Вестник Евразийской науки / The Eurasian Scientific Journal https://esj.today

2019, №3, Том 11 / 2019, No 3, Vol 11 https://esj.today/issue-3-2019.html

URL статьи: https://esj.today/PDF/30ITVN319.pdf

Ссылка для цитирования этой статьи:
Афанасьев В.В., Благий В.А., Воробьев А.А. Алгоритм перераспределения неопределившихся респондентов на основе мультиномиальной логистической регрессии // Вестник Евразийской науки, 2019 №3, https://esj.today/PDF/30ITVN319.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

For citation:
Afanasev V.V., Blagij V.A., Vorobjev A.A. (2019). Algorithm of redistribution of undecided respondents on the basis of multinomial logistic regression. *The Eurasian Scientific Journal*, [online] 3(11). Available at: https://esj.today/PDF/30ITVN319.pdf (in Russian)

УДК 004.02

ГРНТИ 04.15.41

Афанасьев Вадим Владимирович
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Орёл, Россия
Сотрудник
Кандидат технических наук
E-mail: affa@mail.ru

Благий Владимир Александрович
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Орёл, Россия
Сотрудник
E-mail: blgij1@yandex.ru

Воробьев Андрей Анатольевич
ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Орёл, Россия
Сотрудник
Кандидат технических наук, доцент
E-mail: awa@mail.ru

Алгоритм перераспределения неопределившихся респондентов на основе мультиномиальной логистической регрессии

Аннотация. В статье проведен анализ существующих подходов, методов и средств обработки социологической информации, на основании которого сделан вывод о необходимости повышения точности показателей оценки исследуемой социологической ситуации, влиянии на результат социологического исследования неопределившихся респондентов, необходимости выбора методов обработки результатов социологических исследований с учетом шкал, применяемых для измерения ответов на вопросы в исследуемых анкетах.

Проведен анализ существующих методов обработки результатов социологических исследований, в результате которого был выбран аппарат логистической регрессии в соответствии с направлениями совершенствования подходов к оценке социально-экономической и общественно-политической ситуации в регионе.

Авторами была сформулирована гипотеза о возможности использования мультиномиальной логистической регрессии в качестве универсального (базового) метода

исследования социологических данных, которая в дальнейшем была подтверждена в ходе поставленного эксперимента по множеству статистических показателей.

Для решения задачи оценки влияния факторов и перераспределения неопределившихся респондентов в работе был предложен алгоритм, входными данными для которого является количество неопределившихся респондентов, а выходными – прогнозное значение распределения неопределившихся респондентов по категориям зависимой переменной с учётом перераспределения.

Для подтверждения возможности повышения качества аналитических выводов был проведен ряд экспериментов с использованием предложенного алгоритма. Результаты экспериментальной оценки, полученные с помощью SPSS Statistics, позволили сделать следующие выводы: существующая практика "отбрасывания" неопределившихся респондентов существенным образом влияет на качество оценок социально-экономической и общественно-политической обстановки в регионе. Предлагаемый алгоритм построения прогнозных оценок перераспределения неопределившихся респондентов позволяет нивелировать влияние этого показателя и обеспечить тем самым повышение качества прогнозов.

Ключевые слова: алгоритм; социологическое исследование; мультиномиальная логистическая регрессия; методы обработки информации; неопределившийся респондент; статистический эксперимент; прогнозная оценка

Введение

В политической и социально-экономической жизни современного общества все большую роль приобретают результаты социологических исследований. Будучи предметом широкого освещения средствами массовой информации, они становятся средством формирования симпатий или антипатий к персоне или общественной структуре, на них ссылаются специалисты-аналитики, призывающие социум поверить в обоснованность или, наоборот, необоснованность тех или иных социально значимых утверждений.

Бытующее мнение о мажоритарной оценке результатов социологического опроса, как основном способе проведения социологического измерения, как правило внушает небезосновательные сомнения в качестве такого исследования.

Социологическое исследование – "это сбор новых фактов и их интерпретация в терминах выбранной или построенной в соответствии с поставленной задачей теоретической модели с помощью методов, адекватных операциональным определениям свойств конструктов, лежащих в основании этой модели" [1].

Анализ существующих подходов, методов и средств обработки социологической информации

В настоящее время оценка социально-экономической и общественно-политической ситуации формируется в основном с использованием результатов социологических исследований (опросов). В результате выявляются и рассматриваются некоторые показатели, например, – оценка населением состояния социально значимых сфер и отраслей экономики, таких как сельское хозяйство, промышленность, образование, здравоохранение, а также обеспеченность доступным жильем и жилищно-коммунальными услугами. При этом ответы на вопросы анкеты представлены в категориальных (не метрических) шкалах двух типов: номинальной и порядковой.

Пример вопроса в номинальной шкале: **Укажите свой пол.**

Варианты ответов:

- мужской;
- женский.

Пример вопроса в порядковой шкале: **Как Вы оцениваете экономическую ситуацию в населенном пункте (муниципальном районе) Вашего проживания?**

Возможные варианты ответов:

- очень хорошее;
- хорошее;
- удовлетворительное;
- плохое;
- очень плохое;
- затрудняюсь ответить.

Кроме того, при анкетировании часто используются вопросы с множественными ответами, для анализа которых предполагается использовать метод множественной дихотомии, когда каждый ответ будет представляться отдельной переменной в дихотомической шкале.

Пример вопроса с множественными ответами в дихотомической шкале: **Что из перечисленного произошло в последние 2–3 месяца в субъекте Российской Федерации, где Вы живете, и затронуло лично Вас в наибольшей степени?**

Возможные варианты ответов (с возможностью выбора нескольких):

- рост цен на продукты питания;
- рост тарифов ЖКХ;
- рост тарифов на проезд на общественном транспорте;
- закрытие предприятий в регионе;
- рост цен на лекарства;
- сокращения персонала на предприятиях региона;
- перевод работников на неполную рабочую неделю;
- ничего из перечисленного не наблюдал(а), меня это не затронуло.

Как показывает практика, показатели оценки социально-экономической и общественно-политической ситуации в субъекте рассчитываются на основе разности положительных (например, сумма частот вариантов ответов "очень хорошее", "хорошее" и "удовлетворительное") и отрицательных (например, сумма частот вариантов ответов "плохое" и "очень плохое") оценок для примера вопроса в порядковой шкале. Заметим, что мнение неопределившихся респондентов при этом не учитывается. При этом повторные опросы, проводимые в тех же регионах и с той же тематикой, показывают, что количество неопределившихся респондентов растет и требует дополнительного учёта при анализе и прогнозировании социально-экономической и общественно-политической ситуации в регионе [2].

С другой стороны, в результате изучения материалов по тематике было выявлено следующее: при проведении анализа результатов социологического исследования обоснование выводов предполагает использование статистических данных, получаемых из различных справочников и сведений по региону, не опираясь на результаты социологического опроса. Следовательно, выводы делаются на основании экспертных оценок, а их аналитическое (математическое) подтверждение отсутствует [3].

Таким образом, обозначенная проблематика проведения социологических исследований, существующая в настоящее время, позволяет сформулировать следующие направления совершенствования подходов к оценке социально-экономической и общественно-политической ситуации в регионе:

1. для повышения точности показателей оценки социально-экономической и общественно-политической ситуации в регионе необходимо предложить способ аналитического решения проблемы учета неопределившихся респондентов в условиях роста их количества;
2. поскольку использование только экспертных методов для выявления факторов влияния на учитываемые показатели однозначно снижает доверие к выводам, формулируемым аналитиками, необходимо разработать методику, позволяющую выполнять эту работу с большей степенью достоверности;
3. выбор методов обработки результатов социологических исследований для решения рассмотренных выше двух проблем необходимо осуществлять по возможности с учетом шкал, применяемых для измерения ответов на вопросы в исследуемых анкетах.

Анализ существующих методов обработки результатов социологических исследований

В работе предлагается осуществить выбор методов обработки результатов социологических исследований с учётом имеющихся ограничений – ответы на вопросы измеряются в номинальной, дихотомической или порядковой шкалах, с учетом увеличивающегося во времени количества неопределившихся респондентов и, как следствие, повышения качества выводов по оценке показателей, например, показатели оценки социально-экономической и общественно-политической ситуации в регионе.

Известно, что анализ данных в рассмотренных выше шкалах лучше проводить при помощи таблиц сопряженности (значения одной переменной образуют строки, а значения другой – столбцы таблицы). В статистических исследованиях особенно большой интерес представляет гипотеза о независимости переменных друг от друга, для которых строится таблица сопряженности. Для принятия или отклонения гипотезы о независимости переменных применяют критерий χ^2 Пирсона, при котором проверяется, есть ли значимое различие между наблюдаемыми и ожидаемыми частотами. В зависимости от его значения, например, если значимость χ^2 Пирсона равна нулю (наблюдаемые частоты совпадают с ожидаемыми), то принимается решение о независимости признаков. Использование критерия χ^2 Пирсона позволяет определить есть ли связь между переменными или нет, однако не позволяет определить силу этой связи, вид и её направленность.

Для определения силы связи, вида и её направленности применяют корреляционный анализ. Однако данный вид анализа нельзя применять в качестве оценки зависимости между переменными, если эти переменные принадлежат к номинальной шкале и имеют больше двух категорий, потому что между их кодировками невозможно установить порядкового отношения и, следовательно, они не могут быть расположены в определенном, рационально объяснимом

порядке. Для оценки корреляций между переменными, принадлежащими к порядковой шкале, могут применяться коэффициенты Спирмена, Кендалла или Гудмена-Краскела. Тем не менее, применение различных коэффициентов корреляции не смогут решить проблему повышения точности показателей оценки социально-экономической и общественно-политической ситуации в регионе при росте количества неопределившихся [4].

Для решения проблемы учета мнения большого количества неопределившихся респондентов могут выступать подходы, основанные на различных модификациях метода множественного регрессионного анализа. Так как в качестве исследуемых переменных выступают категориальные переменные, то допустимыми методами выступают: CHAID-анализ (деревья классификации) и логистическая регрессия. Возможность работы с пропущенными данными подробно рассмотрена в [5], но, учитывая ограниченность использования метода CHAID при большой доле пропусков, то в работе предлагается подход, основанный на логистической регрессии. С её помощью можно построить уравнения регрессии и осуществлять прогнозирование (распределение) неопределившихся с учетом половозрастных и других характеристик респондентов, а также формировать оценки влияния факторов на различные показатели оценки социально-экономической и общественно-политической ситуации в регионе.

Таким образом, в результате анализа существующих методов обработки результатов социологических исследований был выбран аппарат логистической регрессии в соответствии с направлениями совершенствования подходов к оценке социально-экономической и общественно-политической ситуации в регионе, сформулированными ранее.

Анализ возможностей обработки результатов социологических исследований с помощью различных видов логистического регрессионного анализа

Логистическая регрессия предназначена для изучения причинных связей между категориальными переменными. Зависимая переменная может быть только категориальной. Независимые переменные могут быть двух видов – категориальные и количественные. Категориальные независимые переменные в уравнении логистической регрессии называют факторами, количественные – ковариатами. Для построения уравнений логистической регрессии используется метод максимального правдоподобия.

Модели логистической регрессии различаются по виду зависимой переменной: бинарная (зависимая переменная с двумя градациями), мультиноминальная (зависимая переменная номинальная с числом градаций больше двух), порядковая (зависимая переменная порядковая) [4].

В работе была выдвинута гипотеза о возможности использования одного из трех рассмотренных выше методов логистической регрессии в качестве базового (универсального). Для этого был проведен эксперимент по сравнению результатов, получаемых при использовании мультиномиальной логистической регрессии (МЛР) и порядковой регрессии (ПР).

В качестве зависимой переменной был рассмотрен вопрос: "Как Вы оцениваете свое материальное положение?", а в качестве независимых переменных были выбраны "возраст респондентов" и "образование респондентов".

Сравнение полученных результатов представлено ниже с использованием ряда статистических коэффициентов:

1. Совокупный абсолютный остаток Пирсона, показывающий отклонение прогнозных значений от фактических:

- для МЛР – 7,084;
- для ПР – 8,409.

2. Критерии подгонки (тест отношения правдоподобия) для мультиномиальной логистической регрессии и порядковой регрессии представлены в таблицах 1 и 2 соответственно:

Таблица 1

Результаты критерия подгонки для МЛР

	-2 Log-правдоподобие	χ^2	Уровень значимости
Только свободный член	280,080		
Окончательная	232,567	47,514	0,045

Таблица 2

Результаты критерия подгонки для ПР

	-2 Log-правдоподобие	χ^2	Уровень значимости
Только свободный член	280,080		
Окончательная	268,957	11,123	0,267

3. Псевдо R-квадрат, определяется как доля объясненной линейной регрессией суммы квадратов (дисперсии) в общей сумме квадратов (дисперсии) зависимой переменной. Для логистической регрессии имеет приставку "псевдо", так как рассчитать его корректное значение невозможно.

Результаты значения псевдо R-квадрат для МЛР:

- псевдо R-квадрат Кокса и Снелла – 0,62;
- псевдо R-квадрат Нэйджелкерка – 0,72;
- псевдо R-квадрат МакФаддена – 0,39.

Результаты значения псевдо R-квадрат для ПР:

- псевдо R-квадрат Кокса и Снелла – 0,20;
- псевдо R-квадрат Нэйджелкерка – 0,22;
- псевдо R-квадрат МакФаддена – 0,09.

В результате сравнения по совокупному абсолютному остатку Пирсона (чем меньше значение, тем лучше), критериям подгонки модели (-2 Log-правдоподобие (чем меньше значение, тем лучше), уровню значимости критерия χ^2 Пирсона (значения > 0,05 показывают статистическую незначимость) и псевдо R-квадрату (чем больше значение, тем лучше) можно сделать вывод, что мультиномиальная логистическая регрессия является более точным методом и, следовательно, может использоваться в качестве основного метода для дальнейших исследований.

Таким образом, в результате анализа различных видов логистической регрессии для исследуемых переменных (ответов, измеряемых в порядковой, номинальной или дихотомической шкалах) была сформулирована гипотеза о возможности использования мультиномиальной логистической регрессии в качестве универсального (базового) метода исследования социологических данных. В дальнейшем в ходе поставленного эксперимента по множеству статистических показателей данная гипотеза была подтверждена, т. е. эксперимент показал, что модель мультиномиальной логистической регрессии является более точной, чем модель порядковой регрессии. Что касается бинарной логистической регрессии, то она была

исключена из дальнейшего использования, поскольку в работе не рассматривался подход с использованием фиктивных переменных, т. е. дихотомизации категориальных переменных. Дихотомизация категориальных переменных не рассматривалась по причине снижения оперативности процесса обработки результатов социологических исследований.

Алгоритм оценки влияния факторов и перераспределения неопределившихся респондентов при оценке социально-экономической и общественно-политической обстановки в регионе

Для решения задачи оценки влияния факторов и перераспределения неопределившихся респондентов в работе предлагается следующий алгоритм (рисунок 1):

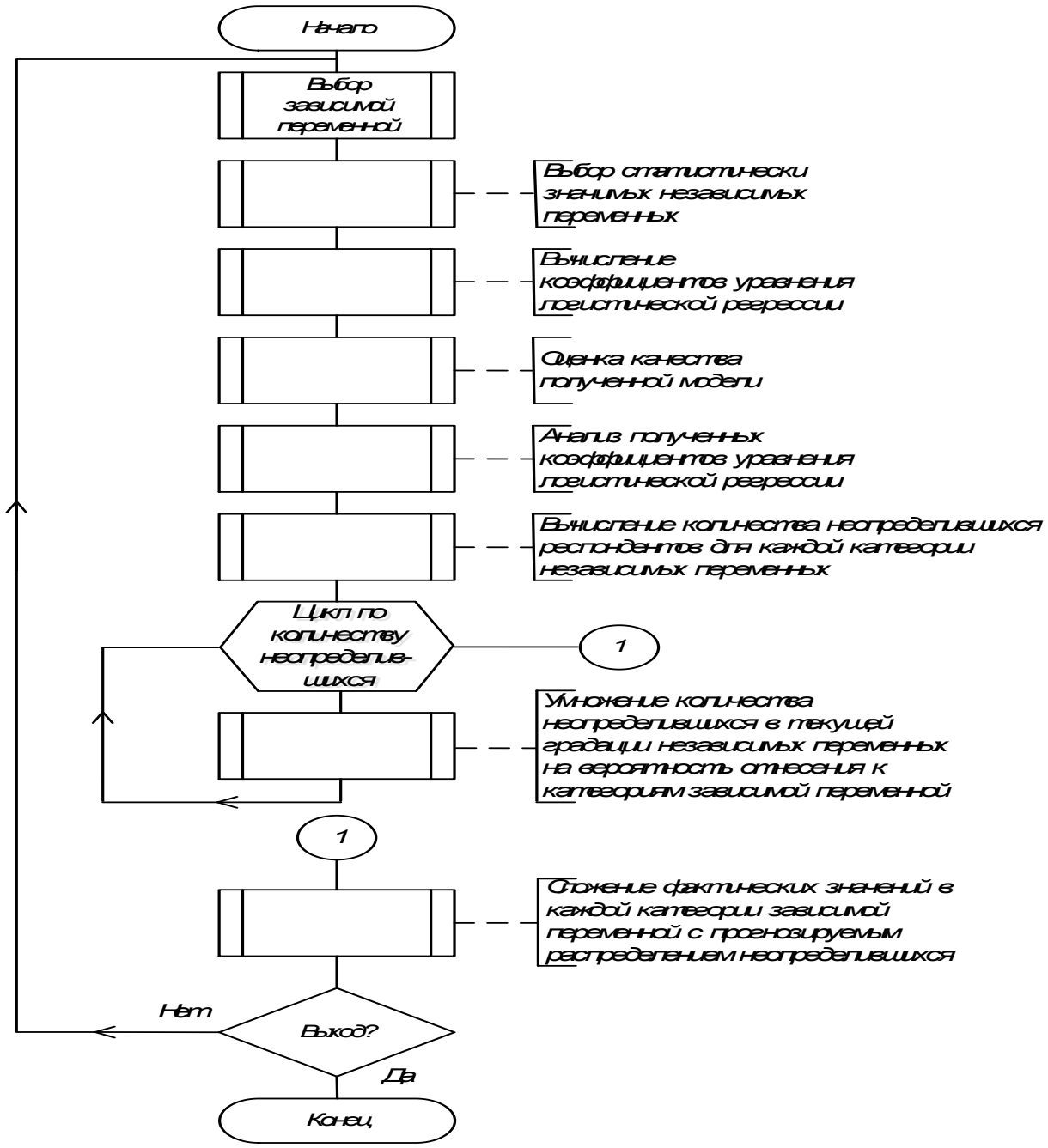

Рисунок 1. *Алгоритм оценки влияния факторов и перераспределения неопределившихся респондентов*

30ITVN319
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Пояснения к рисунку 1:

1. Выбор зависимой переменной. В качестве зависимой переменной используется объясняемая (прогнозируемая) переменная. Результатом данного этапа является выбранная зависимая переменная.

2. Выбор статистически значимых независимых переменных.

2.1. В качестве независимых переменных необходимо использовать переменные, посредством которых необходимо будет объяснять зависимую переменную.

2.2. Расчёт значения χ^2 Пирсона, определение *p*-уровня значимости для предполагаемых независимых переменных, который характеризует вероятность ошибки первого рода (в рассматриваемом контексте под ошибкой первого рода будем понимать вероятности того, что найденная зависимость свойственна только конкретной выборке, а, следовательно, не может быть применена для генеральной совокупности).

2.3. Результатом выполнения данного этапа является получение статистически значимых предикторов (категориальных независимых переменных).

3. Формирование уравнения мультиномиальной логистической регрессии на основе выбранной зависимой переменной и отобранных независимых переменных на предыдущих этапах. Результатом данного этапа являются полученные коэффициенты уравнения логистической регрессии.

4. Оценивание качества рассчитанных моделей. В качестве одного из показателей моделей будут рассматриваться псевдо *R*-квадраты для возможности сравнения моделей между собой.

5. Анализ полученных коэффициентов уравнения регрессии. Для полученных коэффициентов необходимо оценить *p*-уровень значимости каждого фактора для каждой градации зависимой переменной с помощью статистики Вальда [4]. Затем рассчитывается экспонента от этих коэффициентов. Далее рассчитывается вероятность отнесения к категориям зависимой переменной. Входными данными для данного этапа являются коэффициенты уравнения логистической регрессии. Результатом является оценка этих коэффициентов, вычисление экспоненты от этих коэффициентов, расчет вероятности отнесения к категориям зависимой переменной. Если целью исследования являлась оценка влияния факторов на зависимую переменную, то на этом этапе выполнение методики завершается.

6. Перераспределение неопределившихся респондентов.

6.1. Для всех категорий независимых переменных необходимо получить число неопределившихся респондентов. Входными данными являются категории независимых переменных. Выходными данными является количество неопределившихся респондентов по категориям.

6.2. Умножение полученного количества неопределившихся респондентов на соответствующие вероятности отнесения к категориям зависимой переменной. Входными данными для этого этапа является количество неопределившихся респондентов. Выходными данными является перераспределение неопределившихся респондентов.

6.3. Сложение предсказанных вероятностей распределения определившихся респондентов с распределением неопределившихся респондентов по категориям зависимой переменной. Входными данными является перераспределение неопределившихся респондентов. Выходными данными является прогнозное значение распределения респондентов по категориям зависимой переменной с учётом перераспределения неопределившихся.

Для подтверждения возможности повышения качества аналитических выводов был проведен ряд экспериментов с использованием предложенного алгоритма.

Результаты эксперимента по перераспределению неопределившихся респондентов

С использованием предложенного выше алгоритма, был проведен эксперимент по возможности перераспределения неопределившихся респондентов [6; 7].

В качестве зависимой переменной был рассмотрен следующий вопрос анкеты: **"Удовлетворены ли Вы своим материальным положением?"**

Ответы распределились следующим образом:

- удовлетворен: 3,7 %;
- скорее удовлетворен: 13,3 %;
- скорее не удовлетворен: 29,0 %;
- не удовлетворен: 48,5 %;
- затрудняюсь ответить: 5,5 %.

В качестве независимых переменных были выбраны следующие характеристики респондентов:

Пол

- мужской;
- женский.

Ваш возраст

- 18–29 лет;
- 30–49 лет;
- 50–59 лет;
- 60–65 лет;
- 66 лет и старше.

В результате проведенного анализа с помощью IBM SPSS Statistics [8; 9] получен *p*-уровень значимости критерия χ^2 Пирсона равный 0,043, что свидетельствует о том, что взаимосвязь статистически значима и есть смысл её рассматривать в дальнейшем исследовании.

В результате проведения мультиномиального логистического регрессионного анализа получены следующие значения коэффициентов уравнения логистической регрессии (таблица 3).

Для оценивания качества рассчитанной модели приведены результаты значения псевдо *R*-квадрата:

- псевдо *R*-квадрат Кокса и Снелла: 0,47;
- псевдо *R*-квадрат Нэйджелкерка: 0,51;
- псевдо *R*-квадрат МакФаддена: 0,19.

30ITVN319
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Таблица 3

Значения коэффициента *B* уравнения логистической регрессии

Удовлетворены ли вы своим материальным положением?		Коэффициент *B*
Скорее удовлетворен	свободный член	1,269
	мужской	-0,441
	женский	0
	18–29 лет	0,395
	30–49 лет	-0,351
	50–59 лет	19,020
	60–65 лет	1,076
	66 лет и старше	0
Скорее не удовлетворен	свободный член	2,672
	мужской	-0,076
	женский	0
	18–29 лет	-0,627
	30–49 лет	-1,158
	50–59 лет	18,510
	60–65 лет	0,160
	66 лет и старше	0
Не удовлетворен	свободный член	2,934
	мужской	-0,414
	женский	0
	18–29 лет	-0,216
	30–49 лет	-0,660
	50–59 лет	18,993
	60–65 лет	0,464
	66 лет и старше	0
Затрудняюсь ответить	свободный член	0,365
	мужской	-1,190
	женский	0
	18–29 лет	0,482
	30–49 лет	-0,106
	50–59 лет	19,264
	60–65 лет	1,557
	66 лет и старше	0

В таблице 4 приведены результаты анализа полученных коэффициентов уравнения регрессии.

Таблица 4

Значения коэффициентов уравнения логистической регрессии

Удовлетворены ли вы своим материальным положением		Коэффициент *B*	Коэффициент Вальда	Уровень значимости (*p*-уровень)	Exp (*B*)
Скорее удовлетворен	свободный член	1,269	1,170	0,279	3,557
	мужской	-0,441	0,806	0,369	0,644
	женский	0	0	0	0
	18–29 лет	0,395	0,098	0,754	1,485
	30–49 лет	-0,351	0,085	0,770	0,704
	50–59 лет	19,020	214,281	0,000	182 171 694,544
	60–65 лет	1,076	0,612	0,434	2,934
	66 лет и старше	0	0	0	0

Удовлетворены ли вы своим материальным положением		Коэффициент *B*	Коэффициент Вальда	Уровень значимости (*p*-уровень)	Exp (*B*)
Скорее не удовлетворен	свободный член	2,672	6,419	0,011	14,469
	мужской	-0,076	0,027	0,869	0,927
	женский	0	0	0	0
	18–29 лет	-0,627	0,302	0,583	0,534
	30–49 лет	-1,158	1,153	0,283	0,314
	50–59 лет	18,510	250,731	0,000	109 320 001,368
	60–65 лет	0,160	0,016	0,899	1,174
	66 лет и старше	0	0	0	0
Не удовлетворен	свободный член	2,934	7,812	0,005	18,803
	мужской	-0,414	0,843	0,358	0,661
	женский	0	0	0	0
	18–29 лет	-0,216	0,036	0,849	0,806
	30–49 лет	-0,660	0,379	0,538	0,517
	50–59 лет	18,993	269,857	0,000	177 170 044,422
	60–65 лет	0,464	0,136	0,712	1,591
	66 лет и старше	0	0	0	0
Затрудняюсь ответить	свободный член	0,365	0,065	0,799	1,441
	мужской	-1,190	3,974	0,046	0,304
	женский	0	0	0	0
	18–29 лет	0,482	0,098	0,754	1,619
	30–49 лет	-0,106	0,005	0,942	0,899
	50–59 лет	19,264	–	–	232 426 337,924
	60–65 лет	1,557	0,929	0,335	4,745
	66 лет и старше	0	0	0	0

Для перераспределения неопределившихся респондентов в работе были рассмотрены два варианта построения модели регрессии:

Вариант № 1. Перераспределение неопределившихся в соответствии с моделью, учитывающей респондентов, ответивших "затрудняюсь ответить".

В результате проведения мультиномиального логистического регрессионного анализа были получены значения вероятности отнесения к категориям зависимой переменной (рисунок 2) [10]. Затем для соответствующих половозрастных категорий количество неопределившихся респондентов умножено на соответствующие вероятности отнесения к категориям зависимой переменной (таблица 5). Далее уже имеющиеся значения для остальных категорий зависимой переменной суммируются, находятся доли категорий в процентах и рассчитывается разность положительных и отрицательных оценок (таблица 6). Расчёт необходим для формирования показателя удовлетворенности населения своим материальным положением, как разности положительных и отрицательных оценок.

Таблица 5

Перераспределение неопределившихся для мужчин в возрасте 18–29 лет

Прогнозируемая вероятность	Перераспределение для 4 неопределившихся
0,045	0,18
0,153	0,612
0,321	1,284
0,449	1,796

30ITVN319
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Наблюденные и предсказанные частоты

Возраст	Пол	Удовлетвор	Частота			Процент	
			Наблюденные	Предсказанные	Остаток Пирсона	Наблюденные	Предсказанные
18-29	мужской	Удовл	2	2,872	-0,527	3,10%	4,50%
		Скорее удовл	11	9,762	0,43	17,20%	15,30%
		Скорее не удовл	20	20,561	-0,15	31,30%	32,10%
		Не удовл	27	28,768	-0,444	42,20%	44,90%
		Затр отв	4	2,037	1,398	6,30%	3,20%
	женский	Удовл	3	2,128	0,608	4,50%	3,20%
		Скорее удовл	10	11,238	-0,405	14,90%	16,80%
		Скорее не удовл	17	16,439	0,159	25,40%	24,50%
		Не удовл	34	32,232	0,432	50,70%	48,10%
		Затр отв	3	4,963	-0,916	4,50%	7,40%

Рисунок 2. *Прогнозируемые значения вероятности отнесения к категориям зависимой переменной*

Таблица 6

Прогнозируемые значения распределения респондентов по категориям зависимой переменной

Категория зависимой переменной	Абсолютные значения	В процентах	Суммарное значение
Удовлетворен	23,086	3,862	18,022
Скорее удовлетворен	84,639	14,160	
Скорее не удовлетворен	182,982	30,613	81,977
Не удовлетворен	307,022	51,365	

Вариант № 2. Перераспределение неопределившихся в соответствии с моделью, не учитывающей респондентов, ответивших "затрудняюсь ответить".

Отличие второго варианта от первого будет заключаться в используемых данных для расчетов модели. Для этого из рассматриваемой модели исключаются неопределившиеся респонденты до момента её построения. После чего этапы расчетов проводятся аналогично первому варианту. Второй вариант предпочтительнее использовать при малой доле неопределившихся респондентов и в таком случае значения исследуемого показателя будут изменяться незначительно.

Результаты расчётов:

При расчёте показателя самооценки материального положения путем отбрасывания неопределившихся (способ, используемый в настоящее время) показатель равен "минус 60,5 п.п.". При расчёте по варианту 1 равен "минус 63,955 п.п.", а по варианту 2 равен "минус 63,941 п.п.".

Рассмотрим результаты эксперимента на другом примере, в котором в качестве зависимой переменной выбран другой вопрос анкеты, но и присутствует большая доля неопределившихся:

Удовлетворены ли Вы в целом качеством предоставляемых в Российской Федерации услуг туристско-рекреационного комплекса?

- полностью удовлетворен 3,7 %
- скорее удовлетворен 29,2 %
- скорее не удовлетворен 14,5 %
- полностью не удовлетворен 6,3 %
- затрудняюсь ответить 46,3 %

30ITVN319
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

В качестве независимых переменных также выбраны пол и возраст.

Значение *p*-уровня значимости χ^2 Пирсона равен 0,039, что также свидетельствует о том, что взаимосвязь статистически значима и есть смысл их интерпретировать. При использовании следующих этапов методики перераспределения неопределившихся респондентов получены следующие результаты:

При расчёте показателя путем исключения значений неопределившихся показатель удовлетворенности туристско-рекреационным комплексом равен "12,1 п.п.". При расчёте по варианту 1 (модель с учётом неопределившихся) равен "22,891 п.п.", а по варианту 2 (модель без учёта неопределившихся) равен "24,526 п.п.".

Заключение

В работе сформулирована и апробирована методика, которая была положена в основу алгоритма перераспределения неопределившихся в области оценки социально-экономической и общественно-политической обстановки в регионе. В ходе экспериментальной оценки получились следующие результаты: при небольшом процентном отношении неопределившихся значение исследуемого показателя может отличаться на несколько единиц и отличаться существенно в случае большого количества неопределившихся респондентов (отличие более чем в 2 раза при 46,3 процентов неопределившихся). Кроме того, результаты эксперимента показывают, что способ, используемый в настоящее время при расчёте показателей, основанных на социологических исследованиях, заключающийся в исключении неопределившихся респондентов теряет свою актуальность. Для подтверждения эффективности предложенного алгоритма построения прогнозных оценок при большом количестве неопределившихся респондентов была проведена его апробация на базе социологических исследований по выборной тематике, что позволило обеспечить повышение качества прогнозов.

ЛИТЕРАТУРА

1. Толстова, Ю.Н. Качественная и количественная стратегии. Эмпирическое исследование как измерение в широком смысле / Ю.Н. Толстова, Е.В. Масленников // Социс. – 2000. – № 10. – С. 102.
2. Бослаф, С. Статистика для всех. / Пер. с англ. П.А. Волкова, И. М. Флямер, М.В. Либерман, А.А. Галицына. – Москва: ДМК Пресс, 2015. – 586 с.: ил.
3. Магнус, Я.Р., Катышев, П.К., Пересецкий, А.Л. Эконометрика. Начальный курс: Учебник – 6-е изд., перераб. и доп. – Москва: Дело, 2004. – 576 с.
4. Толстова, Ю.Н. Математическая статистика для социологов: учебник и практикум для СПО / Ю.Н. Толстова. – 2-е изд., испр. и доп. – Москва: Издательство Юрайт, 2018. – 258 с. – Серия: Профессиональное образование.
5. Жучкова, С.В., Ротмистров, А.Н. Возможность работы с пропущенными данными при использовании CHAID: результаты статистического эксперимента / Социология: методология, методы, математическое моделирование. Научный журнал Российской академии наук №46, 2018. – 202 с.
6. Черткова, Е.А. Статистика. Автоматизация обработки информации: учеб. пособие для вузов / Е.Л. Черткова. – 2-е изд., испр. и доп. – Москва: Издательство Юрайт, 2018. – 195 с. – Серия: Университеты России.
7. Бююль, А., Цёфель П. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: Пер. с нем. / Ахим Бююль, Петер Цёфель – Санкт-Петербург: ООО "ДиаСофтЮП", 2005 – 608 с.
8. Наследов, А. IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных / А. Наследов. – Санкт-Петербург: Питер, 2013.
9. Обработка и анализ данных социологических исследований в пакете SPSS 17.0. Курс лекций: учебное пособие / Ш.Ф. Фарахутдинов, А.С. Бушуев. – Тюмень: ТюмГНГУ, 2011. – 220 с.
10. Сидняев, Я.И. Теория планирования эксперимента и анализ статистических данных: учеб. пособие для магистров / Н.И. Сидняев. – 2-е изд., перераб. и доп. – Москва: Издательство Юрайт, 2018.

Afanasev Vadim Vladimirovich
The academy of the federal guard service of the Russian Federation, Orel, Russia
E-mail: affa@mail.ru

Blagij Vladimir Aleksandrovich
The academy of the federal guard service of the Russian Federation, Orel, Russia
E-mail: blagij1@yandex.ru

Vorobjev Andrej Anatol′evich
The academy of the federal guard service of the Russian Federation, Orel, Russia
E-mail: awa@mail.ru

Algorithm of redistribution of undecided respondents on the basis of multinomial logistic regression

Abstract. The article analyzes the existing approaches, methods and means of processing sociological information, on the basis of which it is concluded that it is necessary to improve the accuracy of indicators of evaluation of the studied sociological situation, the impact on the result of sociological research of undecided respondents, the need to choose methods of processing the results of sociological research, taking into account the scales used to measure the answers to questions in the questionnaires.

The analysis of the existing methods of processing the results of sociological research, as a result of which the apparatus of logistic regression was chosen in accordance with the directions of improving approaches to assessing the socio-economic and socio-political situation in the region.

The authors formulated a hypothesis about the possibility of using multinomial logistic regression as a universal (basic) method of sociological data research, which was later confirmed in the course of the experiment on a set of statistical indicators.

To solve the problem of assessing the impact of factors and redistribution of undecided respondents, an algorithm was proposed in the work, the input data for which is the number of undecided respondents, and the output –the forecast value of the distribution of undecided respondents by categories of the dependent variable, taking into account the redistribution.

To confirm the possibility of improving the quality of analytical findings, a number of experiments were conducted using the proposed algorithm. The results of the experimental evaluation, obtained with the help of SPSS Statistics, allowed us to draw the following conclusions: the existing practice of "discarding" undecided respondents significantly affects the quality of assessments of the socio-economic and socio-political situation in the region. The proposed algorithm for constructing predictive estimates of redistribution of undecided respondents allows to neutralize the impact of this indicator and thus improve the quality of forecasts.

Keywords: the algorithm; a case sociological study; a multinomial logistic regression; information processing methods; undecided respondents; statistical experiment; predictive assessment