ФИО полностью	Ротмистров Алексей Николаевич Rotmistrov Alexei Nikolaevich
Место работы	НИУ ВШЭ NRU HSE
Должность	Доцент Docent
- Ученая степень - Звания	К.соц.н.
Электронная почта	alexey.n.rotmistrov@gmail.com

ФИО полностью	Шулус Алексей Апполинариевич Shulus Alexei Appolinarievich
Место работы	ГУУ SUM
Должность	Профессор Professor
- Ученая степень - Звания	Д.э.н. Профессор
Электронная почта	shulus@bk.ru

Специальность по которой ведется исследование (только на русском)	Социология
Название статьи (строчными буквами)	Регрессионное моделирование паттернов посещения детских досуговых центров московскими семьями с детьми Regression modeling of patterns of Moscow families' with children visiting leisure centers for children
Аннотация / Abstract (до 100 слов / не более 1000 знаков)	В статье кратко описано исследование потребления московскими семьями с детьми услуг развивающих и досуговых детских центров, в частности то, как разные характеристики семей с детьми и отдельно родителей влияют на паттерны посещения ими детских досуговых центров; описана суть регрессионного моделирования, рассмотрен феномен статистического взаимодействия и его проявления в регрессионном моделировании. Среди выводов статьи: на паттерны посещения московскими семьями с детьми детских досуговых центров влияют полнота семьи, её уровень жизни, образование и социальный статус родителей; это влияние имеет вероятностные оценки, на основе которых можно строить прогнозы. The article briefly describes a study of Moscow families's with children consumption of services of developing centers' and leisure centers' for children, in particular how different characteristics of families with children and the parents' characteristics influence the patterns of their visiting leisure centers for children; describes the essence of regression modeling, considers the phenomenon of statistical interaction and its effects in regression modeling. Among

	conclusions of the article: completeness of the family, its standard of living, education and social status of the parents affect on patterns of Moscow families' with children visiting leisure centers for children; this influence is probabilistically assessed, these assessments can be a basement for predictions.
Ключевые слова / Key words (5-10 слов)	Семья, досуг, социализация, паттерн, моделирование, регрессия, предиктор, взаимодействие, вероятность
	Family, leisure, socialization, patterns, modeling, regression, predictor, statistical interaction, probability
Обязательно! ФИО, должность, степень, место работы рецензента	

Ротмистров А.Н., Шулус А.А.

Регрессионное моделирование паттернов посещения детских досуговых центров московскими семьями с детьми

Аннотация*: В статье кратко описано исследование потребления московскими семьями с детьми услуг развивающих и досуговых детских центров, в частности то, как разные характеристики семей с детьми и отдельно родителей влияют на паттерны посещения ими детских досуговых центров; описана суть регрессионного моделирования, рассмотрен феномен статистического взаимодействия и его проявления в регрессионном моделировании. Среди выводов статьи: на паттерны посещения московскими семьями с детьми детских досуговых центров влияют полнота семьи, её уровень жизни, образование и социальный статус родителей; это влияние имеет вероятностные оценки, на основе которых можно строить прогнозы.*

Ключевые слова*: Семья, досуг, социализация, паттерн, моделирование, регрессия, предиктор, взаимодействие, вероятность*.

Социология – сравнительно молодая и довольно неточная наука. Она в меру своих сил старается выполнять 3 главные общенаучные функции: описание, объяснение и предсказание (прогнозирование). И если с первыми двумя из них она более или менее справляется, то с последней – обычно нет. На ум сразу приходит свежий пример неверного прогноза явки москвичей на выборах мэра. Отклонение эмпирических значений от прогнозируемых было столь удручающим, что глава фонда «Общественное мнение» Александр Ослон выступил с заявлением: «Да, я это признаю – фиаско социологов произошло. Мы не умеем прогнозировать явку <…> На будущее я откажусь от прогнозирования выборов. Наше дело проводить опросы» [6].

Сложность прогнозирования в социальных науках имеет объективные причины. «Сложность соответствующих явлений влечет сложность формализации наших представлений о них. Модели реальности, которые мы фактически строим, используя тот или иной метод анализа данных, оказываются чересчур приблизительными, соответствующие прогнозы не сбываются и т.д. Эти модели настолько субъективны, что исследователь все время рискует получить результаты, плохо отражающие реальность». [4, гл.1.3]. Примером исключительно сложного не то чтобы для прогнозирования – для описания и объяснения явления служит протестное поведение. Регистрация протестного поведения и составление перечня гипотетических его детерминант – большой труд [3].

Однако и в словах А. Ослона есть правда: социологи неумело пользуются математическим аппаратом обработки собранных данных. Другой вопрос: стоит ли научиться пользоваться этим аппаратом более умело или заниматься только опросами?

Предлагаем Вашему вниманию результаты эмпирического исследования на не менее актуальную для России тему, чем выборы мэра. В этом исследовании есть и массовый опрос, и обработка ответов, позволяющая делать довольно точные прогнозы.

1. Краткое описание исследования

Тема воспитания подрастающего поколения ключевая для воспроизводства нашего общества. Известно, что социализация детей в европейских социокультурных системах в основном проходит с участием двух социальных институтов: семьи и школы. Причём если в школе ребёнок социализируется в урочное (в буквальном смысле слова) время, то семье – в остальное время. Как раз в это время родители вольны заниматься ребёнком и его развитием – или не заниматься. Именно паттернам взаимодействия родителей с детьми посвящено большое эмпирическое исследование на тему: «Потребление московскими семьями с детьми услуг развивающих и досуговых детских центров».

Цель: составить детальные описания того, как проводят своё время малолетние дети из московских семей, классифицировать семьи по этому признаку, выяснить какие характеристики семьи в целом и родителей в отдельности предопределяют попадание семьи в тот или иной класс. Соответственно цель распадается на 3 блока, каждый из которых включает разные опции времяпрепровождения детей: начиная с центров дополнительного образования через различные секции и кружки к досуговым центрам.

Генеральная совокупность – московские семьи (в пределах «старой» Москвы). Эмпирическа база – представители московских семей (мамы, папы, бабушки, дедушки). Выборочная совокупность – 500 семей.

В данной статье мы взяли только один блок задач: смоделировать паттерны посещения московскими семьями с детьми детских досуговых центров в зависимости от различных характеристик их родителей и семей; и только 328 семей – которые имеют мало пропусков по переменным, относящимся к интересующему нас блоку. Эти переменные таковы. Зависимая: «Часто ли Вы с ребенком (детьми) посещаете досуговые детские мероприятия»: «Редко (раз в месяц и даже реже)» и «Часто (чаще раза в месяц)». Потенциальными предикторами выступают 6 категориальных переменных: «Ваше семейное положение», «Материальное положение Вашей семьи», «Ваше образование», «Ваша должность», «В каком административном округе Москвы Вы проживаете?», «Насколько далеко от центра проживаете?».

Поставленную задачу мы решили методом бинарного логистического регрессионного моделирования с привлечением статпакета SPSS. Мы ссылаемся на алгоритмы статпакета и результаты там, где без этого не обойтись.

2. Краткое описание регрессионного моделирования

Регрессионное моделирование – активно развивающийся класс методов. Они находятся на стыке анализа данных и моделирования явлений. Корень регрессионного моделирования – уравнение регрессии. В классическом виде оно выглядит так (1):

$$Y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n \quad (1),$$

где a_0 – константа, x_i – независимые переменные, или предикторы, a_i – коэффициенты при них, а Y – зависимая переменная.

В анализе социологических данных предикторы и зависимая переменная – некоторые признаки респондентов или иных изучаемых эмпирических объектов, а связь между правой и левой сторонами уравнения обычно не динамическая (функциональная), а статистическая; поэтому слева обычно стоит не просто зависимая переменная, а её выборочная оценка математического ожидания (т.е. среднее арифметическое значение). Например, зная возраст автовладельца, его доход и цену его нынешнего автомобиля, можно попробовать оценить примерную цену нового автомобиля (в случае наличия у автовладельца желания сменить автомобиль).

Регрессионная модель считается качественной, во-первых, если предикторы «объясняют» большую долю вариации зависимой переменной. Другими словами, насколько знание возраста автовладельца, его дохода и цены его нынешнего автомобиля позволяет повысить точность прогноза цены нового автомобиля по сравнению с ситуацией незнания этих параметров, настолько регрессионная модель качественна. Оценке качества модели служит параметр, называемый R^2, или его аналоги. Во-вторых, качественная регрессионная модель такая – в которой все предикторы статистически значимы. Это означает, что коэффициенты a_i не равны нулю ни в выборочной совокупности, ни в генеральной совокупности. Для проверки статистической значимости существует специальный математико-статистический инструментарий.

Здесь мы не будем углубляться в детали регрессионного моделирования: каков математико-статистический инструментарий проверки статистической значимости, каковы показатели объясняющей способности модели, каковы требования, предъявляемые ею к данным. Эти детали хорошо изложены в [9, ch.5].

Значимость регрессионного моделирования для социологов становится совершенно очевидной, когда речь идёт о такой его ветви, как логистическое регрессионное моделирование. Дело в том, что слева в этой разновидности регрессионного моделирования находится не метрическая или хотя бы интервальная переменная, не слишком характерная для социологии, а вероятность интересующего исследователя события (например, того, что респондент купит ту или иную модель автомобиля). Если быть точным, то слева стоит отношение вероятности того, что интересующее событие произойдёт, к вероятности, что оно не произойдёт; причём от этого отношения с целью нормировки взят логарифм. Таким образом, в логистическом виде уравнение регрессии (1) превращается в:

$$\log(p/q) = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n \qquad (2),$$

где p – вероятность того, что интересующее событие произойдёт, q – вероятность того, что интересующее событие не произойдёт, log – логарифм по любому основанию, обычно в качестве основания выступает экспонента.

Логистическое регрессионное моделирование с большинства точек зрения сходно с классическим. Его особенности изложены, напр., в [8, ch.9.3] и в контексте настоящей статьи они не принципиальны.

Казалось бы, методам регрессионного моделирования не один десяток лет, они нашли своё воплощение в различных статистических пакетах – какие здесь могут быть проблемы? По крайней мере, методологические? В нашем случае имеют место две такие проблемы.

3. Краткое описание проблем регрессионного моделирования, актуальных для нашего исследования

Эти проблемы: категориальность предикторов и статистическое взаимодействие между ними. Категориальные переменные как таковые нельзя использовать в регрессионном моделировании в качестве предикторов. Чтобы обойти этот запрет, диктуемый математической статистикой, в анализе данных разработана процедура дихотомизации: каждая номинальная или порядковая переменная «рассыпается» на новые дихотомические переменные. Каждая новая дихотомическая переменная соответствует одной категории исходной категориальной переменной и кодируется «да/нет» (1/0). Новые дихотомические переменные выступают в роли предикторов регрессионного уравнения. Содержательно это оправданно в большинстве случаев номиналистичностью социологических шкал. Оправданно это и в части применения метода наименьших квадратов (один из ключевых элементов большинства регрессионных моделей), поскольку он как раз предполагает расчёт средних арифметических значений предикторов и зависимой переменной. (см. [5])

Описание проблемы статистического взаимодействия начнём с предостережений, которые даёт в своём учебнике К. Доугерти пользователям регрессионного моделирования: «Свойства оценок коэффициентов регрессии в значительной мере зависят от правильности

спецификации модели. Результаты неправильной спецификации переменных в уравнении могут быть в обобщенном виде выражены следующим образом. 1. Если опущена переменная, которая должна быть включена, то оценки коэффициентов регрессии, вообще говоря, хотя и не всегда, оказываются смещенными. Стандартные ошибки коэффициентов и соответствующие тесты в целом становятся некорректными. 2. Если включена переменная, которая не должна присутствовать в уравнении, то оценки коэффициентов регрессии будут несмещенными, однако, вообще говоря (хотя и не всегда), – неэффективными. Стандартные ошибки будут в целом корректны, но из-за неэффективности регрессионных оценок они будут излишне большими» [1, с.166]. Далее автор подробно, с примерами и упражнениями раскрывает эти два пункта. Несомненно, его предупреждения актуальны для исследователей не только экономических явлений. Но проблема смущённости и неэффективности регрессионной модели проистекает не только из неучтённой корреляции включённых и не включённых в модель предикторов – на чём акцентирует своё внимание К. Доугерти. Эта проблема шире. Оба описанных К. Доугерти изменения набора предикторов даже при отсутствии внутри набора какой-либо корреляции могут повлиять на поведение всех предикторов этого набора. Может незначимый коэффициент стать значимым? Да. Может, наоборот, значимый коэффициент стать незначимым? Да. Может величина коэффициента измениться? В разы. Может коэффициент поменять свой знак? Запросто. Это его величество взаимодействие. Общее определение феномена взаимодействия дано в [4, гл.2.2.1], предложения по его детализации и учёту – в [7, p. 63-66].

Для учёта взаимодействия нами разработан специальный алгоритм. Он в частности описан в [2].

4. Модель паттернов посещения московскими семьями с детьми детских досуговых центров

Применив рекомендации по учёту взаимодействия, мы смогли построить две качественные модели.

Первая модель:

- оценка качества модели (аналог R^2) равна 37,5%. Для социологических моделей вполне приемлемо;
- 5 предикторов, значимых на 5-процентном уровне значимости.

Таблица 1.

Предикторы первой модели

		B	S.E.	Wald	df	Sig.	Exp(B)
«Разведен(а)»	PQR_3	-2,92	0,34	73,92	1	0,00	0,05
«Заработков хватает на все, кроме покупки недвижимости»	PR1_5	-2,17	0,20	124,09	1	0,00	0,12
«Учёная степень».	PR2_6	-3,63	0,48	56,53	1	0,00	0,03
«Западный административный округ»	R1_7	-2,21	0,28	61,71	1	0,00	0,11
«Московская область»	R1_11	-3,17	0,24	172,67	1	0,00	0,04
	Constant	5,18	0,24	465,68	1	0,00	176,81

Интерпретация:

- если семья имеет доход ниже или выше, чем «Заработков хватает на все, кроме покупки недвижимости», живёт в любом округе Москвы, кроме Западного, родители не разведены, не имеют учёной степени, то вероятность, что они будут водить детей

на досуговые детские мероприятия чаще, чем раз в месяц, равна $P = 0{,}99$. Поскольку соотношение шансов $P/(1-P) = 176{,}81$ (строка таблицы 1, соответствующая константе). Назовём это состояние «status-quo». Теперь рассмотрим учтённые моделью потенциальные изменения (каждое изменение предполагает, что остальные признаки соответствуют status-quo):

- если материальное состояние семьи меняется, то вероятность снижается до 0,95.
- если семья переезжает в Западный административный округ, то вероятность снижается до 0,88.
- если родители разводятся, то вероятность снижается до 0,91.
- если кто-то из родителей получает учёную степень, то вероятность снижается до 0,83.

Как видим, все предикторы в случае их «активации» (т.е. когда респонденты обладают этими свойствами) снижают вероятность того, что родители будут водить детей на досуговые детские мероприятия чаще, чем раз в месяц.

Разумеется, модель доступна для более глубокой интерпретации. Например, интерпретации сочетаний предикторов (без перемножений).

Вторая модель:

- оценка качества модели (аналог R^2) равна 37,5%;
- 10 предикторов, значимых на 5-процентном уровне значимости.

Таблица 2.

Предикторы второй модели

		B	S.E.	Wald	df	Sig.	Exp(B)
«Разведен(а)»	PQR_3	-3,16	0,36	75,24	1	0,00	0,04
«Заработков хватает на все, кроме покупки недвижимости»	PR1_5	-2,45	0,22	128,59	1	0,00	0,09
«Высшее образование»	PR2_5	1,07	0,23	22,40	1	0,00	2,92
«Руководитель»	PR4_1	-2,12	0,29	51,75	1	0,00	0,12
«Менеджер»	PR4_2	-0,85	0,29	8,75	1	0,00	0,43
«Служащий офиса»	PR4_3	-1,02	0,27	14,21	1	0,00	0,36
«Западный административный округ»	R1_7	-2,63	0,32	67,10	1	0,00	0,07
«Восточный административный округ»	R1_9	-1,73	0,41	17,44	1	0,00	0,18
«В пределах Третьего кольца»	R2_2	3,20	0,44	52,94	1	0,00	24,41
«В пределах МКАД»	R2_3	2,81	0,27	107,94	1	0,00	16,66
	Constant	3,12	0,26	142,19	1	0,00	22,65

Интерпретация:

- если семья имеет доход ниже или выше, чем «Заработков хватает на все, кроме покупки недвижимости», живёт в любом округе Москвы, кроме Восточного и Западного, в пределах Садового кольца или, наоборот, за МКАД, родители не разведены, не имеют высшего образования, не руководители, не менеджеры и не служащие офиса, то вероятность, что родители будут водить детей на досуговые

детские мероприятия чаще, чем раз в месяц, равна $P = 0{,}96$. Поскольку соотношение шансов P/(1-P) = 22,65 (строка таблицы 2, соответствующая константе). Теперь рассмотрим учтённые моделью потенциальные изменения (каждое изменение предполагает, что остальные признаки соответствуют status-quo):

- если материальное состояние семьи меняется, то вероятность снижается до 0,66.
- если семья переезжает в Западный административный округ, то вероятность снижается до 0,62.
- если семья переезжает в Восточный административный округ, то вероятность снижается до 0,8.
- если семья переезжает на новое место жительства между Садовым и Третьим кольцами, то вероятность вырастает до 0,99.
- если семья переезжает на новое место жительства между Третьим кольцом и МКАД, то вероятность вырастает чуть меньше.
- если родители разводятся, то вероятность снижается до 0,5.
- если кто-то из родителей получает высшее образование, то вероятность вырастает до 0,99.
- если кто-то из родителей становится руководителем, то вероятность снижается до 0,73.
- если кто-то из родителей становится менеджером, то вероятность снижается до 0,91.
- если кто-то из родителей становится служащим офиса, то вероятность снижается до 0,89.

Как мы видим, смена семьёй места жительства относительно центра города и получение высшего образования кем-то из родителей в среднем способствуют частому посещению их детьми досуговых детских мероприятий. Остальные признаки семьи и родителей, учтённые моделью, действуют в обратном направлении.

Следует ясно понимать, что это скорее риторический приём – рассматривать изменения в динамике. Мы динамику не измеряли. Поэтому его выводы должны быть сформулированы более сухо: в терминах отнесения детей, чьи родители имеют такие-то социально-демографические и иные характеристики и чьи семьи имеют такие-то социально-демографические и иные характеристики, в категорию часто посещающих досуговые детские мероприятия или в категорию редко посещающих. Именно поэтому логистическое регрессионное моделирование наряду с прочими задачами решает и задачу классификации объектов наблюдения. Фактически, это конечная задача логистического регрессионного моделирования – в отличие от более базового линейного регрессионного моделирования, конечной задачей которого выступает построение регрессионного уравнения.

Литература

1. Доугерти, К. Введение в эконометрику. М. ИНФРА-М, 1999. – 402 с.
2. Ротмистров А.Н., Шулус А.А. Проблема взаимодействия предикторов в регрессионном моделировании на примере исследования посещения московскими семьями развлекательных центров // Науковедение, 2013, 5 (18) выпуск // URL: http://naukovedenie.ru/
3. Ротмистров А.Н. Сравнительный анализ факторов студенческого протестного движения в России на рубеже XIX-XX и в начале XXI века // Высшее образование сегодня. 2009. № 1. С. 36-41 // URL: http://www.hetoday.org/arxiv/2009/arxiv_0109.html
4. Толстова Ю.Н. Анализ социологических данных. Научный мир, 2000. – 352 с. // URL: http://socioline.ru/pages/tolstova-yun-analiz-sotsiologicheskih-dannyh
5. Толстова Ю. Н. Измерение в социологии. КДУ, 2009. – 291 с., гл.13.3.1, URL: http://socioline.ru/pages/yun-tolstova-izmerenie-v-sotsiologii

6. Фонд "Общественное мнение" не будет больше прогнозировать результаты выборов // Электронный ресурс ИТАР-ТАСС. 2013. 12 сентября // URL: http://www.itar-tass.com/c1/874380.html (Дата обращения: 26.09.2013)
7. Hosmer, D. W., and S. Lemeshow. 2000. Applied Logistic Regression, 2nd ed. New York: John Wiley and Sons.
8. Agresti, A. An introduction to categorical data analysis. John Wiley & Sons, 1996. – 296 с.
9. Agresti, A. Statistical methods for the social sciences. Upper Saddle River Pearson Education International, 2009. – 609 с., ch.5 // URL: http://bookfi.org/book/833357