Зангиева Ирина

выпускница 2008 года магистерской программы

«Прикладные методы социального анализа рынков» ГУ-ВШЭ.

РЕШЕНИЕ ПРОБЛЕМЫ НЕПОЛНОТЫ ДАННЫХ МАССОВЫХ ОПРОСОВ

ВВЕДЕНИЕ. ПОСТАНОВКА ПРОБЛЕМЫ

Часто при принятии важных стратегических решений, касающихся например, рынка труда, используются результаты социологических исследований, основанных на массовых опросах. Самыми известными исследованиями являются: НОБУС (Национальное обследование благосостояния домохозяйств и участия в социальных программах, 2003 год), РМЭЗ (Российский мониторинг экономики и здоровья 1996-2006 годы), ОНПЗ (Опрос население по проблемам занятости, ежеквартально, начиная с 1995 года).

Во всех перечисленных исследованиях могут быть пропуски в данных, особенно на вопросы, касающиеся доходов. Подобные пропуски могут возникнуть по вине исследователя (неграмотно составленная анкета, обилие сенситивных вопросов, не достаточно подготовленные интервьюеры). Но даже если со стороны исследователя ошибок в организации и проведении полевого этапа допущено не было, то недостающая информация (не ответы на отдельные вопросы) возникнуть по вине респондента из-за:

1. социально – психологических характеристик;
2. некомпетентности в теме исследования;
3. неконтактности;
4. не желания отвечать.

Чаще всего пропуски возникают при ответе на сенситивные вопросы: о доходе, некоторых предпочтениях и отдельных поведенческих аспектах, которые, по мнению респондента, могут отличаться от социально одобряемых. Поэтому респондент предпочтет скорее не ответить на вопрос, чем высказать «неправильное» мнение или выбрать «неправильный» вариант ответа.

Наличие пропусков в данных, так же как и анализ только полных наблюдений (после исключения наблюдений с пропусками), может привести к получению смещенных результатов, и как следствие к искажению выводов, которые могут быть сделаны по результатам исследования и принятию неверных стратегических решений.

Данную проблему исследователи решают по – разному. Некоторые просто исключают из рассмотрения наблюдения с пропущенными данными.

Другие исследователи подходят к решению проблемы пропущенных данных более рационально. Они стремятся на этапе первичной обработки данных заполнить пропуски в уже имеющихся данных, для того чтобы восстановить исходную зависимость.

Мы, придерживаемся второй точки зрения, а исследование, описываемое в данной статье, было посвящено проблеме неполноты данных, и прежде всего, способам восстановления пропущенных значений.

Перед нами стояла **основная цель** – систематизировать основные подходы и методы заполнения пропусков в данных и продемонстрировать на практике использование 2-х наиболее универсальных методов импутирования (заполнения пропусков в данных).

1.СУЩЕСТВУЮЩИЕ МЕТОДЫ ЗАПОЛНЕНИЯ ПРОПУСКОВ В ДАННЫХ

Существует множество способов заполнения пропусков (ремонта выборки) уже после этапа сбора данных: заполнение средним значением, пропорциональное размещение наблюдений с пропущенными данными по уже имеющимся градациям шкалы, расчет возможного значения при помощи регрессионной модели и так далее. [Давыдов, Крыштановский, 2006, стр. 231-240]. Заполнение пропусков позволяет не только получить дополнительную информацию (предсказанные значения), но и сохранить уже имеющуюся, часто очень важную и полученную ценой значительных усилий информацию, за счет сохранения наблюдений изначально содержавших пропуски.

Помимо очевидных достоинств, импутирование, как способ решения проблемы недостающей информации имеет несколько недостатков, которые нельзя не учитывать:

1. Использование для предсказания пропусков имеющихся полных данных искажает структуру результирующих данных (после импутирования), которая смещается в сторону структуры только полных наблюдений;
2. Искусственная подстановка пропусков вносит в массив определенную долю искусственных данных, которые в свою очередь приводят к смещению значимости получаемых на их основе результатов. [Rubin, 1987, p.64-66].

Возможно, что, модели, построенные по импутированным данным, будут менее точными по сравнению с идеальной моделью, построенной только на полных наблюдениях. Потери в их точности будут зависеть от качества предсказания отсутствующих значений. Но, зато, потеряв в точности, можно выиграть в репрезентативности результатов.

При выборе конкретного метода восстановления пропущенных значений следует учитывать, что, так как алгоритмы заполнения пропусков не универсальны, возможности применения того или иного способа заполнения пропусков зависят от метода анализа данных, который планируется использовать в дальнейшем.

Наиболее популярные из существующих методов импутирования, в наиболее полной и подробной классификации Р. Литла, отражены на следующей схеме (Рис.1).

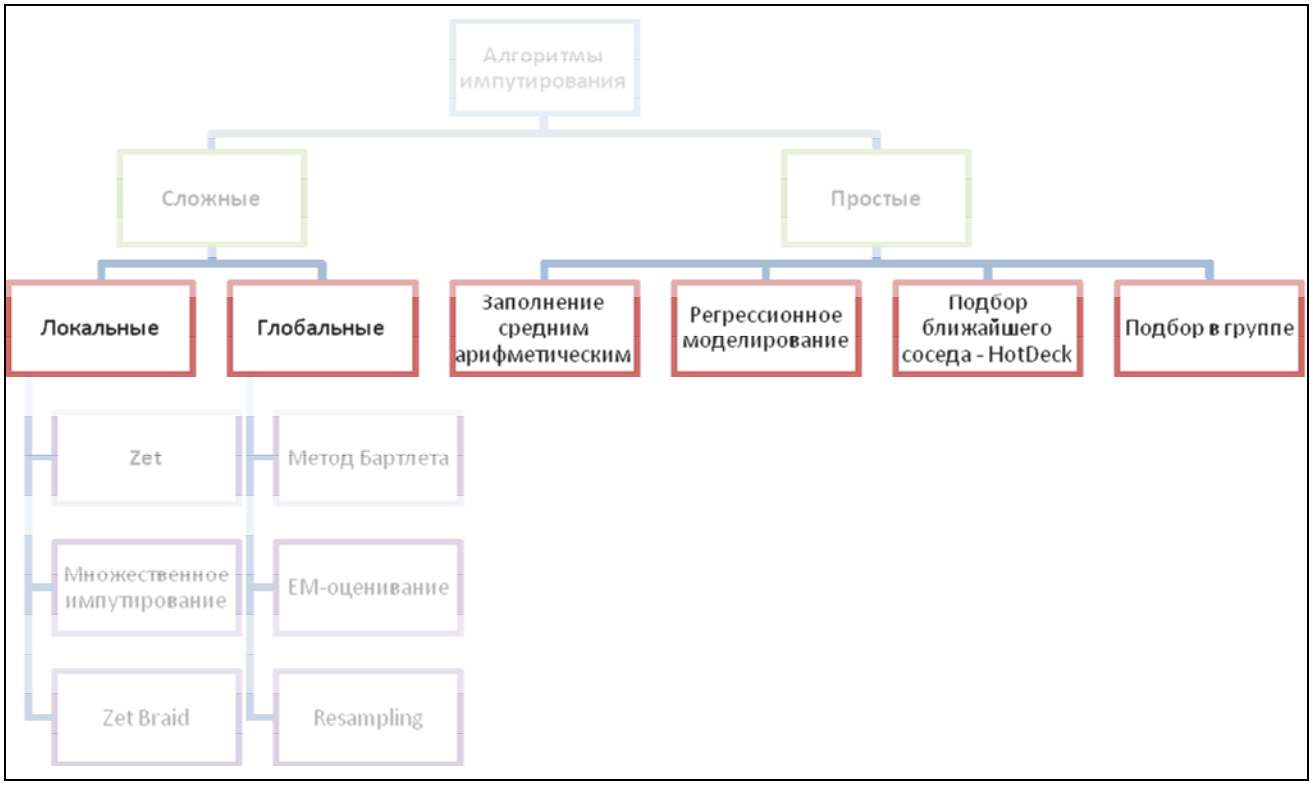

Рис.1. Классификация алгоритмов импутирования. [Злоба Е., Яцкив И., 2004, стр. 52 – 54]

Охарактеризуем каждую группу методов:

Простые алгоритмы – неитеративные алгоритмы, основанные на простых арифметических операциях, расстояниях между объектами, регрессионном моделировании. К ним относится заполнение пропусков средним арифметических, регрессионное моделирование пропусков, метод HotDeck и подбор в группе.

Сложные алгоритмы – итеративные алгоритмы, предполагающие оптимизацию некоторого функционала, отражающего точность расчета подставляемых на место пропуска значений. Их можно разделить на глобальные и локальные.

Глобальные алгоритмы - в оценивании (предсказании) каждого пропущенного значения участвуют все объекты рассматриваемой совокупности: метод Бартлета, EM - оценивание и Resampling.

Локальные алгоритмы – в оценивании (предсказании) каждого пропущенного значения участвуют полные наблюдения, находящиеся в некоторой окрестности предсказываемого объекта**.** К данной группе относятся алгоритмы Zet и Zet Braid

Кратко рассмотрим отдельные методы, импутирования, входящие в состав перечисленных групп.

Заполнение средним и подбор внутри групп.

Метод заполнения средним значением, предполагает, что все пропущенные значения заменяются средним значением данного признака, рассчитанным по имеющимся данным.

Подбор внутри групп предполагает, что вся совокупность объектов разбивается на группы по определенному признаку, внутри каждой группы для заполнения пропусков используются только присутствующее в ней значения. [Злоба Е., Яцкив И., 2004, стр. 52 – 54].

МЕТОД HOT DECK.

Hot Deck используется в одномоментных исследованиях, и представляет собой подстановку вместо пропуска значения по данной переменной у наиболее близкого объекта с полной информацией. Причем подпор может осуществляться как из всей совокупности полных наблюдений, так из ее некоторой подгруппы – кластера, к которому принадлежит целевой объект.

Для заполнения пропуска по данной характеристике у целевого объекта, используется значение данной характеристики у объекта, ближайшего к целевому (расстояние до которого от целевого объекта меньше, меньше чем до всех остальных объектов).

Тип функции расстояния для определения наблюдения, ближайшего к целевому (с пропуском), выбирается исходя из типа используемых данных, представлений исследователя о характере связи между переменными и задач каждого конкретного исследования. [Kalton,1983, p. 11-16].

МЕТОД БАРТЛЕТА

Метод Бартлета состоит из двух этапов: подстановке вместо пропусков начальных значений на первом этапе и проведении на втором этапе ковариационного анализа целевой переменной и дихотомического индикатора полноты наблюдения по целевой переменной. Индикатор полноты наблюдения всегда равен 0, за исключением одного единственного случая, *i–e* значение является целевой переменной пропущено, только в этом случае он принимает значение 1. [Злоба Е., Яцкив И., 2004, стр. 55-56].

АЛГОРИТМ ZET

Суть алгоритма Zet заключается в подборе для каждого пропуска импутируемого значения не из всей совокупности полных наблюдений, а из некоторой ее части, называемой компонентной матрицей. Она состоит из компонентных строк и столбцов. Компонентность некоторой строки или объекта представляет собой величину обратнопропорциональную декартовому расстоянию до целевой строки (неполного наблюдения с пропуском) в пространстве, оси которого заданы переменными – рассматриваемыми характеристиками объектов.

По данным компонентной матрицы затем строится функциональная зависимость прогнозируемого значения от соответствующего значения в компетентной матрице, на основе которой, затем, прогнозируется значение пропуска. [Снитюк В.Е., 2008, стр. 5-6].

ZETBRAID.

Основное отличие и одновременно достоинство алгоритма **ZetBraid** (плетение) от алгоритма Zet заключается в том, что в нем заложен аппарат для объективного определения размерности компонентной матрицы.

В процессе работы алгоритма происходит последовательный поочередный отбор компетентных строк и компетентных столбцов. При каждом новом отборе строки или столбца формируется новая компетентная матрица. По заданному критерию определяется ее эффективность при прогнозировании пропусков. [Алгоритм ZetBraid, 2008.].

RESAMPLING

В итеративном алгоритме Resampling строки, содержащие пропущенные данные заменяют случайно подобранными строками из матрицы полных наблюдений. Затем строится регрессионное уравнение для предсказания отсутствующего значения.

Процедура построения регрессионного моделирования повторяется несколько раз. После определенного количества повторений значения полученных регрессионных коэффициентов усредняют и получают окончательное решение, дающее максимальную точность прогноза пропущенного значения. [Злоба Е., Яцкив И., 2004, стр. 56].

МНОЖЕСТВЕННОЕ ИМПУТИРОВАНИЕ.

Метод множественного импутирования был разработан Дональдом Рубиным в 1970 –х годах 20 века. С точки зрения Рубина приписывание каждому пропуску нескольких потенциальных значений призвано отразить степень неопределенности, с которой осуществляется импутирование. Сейчас этот метод является наиболее перспективным и реализован в специализированном программном обеспечении, к сожалению, в большей части коммерческом. [Horton, Lipsitz. 2001, p.246-248].

Техника множественного импутирования предусматривает подстановку сразу нескольких значений на место каждого из пропусков. Существенный разброс этих значений свидетельствует о неопределенности модели и не позволяет сделать однозначный вывод об их типе и причине возникновения.

Данные с каждым набором заполненных пропусков сохраняются в отдельный массив, каждый из которых затем анализируется как состоящий только из полных наблюдений. [Rubin, 1996, p. 473-489, Lipsitz S. R, Lue Ping Zhao, Molenberghs G. A.1998. p.138-140, Schafer 1999, Shulte Nordholt 1998].

Следующие 2 метода заполнения пропусков ЕМ – оценивание и регрессионное моделирование пропусков были использованы на практике и поэтому в статье они описаны более подробно. Для практического использования эти методы выбраны в силу своей универсальности, которая выражается в том, что их можно использовать для восстановления значений переменных, измеренных не только интервально, но и на более низком уровне.

ЕМ – ОЦЕНИВАНИЕ

Метод максимизации ожиданий (ЕМ –expectation maximization) , в некоторых источниках так же называемый ЕМ – оцениванием, позволяет не только восстанавливать пропущенные значения с использованием двухэтапного итеративного алгоритма, но и оценивать средние значения, ковариационные и корреляционные матрицы для количественных переменных.

ЕМ – алгоритм, в самом общем смысле представляет собой итерационную процедуру, предназначенную для решения задач оптимизации некоторого функционала, через аналитический поиск экстремума целевой функции.

Этот алгоритм реализуется в 2 этапа. Первые буквы названий, которых образуют общую аббревиатуру алгоритма:

Этап Е

На **первом этапе Е (expectation)** по совокупности имеющихся абсолютно полных или частично (по целевой переменной) полных наблюдений рассчитываются условные ожидаемые значения целевой переменной для каждого неполного наблюдения. Затем после

получения массива полных наблюдений, оцениваются основные статистические параметры: меры средней тенденции и разброса, показатели взаимной корреляции и ковариации переменных.

В случае работы с неполными даными на Е – этапе определяется функция условного математического ожидания логарифма полной функции правдоподобия при известном значении целевой перменной X $Q(\Theta;\Theta^{(m)})$:

$$Q(\Theta;\Theta^{(m)}) = E_{\Theta(m)}[\log f_{\Theta}(X,Y)\big|X]. \quad (1)$$

Когда имеют дело с полным наблюдением, у которого характеристика X принимает значение x, выражение (1) для вычисления значений функции $Q(\Theta;\Theta^{(m)})$ принимает вид: $Q(\Theta;\Theta^{(m)}) = \int_{R^m} [\log f_{\Theta}(x,y)] f_{\Theta^{(m)}}(y|x)\mu_Y(dy)$. После определения этой вида этой функции начинается вторйо этап работы алгоритма – М этап.

Этап М

На **втором этапе М (maximization)**, задача алгоритма максимизировать степень взаимного соответствия ожидаемых и реально подставляемых данных, а также соответствия структуры импутированных данных структуре данных полных наблюдений.

В классическом варианте алгоритма, формально задачу по максимизации ожидания можно выразить следующим образом: $\Theta^{(m+1)} = \arg\max_{\Theta} Q(\Theta;\Theta^{(m)})$. Здесь Θ обозначает рассчитанное ожидаемое условное значение, отсутствующей характеристики для некоторого наблюдения. [Королев В.Ю., 2007, стр. 12-13].

РЕГРЕССИОННОЕ МОДЕЛИРОВАНИЕ ПРОПУСКОВ

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

1. На первом этапе по совокупности полных наблюдений отстраивается регрессионная модель, и оцениваются коэффициенты в уравнении, где в качестве зависимой переменной выступает целевая переменная – пропущенные значения по которой необходимо восстановить;
2. Затем по полученному на предыдущем этапе уравнению, в которое подставляются известные значения независимых переменных предикторов, для каждого целевого объекта 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.

Выбор регрессионной модели для расчета пропущенных значений переменной, определяется уровнем измерения целевой зависимой переменной (значения которой необходимо восстановить) и независимых переменных, по которым будут предсказываться отсутствующие значения. [Крыштановский А.О., 2006, стр. 182-191].

Помимо теоретического описания различных методов импутирования важно понять, как они работают на практике. Кроме того, всем кто столкнулся или просто интересуется проблемой неполноты данных, будет полезно увидеть численные оценки потерь информации из-за неполноты наблюдений и качества импутирования в зависимости типа целевой переменной и исходной доли пропусков. Для того чтобы получить эти оценки в эмпирической части исследования был реализован методический эксперимент.

Основной задачей эксперимента, помимо оценки потерь информации, вследствие неполноты данных, является сравнение качества импутирования двумя методами: EM - алгоритмом и регрессионным моделированием в зависимости от типа импутируемых переменных и количества пропусков в данных.

2.МЕТОДИЧЕСКИЙ ЭКСПЕРИМЕНТ. СРАВНЕНИЕ ЭФФЕКТИВНОСТИ EM - ИМПУТИРОВАНИЯ И РЕГРЕССИОННОГО МОДЕЛИРОВАНИЯ ПРОПУСКОВ (МЕТОДИЧЕСКИЙ ЭКСПЕРИМЕНТ).

Методический эксперимент был реализован в рамках решения содержательной задачи исследования – анализа российского сектора неформальной занятости. В рамках анализа неформальной занятости строилась логистическая регрессионная модель, оценивающая влияние различных социально – демографических характеристик респондента на тип его занятости: формальный или неформальный.

Регрессионная модель выглядит следующим образом. Зависимая переменная – «Тип занятости»: формальная или неформальная. Переменная дихотомическая.

Независимые переменные:

1. Возраст – интервальная переменная;
2. Пол – номинальная переменная;
3. Уровень образования – порядковая переменная;
4. Совмещение индивидом работы с учебой – номинальная, дихотомическая переменная;
5. Характер собственности фирмы (организации), в которой работает индивид – номинальная переменная;
6. Категория должности индивида – порядковая переменная;
7. Наличие у индивида дополнительной работы – номинальная дихотомическая;
8. Заработная плата на основной работе – интервальная переменная;
9. Получение любых социальных трансфертов – номинальная дихотомическая переменная.

В качестве источника данных было использована часть, посвященная только занятому населению, массива данных Национального обследования благосостояния домохозяйств и участия в социальных программах (НОБУС), соответствующие опросу отдельных респондентов (не домохозяйств) осенью 2003 года, представляющие собой репрезентативную всероссийскую выборку.

Следует, однако, заметить, что изначально в анализ были включены только полные (по переменным, включенным в регрессионную модель переменным) наблюдения. Это было сделано исходя из методических соображений, для получения идеального массива данных, не содержащего пропусков. Вынуждено, по причине неполноты из массива НОБУСа было удалено около 10% наблюдений.

Регрессионная модель, построенная на этом массиве, будет эталоном для сравнения с ней моделей, полученных после заполнения пропусков рассматриваемыми методами. Более эффективным будет тот метод импутирования, который обеспечит большее приближение данных после импутирования и построенной на них регрессионной модели к идеальным показателям.

Для того чтобы оценить потери информации и качество импутирования, в зависимости от количества исходных пропусков, из идеального массива было создано 4

отдельных массива, с разным количеством искусственно внесенных (путем сознательного удаления у некоторых наблюдений известных значений интересующих нас переменных) случайных пропусков.

1. Репрезентативные пропуски – структура (пропорции) искусственных пропусков совпадают с пропорциями пропусков по каждой из целевых переменных в исходном массиве НОБУСа;
2. 20% пропусков по каждой целевой переменной;
3. 40% пропусков по каждой целевой переменной;
4. 60% пропусков по каждой целевой переменной.

Пропуски были внесены только в те независимые переменные модели, по которым в исходном массиве НОБУСа были неответы.

На подготовительном этапе методического эксперимента, для того, чтобы определить, сколько пропусков и в какие переменные нужно внести, чтобы соблюсти их исходную структуру, с помощью процедуры Missing Value Analysis, реализованной в во всех версиях статистического пакета SPSS, начиная с 11-ой, все переменные были проанализированы на предмет масштабов, структуры пропусков и возможных схем сопряженности (совместных пропусков) между пропусками в отдельных переменных.

По результатам MVA не было выявлено зависимости между наличием пропусков в данных и отдельными характеристиками респондентов, так же как и не было найдено схем совместных не ответов на вопросы. Поэтому все пропуски были внесены в массив полностью случайно с помощью генератора случайных чисел. Это сняло необходимость проводить диагностику характера пропусков и позволяет использовать для импутирования EM - алгоритм и регрессионное моделирования, применимые для заполнения только полностью случайных пропусков.

После случайного удаления припусков по каждой из этих переменных в каждом из 4 массивов пропуски были заполнены при помощи сравниваемых методов импутирования: EM - алгоритма и регрессионного моделирования. В итоге было получено 8 массивов полных наблюдений, с разной долей искусственно подставленных значений.

Затем, на каждом из 8 массивов была отстроена логистическая регрессионная модель влияния социально – демографических характеристик на характер занятости респондента. Полученные модели сравнивались с эталонной моделью, чтобы понять какой из двух методов обеспечивает максимальную точность оценок, при заданной доле пропусков в данных.

В первую очередь были оценены потери информации, возникшие в результате исключения из регрессионного анализа неполных наблюдений.

Таблица 1

Потери информации при построении логистической регрессионной модели в зависимости от количества пропусков в данных

	% правильных предсказаний в модели				
Переменная	**Данные без пропусков (идеальный массив)**	**Массив с репрезентативными пропусками**	**20%**	**40%**	**60%**
Формальная занятость	**97%**	97%	96%	96%	96%

Неформальная занятость	**48%**	47%	56%	60%	72%
Общее качество модели	**92%**	91%	91%	92%	93%
Из анализа исключено по причине наличия пропусков (% объектов)	**0%**	23%	68%	92%	99%

Пороговое значение вероятности для логистической модели = 0,5.
Всего объектов в массиве полных наблюдений 41447

На первый взгляд кажется, что с ростом количества пропусков в данных ситуация только улучшается, качество модели повышается, так как при наличии 60% пропусков по каждой переменной процент верно предсказанных случаев неформальной занятости (72%) вырос на 24%, по сравнению с данными без пропусков (48%). А процент правильного предсказания формальной занятости с ростом количества пропусков практически не меняется и остается на очень высоком уровне 96%-97%.

Но, с другой стороны, если обратиться к последней строке таблице, в которой приведена доля исключенных в каждом случае из анализа объектов, видно, что при 60% в каждой переменной пропусков, когда качество предсказания максимально, репрезентативность модели практически нулевая, так как из анализа исключено 99% объектов и модель было построена на 444 наблюдениях. Конечно, если бы 444 человека составляли всю исследуемую совокупность, то качество модели можно было бы считать очень хорошим и использовать ее для анализа взаимосвязи типа занятости и социально демографических характеристик. Но, эти выводы будут справедливы только для 1% даже не обследуемой, а только выборочной совокупности (для 444 из 41447 человек), и их ценность для социолога как исследователя проблемы неформальной занятости стремится к 0, потому что они являются нерепрезентативными по отношению к обследуемой совокупности.

Теперь очевидно, что необходимо, хотя бы с некоторым приближением, восстановить недостающую информацию, чтобы уменьшить масштабы исключения объектов из анализа.

Возможно, что модели, построенные по импутированным данным, будут менее точными по сравнению с идеальной моделью. Потери в их точности будут зависеть от качества предсказания отсутствующих значений, которое можно будет оценить, так как истинные значения по всем характеристикам для каждого объекта известны.

Но, зато, потеряв в точности, можно выиграть в репрезентативности результатов. Предсказательное качество модели и ее репрезентативность, как видно из таблицы 1 находятся в обратной зависимости. Однако каждая из этих характеристик одинакова, важна для интерпретации исследовательских результатов и исследователю необходимо определить для себя их оптимальное сочетание, чтобы оценить необходимость импутирования данных.

Теперь перейдем к результатам эксперимента, направленного на сравнение эффективности восстановления пропущенных значений методами EM - оценивания и регрессионного моделирования в зависимости от доли восстанавливаемых значений. Сравнению эффективности этих методов для импутирования количественных, порядковых и номинальных переменных посвящены следующие 2 таблицы.

Таблица 2

Качество импутирования значений интервальных переменных в зависимости от количества импутируемых значений

		Специфический стаж		Заработная плата	
		Среднее	**Дисперсия**	**Среднее**	**Дисперсия**
Данные без пропусков		**5,96**	**14,42**	**3821,24**	**8369635**
ЕМ-алгоритм	5%	5,95	13,48	3812,7	7628057
	20%	5,99	12,43	3710,3	6724276
	40%	6,13	9,3	3682	5212286
	60%	6,18	7,22	3754,7	3828585
Импутирование при помощи регрессионных уравнений	5%	6,29	14,1	3812,9	7726731
	20%	5,26	13,4	3612,2	6757558
	40%	4,76	10,94	3532,4	5287328
	60%	4,09	8,34	3457,1	3898290

Таблица 3

Качество импутирования значений порядковых и номинальных в зависимости от количества импутируемых значений

	Доля верных предсказаний							
	ЕМ-алгоритм				**Регрессионное моделирование**			
Переменная	**5%**	**20%**	**40%**	**60%**	**5%**	**20%**	**40%**	**60%**
Форма собственности фирмы (номинальная)	35%	25%	14%	5%	55%	46%	35%	35%
Категория должности (порядковая)	20%	14%	11%	13%	48%	33%	25%	19%
Наличие дополнительной работы (номинальная)	97%	97%	97%	95%	98%	97%	95%	95%
Изменеие дисперсии значений (по	-3%	-16%	-35%	-52%	+1%	-5%	-26%	-40%

Переменная	Доля верных предсказаний							
	EM-алгоритм				Регрессионное моделирование			
	5%	20%	40%	60%	5%	20%	40%	60%
сравнению с идеальным случаем)								

При минимальной доле искусственно рассчитанных данных (5%) значения среднего специфического стажа и средней заработной платы отклоняются от истинного значения крайне незначительно. И этими отклонениями можно пренебречь.

С ростом доли искусственных значений точность прогнозирования при расчете среднего для регрессионного моделирования в случае переменных с большой дисперсией (например, доход) ниже, чем у EM – алгоритма. А для переменных с незначительной дисперсией (в нашем случае это специфический трудовой стаж) однозначного вывода в пользу одного из методов сделать нельзя, так как изменения среднего значения для двух методов разнонаправлены: EM - алгоритм приводит к завышению среднего значения стажа, а регрессионное моделирование к его занижению.

Прежде всего, необходимо отметить, что и EM - алгоритм и регрессионное моделирование даже при минимальном количестве заполняемых пропусков уменьшают дисперсию значений переменных. И если при 5% пропусков дисперсии занижается не значительно (EM - алгоритм занижает дисперсию в среднем на 8%, а регрессионное моделирование на 5%), при EM - импутирования 20% пропусков теряется 17% дисперсии, 40% пропусков – 37% дисперсии, 60% пропусков – 52% дисперсии, а при регрессионном моделировании теряется 13%, 30% и 48% дисперсии значений интервальных переменных, соответственно.

Поэтому, если перед исследователем стоит задача оценить степень неоднородности обследуемой совокупности по некоторой интервальной переменной, например по доходу, заполнение пропусков может привести к существенному занижению оценок, по сравнению с ситуацией включения в анализ только полных наблюдений.

Что касается изменений, которые происходят со значениями регрессионных коэффициентов и их значимостью после заполнения пропусков, в зависимости от доли импутируемых значений и метода заполнения пропусков то, по результатам эксперимента получается, что наличие пропусков в данных, даже заполненных, прежде всего, значительно понижает уровень значимости регрессионных коэффициентов.

В идеальной (эталонной) логистической модели все коэффициенты, кроме коэффициента при фиктивной переменной « отсутствие начального образования» значимы на уровне 100%.

Если построить эту же модель на массиве данных с репрезентативными долями пропусков по каждой переменной (совокупная доля пропусков 5%), из анализа будут исключено 23% объектов, но зато все коэффициенты, полученные на оставшихся 77% наблюдений, сохранят свою значимости на уровне 100%.

При импутированных ЕМ - алгоритмом 20% пропусков в данных, теряют значимость 60% коэффициентов (19 из 32) , 40% пропусков - 53%(17 из 32), 60% пропусков – 56% (18 из 32).

При регрессионном моделировании 20% данных в массиве теряют значимость 56% коэффициентов (18 из 32), 40% данных – 60% (19 из 32), 60% данных – 53% (17 из 32).

Если в массиве данных отсутствует пятая часть данных, и исследователя не устраивает то, что модель репрезентирует оставшиеся 80% объектов (20% объектов исключаются из анализа при совокупной доле пропусков 5%), и он принимает решение об импутировании пропусков, то используя оба метода импутирования, он рискует получить в результате модель, в которой половина всех коэффициентов не значима. А так, как нам точно известно, что на самом деле данные коэффициенты значимы, исследователь вынужден не рассматривать половину, якобы не значимых, взаимосвязей.

А те коэффициенты, которые значимы, не совпадают с истинными значениями. Значения коэффициентов, рассчитанных на данных с заполненными обоими методами пропусками, в 1,5 раза меньше истинных.

Получается, что если в данных больше 5% искусственных значений, существенно уменьшается значимость регрессионных коэффициентов. В этом случае, при незначительном выигрыше в репрезентативности результатов (в нашем случае 23%), исследователь может значительно больше проиграть в статистической значимости результатов. Поэтому, если пропусков относительно не много, лучше их не заполнять и строить логистические модели на полных наблюдениях, чем получать совершенно не значимые результаты на восстановленных данных.

С точки зрения изменения средних значений, количества точных подстановок, потерь дисперсии, для логистических моделей между долей исходных пропусков и качеством импутирования наблюдается обратная взаимосвязь: с ростом количества пропусков точность импутирования снижается. Этот вывод вполне логичен, если принять во внимание то, что при увеличении пропусков в данных уменьшается совокупность полных наблюдений, которая используется качестве данных для прогнозирования отсутствующих значений. Получается, что необходимо предсказывать больше пропусков, используя для этого меньше данных. Сохранить при этом прежний уровень точности не возможно.

ЗАКЛЮЧЕНИЕ.

Конечно, мы не претендуем на законченность и полный охват, проведенного анализа теоретических и методических аспектов заполнения пропусков в данных, в частности, проблемы неполноты информации в целом.

Можно надеяться, что данная работа, может быть полезна как практическое руководство по использованию некоторых методов импутирования и как краткий обзор литературы, посвященной данной проблеме.

Нам так же кажется, что в свете названных достоинств, и возможностей, которые дают социологу методы заполнения пропусков, они требуют дальнейшего исследования, и должны занять достойную нишу в исследовательской практике современного социолога.

К достоинствам описанного исследования можно отнести разработку следующих практических рекомендаций по использованию ЕМ - импутирования и регрессионного моделирования пропусков:

1. В случае переменных, характеризующихся большой дисперсией значений, и при увеличении доли пропусков в данных ЕМ - алгоритм является более эффективным методом импутирования, чем регрессионное моделирование;
2. Для случая данных, однородных по некоторой количественной характеристике, складывается неопределенная ситуация. Регрессионное заполнение пропусков приводит к недооцениванию среднего, а ЕМ – импутирование – к его переоценке;
3. При регрессионном моделировании теряется, в среднем, на 10% меньше дисперсии, чем при ЕМ - импутировании. Поэтому с точки зрения оценки степени однородности совокупности использование регрессионного моделирования для заполнения пропусков более предпочтительно.
4. Если для достижения исследовательских задач необходимо оценить степень однородности исследуемой совокупности и пропусков в данных немного, следует ограничиться анализом полных наблюдений, так как при заполнении пропусков существует риск существенного занижения оценок.

Конечно, в исследовании были получены отчасти негативные практические результаты (не значимость части регрессионных коэффициентов, полученных на импутированных данных, занижение истинной дисперсии в результате импутирования и т.д.), но они справедливы только для использованного логистического регрессионного анализа. В случае других методов результат может быть иным. Анализу применимости алгоритмов импутирования для каждого аналитического метода можно посвятить отдельную работу, что возможно, будет сделано автором в дальнейшем.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

1. *Алгоритм ZetBraid* //Информационные интеллектуальные системы. Вып.40, 2008//http://iissvit.narod.ru/rass/vip40.htm
2. *Злоба Е., Яцкив И.* Статистические методы восстановления пропущенных данных // Computer Modeling & New Technologies.; Vol. 6.2004.; Стр.55 – 56.
3. *Королев В.Ю. ЕМ* – алгоритм, его модификации и их применение к задаче разделелния смесей вероятностных распределений. Теоретический обзор. М.:2007. 102 стр.
4. *Крыштановский А.О.* Анализ социологических данных с помощью пакета SPSS.:М. ГУ-ВШЭ. 2006. 263 стр.
5. *Литтл Р.Дж.А., Рубин Д.Б.* Статистический анализ данных с пропусками. Финансы и статистика.: Москва, 1991; 430 стр.
6. *Снитюк В.Е.,* Эволюционный метод восстановления пропусков в данных. 2008 // http://iissvit.narod.ru/index_a.htm;
7. *Horton N. J; Lipsitz S.R.* Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. // The American Statistician, Vol. 55, No. 3. (Aug., 2001), P. 244-254 // http://links.jstor.org/sici?sici=0003-1305%28200108%2955%3A3%3C244%3AMIIPCO%3E2.0.CO%3B2-J
8. *Kalton, G. , Kasprzyk, D.* The treatment of missing survey data. // Survey Methodology, № 12, 1986. P. 1-16.
9. *Lipsitz S. R; Lue Ping Zhao; t Molenberghs G. A.,* Semiparametric Method of Multiple Imputation // Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol.

60, No. 1.1998, P. 127-144. // http://links.jstor.org/sici?sici=1369-7412%281998%2960%3A1%3C127%3AASMOMI%3E2.0.CO%3B2-5
10. *Rubin, D.B.* Multiple Imputation for Nonresponse in Surveys. Ney York: Willey, 1987. P. 64-66;
11. *Rubin, D.B.* Multiple imputation after 18+ years. Journal of the American Statistical Association, № 91, 1996. P. 473-489.
12. *Schafer J.L.; Schenker N.* Inference with Imputed Conditional Means // Journal of the American Statistical Association, Vol. 95, No. 449. 2000 P.. 144-154.// http://links.jstor.org/sici?sici=0162-1459%28200003%2995%3A449%3C144%3AIWICM%3E2.0.CO%3B2-G
13. *Schafer, J. L.* Multiple Imputation: A Primer," Statistical Methods in Medical Research, Vol. 8, 1999. P. 3-15.
14. *Schulte Nordholt E.* Imputation: Methods, Simulation Experiments and Practical Examples // International Statistical Review / Revue Internationale de Statistique, Vol. 66, No. 2. 1998, P. 157-180. // http://links.jstor.org/sici?sici=0306-7734%28199808%2966%3A2%3C157%3AIMSEAP%3E2.0.CO%3B2-W
15. *SPSS Missing Value Analysis 12.0* // Заполнение пропущенных значений для повышения информативности данных и построения адекватных моделей // 2008. http://www.spss.ru/products/missing_value/mva12.pdf;