Э.Д. Понарин, А.В. Лисовский, Ю.А. Зеликова
(Санкт-Петербург)

МОДЕЛИ ДЛЯ ПУАССОНОВСКИХ ЗАВИСИМЫХ ПЕРЕМЕННЫХ: МОЖНО ЛИ ПРОГНОЗИРОВАТЬ РЕЗУЛЬТАТИВНОСТЬ ФУТБОЛЬНЫХ МАТЧЕЙ?

Статья описывает общий случай пуассоновской регрессии и указывает на ее отличия от логарифмически линейных моделей для анализа таблиц сопряженности. Пуассоновская модель применяется к анализу количества забитых мячей в чемпионате России. На данном примере обсуждается отличие пуассоновской модели от линейной регрессии по методу наименьших квадратов.

Ключевые слова: пуассоновская регрессия, логарифмически линейные модели, футбол.

Введение

Данная статья посвящена анализу процессов, в которых зависимая переменная представляет собой число событий, произошедших за фиксированный период времени, при условии, что эти события происходят с некоторой фиксированной средней

[1] **Эдуард Дмитриевич Понарин** – заведующий лабораторией сравнительных социальных исследований НИУ ВШЭ, профессор ф-та социологии Санкт-Петербургского филиала НИУ ВШЭ, PhD University of Michigan. E-mail: ponarin13@gmail.com.

Александр Владимирович Лисовский – кандидат психологических наук, доцент ф-та социологии Санкт-Петербургского филиала НИУ ВШЭ. E-mail: sliss54@hse.spb.ru.

Юлия Александровна Зеликова – кандидат социологических наук, доцент ф-та социологии Санкт-Петербургского филиала НИУ ВШЭ. E-mail: juliazelikova@hotmail.com.

интенсивностью и независимо друг от друга [1]. Такие процессы, связанные со «счетными переменными», измеряемыми как количество определенных событий за данный промежуток времени, представляют интерес для социологов. Примерами могут служить: количество посещений театров или кино за последние двенадцать месяцев, количество детей в семье (событие – это рождение ребенка), участие студентов в научных конференциях за четыре года обучения (сколько раз), количество пересдач зачетов и экзаменов за одну сессию (от нуля и выше), активность участия в протестных акциях (сколько раз), число правонарушений, совершенных подростками за определенный период.

Пуассоновские модели широко используются в западной социологии при исследовании организационных процессов, противоправного и сексуального поведения [2; 3; 4]. Авторам не удалось обнаружить публикаций российских социологов, использующих пуассоновские модели, хотя такие модели довольно часто применяются в отечественных публикациях в сферах экономики, медицины, спорта. Исключение составляют работы Ю.Н. Толстовой и А.В. Рыжовой, а также Д.А. Трофимова, которые посвящены анализу таблиц сопряженности с использованием логлинейных моделей [5; 6]. Логлинейные модели для анализа таблиц сопряженности являются частным случаем пуассоновской модели. В этом частном случае зависимой пуассоновской переменной становится совместное распределение мультиномиальных переменных, т.е. количество наблюдений в ячейках таблицы сопряженности.

Общий случай пуассоновской регрессии в отечественной социологии специально не обсуждался. Между тем социологи довольно часто работают с пуассоновскими переменными. На первый взгляд пуассоновские модели могут показаться тождественными обычным линейным регрессионным моделям, рассчитанным методом наименьших квадратов, и в этом одна из причин невнимания к ним. В данной статье авторы раскрывают причины, по которым корректнее использовать пуассоновскую модель для подобных переменных.

Задачи данной работы:

– представить общий случай пуассоновской модели;

– показать на примере футбольных матчей чемпионата России, как пуассоновскую модель можно использовать для предсказания исхода пуассоновского процесса;

– обсудить преимущества пуассоновской модели по сравнению с линейной регрессией, рассчитанной по методу наименьших квадратов.

Пуассоновская модель: общие сведения

Распределение Пуассона – вероятностное распределение, моделирующее случайную дискретную величину, представляющую собой число повторяющихся событий при том, что данные события происходят с фиксированной средней интенсивностью и независимо друг от друга. Основной параметр распределения Пуассона – λ (лямбда), которая одновременно выступает и дисперсией, и математическим ожиданием этого распределения. Приблизительное равенство средней и дисперсии счетной переменной *y* в выборочной совокупности – важнейшее условие применимости пуассоновских моделей.

На графиках ниже представлены распределения двух переменных, соответствующих формуле Пуассона. У первого параметр λ равен 2, а у второго – 10. Из второго графика видно, что при больших значениях параметра λ распределение Пуассона по форме напоминает нормальное распределение; характерную особую форму оно приобретает только при малых λ (*рис*. 1, график слева).

Важно понимать, что, в отличие от номинальных (категориальных) переменных пуассоновские переменные – это дискретные переменные, характеризующие количество повторений события, и их цифровые значения нельзя заменять буквенными «ярлыками». Кроме того, в отличие от мультиномиального распределения, представляющего ограниченное число возможных категорий некоего

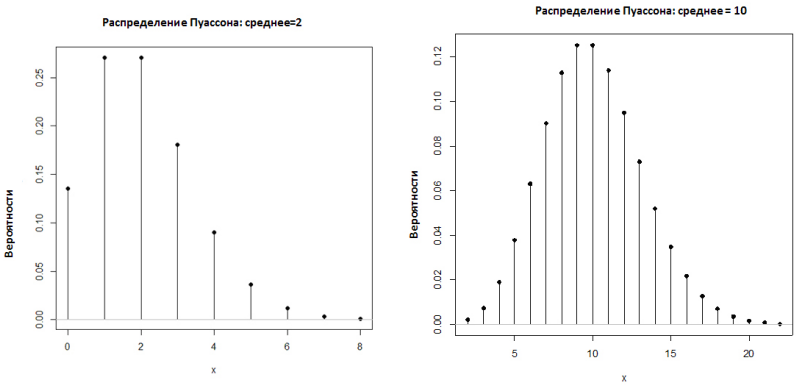

Рис. 1. Графики примеров распределения Пуассона

фактора, пуассоновское распределение в принципе не ограничено с правой стороны. Отличие пуассоновских переменных от интервальных переменных связано с тем, что пуассоновские переменные – дискретные, целочисленные и неотрицательные.

Распределение на графике слева (см. *рис*. 1) может быть, например, количеством фруктов, съеденных индивидами в течение дня, а на графике справа (см. *рис*. 1) – количеством посещений кинотеатра в течение последних трех лет.

Впервые пуассоновская модель была предложена для входящих телефонных звонков на коммутатор. Далее эта модель применялась при анализе надежности оборудования и предсказании его поломки [7; 8], в медицинских исследованиях для анализа роста колонии бактерий в чашке Петри, предсказании летального исхода болезни при различных условиях [9; 10], а также в экономике и социологии для анализа разнообразных счетных данных, например, для прогнозирования поломок оборудования, или зависимости количества «доз» выпитых напитков от различных параметров, а также для анализа разнообразных демографических данных [11; 12].

В обобщенных линейных моделях для обеспечения связи зависимой переменной с линейным выражением вида $\beta_0 + \beta x + \varepsilon$ используют

преобразующие функции (*link function*). В случае пуассоновской регрессии этой функцией является натуральный логарифм – ln *y*. Поэтому пуассоновскую регрессию называют логарифмически линейной (отсюда и английский термин *log-linear models*).

Пуассоновская регрессия вида $\ln y = \beta_0 + \beta x + \varepsilon$ используется для описания зависимости пуассоновской переменной от множества переменных $x_1 \ldots x_k$. В этом случае пуассоновская регрессия сходна с другими видами регрессии; единственное существенное отличие – использование техники офсета (этот термин обычно переводится на русский язык как «компенсация»).

Офсет – прием учета экспозиции, – как правило, продолжительности наблюдения за данным респондентом или случаем, т. е. переменной, влияющей на среднюю (ожидаемую) величину пуассоновской переменной. Например, если *y* – это число посещений кинотеатра разными людьми, то такой величиной может быть длительность периода, за который фиксируется это число для данного человека: чем дольше период, тем больше ожидаемая величина зависимой переменной. Учет экспозиции позволяет включать в модель респондентов или случаи с различной экспозицией, так как зависимая переменная становится сходной с пропорцией или скоростью (*y*/*t*). Экспозиция в пуассоновских моделях может быть связана не только со временем: например, если мы моделируем количество удовлетворенных жалоб в данном судебном округе, тогда экспозицией будет общее число жалоб (как удовлетворенных, так и неудовлетворенных), поданных в данном округе.

Пусть в общем случае экспозиция для каждого случая замерена переменной *t*. Тогда зависимость пуассоновской переменной *y* от некоей переменной *x* можно выразить следующим образом:

$$\ln(y/t) = \beta_0 + \beta x + \varepsilon. \quad (1)$$

Поскольку логарифм частного равен разности логарифмов, то из выражения (1) следует выражение (2):

$$\ln y = \beta_0 - \ln t + \beta x + \varepsilon. \quad (2)$$

Во втором уравнении офсет – это коэффициент при ln *t*, который фиксирован и равен единице (т. е. у него как бы нет коэффициента). Технически это одна из переменных из множества $x_1 \ldots x_k$, но, как объяснялось выше, у нее особая роль. Офсет – не обязательный элемент пуассоновской регрессии: например, он не используется, если экспозиция для всех наблюдений одинакова.

Пример использования пуассоновской модели для предсказания результативности футбольных матчей

Футбол – важный социальный феномен, один из самых популярных в мире видов спорта. На стадионах каждый год за футбольными матчами наблюдают миллионы зрителей, и еще больше болельщиков смотрят телевизионные трансляции. Мы использовали статистику футбольных матчей чемпионата России 2011–2012 гг. с сайта *championat.com*[1]. Количество голов, забитых участниками футбольных матчей, – это пуассоновская переменная. Причем в каждом футбольном матче таких пуассоновских переменных три: это голы, забитые хозяевами, на поле которых проводится матч, голы, забитые командой гостей, а также общая результативность матча – сумма голов, забитых и хозяевами и гостями. Именно голы определяют результат игры и интересуют болельщиков, но нередко игры заканчиваются и без забитых голов (нулевые ничьи). Интересно, что начиная с Чемпионата Европы по футболу 1996 г., чтобы повысить результативность матчей победителям стали присуждать три очка, а не два, как ранее (за ничью команда получает одно очко, а за поражение – ноль). Тем не менее в каждой десятой игре Чемпионата России по футболу 2011–2012 гг. зрители так и не увидели забитых голов. Важно отметить, что

[1] См.: http://www.championat.com/football/_russiapl/288/calendar/tour.html.

все три переменные, связанные с количеством голов, забитых в матче, соответствуют условиям пуассоновских переменных: среднее количество голов в матче практически равно дисперсии, что соответствует допущениям пуассоновской модели. В Чемпионате России 2011-2012 гг. эти соотношения показаны в *табл.* 1.

Таблица 1

ОПИСАТЕЛЬНАЯ СТАТИСТИКА ЗАВИСИМЫХ ПЕРЕМЕННЫХ

Результативность	Значение			Дисперсия
	Min	Max	Среднее	
Голы хозяев	0	6	1,34	1,34
Голы гостей	0	5	1,03	1,14
Голы хозяев плюс голы гостей	0	8	2,37	2,22

Среднее пуассоновской переменной, как отмечалось выше, должно быть равно дисперсии. В нашем случае разница между средним и дисперсией не превышает 0,15, что явно приемлемо с точки зрения случайной ошибки выборки.

Поскольку продолжительность футбольных матчей примерно одинакова – 90 минут, к которым судья может добавить несколько минут, но обычно не более пяти в каждом тайме – то для этих данных можно использовать пуассоновскую модель без офсета.

В 2011 г. Федерация футбола России приняла решение о переходе с системы «весна–осень» на систему «осень–весна», и благодаря этому в полтора раза возросло количество футбольных матчей. Обычно Чемпионат России проводится в два круга, и каждая из команд играет 30 матчей (всего в чемпионате 240 матчей), но для того чтобы чемпионат 2011–2012 гг. закончился весной, в нем проводилось еще два укороченных тура. По итогам двух первых кругов команды разбивались на две восьмерки. В первую вошли команды, занявшие места с 1 по 8-е и они разыгрывали призовые места и право участвовать в европейских турнирах (Лиге

чемпионов и Лиге Европы), а во второй восьмерке команды, занявшие места с 9 по 16-е, определяли две команды, переходившие в низшую лигу. Таким образом, общее количество матчей (наблюдений) увеличилось на 112 матчей (всего было сыграно 352). Дополнительные круги могли повлиять на турнирную мотивацию: у команд-лидеров второй восьмерки, которым не угрожал вылет в низшую лигу, и команд-аутсайдеров первой восьмерки, у которых не было шансов занять призовое место или пройти в Еврокубки, было меньше стимулов бороться за победу, чем у соперников.

Независимые переменные[1]

В наших моделях использовалось два типа независимых переменных: априорные, значения которых известны до начала матча, и апостериорные – их значения становятся известными только после окончания матча. Очевидно, что с точки зрения содержательного анализа важнее априорные переменные, однако включение в наши модели апостериорных переменных позволило увеличить количество независимых переменных и показать применение пошагового моделирования в пуассоновской регрессии.

Априорные независимые переменные

Финансовые

«Бюджет команды гостей», *«бюджет команды хозяев»*, *«разница стоимости составов хозяев и гостей»* (большие значения, если состав хозяев дороже).

Бюджет и стоимость состава сильно коррелируют друг с другом (коэффициент корреляции Пирсона равен 0,895), однако стоило использовать обе переменные, так как стоимость состава отражает в том числе «утопленные инвестиции»: затраты на до-

[1] Данные о статистических параметрах независимых переменных приводятся в приложении 1.

рогих игроков, которые могут не участвовать в играх[1], а бюджет команды – отражает потенциальные возможности команды для покупки новых игроков. Бюджет самой богатой команды чемпионата превышал бюджет самой бедной почти на порядок – в восемь с половиной раз, то же верно и для стоимости составов.

Состав и тренеры

В дополнение к стоимости составов в моделях использовалась переменная «*количество бразильских игроков*» – бразильцы признаются техничными игроками и их присутствие может считаться одним из показателей качества состава и влиять на результативность.

Категориальные переменные «*тренер хозяев*» и «*тренер гостей*» принимали значение «1», если тренер – россиянин или гражданин страны, входящий в СНГ, и «2», если тренер – гражданин другой страны.

«Круг» – уже отмечалось, что первый и второй (полные) круги и третий и четвертый (игры в восьмерках) отличались не только по продолжительности, но и по характеру турнирной мотивации. Чтобы значения переменной были сравнимы по продолжительности, они кодировались так: «1» – первый круг, «2» – второй круг, «3» – третий и четвертый круги.

«Качество поля» – эта переменная кодировалась так: «0» – плохие поля (март, апрель, ноябрь), «1» – хорошие поля (с мая по октябрь).

«Количество зрителей» – вводилось количество зрителей, присутствовавших на матче. Минимальное значение: 2 300, а максимальное – 58 600.

«Судьи» – категориальная переменная. В модели включались бинарные переменные для каждого из судей, обслуживших не менее 17 матчей (т.е. 80% всех сыгранных матчей).

[1] Например, футболист сборной России Андрей Аршавин последние полтора года почти не играет в основном составе лондонского «Арсенала», но его заработная плата – одна из самых высоких в команде.

Апостериорные независимые переменные

«Удары по воротам – хозяева» и «*удары по воротам – гости»*.

«Удары в створ ворот – хозяева» и «*удары в створ ворот – гости»*.

Удары по воротам и в створ ворот, произведенные в одном матче, сильно коррелируют друг с другом. Для хозяев коэффициент корреляции Пирсона – +0,70, а для гостей – +0,72.

«Процент владения мячом хозяевами» – очевидно, что аналогичный показатель для гостей в модель было включать нельзя, так как корреляция этих двух переменных – минус единица.

«Количество предупреждений – хозяева», *«количество предупреждений – гости»*. Так как удаления игроков с поля происходят редко, они кодировались как два предупреждения (что логически оправдано, так как игрок, получивший два предупреждения в одном матче, удаляется с поля).

Спецификация и анализ моделей

В статье приводятся модели для двух зависимых переменных: голы хозяев и голы гостей[1]. Мы предполагали, что модели для голов хозяев и голов гостей будут различаться.

Использовалась процедура *glm* (обобщенные линейные модели) статистической среды *R* и рассчитывались модели для пуассоновской зависимой переменной.

На первом шаге рассчитывались модели для каждой из независимых переменных по отдельности. Это было необходимо, поскольку, учитывая сильные корреляции некоторых независимых переменных, мы ожидали проявления мультиколлинеарности. Важно было оценить «максимально возможные» эффекты для каждой из независимых переменных и их знаки.

[1] Модели для общей результативности (голы хозяев плюс голы гостей) не включены в статью, но они были рассчитаны, и авторы вышлют их заинтересованным читателям.

На втором шаге в модель вводились все априорные независимые переменные, а потом с помощью алгоритма пошагового выбора «вперед и назад» находилась наилучшая модель по Байесовскому информационному критерию (*BIC*).

На третьем шаге в модель добавлялись апостериорные независимые переменные, и снова находилась оптимальная модель.

На четвертом шаге полученные результаты проверялись на устойчивость: были исключены данные для третьего и четвертого круга (игры в восьмерках) и строились модели для априорных переменных только для игр первого и второго круга.

На пятом этапе проверялась гипотеза о статистическом взаимодействии качества поля и влияния бразильских игроков на результативность. Гипотеза предполагала, что бразильские футболисты повышают результативность матча при игре на хорошем поле. Плохое состояние поля нивелирует их техническое мастерство.

В нулевом столбце *табл*. 2 представлены парные модели с единственной независимой переменной. У всех значимых коэффициентов – предсказанные и ожидаемые с точки зрения логики знаки. Повышают количество голов, забитых хозяевами: бюджет хозяев, разница стоимости составов команд (в пользу хозяев), общее количество ударов по воротам и в створ ворот хозяев. Снижают – бюджет команды гостей, и количество ударов в створ ворот команды гостей.

В модель (1), представленную в следующем столбце *табл*. 2, были включены все независимые переменные: и априорные, и апостериорные. В этой модели значимы три переменные. Однако возникла проблема мультиколлинеарности: в дополнение к уже указанным проблемам с финансовыми переменными, сильно коррелируют «удары по воротам» ($VIF = 1{,}8$)[1] и «удары в створ ворот»

[1] *VIF* (фактор инфляции ошибок) – в большинстве руководств по статистике указывается, что проблемой могут быть значения *VIF*, приближающиеся к пяти, однако при невысокой связи зависимой переменной с зависимыми переменными и сильно коррелированными независимыми переменными (а это наш случай) проблемой могут быть и более низкие значения *VIF*.

Таблица 2

РЕГРЕССИОННЫЕ МОДЕЛИ; ЗАВИСИМАЯ ПЕРЕМЕННАЯ – ГОЛЫ ХОЗЯЕВ

Номер модели	0	1	2	3	4
	Модели				
Голы хозяев	парные	полная	пошаговая	пошаговая	пошаговая
Константа (смещение)		0,078	0,271***	–0,192	–0,555***
		(0,498)	(0,047)	(0,147)	(0,111)
Бюджет хозяев	0,004***	0,002			
	(0,001)	(0,002)			
Бюджет гостей	–0,003*	0,001			
	(0,001)	(0,002)			
Тренер хозяев	0,053	–0,004			
	(0,098)	(0,106)			
Тренер гостей	–0,130	–0,058			
	(0,100)	(0,104)			
Качество поля	0,136	0,063			
	(0,095)	(0,103)			
Разница стоимости составов	0,002***	0,001	0,002***	0,001*	
	(0,0005)	(0,001)	(0,001)	(0,0003)	
Количество зрителей	0,0002	–0,000			
	(0,0006)	(0,001)			

Примечание. В ячейках представлены нестандартизованные регрессионные коэффициенты; их стандартные ошибки даны в скобках. Звездочки соответствуют уровню значимости 5% (*), 1% (**) и 0,1% (***).

Продолжение табл. 2

Номер модели	0	1	2	3	4
	Модели				
Голы хозяев	парные	полная	пошаговая	пошаговая	пошаговая
Количество бразильцев – хозяева	0,038	0,083*			
	(0,028)	(0,033)			
Количество бразильцев – гости	–0,012	–0,005			
	(0,030)	(0,031)			
Круг (1, 2, 3+4)	–0,033	0,017			
	(0,057)	(0,062)			
Удары хозяев	0,043***	*–0,049***		*–0,051***	
	(0,010)	(0,017)		(0,015)	
Удары гостей	–0,015	0,018			
	(0,012)	(0,018)			
Удары в створ, хозяева	0,147***	0,197***		0,193***	0,147***
	(0,016)	(0,025)		(0,024)	(0,016)
Удары в створ, гости	–0,045*	–0,044			
	(0,021)	(0,032)			
Процент времени владения мячом – хозяева	0,002	–0,015			
	(0,006)	(0,008)			
Предупреждения – хозяева	–0,024	–0,006			
	(0,030)	(0,032)			

Окончание табл. 2

Номер модели	0	1	2	3	4
	Модели				
Голы хозяев	парные	полная	пошаговая	пошаговая	пошаговая
Предупреждения - гости	0,054	0,047			
	(0,024)	(0,026)			
Псевдо-R^2 *Нагелькерке*		0,392	0,077	0,339	0,292
Лог. правдоподобия		–464	–510	–473	–480
AIC		964	1024	954	964
BIC		1034	1032	969	972
N		352	352	352	352

(*VIF* = 1,8). Коэффициент для «количества бразильцев в составе», незначимый в простой парной модели, увеличивается и становится значимым, коэффициент для переменной «хозяева – количество ударов по воротам» меняет знак, а коэффициент для переменной «хозяева – количество ударов в створ ворот» сохраняет верный знак, но увеличивается в полтора раза.

Модель (2) была построена в результате пошагового отбора априорных переменных. Единственной значимой переменной остается разница стоимости составов. Показатели пригодности модели, в том числе *BIC* показывают, что эта модель немного лучше модели 1.

Затем мы рассчитали «наилучшую» модель с помощью процедуры пошагового отбора по всем независимым переменным (модель 3). У этой модели самый лучший *BIC* – 969, однако в ней сохранился неверный знак для регрессионного коэффициента показателя «хозяева – удары по воротам» и завышен коэффициент для показателя «хозяева – удары в створ ворот»; как уже упоминалось выше, это вызвано довольно сильной корреляцией этих переменных. Единственный регрессионный коэффициент, вызывающий полное доверие, – это коэффициент для показателя «разница стоимости составов».

На последнем шаге (модель 4) снова использовался пошаговый алгоритм отбора лучшей модели, но переменная «удары хозяев» была предварительно исключена. В этом случае «разница бюджета» становится незначимой и в итоге оптимальная модель – это модель всего с одной объяснительная переменной – «удары в створ – хозяева».

Мы также рассчитывали отдельно модель, не показанную в *табл.* 2 и включающую все априорные независимые переменные, – в этом случае все коэффициенты оказались незначимы. Причина – в мультиколлинеарности трех финансовых переменных: бюджет гостей (*VIF* = 2,8), бюджет хозяев (*VIF* = 2,8) и разница стоимости составов (*VIF* = 4,2).

Если бы нашей целью было предсказание результата матча до его начала, то мы могли бы основываться на модели 2 как итоге пошагового отбора исключительно априорных переменных. Однако по статистикам пригодности модели видно, что включение апостериорных переменных повышает точность моделей.

Уже на уровне парных зависимостей, представленных в первом столбце *табл*. 3, обнаруживается разница между моделями для голов, забитых хозяевами и гостями. Эффекты для бюджета хозяев и гостей, разницы стоимости составов и ударов по воротам и в створ ворот остаются значимыми и с предсказуемыми знаками, но появляются две новые значимые переменные. Первая – это качество полей: чем оно выше, тем больше голов забивают гости, вторая – предупреждения, полученные хозяевами (чем чаще судьи предупреждают хозяев, тем больше забивают гости).

Как далее видно из *табл.* 3, в модель (1) включены все независимые переменные, как априорные, так и апостериорные. Из априорных переменных значимы две, причем с предсказуемыми знаками: «качество поля» – чем оно лучше, тем больше гости забивают голов, а также «количество бразильцев в составе». С последней переменной есть проблемы: хотя знак этой переменной соответствует нашим теоретическим ожиданиям, однако в парной модели соответствующий регрессионный коэффициент был ниже и оказался незначим. Среди апостериорных переменных также значимы две: «количество ударов в створ ворот – гости» (напомним, что сходная переменная «количество ударов в створ ворот – хозяева» была значима, когда моделировались голы хозяев), а вот коэффициент для переменной – «процент владения мячом – хозяева» приобретает парадоксальный знак, отличный от оценки в парной модели (в которой этот коэффициент незначим). Иными словами, если принять эту модель, то окажется, что чем дольше хозяева владеют мячом, тем больше голов забивают гости. Мы склонны считать интерпретировать этот артефакт как эффект мультиколлинеарности.

Таблица 3

РЕГРЕССИОННЫЕ МОДЕЛИ; ЗАВИСИМАЯ ПЕРЕМЕННАЯ – ГОЛЫ ГОСТЕЙ

Номер модели	0	1	2	3	4
	Модели				
Голы гостей	парные	полная	пошаговая	пошаговая	пошаговая
Константа (смещение)		–3,311***	–0,307**	–2,725***	–1,065***
		(0,601)	(0,095)	(0,470)	(0,170)
Бюджет хозяев	–0,004**	–0,004		–0,006***	–0,004*
	(0,001)	(0,003)		(0,002)	(0,002)
Бюджет гостей	0,007***	0,004		0,005***	0,004**
	(0,001)	(0,002)		(0,001)	(0,001)
Тренер хозяев	–0,124	0,023			
	(0,116)	(0,121)			
Тренер гостей	0,198	0,030			
	(0,108)	(0,120)			
Качество поля	0,450***	0,381**	0,435***	0,347**	0,340**
	(0,113)	(0,122)	(0,113)	(0,114)	(0,114)
Разница стоимости составов	–0,004***	–0,001	–0,004***		
	(0,001)	(0,001)	(0,001)		
Количество зрителей	0,001	–0,001			
	(0,001)	(0,001)			

Примечание. В ячейках представлены нестандартизованные регрессионные коэффициенты; их стандартные ошибки даны в скобках. Звездочки соответствуют уровню значимости 5% (*), 1% (**) и 0,1% (***).

Продолжение табл. 3

Номер модели	0	1	2	3	4
	Модели				
Голы гостей	парные	полная	пошаговая	пошаговая	пошаговая
Количество бразильцев – хозяева	–0,025	–0,043			
	(0,034)	(0,038)			
Количество бразильцев – гости	0,057	0,071*			
	(0,032)	(0,035)			
Круг (1, 2, 3+4)	0,030	0,099			
	(0,065)	(0,075)			
Удары хозяев	–0,018	–0,002			
	(0,012)	(0,019)			
Удары гостей	0,080***	–0,011			
	(0,013)	(0,020)			
Удары в створ – хозяева	–0,021	0,004			
	(0,020)	(0,030)			
Удары в створ – гости	0,213***	0,204***		0,191***	0,179***
	(0,021)	(0,034)		(0,023)	(0,023)
Процент владения мячом – хозяева	–0,004	0,036***		0,031***	
	(0,007)	(0,009)		(0,008)	

Окончание табл. 3

Номер модели	0	1	2	3	4
	Модели				
Голы гостей	парные	полная	пошаговая	пошаговая	пошаговая
Предупреждения – хозяева	0,087**	0,062			
	(0,032)	(0,034)			
Предупреждения – гости	–0,045	0,004			
	(0,029)	(0,033)			
Псевдо-R^2 *Нагелькерке*		0,479	0,222	0,447	0,405
Лог. правдоподобия		–401	–444	–407	–415
AIC		839	895	827	840
BIC		908	907	850	859
N		352	352	352	352

Модель (2) рассчитана пошаговым методом только для априорных переменных; незначимые переменные из нее исключены. Значимы две переменных с предсказуемыми знаками: «качество поля» (качество выше – гости забивают чаще) и «разница стоимости состава» с логичным знаком: чем больше разница стоимости составов в пользу хозяев, тем меньше гости забивают голов.

Модель (3) рассчитана пошаговым методом для всех переменных, как априорных, так и апостериорных; незначимые переменные из нее исключены. Качество поля сохраняет значимость, кроме того, в модель вошли с предсказуемыми знаками «бюджет хозяев» и «бюджет гостей». Обе эти переменные значимы в парных моделях, однако коэффициент для второй из них завышен в полтора раза в сравнении с аналогичным коэффициентом в парной модели (соответственно, –0,004 в парной модели и –0,006 в модели 3). Ситуация с двумя апостериорными переменными уже обсуждалась при интерпретации модели 3: логичный знак для «ударов в створ ворот» и парадоксальный для «процента владения мячом».

В модели (4) мы исключили «процент владения мячом», так как регрессионный коэффициент в парной модели для этой переменной был незначим. Оценка регрессионного коэффициента для «бюджета хозяев» теперь корректнее – это коэффициент уже такой же, как в парной модели, а не завышен, как это было в предыдущей модели. Такое решение, однако, приводит к некоторому ухудшению *BIC* (с 850 до 859), но мы находим это приемлемым.

Проверка моделей на устойчивость

Возникает вопрос, насколько уникальны или типичны представленные выше модели с учетом необычной длительности Чемпионата 2011–2012 гг., в котором было два дополнительных укороченных круга? Для проверки мы исключили данные для третьего и четверного круга (т.е. количество игр стало обычным для типичного двухкругового турнира). Представленные ниже мо-

дели показывают, что полученные нами результаты (для наиболее интересных – априорных переменных) – устойчивы.

Только одна переменная значимо влияла на количество голов хозяев: разница стоимости составов. Чем выше была эта разница в пользу хозяев, тем больше они забивали голов. Та же переменная влияла на количество голов гостей (с обратным знаком), кроме того, гости забивали больше хозяев на хороших полях. Наши модели показывают, что количество голов гостей (псевдо-R-квадрат по Нагелькерке равен 0,267) предсказывается лучше, чем количество голов хозяев (псевдо-R-квадрат 0,131) (*табл.* 4).

В целом анализ отдельно для матчей, сыгранных по традиционному двухкруговому регламенту, демонстрирует устойчивость наших моделей для всех четырех кругов. Направление и размерность коэффициентов сохранились.

Таблица 4

РЕГРЕССИОННЫЕ МОДЕЛИ; ЗАВИСИМЫЕ ПЕРЕМЕННЫЕ КОЛИЧЕСТВА ГОЛОВ В ПЕРВОМ И ВТОРОМ КРУГЕ

Зависимая переменная	Голы хозяев	Голы гостей
Константа (смещение)	0,282***	–0,461**
	(0,057)	(0,159)
Разница стоимости составов (хозяева минус гости)	0,003***	–0,005***
	(0,001)	(0,001)
Качество поля (плохое или хорошее)		0,546**
		(0,171)
Псевдо-R^2 *Нагелькерке*	0,131	0,267
Лог. правдоподобия	–355	–304
AIC	714	614
BIC	721	624
N	240	240

Примечание. В ячейках представлены нестандартизованные регрессионные коэффициенты; их стандартные ошибки даны в скобках. Звездочки соответствуют уровню значимости 5% (*) , 1% (**) и 0,1% (***).

Эффект взаимодействия: бразильцы в составе и качество поля

Мы предполагали, что мастерство бразильцев, играющих в составах некоторых российских команд, может приводить к росту результативности. Как показывают приведенные ниже модели, такой эффект был обнаружен для игр второго круга, когда команды встречаются в основном на хороших полях и только для количества голов гостей (но не хозяев) (*табл.* 5).

Таблица 5

РЕГРЕССИОННЫЕ МОДЕЛИ; ЗАВИСИМАЯ ПЕРЕМЕННАЯ – РЕЗУЛЬТАТИВНОСТЬ ГОСТЕЙ: ЭФФЕКТЫ КАЧЕСТВА ПОЛЯ И КОЛИЧЕСТВА БРАЗИЛЬСКИХ ФУТБОЛИСТОВ

	Модель		
	пошаговая (1)	парная (2)	парная (3)
Константа (смещение)	–0,074	0,100	–0,004
	(0,129)	(0,090)	(0,122)
Разница бюджетов (хозяева минус гости)	–0,004***	–0,004***	
	(0,001)	(0,001)	
Количество бразильцев в составе гостей	0,104*		0,111*
	(0,050)		(0,049)
Псевдо-R^2 *Нагелькерке*	0,260	0,221	0,056
Лог. правдоподобия	–159,824	–161,816	–169,557
AIC	326	328	343
BIC	334	333	349
N	120	120	120

Примечание. В ячейках представлены нестандартизованные регрессионные коэффициенты; их стандартные ошибки даны в скобках. Звездочки соответствуют уровню значимости 5% (*), 1% (**) и 0,1% (***).

Модель 1 – это результат пошагового отбора оптимальной модели для матчей, сыгранных во втором круге. Значимы две переменные: «разница стоимости составов» и «количество бразильцев в заявке команды». Если состав хозяев дороже состава гостей, это снижает результативность гостей, а количество бразильцев в составе – ее повышает. Модели 2 и 3 оценивались для проверки: во второй модели единственная независимая переменная – разница стоимости составов, а в третьей – количество бразильцев в составе. Размер и знаки регрессионных коэффициентов соответствуют ожиданиям и значениям, полученным в первой модели. Интересно, что и *AIC* и *BIC* для моделей 1 и 2 очень близки, но *AIC* чуть лучше для модели 1, а более строгий критерий *BIC* – для модели 2. Мы все же склонны выбрать модель 1, так как у нее лучше также и псевдо-R^2 Нагелькерке.

Пуассоновские регрессии и регрессии по методу наименьших квадратов

Мы предлагаем теперь сравнить результаты пуассоновской регрессии и обычной линейной регрессии. Как известно, линейная регрессия по методу наименьших квадратов строится на определенных допущениях. В частности, предполагается, что зависимая переменная распределена нормально; регрессионные остатки тоже распределены нормально и гомоскедастичны, т.е. их дисперсия не зависит от значений зависимой переменной и независимых переменных. Эти допущения, очевидно, не обоснованы в том случае, если зависимая переменная имеет распределение Пуассона.

Оценки коэффициентов очень близки; однако приведенные ниже графики остатков показывают, что для использованных зависимых переменных не выполняются допущения о нормальности и гомоскедастичности остатков, обязательные для использования простой модели линейной регрессии по методу наименьших квадратов. Преимущество пуассоновских моделей заключается в том, что для них не требуется соответствия указанным выше допущениям.

Таблица 6

СРАВНЕНИЕ ПУАССОНОВСКОЙ И ОБЫЧНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ

	Модель	
	Пуассоновская пошаговая	По методу наименьших квадратов
Константа (смещение)	–0,307**	0,775***
	(0,095)	(0,081)
Качество поля	0,435***	0,438***
	(0,113)	(0,106)
Разница стоимости составов	–0,004***	–0,004***
	(0,001)	(0,001)
Псевдо-R^2 *Нагелькерке*	0,222	
Лог. правдоподобия	–444	
AIC	895	
BIC	907	
N	352	
Скорректированный R^2		0,152

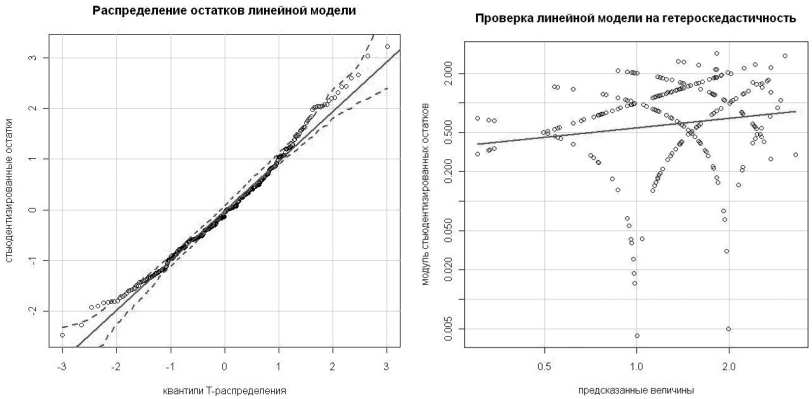

Рис. 2. Графики распределения остатков

Обсуждение результатов

Во-первых, оценки коэффициентов в пуассоновских моделях близки к аналогичным оценкам в обычной линейной регрессии, рассчитанной по методу наименьших квадратов, однако корректнее использовать именно пуассоновские модели, которые не требуют нормального распределения и гомоскедастичности остатков.

Во-вторых, количество зрителей (наиболее «социологическая» из использовавшихся в наших моделях независимых переменных) во всех наших моделях оказывается незначимым, т.е. не влияет на результативность. Это не совсем согласуется с заявлениями футболистов, что «играют они для зрителей». Вероятно, важнее региональные различия: в одних городах команды собирают почти полные стадионы, а в других – стадионы почти всегда пустые. К тому же, зрители могут ошибаться, предполагая, что игра будет результативной и интересной, когда покупают билеты. Кроме того, это может отражать специфику российского футбола, в котором многие команды финансируются на средства институциональных и частных спонсоров, не слишком заинтересованных в прибыльности футбольного бизнеса, которая требует привлечения зрителей на стадионы. На результативность также не влияет, какой именно из судей судил матч (ни одна из соответствующих бинарных переменных не оказалась значимой).

В-третьих, как и в любых регрессионных моделях проявились серьезные проблемы с мультиколлинеарностью. Мы решали здесь эти проблемы, удаляя те переменные во множественных моделях, которые были незначимы в парных моделях.

Итак, еще раз подчеркнем, что применение пуассоновских моделей рекомендуется, когда используются счетные, дискретные зависимые переменные, соответствующие распределению Пуассона. Такие модели могут заинтересовать не только социологов, но и политологов и сотрудников компаний, работающих в сфере

прикладных социальных и маркетинговых исследований (например, модели оценки вероятности покупок товаров и услуг).

Заключение

Мы показали, что пуассоновские модели могут применяться для моделирования результативности футбольных матчей. Финансовые возможности команд значимо предсказывают их результативность. Кроме того, был обнаружен интерактивный эффект: присутствие в составе команд бразильцев позитивно сказывается на результативности команды, но только на хороших полях.

Было показано преимущество пуассоновских моделей в сравнении с линейными моделями, рассчитанными по методу наименьших квадратов, поскольку пуассоновские модели корректны и в тех случаях, когда не выполняются допущения о нормальности и гомоскедастичности остатков.

ЛИТЕРАТУРА

1. *Carter Hill R., Griffiths W.E., Lim G.C.* Principles of Econometrics, 4th ed. N.Y.: John Wiley, 2011.

2. *Matsueda R.L., Kreager D.A., Huizinga D.* Deterring Delinquents: A Rational Choice Model of Theft and Violence // American Sociological Review. 2006. Vol. 71(1). P. 95–122.

3. *Maimon D., Kuhl D.C.* Social Control and Youth Suicidality: Situating Durkheim's Ideas in a Multilevel Framework // American Sociological Review. 2008. Vol. 73(6). P. 921–943.

4. *Kornrich S., Brines J., Leupp K.* Egalitarianism, Housework, and Sexual Frequency in Marriage // American Sociological Review. 2013. Vol. 78(1). P. 26–50.

5. *Толстова Ю.Н., Рыжова А.В.* Анализ таблиц сопряженности: использование отношения преобладаний и логлинейных моделей // Социология: методология, методы, математические модели. 2003. № 16.

6. *Трофимов Д.А.* Логлинейный анализ таблиц мобильности: обзор основных моделей // Социология: методология, методы, математическое моделирование. 2008. № 26.

7. *Ascher H., Feingold H.* Repairable Systems Reliability: Modelling, Inference, Misconceptions and Their Causes. N.Y.: Marcel Dekker, 1984.

8. *Crow L. H.* Reliability Analysis for Complex, Repairable Systems // Reliability and Biometry / Eds. F. Proschan, R.J. Serfling. Philadelphia: SIAM, 1974. P. 379–410.

9. *Bartoszynski R., Brown B. W., McBride C. M., Thompson J. R.* Some Nonparametric Techniques for Estimating the Intensity Function of a Cancer Related Nonstationary Poisson Process // The Annals of Statistics, 1981. P.1050–1060.

10. *Gail M.H., Santner T.J., Brown C.C.* An Analysis of Comparative Carcinogenesis Experiments Based on Multiple Times to Tumor // Biometrics. 1980. Vol. 36. P. 255–266.

11. *Heckman J.J., Singer B.* Social Science Duration Analysis // Longitudinal Analysis of Labour Market Data / Eds. J.J. Heckman, B. Singer. Cambridge, U.K.: Cambridge Univ. Press, 1985. P. 39–110.

12. *Tuma N.B., Hannan M.T.* Social Dynamics: Models and Methods. San Diego: Academic Press, 1984.

13. *Osgood D.W.* Poisson-Based Regression Analysis of Aggregate Crime Rates // Journal of Quantitative Criminology. 2000. Vol. 16. P. 21–43.

Приложение 1

Описательная статистика для независимых переменных

Таблица 7

АПРИОРНЫЕ НЕЗАВИСИМЫЕ ПЕРЕМЕННЫЕ
(информация, известная до начала матча)

Интервальные и порядковые переменные				
	Значение			Стандартное отклонение
	Min	Max	Среднее	
Бюджет команды хозяев (в млн долл. США)	21	165	59,8	39,3
Бюджет команды гостей (в млн долл. США)	21	165	59,8	39,3
Разница стоимости составов (хозяева минус гости в млн долл. США)	–204	+204	0,0	83,2
Количество зрителей на матче	1 500	58 600	12 900	7 600,6

Окончание табл. 7

Интервальные и порядковые переменные				
	Значение			Стандартное отклонение
	Min	Max	Среднее	
Количество бразильцев в составе команды в заявке на чемпионат – хозяева	0	5	1,6	1,6
Количество бразильцев в составе команды в заявке на чемпионат – гости	0	5	1,6	1,6

Таблица 8

АПРИОРНЫЕ НЕЗАВИСИМЫЕ ПЕРЕМЕННЫЕ
(информация, известная до начала матча), категориальные переменные

Показатель	Кодировка		
Хорошие или плохие поля	0 – плохие поля (март, апрель, ноябрь)	1 – хорошие поля (с мая по октябрь)	
Тренер хозяев – из России или СНГ или иностранец	0 – россиянин или из стран СНГ (68% команд)	1 – иностранец (32% команд)	
Тренер гостей – из России или СНГ или иностранец	0 – россиянин или из стран СНГ (68% команд)	1 – иностранец (32% команд)	
Круг	1 – первый круг	2 – второй круг	3 – третий и четвертый круг*
Судьи**			

Примечания.
* В первом и втором круге каждая команда играла по 15 матчей, в третьем и четвертом круге – команды были разбиты на две восьмерки по результатам первых двух кругов и играли по семь игр (7 в третьем, и 7 в четвертом) с командами из своей восьмерки.
** Закодированы 13 судей, каждый из которых судил не менее 17 матчей. В совокупности они обслужили 81% матчей, сыгранных в чемпионате.

Таблица 9

АПОСТЕРИОРНЫЕ НЕЗАВИСИМЫЕ ПЕРЕМЕННЫЕ
(информация, известная после окончания матча),
интервальные переменные

	Значение			Стандартное отклонение
	Min	Max	Среднее	
Удары по воротам – хозяева	0	24	12,2	4,4
Удары по воротам – гости	0	2	9,9	3,9
Удары в створ – хозяева	0	14	5,2	2,6
Удары в створ – гости	0	11	4,1	2,3
Хозяева – процент владения мячом	33	75	52	7,6
Гости – процент владения мячом	25	67	48	7,6
Хозяева – предупреждения и удаления*	0	10	2,3	1,5
Гости – предупреждения и удаления*	0	9	2,9	1,9

*Примечание.** Предупреждения кодировались как 1 балл, удаления – как 2 балла.