ТЕОРИЯ И МЕТОДОЛОГИЯ

П.А. ПОПОВА, А.Н. РОТМИСТРОВ

РЕГРЕССИЯ С КАТЕГОРИАЛЬНЫМИ ПРЕДИКТОРАМИ: КРИТИКА ПРИМЕНЕНИЯ ФИКТИВНЫХ ПЕРЕМЕННЫХ И ЛОГЛИНЕЙНЫЙ АНАЛИЗ КАК АЛЬТЕРНАТИВНЫЙ ПОДХОД

Аннотация. Статья посвящена методологии выявления категориальных (номинальных и ранговых) предикторов. Применение логистической регрессии предполагает добавление фиктивных переменных, что утяжеляет модель и затрудняет оценку её качества. Авторы предлагают альтернативный регрессии метод поиска детерминант — логлинейный анализ. На данных «Электоральной панели 2011–2012» Всероссийского центра изучения общественного мнения (ВЦИОМ) сравниваются результаты двух указанных методов и качество полученных моделей; зависимой переменной выступил протестный потенциал, а набор гипотетических детерминант был составлен из переменных социально-экономического блока панели. В итоге получено подтверждение, что логлинейный анализ может применяться как метод, альтернативный регрессионному анализу, и может давать результаты, превосходящие результаты регрессионного анализа.

Ключевые слова: детерминанты протестного потенциала; категориальные предикторы; логистическая регрессия; логлинейный анализ.

Для цитирования: *Попова П.А., Ротмистров А.Н.* Регрессия с категориальными предикторами: критика применения фиктивных переменных и логлинейный анализ как альтернативный подход // Социологический журнал. 2016. Том 22. № 3. С. 8–31. DOI: 10.19181/socjour.2016.22.3.4583

Попова Полина Артемовна — магистрант департамента социологии НИУ ВШЭ, менеджер Лаборатории экономико-социологических исследований НИУ ВШЭ. **Адрес:** 101000, Москва, ул. Мясницкая, д. 9/11, оф. 530, НИУ ВШЭ. **Телефон:** 8 (916) 906-43-45.
Электронная почта: papopova@hse.ru

Ротмистров Алексей Николаевич — кандидат социологических наук, доцент кафедры методов сбора и анализа социологической информации НИУ ВШЭ. **Адрес:** 101000, Москва, ул. Мясницкая, д. 9/11, оф. 530, НИУ ВШЭ. **Телефон:** +7 (926) 148-54-47.
Электронная почта: alexey.n.rotmistrov@gmail.com

Введение и постановка проблемы

Один из основных классов задач в социологии — это поиск связей между признаками. Как отмечает Ю.Н. Толстова, анализ связей ценен для социолога не сам по себе: социолог посредством него пытается проверить наличие причинно-следственных связей, которые невозможно выявить вопросами «в лоб». Для выявления причинно-следственных связей предназначены эксперимент и регрессионное моделирование. Эксперимент труднее организовать, но он позволяет доподлинно установить, что является причиной, а что следствием. Регрессионное моделирование несравненно проще эксперимента в организационном плане, поскольку не предъявляет специальных требований к дизайну сбора данных (кроме технического требования размера выборки, достаточного для работы с желаемым числом независимых переменных); однако в выявляемой связи оно, к сожалению, не позволяет доподлинно установить, что является причиной, а что следствием [15, с. 164].

В статье речь идет о логистической регрессии (далее — ЛР), которая предсказывает вероятности появления зависимой переменной при определённых независимых признаках (предикторах), выраженных в категориальных переменных. Действительно, так как регрессия работает с интервальными предикторами, включение категориальных предикторов требует создания фиктивных переменных, что сопряжено с рядом «подводных камней» [13], о которых социолог может не подозревать, но которые влияют на качество итоговой модели. В статье разбираются ошибки применения регрессии к категориальным предикторам и возможные пути их устранения.

Основные ошибки построения регрессии с категориальными предикторами

Первая ошибка тривиальна; она состоит в том, что авторы, оценивая качество полученной модели, отсеивают незначимые предикторы, но не перестраивают модель заново, а просто отбирают в итоговое уравнение переменные со значимыми коэффициентами (давать ссылки в данном случае считаем неэтичным, но такие работы есть). При таком подходе существует риск упустить из виду присутствие в системе предикторов, для которой была рассчитана модель, многомерных связей и взаимодействий. Поскольку последние нельзя разложить на ряд парных связей [15, с. 228–234], проверка предикторов на мультиколлинеарность может их не выявить. Бывает, что в многомерных связях предикторы и зависимая переменная участвуют не как целые переменные, а лишь отдельными своими значениями (что особенно актуально для категориальных предикторов); такую ситуацию мы называем *взаимодействием*. При наличии в системе предикторов многомерных связей и взаимодействий отобранные значимые в первоначальной модели предикторы могут потерять свою значимость, будучи лишенными контекста исключенных незначимых предикторов; про-

верить контекстуальную (не)значимость исключенных предикторов можно, только перестроив новую модель. Практика показывает, что в перестроенной модели может измениться все: ее качество, характер зависимости между предикторами и зависимой переменной, незначимые ранее предикторы могут стать значимыми, и наоборот [13, с. 167].

Как видно, отбор предикторов для итоговой регрессионной модели — не механическая задача. Для ее решения нужно применить сочетание методов и алгоритмов (1) поиска многомерных связей и взаимодействий между предикторами и зависимой переменной, (2) автоматического пошагового добавления в модель и исключения из нее предикторов с пересчетом оценок качества модели на каждом шаге. В математической статистике, анализе данных и статпакетах есть подходящие для решения этих задач методы и алгоритмы, но крайне мало работ, посвященных осмыслению опыта их применения именно в регрессионном моделировании. Так, при поиске “predictors selection in regression” в категориях социальных наук в базе данных Scopus можно обнаружить всего 27 релевантных методологических работ. Причем основные заслуги в этой области принадлежат математикам, чьих работ социологи часто сторонятся ввиду сложности и глубины подачи материала. Ниже говорится о методе, который может успешно дополнять регрессионное моделирование в поиске многомерных связей и взаимодействий между предикторами и зависимой переменной. Отдельно следовало бы провести сравнительный анализ стандартных (для статпакетов) алгоритмов автоматического пошагового добавления в модель предикторов и их исключения из нее, однако в силу обширности эта тема требует отдельной статьи.

Вторая ошибка возникает в результате известной процедуры подготовки [15; 19] категориальных предикторов для использования в регрессии (которая исходно предназначена для работы с интервальными предикторами). Действительно, нельзя помещать категориальные предикторы в регрессию без специальной подготовки данных, в противном случае параметры регрессии будет невозможно проинтерпретировать [24] (например, непонятно, как интерпретировать для номинальных недихотомических предикторов заложенные во многих видах регрессии линейные связи). Процедура подготовки состоит в разложении каждого категориального предиктора на ряд дихотомических (называемых «фиктивными», “dummy”) переменных, число которых всегда получается на единицу меньше числа значений (категорий) исходной переменной. Процедура разложения на фиктивные переменные обосновывается тем, что дихотомические переменные можно считать частным случаем интервальных, а математические статистики для дихотомической шкалы достаточно легко интерпретируются [15, с. 306–309]. Однако в последнее время появляются методологические аргументы против этой процедуры:

- создание фиктивных переменных занижает величину парной связи и уменьшает статистическую мощность выборки для выявления истинного эффекта независимой переменной в генеральной совокупности [27, p. 265]. При этом чем больше преобразуется предикторов и их категорий, тем ниже шансы получить качественную модель — с высокой прогностической способностью и значимыми предикторами [27, p. 266–267];

- дихотомизация приводит к неверной оценке значимости регрессионных коэффициентов в логистической регрессии (на основе статистики Вальда [23]), что впоследствии приводит к неправильному выбору переменных для итогового уравнения и неверным результатам [27, p. 270–271]. Дело в том, что при больших величинах регрессионного коэффициента стандартная ошибка становится столь большой, что статистика Вальда может быть недооценена. Это обстоятельство увеличивает вероятность совершить ошибку второго рода: отказаться от предиктора, который в реальности вносит большой вклад в объяснение зависимой переменной.

Получается, что изменение структуры связей из-за искусственной дихотомизации многоместных категориальных переменных и неточность в оценке статистической значимости связей порождают риск неправильно подобрать модель и соответствующее ей регрессионное уравнение.

В исследованиях последних лет приводятся дополнительные аргументы против использования фиктивных переменных [21, p. 234]:

- тестирование гипотез для классических дамми-переменных справедливо лишь тогда, когда дисперсии категорий равны, что в реальном эмпирическом исследовании маловероятно;

- неинтервальные переменные должны быть приблизительно равномерно распределены, в противном случае выбор контрольной группы может повлиять на набор переменных в итоговом уравнении и, следовательно, на содержание результирующей модели;

- создание фиктивных переменных повышает риски мультиколлинеарности, поскольку в силу самой процедуры их создания между ними устанавливается сильная отрицательная связь.

Таким образом, ставится под сомнение правомерность применения ЛР для категориальных предикторов. Процедура создания дамми-переменных удобна и имеет определенные методологические основания, но и контраргументы выглядят довольно убедительно. Можно ли по категориальным предикторам строить прогнозы, не прибегая к созданию фиктивных переменных при регрессионном моделировании? Ниже предложим метод, менее популярный, чем регрессия, но способный решать те же задачи без издержек, свойственных ЛР.

Альтернативный метод поиска детерминант: логлинейный анализ

Логлинейный анализ (далее — ЛЛА) более всего подходит для сравнения с ЛР, поскольку он:

- логически допускает оперирование понятиями зависимой и независимой переменных;

- работает как с количественными, так и с категориальными шкалами, не требуя преобразовывать последние в фиктивные переменные;

- позволяет не только находить многомерные связи и взаимодействия, но и определять детерминанты явлений.

Идея, что регрессионный анализ плохо «справляется» с нелинейными связями между категориальными переменными и что, как следствие, имеет смысл обращаться к альтернативным регрессии методам, коренится в нашей статье о проблемах построения нелинейных регрессионных моделей в социологии [13]. ЛЛА применяется в российской социологии гораздо реже, чем ЛР [16, с. 162], что, вероятно, связано с более трудоемкой процедурой. ЛЛА был создан, прежде всего, для тех случаев, когда и зависимая, и независимые переменные являются номинальными [25, p. 341]. Есть разные взгляды на место ЛЛА в статистике: одни авторы полагают, что он близок логистическому моделированию [20], другие отмечают, что он ближе корреляционному анализу, так как не имеет функции предсказания какого-либо явления, а применяется для отыскания связей [19]. В действительности ЛЛА позволяет искать многомерные связи как (1) между равнозначными переменными, так и (2) между многими предикторами и одной зависимой переменной, поэтому он вполне сравним с ЛР.

ЛЛА является углубленным методом изучения таблиц сопряженности, где связь между двумя и более категориальными переменными анализируется с помощью нахождения натурального логарифма частот ячеек [22, p. 1]. Основное преимущество этого метода — возможность исследовать не только двумерные, но и многомерные связи и взаимодействия между переменными. Взаимодействия и связи интерпретируются посредством эффектов, которые представляют собой отношение шансов (как и в ЛР), а гипотезы о независимости признаков проверяются логарифмированием отношений правдоподобия на основе статистики хи-квадрат [17]. Предназначение логлинейной модели — описать все связи и взаимодействия между отобранными переменными при контроле остальных переменных [26].

Априорные критерии сравнения методов

Теперь последовательно рассмотрим важнейшие характеристики ЛР и ЛЛА, сгруппированные в 10 критериев сравнения.

1. *Главная функция метода*. Прямое предназначение ЛР — прогнозировать значения зависимой переменной; ЛЛА нацелен на поиск многомерных связей и взаимодействий между предикторами. На деле оба метода позволяют и находить связи между предикторами, и прогнозировать изменения зависимой переменной.

2. *Контрольный профиль*. Контрольным профилем мы называем одно или несколько сочетаний значений изучаемых признаков. В ЛР аналогом контрольного профиля является контрольная группа. Однако во многих методах, работающих с таблицами сопряженности, в том числе и в ЛЛА, принято профилем называть некоторую специально выбранную строку или столбец таблицы. ЛР с категориальными предикторами в этом отношении уподобляется ЛЛА, поэтому в контексте темы данной статьи считаем уместным и целесообразным для обоих методов применять термин «контрольный профиль».

В мультиномиальной ЛР контрольных профилей столько же, сколько сочетаний между собой образуют значения предикторов; каждый контрольный профиль включает сочетание значений предикторов с первым или последним (зависит от выбора опции) выбранным исследователем значением зависимой переменной. В бинарной ЛР контрольный профиль по умолчанию включает сочетание нулевых значений предикторов с нулевым значением зависимой переменной. В ЛЛА как таковом контрольный профиль включает сочетание последних (по кодировке) значений изучаемых признаков; скажем, если переменная «протестный потенциал» имеет четыре значения, закодированные натуральными числами от 1 до 4, а переменная «отсутствие уверенности в завтрашнем дне» имеет два значения, закодированные числами 0 и 1, то контрольный профиль автоматически составится сочетанием максимальных значений: «4» из первой переменной и «1» из второй переменной. В ЛЛА с назначенной зависимой переменной контрольный профиль включает сочетание последних (по кодировке) значений предикторов и последнего (по кодировке) значения зависимой переменной. В такой разновидности ЛЛА, как логит-регрессия, контрольный профиль включает все сочетания предикторов с первым или последним (зависит от выбора опции) значением зависимой переменной (как и в мультиномиальной ЛР).

Из сказанного следует два вывода. Во-первых, сравниваемые методы в общем случае чувствительны к выбору контрольной группы (контрольного профиля), что является их общим недостатком. При этом нам не известны какие-либо источники, в которых были бы строго прописаны правила выбора контрольной группы для ЛР и контрольного профиля для ЛЛА. Основываясь на собственных наблюдениях и поиске их теоретического обоснования (впрочем, пока не отраженных в публикациях), мы рекомендуем для ЛЛА выбирать контрольный профиль, максимально приближенный к средней геометрической по выборке. Именно такой профиль мы выбрали в качестве контрольного в нашем исследовании детерминант протестного потенциала (см. табл. 1)[1]; на него же мы ориентировались (для сравнимости результатов), выбирая контрольную группу для ЛР.

---

[1] Переменные взяты из проекта: «Электоральная панель 2011–2012» // ВЦИОМ [электронный ресурс]. Дата обращения 15.08.2016. URL: <http://politpanel.wciom.ru/>.

Таблица 1

Сочетание значений изучаемых переменных, выбранных в качестве контрольного профиля для ЛЛА

Переменная	Значение
Ступень на лестнице материального достатка (wealth_ladder): На какой из десяти ступенек «лестницы материального достатка» Вы поместили бы себя, если самых обеспеченных принять за десятую ступеньку, а наименее обеспеченных — за первую?	1
Ступень на лестнице положения в обществе (soc_ladder): На какой из ступенек «лестницы положения в обществе» Вы поместили бы себя, если имеющих самое высокое положение принять за десятую ступеньку, а самое низкое — за первую?	1
Уверенность в завтрашнем дне (conf_tom): Чувствуете ли Вы уверенность в завтрашнем дне?	безусловно, не чувствую
Оценка материального положения (wealth): Как бы Вы оценили в настоящее время материальное положение Вашей семьи?	очень плохое
Самоотнесение к социальной группе (soc_group): На Ваш взгляд, к какой социальной группе Вы относитесь?	верхний средний класс и верхний класс
Удовлетворенность жизнью (life_satisf): Если говорить в целом, то в какой мере Вас устраивает сейчас жизнь, которую Вы ведете?	совершенно не устраивает
Протестный потенциал, 8 волна, группированная переменная_1	выраженный потенциал

Во-вторых, ЛЛА позволяет дифференцировать вероятности профилей, включающих значение зависимой переменной, принадлежащее и к контрольному профилю, чего не позволяет ЛР. То есть ЛЛА благодаря иной организации контрольного профиля дает более детальную картину прогнозируемых вероятностей, чем ЛР[2]. Чтобы получить такую же детализацию в ЛР, необходимо построить множество бинарных ЛР как на всей выборке, так и на подвыборках, заданных каждой парой значений зависимой переменной.

3. *Требования к данным*. ЛР требует равномерности распределения зависимой переменной, поскольку прогноз в ней модальный, то есть основан на модальном (по ожидаемой частоте) значении зависимой переменной внутри каждого сочетания предикторов. Следовательно, даже если ожидаемые и эмпирические частоты полностью совпадут, прогнозироваться будет наиболее часто встречающееся значение зависимой переменной внутри каждого сочетания предикторов. Следовательно, чем дальше распределения зависимой переменной от

[2] Эта тема заслуживает отдельной статьи, пока же см. иллюстрацию: Видео авторов данной статьи [электронный ресурс]. Дата обращения 15.08.2016. URL: <https://youtu.be/H4l8I151UmI>.

равномерности, тем чаще, при прочих равных условиях, будет прогнозироваться преобладающее значение. Другими словами, исходный перекос в пользу одного из значений зависимой переменной может «перебить» выявленные моделью закономерности (так же отдельного рассмотрения заслуживают методы работы с отклоняющимися от равномерности зависимыми переменными — регрессия с фильтром, пропорциональный прогноз и проч.). Поскольку в ЛЛА не интегрирован модальный прогноз, равномерности в распределении зависимой переменной не требуется.

Трудности в анализе могут возникнуть из-за того, что ЛЛА обсчитывает характеристики многомерной таблицы сопряженности, что требует большой оперативной памяти и большой тактовой частоты процессора ПК, поэтому обычно метод принудительно ограничивает число изучаемых признаков (в SPSS не более 10). Впрочем, и при попытке применять все опции мультиномиальной ЛР (в частности пошаговые процедуры) к большому числу изучаемых признаков пользователь, скорее всего, столкнется с остановкой процесса обработки из-за нехватки памяти.

Применение обоих методов требует адекватного объема выборки, который рассчитывается с заданной точностью и надежностью по наименее наполненным ячейкам пересечений. Зависимость наблюдений и мультиколлинеарность предикторов порождают некоторые проблемы при интерпретации результатов в ЛР, тогда как ЛЛА от подобных проблем свободен.

4. *Наличие зависимой переменной.* ЛР обязательно предполагает наличие зависимой переменной; в ЛЛА наличие зависимой переменной необязательно, но при необходимости ею можно назначить как целую переменную (логит-регрессия), так и отдельные ее значения.

5. *Включение категориальных предикторов.* ЛЛА мы выбрали как метод, который без преобразований работает с номинальными и порядковыми предикторами. ЛР работает с интервальными предикторами, а категориальные предикторы требуется преобразовывать в дихотомические.

6. *Учет рангов порядковых предикторов.* Преобразование категориальных предикторов в дихотомические переменные для ЛР приводит к потере информации о порядке между рангами, а ЛЛА этого ограничения лишен.

7. *Оценка качества модели.* В обоих методах используются одни и те же показатели качества получаемых моделей, но процедуры их расчетов технически разные, поэтому сравнение требует преобразований. Наш опыт таких преобразований мы опишем в другой статье, здесь же коснемся лишь необходимых для раскрытия темы данной статьи положений.

Качественной считается модель, которая позволяет генерировать данные, статистически не отличающиеся от эмпирических (по крайней мере, на данный момент это считается консенсусом по поводу всех

моделей, претендующих на обобщение на генеральную совокупность [8, с. 26; 18, p. 87]). И в логистической, и в логлинейной моделях мерой качества служит логарифм отношения правдоподобия, посредством которого производится статистическая оценка суммарного отклонения от ожидаемой частоты каждого сочетания значений изучаемых признаков частоты эмпирической. Однако напрямую логарифмы отношения правдоподобия в ЛР и ЛЛА на одних и тех же данных не сравнимы, поскольку в ЛР виртуальную таблицу сопряженности формируют в том числе фиктивные переменные, тогда как в ЛЛА таблицу сопряженности формируют исходные переменные в неизменном виде; следовательно, логарифмы отношения правдоподобия из ЛР и ЛЛА в общем случае имеют разное число степеней свободы.

Поскольку ЛР нацелена на прогнозирование, причем через модальный прогноз, дополнительными оценками качества выступают таблица классификации (показывает процент правильных предсказаний для каждой категории зависимой переменной при том или ином "Cutoff"; как его выбирать — тема отдельной статьи) и псевдо-R^2.Как мы уже отмечали, неравномерность распределения зависимой переменной обрекает ЛР на производство неудовлетворительных по своим прогностическим способностям моделей, даже если с точки зрения логарифма отношения правдоподобия они качественны.

Наконец, логарифм отношения правдоподобия — статистический инструмент, а таблица классификации псевдо-R^2 — не статистическая.

8. *Механизм поиска связей и взаимодействий.* В логистической модели применяется статистическая (обычно посредством статистики Вальда) оценка значимости линейной связи между каждым предиктором и логарифмированной относительной частотой зависимой переменной. В логлинейной модели используется величина отклонения логарифма отношения правдоподобия от нуля при исключении по иерархическому принципу [22, p. 9] каждой связки переменных, а также статистическая (обычно посредством Z-статистики) оценка значимости логарифмированного отклонения частоты интересующего сочетания признаков от частоты контрольного профиля (о котором мы писали выше). Иными словами, в ЛР поиск связей производится в один этап: ищутся линейные связи между всеми парами предикторов, с одной стороны, и зависимой переменной — с другой; нет привязки к оценке отклонения теоретических частот (рассчитанных на основе регрессионного уравнения) от эмпирических. В ЛЛА механизм поиска связей отталкивается от того, насколько хорошо теоретические частоты (рассчитанных на основе логлинейного уравнения) подогнаны под эмпирические; ищутся все N-мерные связи, без которых эта подгонка окажется статистически неудовлетворительной.

9. *Виды искомых связей.* ЛР ищет только линейные связи, тогда как ЛЛА ищет сколь угодно многомерные пучки (взаимодействия) изучаемых признаков. Таким образом, ЛР ищет связи, являющиеся

по своему виду частным случаем связей, искомых ЛЛА (уместна аналогия с соотношением между коэффициентом корреляции Пирсона и критерием хи-квадрат Пирсона: второй фиксирует все связи, фиксируемые посредством первого, но не наоборот).

10. *Интерпретация результатов*. ЛР дает уравнение, коэффициенты которого интерпретируются либо как вероятности того или иного значения зависимой переменной, либо как отношения шансов. Отношения шансов мы получаем и в ЛЛА, но не только для двухмерного сочетания (тот или иной предиктор и зависимая переменная), но и n-мерного (n-1 предиктор плюс зависимая переменная).

Теперь посмотрим, как скажутся эти сходства и различия методов на сходствах и различиях полученных с их помощью моделей. Преследуя методологическую цель, из круга актуальных социологических тем мы для анализа выбрали область протестного поведения и протестного потенциала, в исследованиях которой в основном применяются категориальные шкалы.

Методология исследования

Содержательная часть работы выполнена на доступных базах данных. Знакомство с работами, где выявляются детерминанты протестного потенциала, показало, что их авторы чаще всего используют регрессионные модели и корреляционный анализ, а в некоторых случаях прибегают к факторному анализу; все эти методы применяются в основном к категориальным переменным (см, например: [5; 7; 11; 14]).

На основе тех же работ мы выделили шесть блоков детерминант протеста: социально-демографические, социально-экономические, политические, информационные и внешние детерминанты, а также установки к протесту как таковому. Далее следовало определиться с временной рамкой. По оценке некоторых авторов [5; 10], уровень протестной активности россиян в XXI веке не очень высок по сравнению с концом XX века: в 2001–2010 гг. ситуация в стране была довольно спокойной и скупой на акции протеста. Однако в 2011 г. сразу после парламентских выборов, прошедших 4 декабря, началась неожиданная волна протестных акций, ставшая «непредсказуемым и эйфорическим опытом как для непосредственных участников уличного движения, так и для исследователей» [2, с. 130]. Д. Громов [4] отмечает, что занимался лонгитюдным исследованием с конца 2005 г. и планировал закончить его в начале 2012-го, так как не ожидал никаких всплесков гражданской активности — уличные акции были малочисленными и локальными. Однако всплеск конца 2011 г. заставил его продолжить работу. Ряд авторов [1–4; 6; 7] отмечают, что изучения требовал не только сам внезапный рост общественного движения, но и его новые черты. Именно в 2011 г. впервые за долгое время доля тех, кто декларировал готовность участвовать в акциях протеста, превысила долю декларировавших неготовность к участию в протестах [11, с. 75]. По этим причинам мы выбрали для анализа период 2011–2013 гг.

По результатам сравнения баз данных нескольких исследовательских компаний мы остановились на электоральной панели ВЦИОМ 2011–2012[3], поскольку она:

- содержит большое число наблюдений, репрезентирующих население РФ;
- непосредственно ориентирована на изучение электорального поведения и гражданского активизма;
- соответствует интересующему нас периоду времени;
- содержит множество детерминант, фигурирующих в теории.

В качестве предикторов мы взяли переменные из социально-экономического блока, руководствуясь теоретической рамкой популярного в российской социологии депривационного подхода ([7; 9; 12] и др.) к определению причин протеста. Зависимой переменной выступил «протестный потенциал», преобразованный из десяти- в четырехбалльную шкалу. Предикторами выступили:

- ступень на лестнице материального достатка (wealth_ladder);
- ступень на лестнице положения в обществе (soc_ladder);
- уверенность в завтрашнем дне (conf_tom);
- оценка материального положения (wealth);
- самоотнесение к социальной группе (soc_group);
- удовлетворенность жизнью (life_satisf)[4].

Тем самым мы выдвинули содержательную гипотезу о том, что эти предикторы влияют на протестный потенциал граждан России.

Наша методологическая гипотеза состоит в том, что благодаря учету многомерных связей и взаимодействий, а также отсутствию необходимости создавать фиктивные переменные из категориальных предикторов ЛЛА даст лучший результат, чем ЛР: ЛЛА позволит учесть больше связей предикторов с зависимой переменной и даст более точную прогностическую модель.

Сравнение результатов ЛР и ЛЛА

Начнем с логистической регрессии. Мы проделали стандартные подготовительные процедуры: преобразовали категориальные

---

[3] Данные размещены в открытом доступе, см.: Проект «Электоральная панель 2011–2012» // ВЦИОМ [электронный ресурс]. Дата обращения 15.08.2016. URL: <http://politpanel.wciom.ru/>.

[4] Детали подготовки переменных к анализу содержатся в видео, специально посвященном данному исследованию, по ссылке: Видео авторов данной статьи [электронный ресурс]. Дата обращения 15.08.2016. URL: <https://youtu.be/H4l8I151UmI>.

предикторы в фиктивные переменные и получили 21 переменную[5]. В контрольную группу отнесли выраженный протестный потенциал, поскольку именно выраженный потенциал был включен нами в контрольный профиль в ЛЛА (согласно обозначенному выше принципу, см. табл. 1) для обеспечения сравнимости результатов с ЛР.

Ко всем предикторам применили процедуру пошагового отбора (Backward LR). Предполагалось включить в модель и двухмерные эффекты взаимодействия всех гипотетических предикторов, но алгоритм пошагового отбора не справился с задачей из-за недостатка оперативной памяти (о чем мы писали выше). Поэтому мы оставили модель только с главными эффектами. Рассмотрим оценки ее качества.

Логарифм отношения правдоподобия равен 0: он рассчитан на последнем шаге Backward, то есть характеризует таблицу сопряженности, сформированную только значимым предиктором и зависимой переменной (см. Приложение: А. Наблюдаемые и предсказанные по логистической модели частоты). Значим же в модели (на уровне 0,95) только предиктор «отсутствие уверенности в завтрашнем дне», влияние которого на протестный потенциал отражено в таблице 2.

Таблица 2

Детерминанты в логистической регрессии

Значение предиктора «уверенность в завтрашнем дне»	Результирующее значение	На сколько %
Отсутствие уверенности в завтрашнем дне	Шанс попасть в группу наименьшего протестного потенциала (категория 1) по сравнению с выраженным протестным потенциалом (категория 4) *выше*	82
Прочее (остальные три категории: безусловно и скорее чувствую, скорее не чувствую уверенности в завтрашнем дне)	Шанс попасть в группу наименьшего протестного потенциала (категория 1) по сравнению с выраженным протестным потенциалом (категория 4) *выше*	265
	Шанс попасть в группу умеренного протестного потенциала (категория 3) по сравнению с выраженным протестным потенциалом (категория 4) *ниже*	31

Вопреки хорошему логарифму отношения правдоподобия, как мы и предполагали в связи с неравномерностью распределения зависимой переменой, наша модель объясняет только наименьший протестный потенциал (самая наполненная категория) и не объясняет небольшой, умеренный и выраженный (см. Приложение: Б. Таблица классификации по логистической модели). О том же свидетельствуют величины псевдо-R^2 (Кокса и Снелла = 0,11, Нагелькерке = 0,12 и МакФаддена = 0,05). Все это говорит о том, что регрессионная модель плохо объясняет реальность. Допускаем, что многие поклонники ЛР на этом этапе сочли

[5] См. видео по указанному адресу: URL: <https:// youtu.be/H4l8I151UmI>.

бы задачу поиска детерминант нерешаемой, сославшись на проблемы с данными, с выборкой или чем-то еще. Мы же полагаем, что удовлетворительной прогнозной модели не получилось по описанным выше причинам: (1) необоснованно завышена оценка ошибки регрессионного коэффициента, поэтому статистика Вальда заставляет отказываться от включения в модель значимых (согласно результатам последующего ЛЛА) предикторов — «ступень на лестнице положения в обществе (soc_ladder)» и «удовлетворенность жизнью (life_satisf)»; (2) наличие четырех наборов фиктивных переменных существенно увеличивает число сочетаний переменных без прироста новой информации и искажает многомерные связи и взаимодействия внутри модели; (3) основанная на модальном прогнозе модель не справляется с неравномерностью распределения зависимой переменной; (4) модель не замечает многомерные связи (см. Приложение: B. Оценки регрессионных коэффициентов в ЛР). Чтобы доказать это, надо выполнить ЛЛА.

Полагаем, что отказ от модального принципа позволил бы в данном случае уточнить прогноз. Рисунок 1 иллюстрирует то, что выявленный ЛР предиктор действительно влияет на распределение вероятностей зависимой переменной, а расчеты подтверждают, что это влияние статистически значимо (то есть доли людей, чувствующих и не чувствующих уверенность в завтрашнем дне, внутри категорий выраженности протестного потенциала статистически различаются, о чем свидетельствуют и критерий хи-квадрат, и ЛЛА). Но это — тема отдельной статьи.

Обработаем этот же блок переменных с помощью ЛЛА (см. упомянутое выше видео по ссылке <http://www.rotmistrov.com/frcst>), который покажет всю структуру взаимосвязей между значениями переменных, а не только парные связи между каждым предиктором и зависимой переменной. Семь переменных — приемлемое число для ЛЛА. Как и логистическую регрессию, логлинейный анализ мы начали с процедуры пошагового отбора, в данном случае — значимых эффектов, причем ориентиром для остановки процедуры здесь также служит логарифм отношения правдоподобия. Но в ЛЛА происходит перебор всех возможных связей между переменными (от двух- до семимерных), а не только заданных пользователем (в нашем случае двухмерных), как в ЛР. Пошаговый отбор выдал модель с двух- и трехмерными эффектами (связями): во-первых, протестный потенциал респондента связан с его уверенностью в завтрашнем дне (как и в ЛР, см. табл. 3), во-вторых, протестный потенциал респондента связан одновременно с его самооценкой социального положения и удовлетворенности жизнью. Получается, что ЛР, включающая предикторы только в виде слагаемых (что вполне характерно для практики ее использования), чересчур упрощает картину. Более того, ЛЛА показал, что эти предикторы связаны двух- и трехмерными связями с остальными четырьмя гипотетическими предикторами, что позволяет при желании углублять исследование, прибегнув к путевому анализу, Structural Equation Modeling или аналогичным методам.

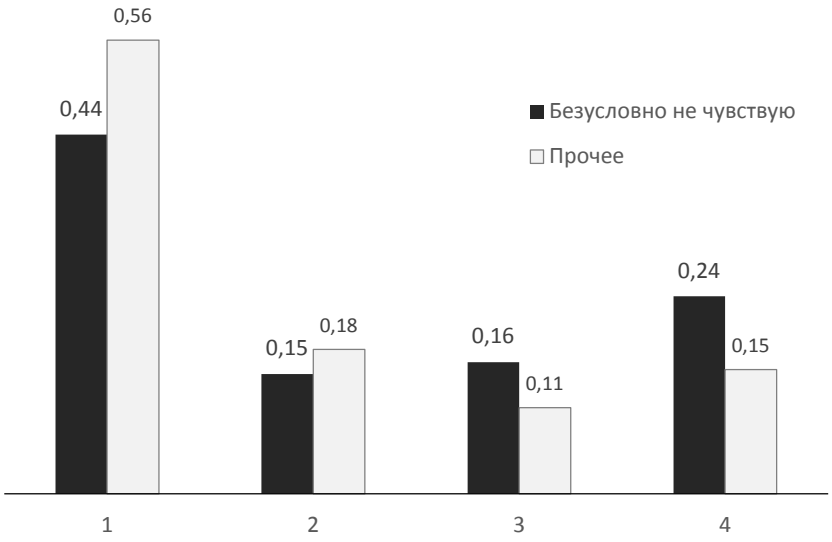

Рис. 1. Вероятность значений протестного потенциала (1 – наименьший, 2 – небольшой, 3 – умеренный, 4 – выраженный) в зависимости от значений предиктора «уверенность в завтрашнем дне (conf_tom)»

Таблица 3

Детерминанты протестного потенциала, выявленные…

… логистической регрессией	… логлинейным анализом
• отсутствие уверенности в завтрашнем дне	• отсутствие уверенности в завтрашнем дне
	• ступень на лестнице социального положения
	• удовлетворенность жизнью

Затем мы перешли от исходной семимерной таблицы (число предикторов плюс зависимая переменная) к одной двух- и одной трехмерной таблицам, формируемым выявленными двух- и трехмерными эффектами соответственно, чтобы выявить значимые микроэффекты взаимодействия. Причем мы воспользовались как ЛЛА с зависимой переменной, так и логит-регрессией. Они дали идентичные результаты. Значимыми в модели (на уровне 0,95) оказались микроэффекты, позволяющие прогнозировать три из четырех уровней протестного потенциала, что заметно лучше, чем результаты работы модального принципа, заложенного в ЛР. Подробнее значимые микроэффекты и их влияние на протестный потенциал отражены в таблице 4.

Таблица 4

Детерминанты в логлинейном анализе

Значение предиктора	Результирующее значение	На сколько %
Отсутствие уверенности в завтрашнем дне	Шанс попасть в группу небольшого протестного потенциала (категория 2) по сравнению с наименьшим протестным потенциалом (категория 1) *ниже*	66
	Шанс попасть в группу умеренного протестного потенциала (категория 3) по сравнению с наименьшим протестным потенциалом (категория 1) *ниже*	63
	Шанс попасть в группу выраженного протестного потенциала (категория 4) по сравнению с наименьшим протестным потенциалом (категория 1) *ниже*	45
Прочее (наличие уверенности в той или иной мере)	Шанс попасть в группу наименьшего протестного потенциала (категория 1) *выше*, чем при отсутствии уверенности в завтрашнем дне	805
Отсутствие уверенности в завтрашнем дне *vs* прочее	Шанс попасть в группу умеренного протестного потенциала (категория 3) при наличии уверенности в завтрашнем дне *ниже* по сравнению с вероятностью попадания в группу наименьшего протестного потенциала (категория 4) при отсутствии уверенности в завтрашнем дне	49
	Шанс попасть в группу выраженного протестного потенциала (категория 4) при наличии уверенности в завтрашнем дне *ниже* по сравнению с вероятностью попадания в группу наименьшего протестного потенциала (категория 4) при отсутствии уверенности в завтрашнем дне	50
Высшая ступень в социуме и средняя удовлетворенность жизнью	Шанс попасть в группу наименьшего протестного потенциала (категория 1) по сравнению с выраженным протестным потенциалом (категория 4) *выше*	229
Средняя ступень (3–5-я) в социуме и средняя удовлетворенность жизнью	Шанс попасть в группу выраженного протестного потенциала (категория 4) *выше*, чем занимая высшую ступень в социуме при той же удовлетворенности жизнью	286–486
Средняя ступень (4-я) в социуме и наивысшая удовлетворенность жизнью	Шанс попасть в группу выраженного протестного потенциала (категория 4) *ниже*, чем занимая высшую ступень в социуме и имея среднюю удовлетворенность жизнью	90
Низкая (2-я) ступень в социуме и низкая удовлетворенность жизнью vs высшая ступень в социуме и средняя удовлетворенность жизнью	Шанс попасть в группу умеренного протестного потенциала (категория 3), занимая низкую ступень (2) в социуме и имея низкую удовлетворенность жизнью, *выше*, чем вероятность попадания в группу выраженного протестного потенциала (категория 4), занимая высшую ступень в социуме и имея среднюю удовлетворенность жизнью	4514

Заключение

Из наших рассуждений мы делаем следующие основные выводы.

1. Согласно обоим методам степень протестного потенциала связана с уверенностью человека в завтрашнем дне и не связана непосредственно с социальной группой, к которой он себя относит, с его материальным положением и его субъективной оценкой материального положения. Согласно ЛЛА степень протестного потенциала связана напрямую не только с уверенностью человека в завтрашнем дне, но и с занимаемой им ступенью социальной лестницы и его удовлетворенностью жизнью, а также косвенно — через эти предикторы — с остальными предикторами.

2. Как и ожидалось, метод, работающий с многомерными связями и взаимодействиями, устойчивый к неравномерности распределения зависимой переменной и не требующий преобразовывать категориальные предикторы в дихотомические переменные, дал более насыщенные предикторами модели с более высокой прогностической способностью. ЛЛА объясняет и предсказывает попадание в три из четырех категорий протестного потенциала с помощью взаимодействия предикторов; в нем мы имеем больше оснований доверять объяснению некоторых уровней протестного потенциала, ориентируясь на его устойчивость к форме распределения зависимой переменной.

Таким образом, считаем содержательную гипотезу частично подтвержденной (в случае ЛР для одного предиктора из шести, в случае логлинейного анализа для трех предикторов из шести), а методологическую гипотезу — полностью подтвержденной на выбранных эмпирических данных. Генерализация полученного методологического вывода возможна на пути искусственной трансформации проанализированных нами данных и применения процедуры Bootstrap[6] или ее аналогов.

Изначальная гипотеза о том, что ЛР обнаружит меньше значимых детерминант, чем остальные методы, нашла свое подтверждение на практике. Как показано на социально-экономическом блоке независимых переменных, ЛР выявляет значимость только одной фиктивной переменной, в то время как ЛЛА обнаруживает больше предикторов.

При этом мы не можем считать, что обе сконструированные модели в полной мере отражают реальную ситуацию, так как оба метода имеют свои ограничения. 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. Нельзя не отметить и то, что зависимая переменная в нашем исследовании имела существенные смещения, что, как и ожидалось, повлияло на результаты. Чтобы убедиться в том, что ЛР и ЛЛА дают разные результаты,

---

[6] Bootsrap — метод, позволяющий многократно генерировать (псевдо)выборки на базе имеющейся выборки и проверять статистические показатели и вероятностные распределения.

необходимо провести дальнейшие исследования на других выборках и других социальных феноменах, имеющие другую структуру переменных и более однородные данные. Также следовало бы проверить эти методы на более обширных выборках, чтобы исключить возможность искажения результатов или снижения качества модели из-за недостаточности выборки.

Выбор между ЛР и ЛЛА следует делать в зависимости от типа шкал предикторов (формальная адекватность метода). При большом числе интервальных предикторов (и малом числе номинальных предикторов) мы имеем все основания использовать ЛР, не заботясь о возможном искажении данных. При большом числе категориальных переменных следует предпочесть ЛЛА.

В качестве направлений, которые могли бы углубить исследование, мы видим: (1) доработку и описание логического алгоритма, позволяющего сделать результаты логистической регрессии и логлинейного анализа более сопоставимыми; (2) доработку и описание альтернатив простому, но неэффективному принципу модального прогноза, заложенному в логистической регрессии; (3) сравнение логистической регрессии и логлинейного анализа с методами деревьев классификаций.

На наш взгляд, перспективна разработка на основе перечисленных методов нового методного комплекса, гибко учитывающего любые особенности исходных данных и продуцирующего на их основе высокоточные прогностические модели. В содержательном аспекте внедрение такого методного комплекса позволило бы охватывать не только непосредственные, но и косвенные предикторы.

ЛИТЕРАТУРА

1. *Баранова Г.В.* Методика анализа протестной активности населения России. 2012 [электронный ресурс]. Дата обращения 03.08.2016. URL: <http://www.isras.ru/files/File/Socis/2012_10/Baranova.pdf>.
2. *Бикбов А.* Методология исследования «внезапного» уличного активизма (российские митинги и уличные лагеря. Декабрь 2011 – июнь 2012) // Laboratorium. 2012. № 2. С. 130–163.
3. *Волков Д.* Протестные митинги в России конца 2011 – начала 2012 гг.: запрос на демократизацию политических институтов // Вестник общественного мнения. 2012. № 2 (112), апрель – июнь. С. 73–86.
4. *Громов Д.В.* «Мы не оппозиция, а народ»: новые черты уличного политического акционизма // Антропологический форум. 2011. № 16. С. 135–153.
5. *Дементьева И.Н.* Социально-экономические и общественно-политические аспекты формирования протестного потенциала в регионе // Мониторинг общественного мнения 2013. № 6 (118). С. 39–50.
6. *Зайцев Д.Г.* Массовый политический протест: проблемы концептуализации и методологии анализа // Политические изменения в глобальном мире: теоретико-методологические проблемы анализа и прогнозирова-

ния / Отв. ред. И.С. Семененко, В.В. Лапкин, В.И. Пантин. М.: ИМЭМО РАН, 2014. С. 124–142.
7. *Кинсбурский А.В.* Социальное недовольство и потенциал протеста // Социологические исследования. 1998. № 10. С. 92–95 [электронный ресурс]. Дата обращения 31.08.2016. URL: <http://ecsocman.hse.ru/data/204/881/1216/014.KINSBOURSKIY.pdf>.
8. *Крыштановский А.О.* Анализ социологических данных с помощью пакета SPSS / А.О. Крыштановский; ГУ–ВШЭ. М.: ИД ГУ–ВШЭ, 2006. — 281 с.
9. *Левада Ю.А.* Человек недовольный: протест и терпение // От мнений к пониманию: Социологические очерки, 1993–2000. М.: МШПИ, 2000. С. 467–488.
10. *Мамонов М.В.* Протестная активность россиян в 2011–2012 гг.: основные тренды и некоторые закономерности // Мониторинг общественного мнения. 2012. № 1 (107). С. 5–22.
11. *Мтиулишвили П.И.* Анализ влияния внешних и внутренних факторов на рост протестных настроений россиян // Мониторинг общественного мнения: Экономические и социальные переменные. 2011. № 4 (104). С. 75–82 [электронный ресурс]. Дата обращения 31.08.2016. URL: <http://wciom.ru/fileadmin/Monitoring/104/2011_104_9_Mtiulishvili.pdf>.
12. *Ольшанский Д.В.* Психология масс. М. [и др.]: Питер, 2002. — 368 с.
13. *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.
14. *Стребков Д.О.* Экономические детерминанты протестного поведения населения России // Экономическая социология. 2000. Т. 1. № 1. С. 48–66.
15. *Толстова Ю.Н.* Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. М.: Научный мир, 2000. — 352 с.
16. *Толстова Ю.Н., Рыжова А.В.* Анализ таблиц сопряженности: использование отношения преобладаний и логлинейных моделей // Социология: 4М. 2003. № 16. С. 150–164.
17. *Трошин Л., Балаш В., Балаш О.* Статистический анализ нечисловой информации. М.: Московский государственный университет экономики, статистики и информатики, 2003. — 67 с.
18. *Agresti A.* An introduction to categorical data analysis. New York: Wiley, 1996. — 400 p.
19. *Agresti A., Finlay B.* Statistical Methods for the Social Sciences. 4th ed. Upper Saddle River, NJ: Prentice Hall, 2009. — 609 p.
20. *Fu C.Y.* Combining loglinear model with classification and regression tree (CART): An application to birth data // Computational Statistics & Data Analysis. 2004. No. 45. P. 865–874. DOI: 10.1016/S0167-9473(03)00092-6
21. *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. DOI:10.1080/02664763.2013.838665

22. *Jeansonne A.* Loglinear models. 2002 [online]. Accessed 31.08.2016. URL: <http://userwww.sfsu.edu/efc/classes/biol710/loglinear/Log%20Linear%20Models.pdf>.
23. *Menard S.* Applied logistic regression analysis. Sage university paper series on quantitative applications in the social sciences, 07–106. 2nd ed. Thousand Oaks, CA: Sage, 1995. — 111 p.
24. *Starkweather J.D.* Categorical Variables in Regression: Implementation and Interpretation. 2010 [online]. Accessed 31.08.2016. URL: <http://www.unt.edu/rss/class/Jon/Benchmarks/CategoricalRegression_JDS_June2010.pdf>.
25. *Tansey R., White M., Long R., Smith M.* A Comparison of Loglinear Modeling and Logistic Regression in Management Research // Journal of Management. 1996. Vol. 22. No. 2. P. 339–358. DOI: 10.1177/014920639602200207
26. *Upton G.* The Exploratory Analysis of Survey Data Using Log-Linear Models // Journal of the Royal Statistical Society. Series D (The Statistician). 1991. Vol. 40. No. 2. Special Issue: Survey Design, Methodology and Analysis. P. 169–182.
27. *Vargha A., Rudas T., Delaney D., Maxwell S.* Dichotomization, Partial Correlation, and Conditional Independence // Journal of Educational and Behavioral Statistics. 1996. Vol. 21. No. 3, Autumn. P. 264–282. DOI: 10.2307/1165272

ПРИЛОЖЕНИЕ

А. Наблюдаемые и предсказанные по логистической модели частоты

Чувствуете ли Вы уверенность в завтрашнем дне?	Протестный потенциал	Частоты			%	
		наблюдаемые	предсказанные	остатки по Пирсону	наблюдаемые	предсказанные
Безусловно не чувствую	наименьший	60	60,0	0,0	0,44	0,44
	небольшой	20	20,0	0,0	0,15	0,15
	умеренный	22	22,0	0,0	0,16	0,16
	выраженный	33	33,0	0,0	0,24	0,24
Прочее	наименьший	547	547,0	0,0	0,56	0,56
	небольшой	174	174,0	0,0	0,18	0,18
	умеренный	104	104,0	0,0	0,11	0,11
	выраженный	150	150,0	0,0	0,15	0,15

Б. Таблица классификации по логистической модели

Наблюдаемый потенциал	Предсказанный потенциал				% корректных предсказаний
	наименьший	небольшой	умеренный	выраженный	
наименьший	587	0	0	0	100
небольшой	185	0	0	0	0
умеренный	121	0	0	0	0
выраженный	181	0	0	0	0
Общий процент	100	0	0	0	54

B. Оценки регрессионных коэффициентов в ЛР

Протестный потенциал, 8 волна, группированная переменная_1a		B	Std. Error	Wald	df	Sig.	Exp(B)
Наименьший потенциал	Intercept	2,452	1,617	2,298	1	0,130	
	wealth_ladder	0,065	0,078	0,706	1	0,401	1,068
	soc_ladder	0,002	0,068	0,001	1	0,978	1,002
	[conf_tom=1]	-0,678	0,260	6,777	1	0,009	0,508
	[conf_tom=2]	0b	.	.	0	.	.
	[wealth_1=0]	0,357	0,568	0,396	1	0,529	1,430
	[wealth_1=1]	0b	.	.	0	.	.
	[wealth_2=0]	0,112	0,509	0,048	1	0,826	1,118
	[wealth_2=1]	0b	.	.	0	.	.
	[wealth_3=0]	0,125	0,521	0,057	1	0,811	1,133
	[wealth_3=1]	0b	.	.	0	.	.
	[wealth_4=0]	0b	.	.	0	.	.
	[wealth_4=1]	0b	.	.	0	.	.
	[soc_group_1=0]	-0,639	0,507	1,593	1	0,207	0,528
	[soc_group_1=1]	0b	.	.	0	.	.
	[soc_group_2=0]	-0,849	0,438	3,760	1	0,052	0,428
	[soc_group_2=1]	0b	.	.	0	.	.
	[soc_group_3=0]	-0,489	0,416	1,382	1	0,240	0,613
	[soc_group_3=1]	0b	.	.	0	.	.
	[soc_group_4=0]	0b	.	.	0	.	.
	[soc_group_4=1]	0b	.	.	0	.	.
	[life_satisf_1=0]	-0,372	0,384	0,940	1	0,332	0,689
	[life_satisf_1=1]	0b	.	.	0	.	.
	[life_satisf_2=0]	-0,100	0,225	0,196	1	0,658	0,905
	[life_satisf_2=1]	0b	.	.	0	.	.
	[life_satisf_4=0]	0,197	0,241	0,670	1	0,413	1,218
	[life_satisf_4=1]	0b	.	.	0	.	.
	[life_satisf_5=0]	-0,285	0,413	0,476	1	0,490	0,752
	[life_satisf_5=1]	0b	.	.	0	.	.
	[life_satisf_6=0]	0b	.	.	0	.	.
	[life_satisf_6=1]	0b	.	.	0	.	.
Небольшой потенциал	Intercept	2,872	2,270	1,601	1	0,206	
	wealth_ladder	-0,010	0,095	0,012	1	0,914	0,990
	soc_ladder	0,047	0,083	0,318	1	0,573	1,048
	[conf_tom=1]	-0,455	0,327	1,935	1	0,164	0,635
	[conf_tom=2]	0b	.	.	0	.	.

	[wealth_1=0]	0,002	0,751	0,000	1	0,998	1,002
	[wealth_1=1]	0b	.	.	0	.	.
	[wealth_2=0]	-0,097	0,684	0,020	1	0,887	0,908
	[wealth_2=1]	0b	.	.	0	.	.
	[wealth_3=0]	-0,468	0,694	0,455	1	0,500	0,626
	[wealth_3=1]	0b	.	.	0	.	.
	[wealth_4=0]	0b	.	.	0	.	.
	[wealth_4=1]	0b	.	.	0	.	.
	[soc_group_1=0]	-1,139	0,783	2,118	1	0,146	0,320
	[soc_group_1=1]	0b	.	.	0	.	.
	[soc_group_2=0]	-1,623	0,705	5,293	1	0,021	0,197
	[soc_group_2=1]	0b	.	.	0	.	.
	[soc_group_3=0]	-1,385	0,686	4,079	1	0,043	0,250
	[soc_group_3=1]	0b	.	.	0	.	.
	[soc_group_4=0]	0b	.	.	0	.	.
	[soc_group_4=1]	0b	.	.	0	.	.
	[life_satisf_1=0]	-0,440	0,452	0,951	1	0,330	0,644
	[life_satisf_1=1]	0b	.	.	0	.	.
	[life_satisf_2=0]	-0,023	0,275	0,007	1	0,934	0,977
	[life_satisf_2=1]	0b	.	.	0	.	.
	[life_satisf_4=0]	0,329	0,300	1,206	1	0,272	1,390
	[life_satisf_4=1]	0b	.	.	0	.	.
	[life_satisf_5=0]	0,400	0,564	0,502	1	0,479	1,491
	[life_satisf_5=1]	0b	.	.	0	.	.
	[life_satisf_6=0]	0b	.	.	0	.	.
	[life_satisf_6=1]	0b	.	.	0	.	.
Умеренный потенциал	Intercept	0,526	2,350	0,050	1	0,823	
	wealth_ladder	0,211	0,106	3,962	1	0,047	1,235
	soc_ladder	-0,050	0,095	0,279	1	0,597	0,951
	[conf_tom=1]	0,076	0,338	0,051	1	0,821	1,079
	[conf_tom=2]	0b	.	.	0	.	.
	[wealth_1=0]	-0,091	0,819	0,012	1	0,912	0,913
	[wealth_1=1]	0b	.	.	0	.	.
	[wealth_2=0]	-0,066	0,748	0,008	1	0,929	0,936
	[wealth_2=1]	0b	.	.	0	.	.
	[wealth_3=0]	-0,132	0,767	0,030	1	0,863	0,876
	[wealth_3=1]	0b	.	.	0	.	.
	[wealth_4=0]	0b	.	.	0	.	.

[wealth_4=1]	0b	.	.	0	.	.
[soc_group_1=0]	-0,494	0,767	0,415	1	0,519	0,610
[soc_group_1=1]	0b	.	.	0	.	.
[soc_group_2=0]	-0,805	0,659	1,490	1	0,222	0,447
[soc_group_2=1]	0b	.	.	0	.	.
[soc_group_3=0]	-0,766	0,630	1,480	1	0,224	0,465
[soc_group_3=1]	0b	.	.	0	.	.
[soc_group_4=0]	0b	.	.	0	.	.
[soc_group_4=1]	0b	.	.	0	.	.
[life_satisf_1=0]	-0,072	0,520	0,019	1	0,889	0,930
[life_satisf_1=1]	0b	.	.	0	.	.
[life_satisf_2=0]	0,167	0,314	0,283	1	0,595	1,182
[life_satisf_2=1]	0b	.	.	0	.	.
[life_satisf_4=0]	0,021	0,326	0,004	1	0,949	1,021
[life_satisf_4=1]	0b	.	.	0	.	.
[life_satisf_5=0]	-0,120	0,574	0,043	1	0,835	0,887
[life_satisf_5=1]	0b	.	.	0	.	.
[life_satisf_6=0]	0b	.	.	0	.	.
[life_satisf_6=1]	0b	.	.	0	.	.

Дата поступления: 11.03.2016.

Sotsiologicheskiy Zhurnal = Sociological Journal
2016. Vol. 22. No. 3. P. 8–31. DOI: 10.19181/socjour.2016.22.3.4583

P.A. Popova, A.N. Rotmistrov
National Research University Higher School of Economics,
Moscow, Russian Federation.

Polina A. Popova — Graduate Student, National Research University Higher School of Economics, manager of Laboratory for Studies in Economic Sociology. **Address:** 9/11, room 530, Myasnitskaya str., 101000, Moscow, Russian Federation. **Phone:** +7 (916) 906-43-45. **Email:** papopova@hse.ru ORCID: 0000-0002-7667-148X

Aleksei N. Rotmistrov — Candidate of Sociological Sciences; Associate Professor, Chair of Methods of Sociological Data Collection and Analysis, National Research University Higher School of Economics. **Address:** 9/11, room 530, Myasnitskaya str., 101000, Moscow, Russian Federation. **Phone:** +7 (926) 148-54-47. **Email:** alexey.n.rotmistrov@gmail.com ORCID: 0000-0003-2386-8710

Regression with Categorical Predictors: Criticizing Dummy-Variable Usage and Log-linear Analysis as an Alternative Approach

Abstract. The focus of this article is the methodological aspect of identifying protest behavior determinants, specifically — categorical (nominal and ordinal) predictors

which can hypothetically explain the potency level of one protest or another. The use of regression for explaining protest potency levels implies transforming categorical predictors into dummy variables. Such a solution makes the model cumbersome and causes trouble when it comes to assessing said model. The authors suggested log-linear analysis as an alternative means of searching for determinants as opposed to regression. The aim of their search was to simultaneously implement the two aforementioned methods and compare them based on 1) the proposed "a priori" criteria and 2) obtained empirical results. The raw data was extracted from VCIOM's "Elektoral'naya Panel 2011–2012", which was then used to compare the results of the two aforementioned methods, as well as the quality of the resulting models. The dependent variable was Protest Potential; the set of hypothetical predictors was composed of variables from the socio-economic block of the "Panel". The results show that there are serious statistical reasons to reconsider methods of determinant identification when working with categorical variables, and log-linear analysis can be a better choice in terms of the quality and the set of determinants.

Keywords: protest potency determinants; categorical predictors; logistic regression; log-linear analysis.

For citation: Popova P.A., Rotmistrov A.N. Regression with Categorical Predictors: critics of Dummy-Variables Usage and Loglinear Analysis as an Alternative Approach. *Sotsiologicheskiy Zhurnal = Sociological Journal.* 2016. Vol. 22. No. 3. P. 8–31. DOI: 10.19181/socjour.2016.22.3.4583

REFERENCES

1. Baranova G.V. Method of protest activity analysis in Russia. 2012 [online]. Accessed 03.08.2016. URL: <http://www.isras.ru/files/File/Socis/2012_10/Baranova.pdf>. (In Russ.)
2. Bikbov A. Methodology of accidental street activism research (Russian street rallies and camps. December 2011 – June 2012). *Laboratorium*. 2012. No. 2. S. 130–163. (In Russ.)
3. Volkov D. Protest meetings in Russia from the end of 2011 till the start of 2012: Inquiry for political institutes democratization. *Vestnik obshchestvennogo mneniya*. 2012, April – June. No. 2 (112). P. 73–86 (In Russ.)
4. Gromov D.V. "We are not opposition, we are folk": New characteristics of political street actions. *Antropologicheskii forum*. 2011. No. 16. P. 135–153. (In Russ.)
5. Dement'eva I.N. Socio-economical and socio-political aspects of protest potency formation in the region. *Monitoring obshchestvennogo mneniya*. 2013. No. 6 (118). P. 39–50. (In Russ.)
6. Zaitsev D.G. Mass political protest: problems of conceptualization and analysis methodology. *Politicheskie izmeneniya v global'nom mire: teoretiko-metodologicheskie problemy analiza i prognozirovaniya*. I.S. Semenenko, V.V. Lapkin, V.I. Pantin (eds). Moscow: IMEMO RAN publ., 2014. P. 124–142. (In Russ.)
7. Kinsburskii A.V. Social discontent and protest potency. *Sotsiologicheskie issledovaniya*. 1998. No. 10. P. 92–95. (In Russ.)
8. Kryshtanovskii A.O. *Analiz sotsiologicheskikh dannykh s pomoshch'yu paketa SPSS*. [Sociological data analysis using SPSS.] Moscow: ID GU–VShE publ., 2006. 281 p.
9. Levada Yu.A. Discontent people: protest and patience. *Ot mnenii k ponimaniyu: Sotsiologicheskie ocherki, 1993–2000*. Moscow: MShPI publ., 2000. P. 467–488. (In Russ.)
10. Mamonov M.V. Protest activity of Russians in 2011–2012: The main trends and some patterns. *Monitoring obshchestvennogo mneniya*. 2012. No. 1 (107). P. 5–22. (In Russ.)
11. Mtiulishvili P.I. Analysis of impact of external and internal factors on the increase of protest moods of Russians. *Monitoring obshchestvennogo mneniya: Ekonomicheskie i sotsial'nye peremennye*. 2011. No. 4 (104). P. 75–82. (In Russ.)

12. Ol'shanskii D.V. *Psihologiya mass*. [Mass psychology.] Moscow: Piter publ., 2002. 368 p.
13. Rotmistrov A.N., Tolstova Yu.N. Problems of nonlinear regression models building in sociology: Nominal scales, effects of synergy, search for effective system of predictors. *Matematicheskoe modelirovanie sotsial'nykh protsessov.* 2014. No. 16. P. 159–178. (In Russ.)
14. Strebkov D.O. Economic determinants of protest behavior of Russians. *Ekonomicheskaya sotsiologiya*. 2000. Vol. 1. No. 1. P. 48–66. (In Russ.)
15. Tolstova Yu.N. *Analiz sotsiologicheskikh dannykh. Metodologiya, deskriptivnaya statistika, izuchenie svyazei mezhdu nominal'nymi priznakami.* [Sociological data analysis. Methodology, descriptive statistics, study of correlations between nominal attributes.] Moscow: Nauchnyi mir publ., 2000. 352 p.
16. Tolstova Yu.N., Ryzhova A.V. Contingency tables analysis: Likelihood ratio and loglinear models usage. *Sotsiologiya: 4M*. 2003. No. 16. P. 150–164. (In Russ.)
17. Troshin L., Balash V., Balash O. *Statisticheskii analiz nechislovoi informatsii*. [Statistical analysis of non-numerical information.] Moscow: Moskovskii gosudarstvennyi universitet ekonomiki, statistiki i informatiki publ., 2003. 67 p.
18. Agresti A. *An introduction to categorical data analysis.* New York: Wiley, 1996. 400 p.
19. Agresti A., Finlay B. *Statistical Methods for the Social Sciences.* 4th ed. Upper Saddle River, NJ: Prentice Hall, 2009. 609 p.
20. Fu C. Y. Combining loglinear model with classification and regression tree (CART): An application to birth data. *Computational Statistics & Data Analysis.* 2004. No. 45. P. 865–874. DOI: 10.1016/S0167-9473(03)00092-6
21. 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. DOI:10.1080/02664763.2013.838665
22. Jeansonne A. *Loglinear models.* 2002 [online]. Accessed 31.08.2016. URL: <http://userwww.sfsu.edu/efc/classes/biol710/loglinear/Log%20Linear%20Models.pdf>.
23. Menard S. *Applied logistic regression analysis. Sage university paper series on quantitative applications in the social sciences, 07–106.* 2nd ed. Thousand Oaks, CA: Sage, 1995. 111 p.
24. Starkweather J.D. *Categorical Variables in Regression: Implementation and Interpretation.* 2010 [online]. Accessed 31.08.2016. URL: <http://www.unt.edu/rss/class/Jon/Benchmarks/CategoricalRegression_JDS_June2010.pdf>.
25. Tansey R., White M., Long R., Smith M. A Comparison of Loglinear Modeling and Logistic Regression in Management Research. *Journal of Management.* 1996. Vol. 22. No. 2. P. 339–358. DOI: 10.1177/014920639602200207
26. Upton G. The Exploratory Analysis of Survey Data Using Log-Linear Models. *Journal of the Royal Statistical Society. Series D (The Statistician)*. 1991. Vol. 40. No. 2. Special Issue: Survey Design, Methodology and Analysis. P. 169–182.
27. Vargha A., Rudas T., Delaney D., Maxwell S. Dichotomization, Partial Correlation, and Conditional Independence. *Journal of Educational and Behavioral Statistics.* 1996, Autumn. Vol. 21. No. 3. P. 264–282. DOI: 10.2307/1165272

Received: 11.03.2016.