12.pdf.

24. *Rubin D.B.* Multiple Imputation after 18+ Years // Journal of the American Statistical Association. 1996. No. 91. P. 473–489.

25. *Horton N. J, Lipsitz S.R.* Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables // The American Statistician. 2001. No. 55(3). P. 244–254.

26. *Schafer J. L.* Multiple Imputation: A Primer // Statistical Methods in Medical Research. 2001. No. 8. P. 3–15.

27. *Allison P.D.* Multiple Imputation for Missing Data: A Cautionary Tale // Sociological Methods and Research. 2000. No. 28. P. 301–309.

28. *Zhang P.* Multiple Imputation: Theory and Method // International Statistical Review / Revue Internationale de Statistique. 2003. No. 71(3). P. 581–592.

29. *Ibrahim J. G.* Incomplete Data in Generalized Linear Models // Journal of American Statistical Association. 1990. No. 85. P. 765–769.

30. *Barnard J., Rubin D.B., Zanutto E.* Lecture Notes of the Short Course on Multiple Imputation for Missing Data. Utrecht, 1997.

31. *Kalton G., Kasprzyk D.* The Treatment of Missing Survey Data // Survey Methodology. 1986. No. 12. P. 1–16.

32. *Allison P.D.* Missing Data // Sage University Papers Series on Quantitative Applications in the Social Sciences. CA.: Sage, 2001. No. 136.