П.А. Попова, А.Н. Ротмистров
(*Москва*)

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ С КАТЕГОРИАЛЬНЫМИ ПРЕДИКТОРАМИ И ЭФФЕКТАМИ ВЗАИМОДЕЙСТВИЯ И CHAID: СРАВНИТЕЛЬНЫЙ АНАЛИЗ НА ЭМПИРИЧЕСКОМ ПРИМЕРЕ

Статья посвящена методологическому аспекту выявления детерминант политического активизма, а именно вариантам работы с категориальными предикторами, гипотетически объясняющими уровень активизма. Применение логистической регрессии к таким предикторам предполагает их преобразование в фиктивные переменные, что «утяжеляет» модель и создает ряд трудностей в оценке ее качества. «Утяжеление» усугубляется в случае желания исследователя рассмотреть эффекты взаимодействия, поскольку процесс регрессионного моделирования не позволяет учесть степень сходства величин коэффициентов дихотомических предикторов и на основании этого не включать в обработку «лишние» комбинации значений этих предикторов. Авторы статьи обосновывают возможность использования в качестве альтернативы регрессии метод поиска детерминант: CHAID. Цель исследования: сравнение двух указанных методов на основании априорно известных их свойств, обоснование некоторых теоретических преимуществ CHAID над логистической регрессией, параллельное применение этих методов к эмпирическим данным, сравнение полученных результатов. Исследование проведено на данных Европейского социального обследования (European Social Survey – ESS) 2012 г. Зависимой переменной выступил «политический

Полина Артемовна Попова – магистрант Национального исследовательского университета «Высшая школа экономик» (НИУ ВШЭ), менеджер Лаборатории экономико-социологических исследований НИУ ВШЭ. E-mail: papopova@hse.ru
Алексей Николаевич Ротмистров – кандидат социологических наук, доцент кафедры методов сбора и анализа социологической информации НИУ ВШЭ. E-mail: alexey.n.rotmistrov@gmail.com.

активизм», а набор гипотетических детерминант был составлен из переменных социально-экономического блока панели.

Введение и постановка проблемы

Важный класс задач в социальных науках – поиск причинно-следственных связей между явлениями. Для этого предназначен эксперимент, поскольку он позволяет контролировать все изучаемые переменные, в том числе внешние. В ситуациях, когда проведение эксперимента затруднительно, наиболее частой заменой выступает статистический анализ, в частности, регрессионный – с поправкой на то, что различие между причиной и следствием не поддается исчерпывающей формализации [1, с. 164]. Если зависимая переменная – категориальная (что особенно характерно для социальных наук), то часто применяется логистический регрессионный анализ (далее – ЛР).

Вкратце, ЛР – это метод, который на основе знания значений независимых переменных (или предикторов, часто обозначаемых «*X*» и рассматриваемых в качестве причин) позволяет предсказывать вероятности значений категориальной зависимой переменной (часто обозначаемой «*Y*» и рассматриваемой в качестве следствия). Основы ЛР можно почерпнуть, например, в [2]. В контексте нашей статьи важно, что ЛР предъявляет ряд требований к предикторам, минимальное из которых состоит в том, чтобы тип шкалы всех предикторов был интервального уровня или выше. В то же время в социологических исследованиях часто встречаются категориальные ранговые переменные. Можно ли включать в ЛР такие переменные в качестве предикторов «как есть»?

Бытует мнение, разделяемое в том числе признанными авторитетами в области анализа данных (см., например: [2]), что почти всегда с такими предикторами можно работать как с интервальными, если они имеют 4 градации и более: «Когда размер

выборки не очень велик, вполне может оказаться, что построенная модель адекватна реальности. Поэтому удобно считать порядковые предикторы интервальными, если построенные на основании модели прогнозы достаточно точны; такие модели легче интерпретировать, а коэффициент порядкового предиктора (взятого в исходном виде) более значим, чем коэффициенты дихотомических предикторов (полученных из исходного рангового предиктора)» [2, p. 14]. Однако есть и противоположное мнение (например: [3]), подтверждаемое специальными исследованиями [4; 5; 6]: категориальные ранговые предикторы можно рассматривать в качестве интервальных, если они распределены симметрично относительно среднего арифметического и если они имеют большое число (от 7) упорядоченных градаций, выражающих степень интенсивности некоторого свойства (скажем, удовлетворенность работой), а не самостоятельные состояния (например, тип трудового договора).

Мы придерживаемся второго мнения как более консервативного и осторожного, поэтому считаем, что если изучать влияние категориальных ранговых предикторов (не удовлетворяющих перечисленным выше требованиям) на некую категориальную зависимую переменную посредством ЛР, то придется их преобразовывать в наборы дихотомических предикторов (т.е. принимающих два значения; будем обозначать эти значения 0 и 1), часто называемых фиктивными.

Иллюстрация: допустим, нас интересует влияние трехрангового предиктора «уровень счастья» на уровень политического активизма. Вместо этого предиктора вводим от 1 до 3 фиктивных предикторов. Если мы считаем содержательно допустимым объединить какие-то значения исходного предиктора, чтобы осталось только 2 группы значений, то можем ограничиться заменой исходного предиктора на 1 дихотомический предиктор; в контексте данной статьи назовем такую дихотомизацию «**неполная**». Неполная дихотомизация проста, но сопряжена с потерей информации (чем больше категорий в исходном предикторе, тем, очевидно, больше

теряется информации). Поэтому обычно каждое значение исходной категориальной переменной превращается в дихотомическую переменную (т.е. в самостоятельный математический конструкт).

Очень счастлив: X_1 = 1, если респондент очень счастлив;
0, если он не очень счастлив.
Умеренно счастлив: X_2 = 1, если респондент умеренно счастлив;
0, если он более или, наоборот, менее счастлив;
Несчастлив: X_3 = 1, если респондент несчастлив;
0, если он счастлив в какой-то мере.

Вместо исходного предиктора получили 3 дихотомических предиктора; введем полезный термин: «**набор дихотомических предикторов**» – это новые дихотомические предикторы, относящиеся к одному и тому же исходному категориальному предиктору. В нашем модельном примере мы получили из исходного 3-рангового предиктора набор из 3-х дихотомических предикторов, т.е. провели полную дихотомизацию, тем самым полностью сохранили информацию об уровне счастья респондентов.

Происходит ли вообще какая-то потеря информации при полной дихотомизации исходного предиктора? Если исходный предиктор ранговый, то теряется информация о порядке между его значениями. Впрочем, в рамках алгоритма ЛР такая потеря информации является технической, а не содержательной. Содержательная интерпретация обеднеть не должна, поскольку часто коэффициенты при предикторах, полученных после «рассыпания» исходного порядкового предиктора, оказываются монотонно упорядоченными, а именно: коэффициенты при таких дихотомических предикторах, расположенных согласно порядку категорий исходного порядкового предиктора, возрастают или убывают монотонно. Если же подобная монотонность не наблюдается, это может быть интерпретировано как индикатор наличия немонотонной связи между рассматриваемым порядковым предиктором и зависимой переменной. Получается, что содержательная интерпретация связи между зависимой переменной и порядковым

предиктором может даже обогатиться после его «рассыпания» на дихотомические переменные.

Другое полезное следствие процедуры дихотомизации исходных категориальных предикторов – возможность выяснять наличие влияния на зависимую переменную не только категориальных предикторов по отдельности, но и комбинаций категорий, «надерганных» из разных категориальных предикторов. Напомним, что влияние одного предиктора на зависимую переменную (*Y*) называется эффектом первого уровня или одномерным эффектом, или главным эффектом – это синонимы; влияние комбинации категорий двух разных предикторов (один из них континуальный или дихотомический, а второй – дихотомический; если оба – дихотомические, то они обязательно принадлежат к разным наборам) на *Y* называется эффектом второго уровня, или 2-мерным эффектом; эффекты второго уровня и выше называются «эффектами взаимодействия».

Применение эффектов взаимодействия в качестве предикторов фактически позволяет в рамках линейного моделирования (к которому относится и ЛР) искать нелинейное влияние предикторов на зависимую переменную. Как известно, линейность методов линейного моделирования проистекает из двух источников: из линейности парных связей, измеряемых коэффициентом корреляции Пирсона, и из линейности многочлена, формирующего регрессионное уравнение линейной регрессии [7]. Линейность – это и преимущество модели, так как подобная модель легко обсчитывается и интерпретируется, и ее ограничение, так как она грубо аппроксимирует условные математические ожидания *Y*. ЛР с эффектами взаимодействия отчасти сохраняет обозначенное преимущество и устраняет указанное ограничение, поскольку использует коэффициент корреляции Пирсона и линейный многочлен, но в отношении дихотомических переменных. Другими словами, если прогностическая сила некой логистической регрессионной модели, включающей только главные эффекты, оказывается не-

удовлетворительной, т. е. существует вероятность получить модель лучше и при этом вполне интерпретируемую посредством включения в нее эффектов взаимодействия. Иллюстрация: в результате применения ЛР с дихотомизированными категориальными предикторами может оказаться, что на уровень активного политического участия значимо влияет не уровень счастья респондента и не уровень институционального доверия как таковые, а только комбинация высокого уровня счастья («очень счастлив») и низкого уровня институционального доверия.

Обоснование трактовки дихотомической шкалы как частного случая интервальной можно найти в [1]. К сожалению, перечисленным неоспоримым преимуществам практики дихотомизации категориальных предикторов для их включения в ЛР сопутствуют неявные, но чреватые негативными последствиями проблемы. Ряд таких проблем, рассмотренных в разных научных источниках, собраны и обобщены в [7; 8]. Кроме того, в [8] на основе теоретических и эмпирических изысканий предложена альтернатива логистической регрессии с дихотомизированными переменными: логлинейный анализ (далее ЛЛА).

Однако мы не считаем работу по осмыслению проблем применения ЛР с фиктивными переменными завершенной, хотя бы потому что и ЛЛА – не «панацея», так как он имеет ряд ограничений, свойственных и ЛР с фиктивными переменными. Рассмотрим их кратко.

И ЛР, и ЛЛА оперируют категориями исходных категориальных предикторов как отдельными математическими конструктами. При этом ЛР требует их физического существования (в виде фиктивных переменных), а ЛЛА не требует, благодаря чему отношение размера одной и той же выборки к числу предикторов обычно ниже для ЛР, чем для ЛЛА. Это отношение важно, когда объем выборки не слишком велик. Как известно, для регрессионного анализа достаточный размер выборки зависит прежде всего от планируемого числа предикторов. Общепринятой формулы зависимости размера выборки m от числа предикторов k пока не существует. Одна из

простейших формул: $m = 10k$. Другая простая формула: $m = 2^k$ [9: 15]. Также желательно принимать во внимание планируемые вероятности ошибок первого и второго рода и характеристики распределения зависимой переменной (соответствующая формула приведена в [10]).

Нужно ли рассматривать все категории исходных категориальных предикторов как отдельные конструкты? Нет, если величины одномерных эффектов (выражаемых в ЛР регрессионными коэффициентами) какой-то группы категорий из одного набора статистически равны. Иллюстрация: если влияние низкого и среднего уровней счастья («несчастлив» и «счастлив умеренно») на зависимую переменную имеет одинаковые величины эффектов, то нет смысла рассматривать эти категории как отдельные конструкты; их логично объединить. Та же логика актуальна и в применении к многомерным эффектам. Ни ЛР, ни ЛЛА «не умеют» работать в этой логике, что ведет к жестким техническим ограничениям: если «загрузить» 5–7 предикторов с 3–5 категориями в ЛЛА, то он обычно отказывается работать или работает неприемлемо долго (скажем, месяц и дольше); если «загрузить» эти предикторы в виде фиктивных переменных их взаимодействий в ЛР, то она обычно отказывается работать уже при эффектах второго уровня.

Понимая пользу ЛР с эффектами взаимодействия, в том числе более чем 2-мерными, социологи предлагают предваряющие ЛР процедуры, способствующие избирательному (а не сплошному) включению в ЛР категорий категориальных предикторов и их комбинаций, например в [11]. На наш взгляд, для этого подходят методы, «умеющие» находить похожие по величинам эффектов категории исходного категориального предиктора и объединять их в один конструкт. Эту особенность таких методов в контексте нашей статьи предлагаем называть «обоснованное совместное рассмотрение категорий предиктора». Этот подход кардинально отличается от подхода неполной дихотомизации, предполагающей волюнтаристское совместное рассмотрение некоторых категорий

категориального предиктора еще до применения алгоритма, т.е. без всякого учета схожести величин эффектов этих категорий.

Таким образом, в данной статье мы хотели бы: 1) снова обратить внимание на условия неадекватности применения ЛР к категориальным предикторам; 2) внести некоторый вклад в рефлексию по поводу этих условий и обосновать преимущества дополнения решения задачи поиска детерминант методом, решающим сходные задачи, но лишенным таких ограничений; 3) проиллюстрировать практическую пользу изучения эффектов взаимодействия категориальных предикторов.

Деревья классификации как альтернативное решение задачи поиска детерминант: суть, варианты, опыт применения

Деревья классификации – один из наиболее подходящих для сравнения с регрессионным моделированием класс методов, поскольку он:

– позволяет искать детерминанты явлений, в том числе и в виде многомерных взаимодействий;

– работает со шкалами любых типов, причем не требует преобразования категориальных шкал в фиктивные переменные;

– реализует подход обоснованного совместного рассмотрения категорий предикторов.

Деревья классификации (далее ДК) – класс однородных методов, сердцевину которых составляет алгоритм автоматического поиска взаимодействий («automatic interaction detector – AID»), т.е. комбинаций значений признаков, объясняющих (в математическом смысле) интересующее исследователя явление. Самые распространенные из методов ДК: CHAID и CRT.

CRT ищет значения предикторов и их комбинации, объясняющие интервальную переменную; категории любого пре-

диктора разбивает строго на 2 группы (по несхожести величин их эффектов). В CHAID зависимая переменная категориальна, «склеивание» значений предикторов не ограничено числом групп. В обоих отношениях CHAID представляется более актуальным для социальных наук и более близким логистической регрессии.

Алгоритм содержит в себе 3 основных элемента [12]:

– выяснение, какие категории каких предикторов дают статистически разные условные распределения зависимой переменной (в этом смысле оказывают значимый эффект на нее) и предсказание модального значения зависимой переменной для каждой образовавшейся комбинации предикторов;

– расщепление выборки на подвыборки согласно этим комбинациям и повторение элемента 1;

– условия остановки расщепления.

Первый элемент ключевой для понимания цели CHAID – поиска групп объектов, однородных по зависимой переменной и по предикторам [1]. Таким образом, метод находит такие взаимодействия предикторов, которые образуют однородные группы, максимизируя при этом разнородность между группами по значениям зависимой переменной. Для оценки статистической значимости однородности/разнородности CHAID использует критерий χ^2 (отсюда и название CHAID) [13]. Безусловное распределение зависимой переменной служит корнем «дерева» (корневым узлом), который впоследствии расщепляется на ряд последующих узлов – родительских (которые также имеют «отростки») и терминальных (конечных) [14]. На каждой итерации (реализуя элементы 1 и 2 алгоритма) CHAID анализирует все включенные в модель предикторы и выбирает тот из них, который лучше всего объясняет различия, т. е. имеет наибольшее значение статистики и, соответственно, наибольшую вероятность наличия связи с зависимой переменной, он становится 1 уровнем «дерева». На второй итерации расщепление происходит уже от этого яруса; CHAID «выращивает» 2-й уровень и так далее, пока алгоритм не

остановится. Критерии его остановки следующие: либо достигается максимальная заданная размерность дерева, либо размер выборки в очередном узле оказывается меньше допустимого, либо расщепление невозможно произвести на заданном уровне значимости (*p-value*). Но эти критерии имеют существенный недостаток – они довольно субъективны, так как исследователю приходится задавать их вручную. Примеры реализации CHAID и его нюансы см.: [1, гл. 2.5.3].

Наше предложение использовать ДК в качестве аналога ДР не оригинально. Например, авторы [15] следуют той же логике: «Одна из главных проблем регрессии – ситуация, когда анализ строится на небольшой выборке или с большим количеством предикторов. Чтобы обойти проблему разреженности данных при построении регрессии, был применен разведывательный анализ методом ДК, с помощью которого выявлялись переменные, наиболее сильно связанные с зависимой переменной. Более того, результаты ДК просто интерпретировать» [15, p. 234].

Мы не первые, кому пришла идея сравнить ЛР и ДК, однако практически отсутствуют систематические сравнения методов, чаще речь идет об описании их преимуществ, недостатков и обосновании того, почему они выбраны для решения некоторых содержательных задач. В качестве исключения можно привести работу Холгерссона и соавторов, где обращение к ДК обосновывается тем, что задачи, которые они ставят в своей статье, являются новаторскими, поэтому для отбора предикторов регрессионной модели была необходимость провести разведочный анализ методом ДК, однако почему именно этим методом – не пояснено [15]. Среди преимуществ ДК авторы отмечают факт нелинейности и непараметричности искомых связей. Существенный недостаток выбранного метода ДК – CHAID – отсутствие объективного критерия остановки расщепления уровней «дерева». Тем не менее CHAID удобен, так как не требует задавать число расщеплений в алгоритме построения «дерева». Авторы предлагают и эвристи-

ческое решение указанной проблемы CHAID'а: подтверждающий регрессионный анализ с использованием только тех предикторов, которые оказались значимы в построенном «дереве». Авторы рассматриваемой статьи делают замечание, с которым мы полностью солидарны: работая с выборкой, имеющей мало наблюдений относительно числа предикторов, нельзя ограничиваться только регрессией, поскольку в силу разреженности данных велик риск получить смещенную регрессионную модель.

В работе Хорнера и соавторов исследуется возможность прогнозировать проявления жестокости [16]. Существенным недостатком регрессии авторы называют тот факт, что она не позволяет учесть, что разные переменные могут предсказывать жестокость для разных подгрупп индивидов (в нашей терминологии здесь, видимо, речь идет о взаимодействии между изучаемыми переменными), как следствие, авторы считают, что ДК намного лучше подходят для клинических испытаний такого типа.

Эти же авторы отсылают читателей к исследованию Лью и соавторов, где сопоставлены CRT и ЛР [17]. Независимыми были выбраны клинические и демографические переменные и переменные, имеющие отношения к правонарушениям (их структура не описана). В целом модели были похожи по выявленным связям, но кросс-валидация показала бóльшую точность регрессионной модели (0,74 против 0,66). По свидетельству авторов, 3 из 11 рассмотренных ими исследований показали одинаковую точность моделей CRT и ЛР, в одном CRT сработал гораздо хуже, в 7 – гораздо лучше (но без необходимой валидации).

Таким образом, мы делаем вывод, что деревья классификации способны стать хорошей альтернативой логистической регрессии: этот метод может дать нам более надежные выводы или послужить частью комплексной методологии, состоящей из ДК в качестве разведывательного анализа и регрессионного моделирования в качестве последующего подтверждения результатов применения ДК.

Критерии сравнения ЛР и CHAID, методологическая гипотеза, эмпирическая база

Мы последовательно рассмотрели важные характеристики ЛР и CHAID и сгруппировали их в 11 критериев априорного (теоретического) сравнения; кратко они отражены в *табл.* 1, туда же добавлен (справочно) столбец из другой нашей статьи [8], посвященной ЛЛА.

Главная функция метода. И ЛР, и CHAID ищут связи между изучаемыми признаками и позволяют прогнозировать значения зависимой переменной. При этом CHAID статистически обоснованно группирует категории независимых переменных.

Наличие и суть контрольной группы. Контрольная (или референтная) группа в регрессионном моделировании – это одна или несколько комбинаций значений изучаемых признаков. В мультиномиальной ЛР контрольных групп столько же, сколько комбинаций значений предикторов, каждая контрольная группа включает комбинацию значений предикторов с первым или последним (зависит от выбранной опции) значением зависимой переменной. В бинарной ЛР контрольная группа по умолчанию включает комбинацию нулевых значений предикторов с нулевым значением зависимой переменной. В CHAID контрольная группа отсутствует.

Из написанного следует, что коэффициенты предикторов ЛР дифференцирует только вероятности наборов комбинаций значений изучаемых переменных, отличающихся только значением зависимой переменной, причем одно из этих значений обязательно должно принадлежать к контрольной группе. Например, поместив в ЛР дихотомические предикторы *X1* и *X2* (со значениями 0 и 1), а также зависимую переменную Y (со значениями 1, 2, 3, из которых последнее относится к контрольной группе), получим модель, коэффициенты которой позволяют судить об отношении вероятности, скажем, *X1=1&X2=1&Y=1* к

вероятности *X1=1&X2=1&Y=3*. Но они не позволяют судить об отношении вероятностей *X1=1&X2=1&Y=1* и *X1=1&X2=1&Y=2* или *X1=1&X2=1&Y=1* и *X1=1&X2=0&Y=1*. Само по себе это не проблема, так как на основании коэффициентов можно рассчитать предсказанные вероятности зависимой переменной для каждой комбинации предикторов (в нашем исследовании они представлены в формате процентов в приложении А). Но добавим сюда тот факт, что в зависимости от выбора контрольной группы коэффициенты при одних и тех же предикторах могут иметь бóльшие или меньшие значения и, как следствие, оказаться статистически значимыми или не значимыми; из чего следует возможность, что в моделях, различающихся исходно только выбранной контрольной группой, в итоге будут значимы коэффициенты при разных предикторах и предсказанные на основании этих разных наборов коэффициентов вероятности зависимой переменной для одной и той же комбинации предикторов тоже будут разными. Это приводит к некоторым трудностям при сравнении логистических регрессионных моделей на одних и тех же данных, но с разными контрольными группами. Поэтому сравнение может оказаться трудоемким, а без сравнения, и при этом имея в этих моделях разные наборы значимых предикторов, исследователь склонен выбрать итоговую модель по своему вкусу, что, разумеется, не повышает надежность результатов анализа данных.

Результаты CHAID не привязаны к контрольной группе и, как следствие, избавлены от описанной проблемы. Да, CHAID позволяет выбрать одно из значений зависимой переменной в качестве целевого, но этот выбор не влияет на характеристики модели, а лишь служит ее детализации и визуализации.

Требования к данным. ЛР требует равномерности распределения зависимой переменной, поскольку прогноз в ЛР модальный, т.е. основан на модальном (по ожидаемой частоте) значении *Y* внутри каждой комбинации предикторов. Следовательно, даже если ожидаемые и эмпирические частоты полностью совпадут,

прогнозироваться будет наиболее часто встречающееся значение *Y* при каждой комбинации предикторов. Следовательно, чем дальше безусловное распределение *Y* от равномерности, тем чаще, при прочих равных, будет прогнозироваться преобладающее значение внутри условных распределений *Y*. Другими словами, исходный перекос в пользу одного из значений зависимой переменной может «перебить» выявленные моделью закономерности. В CHAID модальный прогноз также интегрирован, что имеет те же следствия.

Другой общий аспект для обоих методов – выборка должна иметь адекватные размеры (что, впрочем, однозначно нигде не регламентировано, есть разные варианты вычисления оптимального размера).

Некоторые проблемы интерпретации результатов ЛР возникают, если наблюдения зависимы друг от друга или предикторы мультиколлинеарны. В CHAID эти явления не вызывают проблем интерпретации.

Наличие зависимой переменной. Оба метода обязательно предполагают наличие зависимой переменной.

Включение категориальных предикторов. CHAID был выбран нами как метод, который без преобразований работает с номинальными и порядковыми предикторами. ЛР работает с интервальными предикторами, а включение категориальных требует преобразования их в дихотомические.

«Склеивание» категорий со статистически близкими величинами эффектов. Как мы уже показали во введении, ЛР рассматривает категории категориальных предикторов как отдельные математические конструкты. Одно из главных преимуществ CHAID – «склеивание» категорий одного и того же предиктора, имеющих статистически близкие величины эффектов.

Оценка качества модели (реализация в SPSS). Качественной моделью следует считать позволяющую генерировать данные, статистически не отличающиеся от эмпирических. В ЛР мерой качества служит уже упомянутый логарифм отношения правдопо-

добия (log-likelihood ratio), посредством которого статистически оценивается суммарное отклонение ожидаемой частоты каждой комбинации значений изучаемых признаков от его же эмпирической частоты. Поскольку ЛР нацелена на прогнозирование, причем через модальный прогноз, дополнительными оценками качества выступают таблица классификации (показывает процент правильных предсказаний для каждой категории зависимой переменной) и псевдо-R^2. Как мы уже отмечали, неравномерность распределения зависимой переменной обрекает ЛР на производство неудовлетворительных по своей прогностической силе моделей, даже если, с точки зрения логарифма отношения правдоподобия, они качественны. В CHAID для категориальных зависимых переменных мерой качества служит доля неправильных прогнозов и стандартная ошибка этой доли. Таким образом, если для ЛР таблица классификации выступает вспомогательной мерой качества, то для ДК – главной, в этом отношении ДК более чувствителен, чем ЛР к отклонению распределения зависимой категориальной переменной от равномерности. Подчеркнем также, что логарифм отношения правдоподобия и стандартная ошибка доли неправильных прогнозов – статистический инструмент, таблица классификации и псевдо-R^2 – не статистические. Впрочем, и для CHAID нетрудно вручную рассчитать логарифм отношения правдоподобия.

Механизм поиска линейной связей и взаимодействий. В ЛР работает статистическая (обычно посредством статистики Вальда) оценка значимости линейной связи между каждым предиктором и логарифмированной относительной частотой зависимой переменной. В CHAID критерий χ^2 «ищет» связи между зависимой и независимыми переменными по всей выборке или подвыборке, сформированной материнским узлом.

Виды искомых связей. Исходя из изложенных общих принципов сравниваемых методов и, в частности, из предыдущего пункта, ясно, что ЛР ищет линейные связи, а CHAID сколь угодно многомерные пучки (взаимодействия) между предикторами и за-

висимой переменной. Таким образом, ЛР ищет связи, являющиеся по своему виду частным случаем связей, искомых CHAID. Рассыпание любого предиктора на фиктивные переменные в рамках ЛР в некотором смысле приближает ее к CHAID, но влечет указанные выше негативные последствия.

Интерпретация результатов. В результате ЛР мы получаем уравнение, коэффициенты которого обычно интерпретируются как отношения условных вероятностей Y. В CHAID мы получаем вероятность каждого значения зависимой переменной для *n*-мерной комбинации значений предикторов.

Отбор предикторов. Оба метода предлагают пошаговый отбор предикторов, но в CHAID эта «опция» обязательна, так как является одним из ключевых элементов алгоритма, а в ЛР она не обязательна и, как мы уже отмечали, многие исследователи ею пренебрегают. ДК по этому критерию абсолютно не гибки еще и потому, что суть их алгоритма допускает только отбор путем последовательного включения предикторов в модель (Forward), а ЛР не только допускает отбор путем последовательного включения предикторов в модель или их последовательного исключения (Backward), но и предлагает несколько разновидностей этих процедур.

Казалось бы, пошаговый отбор предикторов – техническая процедура, но она может повлиять на содержание итоговой модели. Поскольку в CHAID эта процедура обязательна, и учитывая, что он исходно нацелен на поиск сколь угодно многомерных взаимодействий между предикторами и зависимой переменной (см. выше п. 9), то его разработчики предусмотрели возможность перебирать не все комбинации значений предикторов в рамках многомерных взаимодействий, чтобы не перегружать оперативную память вычислительной техники. Образно выражаясь, обычно CHAID на каждом следующем шаге «выращивает» не все возможные ветви (многомерные эффекты), а только те, которые на предыдущем шаге показали свою значимость. Получается, на первых шагах «растет» много ветвей, но они «маломерны», поэтому оперативная память

вычислительной техники не перегружается, а на последних шагах «растут» многомерные ветви, но их мало, поэтому оперативная память вычислительной техники опять же не перегружается. В ЛР же, как мы уже подробно написали во введении, если стоит задача проверить статистическую значимость многомерных эффектов взаимодействия, но нет теоретической рамки, ограничивающей их перечень, то вычислительной технике приходится обработать все возможные многомерные эффекты взаимодействия, что часто оказывается непосильной задачей из-за нехватки оперативной памяти. Таким образом, шанс выявить значимые многомерные взаимодействия между предикторами и зависимой переменной гораздо выше при применении CHAID, чем ЛР.

В *табл.* 1 предпринятое сравнение отражено схематично. Эта таблица имеет целью помочь исследователям при выборе из трех рассматриваемых методов, учитывая, что они решают похожие задачи.

На основе проведенного теоретического сравнения мы выдвинули методологические гипотезы: благодаря а) отсутствию необходимости создавать фиктивные переменные из категориальных предикторов и возможности «склеивания» категорий предикторов со статистически близкими величинами эффектов CHAID даст возможность изучить эффекты уровня выше, чем ЛР, б) это, в свою очередь, позволит получить модель с более высокой прогностической силой.

Для проверки гипотез мы прибегли к анализу эмпирической базы Европейского социального исследования[1] 2012 г., которая показалась нам актуальной по причине проработанной методологии сбора данных и обширной выборки (достаточность выборки, как мы писали, выступает одним из ограничений рассматриваемых методов); по причине участия в этой волне (в отличие от более поздних волн) представителей России и по причине наличия в

[1] European Social Survey – ESS, подробнее о ней можно узнать на сайте исследовательского проекта http://www.europeansocialsurvey.org/.

Таблица 1

СРАВНЕНИЕ КРИТЕРИЕВ

Критерий сравнения	Логистическая регрессия	Деревья классификации	Логлинейный анализ
Главная функция	Предсказание	Предсказание	Поиск взаимодействий + предсказание
Наличие контрольной группы и ее суть	Комбинация значений предикторов с первым или последним (зависит от выбранной опции) значением зависимой переменной. Число контрольных групп равно числу комбинаций значений предикторов. В бинарной ЛР контрольная группа по умолчанию включает комбинацию нулевых значений предикторов с нулевым значением зависимой переменной	Отсутствует	В ЛЛА как таковом – это комбинация последних (по кодировке) значений изучаемых признаков. Таким образом, контрольный профиль один. В логит-регрессии контрольный профиль включает все комбинации предикторов с первым или последним (зависит от выбранной опции) значением зависимой переменной (то же, что в ЛР)
Требования к данным	Зависимая переменная равномерно распределена, наблюдения не зависимы друг от друга, предикторы, по возможности, не коррелируют между собой, адекватный размер выборки	Ограниченное число переменных, предсказываемая переменная равномерно распределена, адекватный размер выборки	Ограниченное число переменных, адекватный размер выборки

Продолжение табл. 1

Критерий сравнения	Логистическая регрессия	Деревья классификации	Логлинейный анализ
Зависимая переменная	Обязательна, ею выступает целая переменная	Выбирается целая переменная и отдельная ее категория, которая будет как можно более точно предсказываться моделью	Не обязательна, в ее качестве возможно использование целой переменной или отдельной категории
Включение категориальных переменных	Преобразование в дихотомические переменные	Без изменений	Без изменений
«Склеивание» категорий со статистически близкими эффектами	Нет	Есть	Нет
Оценка качества модели	Логарифм правдоподобия + значимость коэффициентов + псевдо-R^2 + таблица классификации	Таблица классификации + стандартная ошибка доли неправильных прогнозов	Логарифм правдоподобия + значимость микроэффектов

Продолжение табл. 1

Критерий сравнения	Логистическая регрессия	Деревья классификации	Логлинейный анализ
Механизм поиска линейных связей и взаимодействий	Коэффициент линейной связи между каждым предиктором и логарифмированной относительной частотой зависимой переменной	Критерий χ^2 между предсказываемой переменной и каждым предиктором по всей выборке или подвыборке, сформированной материнским узлом	Величина отклонения логарифма отношения правдоподобия от нуля при исключении иерархическим методом каждой связки переменных + оценка значимости логарифмированного отклонения частоты интересующей комбинации изучаемых признаков от частоты контрольного профиля
Виды искомых связей	Линейные связи между предикторами и зависимой переменной	Многомерные пучки (взаимодействия) категорий изучаемых признаков между предикторами и зависимой переменной	Многомерные пучки (взаимодействия) категорий всех изучаемых признаков

Окончание табл. 1

Критерий сравнения	Логистическая регрессия	Деревья классификации	Логлинейный анализ
Интерпретация результатов	Уравнение, коэффициенты которого интерпретируются либо как вероятности того или иного значения зависимой переменной, либо как отношения шансов	Вероятность каждого значения зависимой переменной для n-мерной комбинации значений предикторов	Отношения вероятностей, как и в ЛР, причем не только для 2-мерной комбинации, но и *n*-мерной
Отбор предикторов	Возможны и Backward, и Forward	Только Forward. Процедура по ходу применения сама ограничивает перечень комбинации значений предикторов в рамках многомерных взаимодействий	Только Backward

ней содержательно интересного для нас набора переменных. Этот набор включал:

– 7 переменных, отражающих вопросы с закрытыми дихотомическими ответами про разные стороны политического активизма (контакты с политиками, участие в демонстрациях, в работе партий, иных социальных организациях; участие в публичных кампаниях, в том числе в бойкотах; подпись петиций). Из них была получена зависимая переменная: если респондент давал хотя бы один положительный ответ на эти вопросы, то ему приписывалось значение «1», в противном случае «0»;

– переменные – гипотетические предикторы политического активизма: доверие власти (интегральная, получена как результат факторизации четырех релевантных индикаторов), доверие людям (интегральная, получена как результат факторизации трех релевантных индикаторов), уровень счастья, интерес к политике, просмотр политических передач, возраст, доход, пол, и принадлежность к религиозной конфессии.

Технически изучаемые переменные и в целом база, на наш взгляд, хорошо репрезентрируют типичную для социальных наук ситуацию: зависимая переменная существенно смещена в сторону одной из категорий («Отсутствие проявлений активизма»), большинство предикторов исходно ранговые, многие из них смещены к одному из полюсов, выборка – 1612 представителей России. Чтобы снизить смещенность предикторов, мы объединили предварительно малонаполненные категории с более наполненными. В результате переменные, отражающие доверие, уровень счастья, возраст и доход стали 3-ранговыми, интерес к политике и просмотр политических передач – 4-ранговыми, остальные две – дихотомическими[1].

Почему мы не прибегли к статистическому эксперименту на модельных данных? Мы составили перечень критериев сравнения

[1] Подробно подготовительные и аналитические процедуры, а также результаты представлены на нашем сайте http://www.rotmistrov.com/tr.

рассматриваемых методов на основании их известных свойств, т.е. провели некоторую систематизацию знания об этих методах. Но наша систематизация еще, пожалуй, далека от такого уровня, чтобы на основании ее ясно понимать, какими параметрами модельных данных можно регулировать проверку той или иной методологической гипотезы. Наша систематизация – это пока что скорее памятка для желающих применить ЛР или CHAID.

Сравнение результатов применения изучаемых методов

Как мы уже отмечали, основным статистическим методом поиска детерминант является регрессионное моделирование, поэтому «точкой отсчета» из двух сравниваемых методов мы выбрали именно ЛР. Для ее применения исходные предикторы (кроме дихотомических) подверглись процедуре полной дихотомизации; в результате были получены 18 дихотомических переменных. Они выступили главными эффектами в бинарной ЛР.

ЛР, включающая только главные эффекты, не смогла «предсказать» ни одного положительного значения зависимой переменной («наличие проявлений активизма»). В соответствии с изложенным ранее принципом перехода к нелинейной модели в случае неудовлетворительной линейной модели мы подготовили эффекты 2-го и 3-го уровней в виде комбинаций исходных фиктивных переменных. Не имея априорных предположений, какие из многомерных эффектов будут полезны, нам пришлось подготовить в сумме 2559 одно-, двух- и трехмерных предикторов. В случае сплошного включения этих предикторов в ЛР, вычислительная техника сообщала о нехватке оперативной памяти. Чтобы обойти это препятствие, мы прибегли к процедуре пошагового включения предикторов в ЛР. Почему именно включения (Forward), а не исключения (Backward), ведь опыт показывает, что «при прочих равных» последние дают модели с большей прогностической силой?

Во-первых, процедуры исключения требуют предварительного включения всех предикторов в модель, что в нашем случае оказалось невозможно из-за нехватки оперативной памяти; во-вторых, CHAID основан именно на процедуре включения, поэтому выбор аналогичной процедуры для ЛР повышает степень сравнимости изучаемых методов.

Процедура включения длилась 12 шагов и дала модель с отличной от нуля долей предсказаний положительного значения зависимой переменной. Рассмотрим качество этой модели. Логарифм отношения правдоподобия к 12-му шагу снизился по сравнению с 1-м шагом (*табл. 2*), но остался статистически значимо отличен от нуля для интересующего нас значения зависимой переменной (на 95% доверительном интервале); другими словами, расхождение между предсказанными и реальными условными вероятностями значений *Y* велико.

Таблица 2

СВОДКА МОДЕЛИ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Шаг	–2 логарифм правдоподобия	*R*-квадрат Кокса и Снелла	*R*-квадрат Нагелькерка
1	2951,398	0,021	0,034
12	2653,369[c]	0,108	0,176

О том же свидетельствуют величины псевдо-R^2: Кокса и Снелла, Нагелькерка и МакФаддена (*табл. 2*) и процент правильных прогнозов положительного значения зависимой переменной в таблице классификации: 9,9% (Приложение А).

Содержательные результаты отражены в Приложении Б, отметим самые заметные из них:

– сильно повышает вероятность наличия проявлений политического активизма комбинация таких характеристик, как недоверие власти, высокий интерес к политике и высокий уровень счастья;

– чуть слабее положительный эффект комбинации низкой степени политического интереса, средней частоты просмотра политических передач и молодого возраста;

– самую низкую вероятность политического активизма демонстрируют молодые обеспеченные мужчины.

К сожалению, выявленные закономерности объясняют лишь менее 10% случаев политического активизма. Допускаем (и это можно доказать на примере исследований, эмпирические базы для которых открыты), что многие на этом сочли бы задачу поиска детерминант решенной, сделав лишь поправку, что есть проблемы с данными, выборкой или чем-то еще. Мы же видим здесь действие описанных выше причин: а) наличие 7 наборов фиктивных переменных, которые существенно увеличивают число комбинаций переменных без прироста новой информации и искажают оценку статистической значимости многомерных взаимодействий (указанная ранее проблема относительной мощности выборки); б) как следствие техническая невозможность рассмотреть эффекты выше 3-го уровня; в) модель основана на модальном прогнозе и, как следствие, не справляется с неравномерностью распределения зависимой переменной.

Нельзя не отметить, что применив процедуру 2-х и 3-мерного комбинирования фиктивных переменных, мы фактически перешли не к значениям переменных, а к профилям респондентов (под профилями мы понимаем сочетание различных характеристик). Получается, что логистическая регрессия, построенная на фиктивных переменных с эффектами взаимодействий, приводит к построению модели с профилями. Но чтобы достичь этого, мы, во-первых, должны помнить о необходимости проверки многомерных эффектов, а во-вторых, потратить немало времени на кодирование для построения комбинаций большого числа характеристик.

CHAID тоже строит профили, но делает это во всех смыслах более легким путем, благодаря чему можно найти гораздо более многомерные значимые эффекты. Условием остановки метода мы

задали достижение в материнском узле 10 наблюдений, а в производном от него – 5 наблюдений. Это условие тоже вписывается в стремление обеспечить сравнимость ЛР и CHAID, поскольку полученные посредством ЛР наименее наполненные профили, присутствующие в базе, как раз имели 5 наблюдений.

Теоретически CHAID должен был показать нам всю структуру взаимосвязей между значениями предикторов и зависимой переменной. Действительно, построенное «дерево» имеет 8 уровней (при 9 предикторах это почти максимум), т.е. дает нам значимые 8-мерные эффекты (хотя самые важные из них оказались меньшей размерности), которые, согласно сути CHAID'а, дифференцируют группы людей со статистически разными вероятностями политического активизма.

Рассмотрим качество этой модели. Логарифм отношения правдоподобия мы рассчитали вручную, он на 8% (что статистически значимо на 95% доверительном интервале) ниже аналогичной величины для логистической модели, но, к сожалению, все равно статистически значимо отличен от нуля. Процент правильных прогнозов интересующего нас положительного значения зависимой переменной в таблице классификации: 15% (Приложение А). Таким образом, модель CHAID статистически лучше логистической модели, но все еще далека от совершенства. При этом построение модели CHAID потребовало гораздо меньше времени, чем логистической. Обе модели были проверены на переобученность и показали хорошие результаты.

Перейдём к содержательным результатам (Приложение В). Поскольку в «дереве» около сотни узлов, мы рассматриваем только те 12 из них, которые описывают группы людей с вероятностью проявлять политический активизм выше 0,5. В таких узлах мы видим большое разнообразие сочетаний признаков. Во-первых, отметим, что наиболее сильно связанный с политическим активизмом предиктор – заинтересованность в политике: все 4 категории участвуют в комбинациях предикторов, объясняющих попадание объекта в категорию «Наличие проявлений активизма». Так, высокая сте-

пень заинтересованности в сочетании с высоким уровнем счастья и низким доверием власти дает высокую вероятность попадания в категорию политически активных (узел 14). Это единственная совпадающая с результатом ЛР комбинация предикторов. Примерно ту же вероятность попадания в категорию политически активных дают:

– комбинация умеренной заинтересованности в политике, принадлежность к мужскому полу, среднего доверия людям и низкого – власти, низкого дохода и низкой продолжительности (менее получаса в день) смотрения политических передач (узел 66).

– комбинация слабой заинтересованности в политике, низкого доверия власти, умеренного или высокого ощущения счастья, принадлежности к мужскому полу, высокого дохода, возраста старше 25 лет и средней продолжительности (0,5–1,0 часа в день) смотрения политических передач (узел 82);

– комбинация отсутствия заинтересованности в политике, умеренного или высокого доверия власти, умеренного ощущения счастья, отсутствия принадлежности к религиозным течениям, возраста младше 25 лет, среднего или высокого дохода (узел 85).

Такие комбинации навели нас на гипотезу о том, что они детерминируют содержательно разный политический активизм. Например, последняя, возможно, детерминирует инспирированный властью активизм, а первая – направленный против власти.

Можно ли было на располагаемых данных построить удовлетворительную модель, ведь, казалось бы CHAID рассмотрел предикторы почти всех уровней? Ответ положительный, но он лежит за пределами данной статьи. Модель с гораздо более высокой прогностической силой строится посредством очень трудозатратной комбинации методов линейного моделирования, предполагающих процедуру пошагового исключения предикторов. Из чего мы делаем вывод, что потенциал CHAID'а ограничивает именно «зашитый в него» принцип пошагового включения предикторов. Тем не менее считаем доказанным в контексте нашего эмпирического примера преимущество CHAID над ЛР.

Заключение

Данная работа была сфокусирована на проблеме включения категориальных предикторов в модели, оценивающие детерминанты определенного явления. В первую очередь, эта проблема проявляется в регрессионном моделировании, когда вместо изначальных номинальных и порядковых переменных, исследователь создает набор фиктивных переменных. Как мы показали, в работах многих зарубежных и российских социологов и статистиков фиктивные переменные существенно влияют на результаты применения метода: снижается относительная мощность выборки, упускаются из вида важные многомерные эффекты, вследствие чего меняется в худшую сторону качество модели. При этом альтернативные ЛР методы зачастую даже не рассматриваются.

Предпосылка данной работы заключалась в том, что о применении регрессии еще можно говорить и не бояться значительных искажений в тех случаях, когда категориальных переменных в модели не очень много. Когда же мы рассматриваем такой социальный феномен как политический активизм, почти все вопросы предполагают измерение с помощью категориальных шкал. Следовательно, для корректного изучения феномена мы должны принимать во внимание характер шкал и применять те методы, которые не ограничены линейными связями и не требуют измерения расстояний между объектами. Поэтому в качестве альтернативы ЛР был предложен CHAID, который, как и регрессионное моделирование, позволяет нам объяснить и даже предсказать определенное явление, но при этом работает с категориальными шкалами, не требуя их преобразования.

Мы не первые, кто счел ДК достойной альтернативой регрессионному моделированию, однако мы не встречали работ, включающих разностороннее систематизированное сравнение этих методов. Мы постарались внести свой вклад в такое сравнение, подложив критерии и априорно сравнив по ним рассматриваемые методы. Как и у других авторов, это сравнение оказалось скорее в пользу ДК.

Целью этого исследования, таким образом, было в приближенной к социологической практике ситуации сравнить результаты применения двух методов для выявления детерминант и выявить, действительно ли есть основания считать, что результаты применения регрессии не полностью отражают реальность. Аналитическая работа была проведена на базе ESS 2012 г., зависимой переменной был выбран политический активизм.

Изначальная гипотеза, заключающаяся в том, что ЛР обнаружит меньше значимых детерминант, чем ДК, нашла свое эмпирическое подтверждение. Как мы смогли показать, CHAID действительно находит многомерные эффекты при меньших затратах памяти машины и усилий исследователя (нет необходимости преобразовывать категориальные переменные, утяжеляя модель, и искать, какие эффекты взаимодействия надо учесть – алгоритм делает все самостоятельно). Следовательно, и теоретически, и на практике CHAID дал лучший результат.

Мы не можем считать, что обе сконструированные модели в полной мере отражают реальную ситуацию, так как оба метода имеют свои ограничения. Например, ДК не могут быть построены с применением процедуры пошагового исключения предикторов, которая кажется нам более продуктивной для поиска детерминант, но в случае логистической регрессии, которая может применить эту процедуру, предъявляются еще более высокие требования к количеству предикторов и относительной мощности выборки. Еще одна существенная проблема, с которой нам пришлось столкнуться, – высокая чувствительность обоих методов к форме зависимой переменной: неравномерность распределения по категориям приводит к сложности поиска сочетаний, которые отражают вероятность появления менее наполненного признака. Кроме того, мы проверили нашу гипотезу и методологические предпосылки только на одной базе и на одном социальном явлении, а также на не очень большой выборке. По этим причинам, чтобы быть полностью уверенными в том, что логистическая регрессия

и альтернативные методы дают разные результаты, требуются дальнейшие исследования, посвященные другим социальным феноменам, имеющие другую структуру переменных и более однородные данные. Также следовало бы проверить эти методы на более обширных выборках, чтобы отмести самую возможность, что какие-то искажения в результатах или качестве модели связаны с недостаточностью выборки.

Ближайшими направлениями углубления исследования мы видим доработку и описание альтернатив простому, но неэффективному принципу модального прогноза, заложенному в логистической регрессии. Перспективным направлениям исследования мы видим разработку на основе перечисленных методов нового методного комплекса, гибко учитывающего любые ограничения и особенности сырых данных и продуцирующего на основе этих данных высокоточные прогностические модели. В содержательном аспекте внедрение планируемого методного комплекса позволило бы, как мы уже упоминали, охватить не только непосредственные, но и косвенные предикторы любого явления.

ЛИТЕРАТУРА

1. *Толстова Ю. Н.* Анализ социологических данных. М.: Научный мир, 2000.

2. *Agresti A., Finlay B.* Statistical Methods for the Social Sciences. Pearson/ Prentice Hall: New Jersey, 2009.

3. *Толстова Ю.Н.* Измерение в социологии: курс лекций. М.: ИНФРА-М, 1998.

4. *Bollen K.A., Barb K.H.* Pearson's r and coarsely categorized measures // American Sociological Review. 1981. P. 232–239.

5. *Hawkes R.K.* Effects of Grouping on Measures of Ordinal Association // Sociological Methodology. 1976. Vol. 7. P. 176–194.

6. *O'Brien R.M.* The Use of Pearson's with Ordinal Data // American Sociological Review. 1979. P. 851–857.

7. *Ротмистров А.Н., Толстова Ю.Н.* Проблемы построения нелинейных регрессионных моделей в социологии: номинальные шкалы, синергетические эффекты, поиск эффективной системы предикторов // Математическое моделирование социальных процессов. 2014. № 16. С. 159–178.

8. *Попова П.А., Ротмистров А.Н.* Регрессия с категориальными предикторами: критика применения фиктивных переменных и логлинейный анализ как альтернативный подход // Социологический журнал. 2016. №3. С. 8–31

9. *Серая О.В., Дёмин Д.А.* Оценивание параметров уравнения регрессии в условиях малой выборки // Восточно-Европейский журнал передовых технологий. 2009. Т. 6. № 4(42).

10. *Agresti A.* An Introduction to Categorical Data Analysis. Willey, Hoboken, 2007. Ch.5.5.

11. *Толстова Ю.Н., Шишко И.О.* Использование качественного сравнительного анализа для поиска эффективной системы предикторов в логистической регрессии // Математическое моделирование и информатика социальных процессов: сб. трудов. Вып. 18. М.: Экономинформ, МГУ им. Ломоносова, ф-т выч. математики и кибернетики, 2016. С. 222–242.

12. *Loh W.* Classification and Regression Tree Methods // Encyclopedia of Statistics in Quality and Reliability / Ed. F. Ruggeri, R. Kenett, F. Faltin. Wiley, 2008. P. 315–323.

13. *Ritschard G.* CHAID and Earlier Supervised Tree Methods // Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences / Ed. J. McArdle, G, Ritschard. London: Routledge, 2013. P. 48–74.

14. *Rokach L., Maimon O.* Decision Trees // Data Mining and Knowledge Discovery Handbook. Springer. Fovea. La segmentation, 2010.

15. *Holgerssona H., Nordströma L., Öner Ö., Bollen K., Stine R.* Dummy Variables vs. Category-wise Models // Journal of Applied Statistics. 2014. Vol. 41. No. 2. P. 233–241.

16. *Horner S.B., Fireman G.D., Wang E. W.* The Relation of Student Behavior, Peer Status, Race, and Gender to Decisions about School Discipline Using CHAID Decision Trees and Regression Modeling // Journal of School Psychology 2010. Vol. 48. No. 2. P. 135–161.

17. *Liu Y.Y., Yang M., Ramsay M., Li X. S., Coid J. W.* A Comparison of Logistic Regression, Classification and Regression Tree, and Neural Networks Models in Predicting Violent Re-Offending // Journal of Quantitative Criminology. 2011. Vol. 27. No. 4. P. 547–573.

18. *European Social Survey*. URL: http://www.europeansocialsurvey.org/.

Приложение А

ТАБЛИЦЫ КЛАССИФИКАЦИИ ДЛЯ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ И ДЛЯ CHAID

Наблюдаемые значения				Предсказанные значения		
				Политическая активность		Правильные предсказания, в %
				Нет	Да	
ЛР	Шаг 1	Политическая активность	Нет	2628	0	100,0
			Да	576	0	0,0
		Общий процент				82,0
	Шаг 12	Политическая активность	Нет	2601	27	99,0
			Да	519	57	9,9
		Общий процент				83,0
CHAID		Политическая активность	Нет	2537	91	97,0
			Да	492	84	15,0
		Общий процент				81,8

Приложение Б

ПЕРЕМЕННЫЕ В УРАВНЕНИИ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Характеристики респондента		B	Sig.	Exp(B)
Шаг 12	Высокое доверие власти*высокая заинтересованность в политике*мужчины	2,618	0,002	13,712
	Низкое доверие власти*высокая заинтересованность в политике*высокий уровень счастья	3,337	0,000	28,145
	Низкое доверие людям*высокий доход*возраст младше 25 лет	2,158	0,000	8,655
	Среднее доверие людям*низкий доход*средний возраст	1,025	0,000	2,787
	Высокий доход*возраст младше 25 лет*женщины	–1,871	0,000	0,154
	Высокий доход*пенсионный возраст*средний уровень счастья	1,212	0,000	3,359
	Слабая заинтересованность в политике*средняя продолжительность ТВ-смотрения по политическим темам*возраст младше 25 лет	2,722	0,000	15,210
	Высокий доход*слабая заинтересованность в политике*высокий уровень счастья	1,276	0,000	3,581
	Высокий доход*смотрение политических ТВ-передач менее получаса в день*отсутствие принадлежности к религиозным течениям	1,527	0,000	4,604
	Умеренная заинтересованность в политике	0,942	0,000	2,565
	Среднее доверие властям*высокая заинтересованность в политике	1,981	0,000	7,250
	Отсутствие заинтересованности в политике*средний возраст	–1,273	0,000	0,280
	Константа	–2,198	0,000	0,111

Приложение В

ЗНАЧИМЫЕ КОМБИНАЦИЯ КАТЕГОРИЙ ПРЕДИКТОРОВ В CHAID, ПРЕДСКАЗЫВАЮЩИЕ НАЛИЧИЕ ПОЛИТИЧЕСКОЙ АКТИВНОСТИ

Узел	Да		Характеристики респондента
	N	Доля, в %	
60	13	100,0	Доход ниже среднего, слабая степень счастья, возраст меньше или равен 25 годам, смотрение политических ТВ-передач от получаса до часа в день, слабая заинтересованность в политике
82	7	100,0	Низкое доверие властям, умеренное ощущение счастья, принадлежность к мужскому полу, высокий доход, возраст старше 25 лет, смотрение политических ТВ-передач от получаса до часа, слабая заинтересованность в политике
14	8	80,0	Высокая степень заинтересованности, высокий уровень счастья, низкое доверие власти
80	4	80,0	Смотрение политических ТВ-передач менее получаса, высокая степень счастья, отсутствие принадлежности к религиозным течениям, возраст младше или равен 25 годам, принадлежность к женскому полу, слабая заинтересованность в политике
38	7	77,8	Высокая степень счастья, возраст младше или равен 25 годам, смотрение политических ТВ-передач от получаса до часа в день, слабая заинтересованность в политике
36	13	68,4	Слабая степень счастья, возраст меньше или равен 25 годам, смотрение политических ТВ-передач от получаса до часа в день, слабая заинтересованность в политике

Окончание приложения В

Узел	Да		Характеристики респондента
	N	Доля, в %	
66	4	66,7	Умеренная заинтересованность в политике, принадлежность к мужскому полу, среднее доверие людям и низкое – власти, низкий доход и низкая продолжительность (менее получаса в день) смотрения политических передач
87	4	66,7	Отсутствие принадлежности к религиозным течениям, умеренное и высокое доверие властям, умеренная и высокая степень счастья, принадлежность к мужскому полу, высокий доход, возраст старше 25 лет, смотрение политических ТВ-передач от получаса до часа в день, слабая заинтересованность в политике
77	11	61,1	Высокая степень счастья, отсутствие принадлежности к религиозным течениям, возраст младше или равен 25 годам, принадлежность к женскому полу, смотрение политических ТВ-передач больше часа в день, слабая заинтересованность в политике
50	19	57,6	Смотрение политических ТВ-передач более получаса в день, низкий доход, умеренное доверие людям, принадлежность к мужскому полу, умеренная заинтересованность в политике
85	5	55,6	Отсутствие заинтересованности в политике, умеренное или высокое доверие власти, умеренное ощущения счастья, отсутствие принадлежности к религиозным конфессиям, возраст младше 25 лет, средний или высокий доход
22	22	55,0	Возраст меньше 25 лет, смотрение политических ТВ-передач от получаса до часа в день, слабая заинтересованность в политике

Popova Polina
National Research University Higher School of Economics (NRU HSE), Moscow, papopova@hse.ru

Rotmistrov Aleksei
National Research University Higher School of Economics (NRU HSE), Moscow, alexey.n.rotmistrov@gmail.com

Logistic regression using categorical predictors and interaction effects and CHAID: a comparative analysis based on an empirical example

This article focuses on the methodological aspect of identifying determinants of political activism. More specifically we discuss the ways of including categorical predictors into the model explaining the level of activism. When using regression one may transform such predictors into dummy variables. This popular solution makes the model bulky and causes troubles with assessing this model's quality. Moreover, if a researcher wants to consider interaction effects of the mentioned predictors, the supernumerary combinations of the mentioned predictors values are obtained because regression modeling does not take into account the degree of similarity of the mentioned predictors values' effects. Authors propose using CHAID as an alternative to the mentioned solution. We compare these two methods based on their a priori known properties. We argue that CHAID has some theoretical advantages compared to logistic regression. It the empirical part we implement two methods and compare empirical results. The raw data were extracted from ESS 2012. The dependent variable was political activism and the hypothetical predictors belonged to the socio-economic part of the questionnaire.

Key words: determinants of active political participation, categorical predictors, logistic regression, classification tree method, dummy variables, interaction effects

References

1. Tolstova Yu. N. *Analysis of sociological data* (in Russian). M.: Nauchnyj mir, 2000.
2. Agresti A., Finlay B. *Statistical Methods for the Social Sciences*. Pearson/ Prentice Hall: New Jersey, 2009.
3. Tolstova Yu.N. *Measurement in sociology* (in Russian). M.: INFRA-M, 1998.
4. Bollen K.A., Barb K.H. "Pearson's r and coarsely categorized measures", *American Sociological Review*, 1981, 232–239.

5. Hawkes R.K. “Effects of Grouping on Measures of Ordinal Association”, *Sociological Methodology*, 1976, 7, 176–194.
6. O’Brien R.M. “The use of Pearson’s with ordinal data”, *American Sociological Review*, 1979, 851-857.
7. Rotmistrov A. N., Tolstova Yu. N. “Problems of constructing nonlinear regression models in sociology: nominal scales, synergetic effects, search for an effective predictor system” (in Russian), in: *Mathematical modeling of social processes*, 2014, 16, 159–178.
8. Popova P. A., Rotmistrov A. N. “Regression with categorical predictors: criticism of dummy variables and log-linear analysis as an alternative approach” (in Russian), *Sotsiologicheskij Zhurnal (Sociological Journal)*, 2016, 3, 8–31
9. Seraya O.V., Djomin D.A. “Estimation of parameters of the regression equation in case of a small sample”, *Vostochno-Evropejskij zhurnal peredovyh tehnologij (Eastern European Journal of Advanced Technologies),* 2009, 6 (4).
10. Agresti A. *An introduction to categorical data analysis*. Willey, Hoboken, 2007. Ch.5.5.
11. Tolstova Yu.N., Shishko I.O. “Using qualitative comparative analysis to find an effective predictor system in a logistic regression” (in Russian), in: *Matematicheskoe modelirovanie i informatika social’nyh processov (Mathematical modeling and informatics of social processes)*. Issue 18. M.: MGU, 2016. P. 222–242.
12. Loh W. “Classification and regression tree methods”, Ruggeri F., Kenett R., Faltin F. (eds.) *Encyclopedia of Statistics in Quality and Reliability*. Wiley, 2008. P. 315–323.
13. Ritschard G. “CHAID and earlier supervised tree methods”, in: McArdle J., Ritschard G. (eds.) *Contemporary Issues in Exploratory Data Mining in the Behavioral Science*. London: Routledge, 2013. P. 48–74.
14. Rokach L., Maimon O. “Decision trees”, in: *Data Mining and Knowledge Discovery Handbook*. Springer. Fovea. La segmentation, 2010.
15. Holgerssona H., Nordströma L., Öner Ö., Bollen K., Stine R. “Dummy variables vs. category-wise models”, *Journal of Applied Statistics*, 2014, 41(2). P. 233–241.
16. Horner S. B., Fireman G. D., Wang E. W. “The relation of student behavior, peer status, race, and gender to decisions about school discipline using CHAID decision trees and regression modeling”, *Journal of School Psychology*, 2010, 48(2), 135–161.
17. Liu Y. Y., Yang M., Ramsay M., Li X. S., Coid J. W. “A Comparison of Logistic Regression, Classification and Regression Tree, and Neural Networks Models in Predicting Violent Re-Offending”, *Journal of Quantitative Criminology*, 2011, 27(4), 547–573.
18. European Social Survey. URL: http://www.europeansocialsurvey.org/.