УДК 519.233.5, 51.76

ЛИНЕЙНАЯ И НЕЛИНЕЙНАЯ РЕГРЕССИЯ В БИОЛОГИИ И МЕДИЦИНЕ: ПОДХОД, ОСНОВАННЫЙ НА СКОЛЬЗЯЩЕМ СРЕДНЕМ

А. Н. Вараксин, Ю. В. Шалаумова, Т. А. Маслакова

Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия

В методической статье обсуждаются условия применимости линейной и нелинейной (логистической) регрессии. В качестве одного из методов проверки линейности связи между предиктором X и откликом Y в линейной регрессии и линейности связи предиктора X с логитом отклика Y в логистической регрессии предлагается подход, основанный на скользящем среднем. На конкретных примерах показаны преимущества скользящего среднего по сравнению с обычной стратификацией данных. Подчеркивается важность графического представления результатов скользящего среднего в линейной и (особенно) в логистической регрессии.

Ключевые слова: скользящее среднее; проверка линейности; линейная регрессия; логистическая регрессия.

1. Введение

Одним из условий применимости регрессионных моделей является линейность связи между предиктором *X* и откликом *Y* в линейной регрессии и линейность связи предиктора *X* с логитом отклика *Y* в логистической регрессии. Существуют разнообразные методы для проверки линейности связи, основанные как на визуальной оценке графического представления данных, так и на вычислительных процедурах [1, 2]. Однако далеко не всегда в научных исследованиях действительно проводится процедура проверки линейности, что может оказать влияние на достоверность выводов. Если нарушения условий применимости не устраняются должным образом, то результаты могут быть недостоверными, в частности, может снижаться эффективность статистических тестов и увеличиваться частота ошибок I и II рода [3].

Если обратиться к истории вопроса, то предпосылкой к появлению регрессионного анализа стала разработка метода наименьших квадратов, опубликованного А. М. Лежандром в 1805 г. и К. Ф. Гауссом в 1809 г. для определения орбит тел в Солнечной системе. Пионерной публикацией, включавшей термин «регрессия», стали материалы доклада Ф. Гальтона в журнале Nature 1885 г., описывающие регрессионную связь между ростом отцов и сыновей [4]. Более формальная трактовка этого понятия впервые была сделана К. Пирсоном в работе 1896 г. [5]. Современный регрессионный анализ в значительной мере базируется на публикациях Р. А. Фишера 1920-х гг., объединившего работы Гаусса и Пирсона [6]. После Фишера было разработано множество важных расширений регрессии, включая ло-

гистическую, непараметрическую, байесовскую регрессию и др. [4]. Как для линейной, так и для логистической регрессии одним из ключевых условий применимости статистической модели является линейность [1].

Данная работа посвящена описанию метода проверки линейности связи в линейных и логистических регрессионных моделях посредством подхода, основанного на скользящем среднем.

2. Линейная регрессия

Линейная регрессия применяется для описания статистической связи между двумя количественными переменными *Y* и *X*. Переменная *Y* в линейной регрессии называется объясняемой переменной, значения которой объясняются предсказывающей переменной *X*. Переменную *Y* называют еще откликом (outcome) на воздействие предикторной (предсказывающей) переменной *X*. Предикторную переменную *X* для краткости называют «предиктор» [7, 8].

В первоначальной постановке задачи регрессионного анализа предполагалось, что предиктор *X* принимает ряд фиксированных значений, для каждого из которых имеется некоторое множество значений отклика *Y* (рис. 1а); это так называемый fixed-*X* case [9]. Именно для такого случая регрессионная модель формулировалась в виде:

$$E(Y \mid X) = b_0 + b_1 X, \tag{1}$$

где символ *E* означает математическое ожидание значений *Y*, принадлежащих данному *X*. В практической работе с конкретными данными операция математического ожидания означает расчет среднего значения *Y* при заданном *X*.

Ситуация, изображенная на рис. 1а, чаще соответствует случаю планируемых экспериментов, когда исследователь имеет возможность задавать конкретные фиксированные значения предиктора *X*, исходя из поставленной задачи. Это область биологических экспериментов или клинических испытаний [11–13]. Совсем другая ситуация имеет место при проведении так называемых обсервационных исследований, когда исследователь использует существующие данные, не имея возможности их изменять или на них влиять [14]. К этой категории относятся натурные исследования в биологии или эпидемиологические исследования в медицине [15, 16]. При анализе результатов обсервационных исследований данные для линейного регрессионного анализа имеют вид, изображенный на рис. 1б: предиктор *X* принимает произвольный ряд значений, и для каждого значения *X* имеется (как правило) одно значение отклика *Y*, это так называемый variable-*X* case [9].

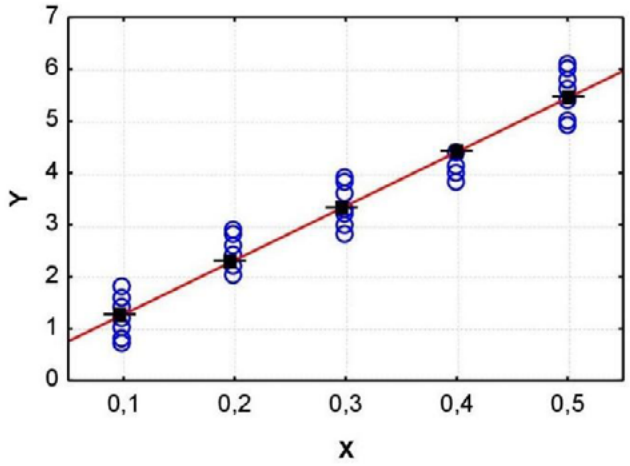

Рис. 1а. Пример зависимости *Y*(*X*) для планируемого эксперимента. Кружки – отдельные наблюдения, квадраты – средние значения *Y* для данного *X*

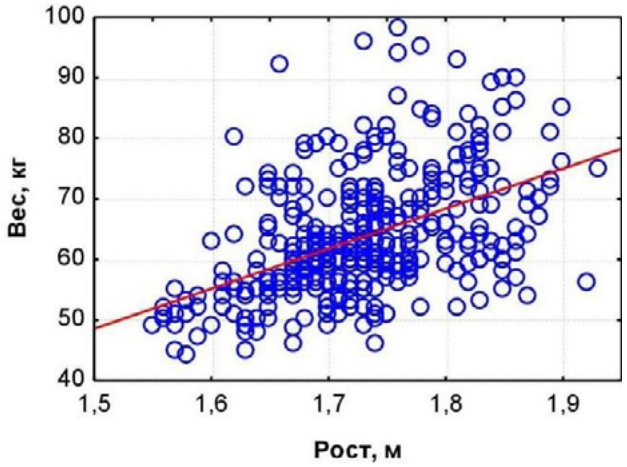

Рис. 1б. Соотношение веса и роста подростков [10]

Линейная регрессия (1) может быть использована только при выполнении ряда условий (условия применимости линейной регрессии):

А) средние значения отклика *Y* линейно связаны со значениями *X*;

Б) распределение значений *Y* для каждого *X* подчиняется нормальному закону $N(\mu_i, \sigma^2)$ со значениями математического ожидания μ_i для значений предиктора X_i и с одинаковыми дисперсиями σ^2 для каждого значения предиктора X_i [7, 8, 17].

Условие (А) относится к применимости модели (1): в линейной регрессии *Y* должен быть линейно связан с *X*. Условие (Б) относится к статистическим свойствам коэффициентов b_0 и b_1 модели (1). В настоящей работе речь идет только о проверке выполнения условия (А).

Ситуации на рис. 1а и 1б по-разному могут быть сопоставлены с условием применимости (А). В ситуации рис. 1а линейность связи может быть оценена визуально (экспертно) [9, 18]; статистическая значимость отклонения от линейности может быть проверена, например, с помощью корреляционного отношения [19]. Ситуация, изображенная на рис. 1б, не позволяет визуально (экспертно) оценить линейность связи *Y* и *X* в той мере, в какой это возможно для ситуации рис. 1а.

В принципе, и здесь возможно применение корреляционного отношения; отметим лишь, что идеология корреляционного отношения предполагает разделение значений предиктора *X* на интервалы (страты), в каждом из которых рассчитываются средние значения *Y* и *X*. После такой стратификации данные рис. 1б становятся похожими на данные рис. 1а. В отличие от данных рис. 1а, где разделение на страты происходит в процессе планирования эксперимента, стратификация данных рис. 1б – субъективный (неоднозначный) выбор биостатистика. Отметим, что экспертная (визуальная) оценка специалистом каких-либо статистических гипотез не является чем-то второстепенным (по сравнению со статистической проверкой гипотез), что подтверждается рекомендациями известных статистиков [9, 18].

Кроме того, в известных статистических пакетах (Statistica, SPSS) не было найдено аналогов проверки линейности связи *Y* с *X* для данных типа рис. 1б через процедуру расчета корреляционного отношения с последующим применением скользящего среднего [20–22]. Анализ научных публикаций последнего десятилетия (2012–2022 гг.) показал лишь небольшое количество работ, содержащих описание линейной регрессии, в которых авторы сообщали о проверке линейности связи

Y и *X* с помощью корреляционного отношения. Запрос на сервисе полнотекстового поиска по научной литературе Google Scholar, включающий термины (и их синонимы, приведенные в скобках): linear regression и linearity test (test of linearity, linearity check, check of linearity, linearity validation, validation of linearity, linearity verification, verification of linearity), показал в результирующей выдаче от 14 до 7 960 документов, в зависимости от формулировки. При добавлении в поисковый запрос термина correlation ratio количество документов варьировалось от 0 до 10. Всего было найдено 13 статей, включавших все три понятия (linear regression, linearity test, correlation ratio), и только в 5 из них корреляционные отношения использовались для проверки линейности связи. Более того, во многих публикациях не проводилась даже визуальная (экспертная) оценка линейности связи *Y* и *X* с помощью диаграмм типа рис. 1б [23–25]; при этом в публикациях [23, 25] приводятся и анализируются значения коэффициентов линейной регрессионной модели (или коэффициентов корреляции Пирсона между *Y* и *X*), которые по определению имеют смысл только при наличии линейной связи типа (1). Между тем в некоторых случаях простейший анализ диаграммы рассеяния может показать нелинейность связи *Y* и *X* и необходимость преобразования переменных для использования линейной регрессии [26].

Внешняя схожесть рис. 1а и 1б не должна вводить в заблуждение: на рис. 1а значения *X* осознанно выбраны для данного эксперимента и отражают объективную реальность. При разделении на страты данных рис. 1б количество страт и их границы – это субъективный выбор биостатистика, осуществляющего анализ данных; очевидно, что различные способы стратификации будут давать различные результаты при проверке линейности связи *Y* и *X*.

Тем не менее математические методы построения линейных регрессионных моделей типа (1) в обоих случаях, изображенных на рис. 1а и рис. 1б, одинаковы (метод наименьших квадратов, метод наибольшего правдоподобия). Для конкретных экспериментальных данных, состоящих из пар значений (x_i, y_i), модель линейной регрессии записывается в виде [8]:

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \qquad (2)$$

где x_i – значение предиктора *X* для *i*-го объекта исследования,

y_i – соответствующее значение отклика *Y*,

b_0 и b_1 – коэффициенты модели линейной регрессии,

ε_i – ошибки интерполяции (модели).

Используя процедуры нахождения коэффициентов модели (2) – метод наименьших квадратов, или метод максимального правдоподобия, получаем расчетные $\hat{y}_i$ значения отклика *Y* в виде:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 \cdot x_i.$$

В рамках линейных моделей (1), (2) изменение *Y* при изменении *X* описывается одним параметром $\hat{b}_1$ для любого *X* (приращение *Y* при изменении *X* на единицу равно $\hat{b}_1$ вне зависимости от значения *X*). При любых других формах связи *Y* и *X*, отличных от (1), (2), изменение *Y* при изменении *X* оказывается разным для разных *X*.

Достаточным (но не обязательно необходимым) условием линейности *Y* по *X* является двумерная нормальность распределения значений *Y* и *X*. Доказать это на реальных биологических или медицинских данных строго не представляется воз-

можным. A. Afifi с соавторами [9] предлагают визуальный способ оценки двумерной нормальности – на диаграмме рассеяния в координатах *X*–*Y* (как на рис. 1б) точки должны заполнять эллипс. Иногда довольствуются проверкой нормальности распределения отдельно *Y* и *X* [7]. Тем не менее проблема установления (проверки) линейности связи *Y* и *X* остается актуальной.

В настоящей работе для проверки условия (А) линейности связи *Y* и *X* для данных типа рис. 1б предлагается использовать методы скользящего среднего. Как будет показано, для данных регрессии типа рис. 1б скользящее среднее позволяет выявить линейность связи *Y* и *X* с применением минимального количества субъективных факторов.

Скользящее среднее для независимых наблюдений

Методы скользящего среднего широко применяются в регрессионном анализе [17, 27]. Наибольшее применение методы скользящего среднего получили при анализе временных рядов, где они используются для сглаживания случайных колебаний изучаемого фактора и выявления трендов [17, 18]. В данной работе методы скользящего среднего применяются для данных независимых наблюдений (не временных рядов) и используются для решения «обратной» (по сравнению с временными рядами) задачи: задачи выявления «структуры» связи между переменными, которая в первичных данных обычно скрыта большой дисперсией обоих переменных *X* и *Y*. Скользящее среднее позволяет уменьшить дисперсию переменных, в результате чего возможная связь между *Y* and *X* становится более явной.

Рассмотрим процедуру построения скользящего среднего для первичных данных, представляющих независимые наблюдения (не временные ряды). Пусть имеются пары значений (x_i, y_i) для *n*-наблюдений переменных *X* и *Y* (i = 1, 2, 3, …, n). Для построения скользящего среднего первичные значения предиктора *X* (x_1, x_2, …, x_n) упорядочиваются по возрастанию, после чего область изменения *X* разделяется на страты; в каждой страте рассчитываются средние значения *Y* и *X*.

В первой страте средние значения предиктора $<X(1)>$ и отклика $<Y(1)>$ определяются формулами:

$$<X(1)> = \frac{1}{n_w}\sum_{i=1}^{n_w} x_i, \quad <Y(1)> = \frac{1}{n_w}\sum_{i=1}^{n_w} y_i, \tag{3}$$

где индекс «1» у средних значений означает номер страты;

n_W – размер «окна» усреднения (window).

Таким образом, в первую страту попадают n_W объектов с минимальными значениями предиктора *X*. Во второй страте средние значения $<X(2)>$ и $<Y(2)>$ определяются формулами:

$$<X(2)> = \frac{1}{n_w}\sum_{i=n_s+1}^{n_w+n_s} x_i, \quad <Y(2)> = \frac{1}{n_w}\sum_{i=n_s+1}^{n_w+n_s} y_i, \tag{4}$$

где n_S – шаг (step) при построении скользящего среднего.

Таким образом, вторая страта получается путем сдвига начала отсчета на величину n_S. Средние значения остальных страт определяются аналогично при сдвиге на n_S для каждой новой страты. В результате первичные значения *X* и *Y* (x_1, x_2, …, x_n; y_1, y_2, …, y_n, всего *n*-значений) заменяются на набор средних значений $<X(N_s)>$

и $<Y(N_s)>$ показателей X и Y в стратах (N_S – номер страты). При этом разброс индивидуальных значений переменных X и Y уменьшается и характер связи между ними становится более явным.

Применение скользящего среднего к данным линейной регрессии

В качестве примера применения скользящего среднего вернемся к данным рис. 1б (статистическая связь веса и роста). На рис. 2а и рис. 2б показаны данные скользящего среднего для показателя «вес» по предиктору «рост» с размером окна $n_W = 25$ наблюдений и с шагом скользящего среднего, равного $n_S = 25$ (получается 31 среднее значение веса и роста) и $n_S = 1$ (351 среднее значение). Для расчета скользящих средних в среде Statistica for Windows использована компьютерная программа [28].

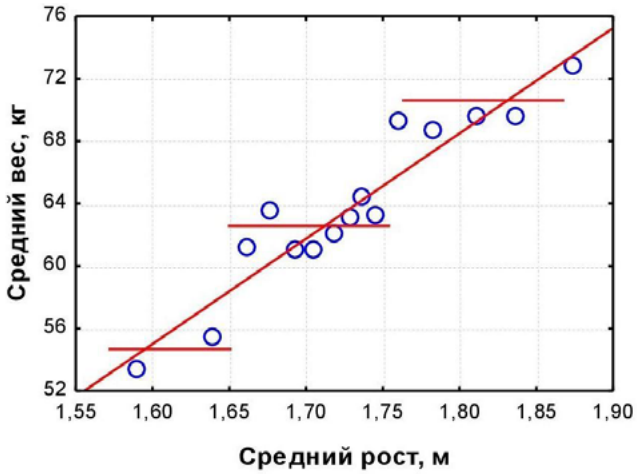

Рис. 2а. Соотношение средних значений веса и роста подростков. Окно усреднения равно $n_W = 25$, шаг $n_S = 25$

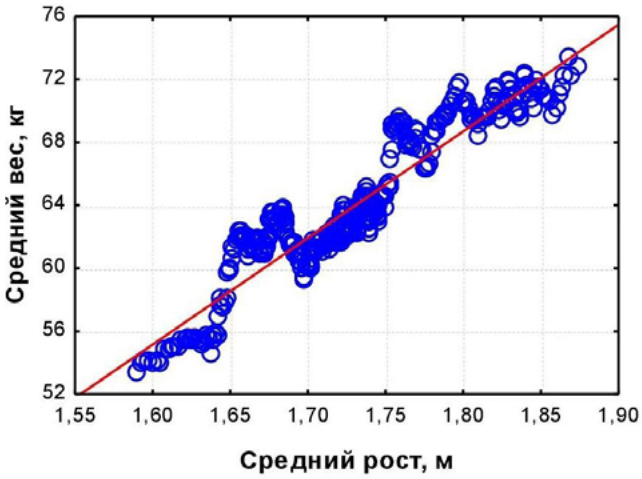

Рис. 2б. Соотношение средних значений веса и роста подростков. Окно усреднения равно $n_W = 25$, шаг $n_S = 1$

На рис. 2а значения предиктора X (рост юношей из рис. 1б) разделены на непересекающиеся страты размером 25 наблюдений, в каждой из которых рассчитаны средние значения роста и веса. Рис. 2а в какой-то мере похож на рис. 1а (для одного значения роста имеем одно *среднее* значение веса), хотя значения роста не расположены здесь так регулярно, как на рис. 1а. Рис. 2б наглядно показывает, что точной линейной связи средних значений веса и роста нет; более того, на рис. 2б можно видеть три области средних значений роста (это области значений роста до 165 см, затем от 165 до 175 см, и наконец область выше 175 см), в каждой из которых средние значения веса слабо меняются при изменении роста, при этом имеются значительные различия веса между областями. Таким образом, рис. 2б показывает более подробную (чем рис. 2а) картину связи средних значений роста и веса. Среди 351 точки на рис. 2б есть, очевидно, та 31 точка, которая показана на рис. 2а.

Скользящее среднее (рис. 2а) дает более детальную картину связи X со средними значениями Y по сравнению с первичными данными (рис. 1б), но уравнения регрессии в обоих случаях похожи (что и следовало ожидать): связь веса и роста на рис. 1б описывается соотношением:

$$\text{вес(кг)} = -50{,}6 + 0{,}661 \cdot \text{рост(см)},$$

а на рис. 2а – соотношением:

$$\text{вес(кг)} = -53{,}0 + 0{,}675 \cdot \text{рост(см)}.$$

Результаты применения процедуры скользящего среднего к реальным биологическим или медицинским данным зависят от выбора размера «окна» усреднения n_W и от выбора шага n_S. Что касается выбора окна n_W, мы предложили подход с использованием кумулятивной функции для Y по X (подробности см. в [29]). Шаг усреднения n_S можно в любом случае выбирать равным единице, как на рис. 2б (получаем максимум информации о средних значениях Y и X). Для «укрупненного» анализа размер шага скользящего среднего n_S можно увеличить, как показано на рис. 2а. В результате такого выбора n_W и n_S неоднозначность результатов скользящего среднего практически устраняется.

3. Логистическая регрессия

Одной из наиболее распространенных моделей нелинейной регрессии является логистическая регрессия. Логистическая регрессия применяется для описания статистической связи между количественным предиктором X и дихотомическим откликом Y (Y принимает два значения: $Y = 0$ и $Y = 1$). Данные такого рода часто встречаются в эпидемиологических исследованиях, когда дихотомический Y кодирует наличие или отсутствие некоторого заболевания, а X – фактор риска возникновения заболевания. Обычно считается, что $Y = 1$ кодирует наличие заболевания, а $Y = 0$ – его отсутствие у конкретного пациента.

В модели логистической регрессии статистическая связь между Y и X предполагается в виде:

$$W(Y=1|X) = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}, \tag{5}$$

где $W(Y = 1 \mid X)$ – вероятность обнаружения в экспериментальных данных значения $Y = 1$ для заданного значения X [9, 30].

При использовании соотношения (5) имеем

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X. \tag{6}$$

Таким образом, как и в случае линейной регрессии, имеем линейную связь, но не между X и Y, а между предиктором X и комплексом $\log\left(\frac{W}{1-W}\right)$, который называют логитом (Y), это условие применимости логистической регрессии. Каких-либо ограничений на тип предиктора X обычно не предъявляется [30]; напомним, что в линейной регрессии для получения нужных статистических свойств коэффициентов регрессии требуется нормальность распределения Y и X.

Известно достаточное (но не обязательно необходимое) условие, при котором выполняется линейное соотношение (6): распределение X в группах $Y = 0$ и $Y = 1$ подчиняется нормальному закону с равными дисперсиями [31]. В остальных случаях линейность связи X с логитом (Y) необходимо проверять.

Аналогично случаю линейной регрессии при выполнении условия линейности логита изменение логита (Y) при изменении X описывается одним параметром, оди-

наковым для любого X. Это параметр – отношение шансов OR, который для изменения X на единицу рассчитывается по формуле:

$$\mathrm{OR} = \exp(b_1),$$

где b_1 – коэффициент модели (6).

На рис. 3а представлены типичные первичные данные для логистической регрессии (количественный X и дихотомический Y), взятые из монографии [30]. Здесь отклик Y представляет заболевание сердечно-сосудистой системы (ССС) у пациентов данного возраста: $Y = 0$ означает отсутствие заболевания, $Y = 1$ – наличие заболевания у данного пациента. Очевидно, что данные типа рис. 3а не позволяют оценить выполнение соотношения (6) даже в той минимальной мере, в какой это возможно для линейной регрессии для данных рис. 1б. Действительно, наличие у отклика Y всего двух значений (нуль и единица) исключает всякую возможность экспертной (визуальной) оценки формы связи Y и X. Кривая логистической регрессии на рис. 3а проведена согласно модели (5), однако нет никакой уверенности, что вероятность иметь заболевание ССС описывается именно такой кривой, если опираться на визуальный анализ нулей и единиц.

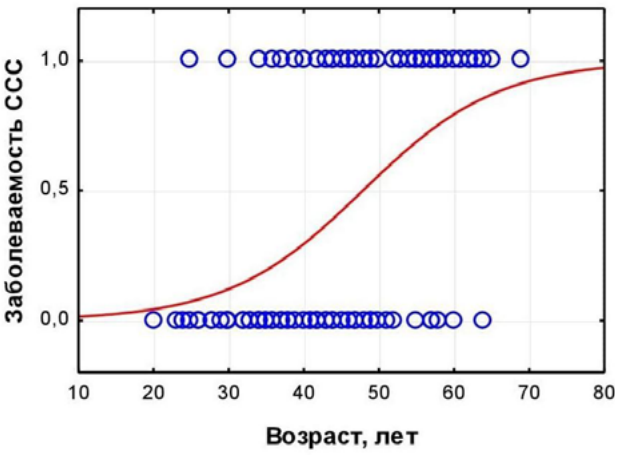

Рис. 3а. Распределение заболеваемости ССС в зависимости от возраста пациента [30]. Сплошная линия – линия логистической регрессии

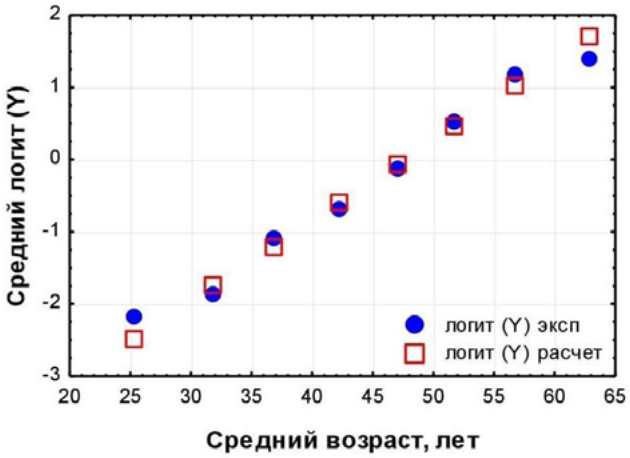

Рис. 3б. Стратификация 100 пациентов на 8 страт по возрастным группам [30]. Синие кружки – средние значения логита заболеваемости ССС и возраста в стратах. Красные квадраты – расчеты методом логистической регрессии

В отличие от линейной регрессии, в логистической регрессии уделяется внимание проверке линейности связи логита (Y) с X, т. е. выполнению соотношения (6). Многие известные статистические пакеты (SAS, SPSS) включают опции, аналогичные процедуре корреляционного отношения в линейной регрессии [20]. Эти опции предполагают разделение предиктора X на непересекающиеся страты (аналогично тому, как это делается при расчете корреляционного отношения в линейной регрессии) и расчет статистических критериев согласия для проверки гипотезы о линейности связи логита (Y) с X (критерий хи-квадрат Хосмера – Лемешова [30], а также индексы R^2 [32, 33]. Пользователи логистической регрессии активно пользуются критериями согласия, однако ограничиваются лишь определением значений критерия согласия и уровня значимости p без графического анализа полученных результатов (иногда данные стратификации представляют в виде таблицы средних значе-

ний *X* и логита (*Y*) в сравнении с результатами логистической регрессии, но без графического представления).

На рис. 3б показано сравнение данных логистической регрессии с данными стратификации предиктора *X*. Предиктор *X* (возраст пациента) в работе [30] был разделен на 8 страт, в каждой страте рассчитаны средние значения возраста и логита (*Y*), которые являются результатом усреднения первичных экспериментальных данных. На рис. 3б приведены также значения логита (*Y*), вычисленные в этих же стратах методом логистической регрессии. В результате перехода от вероятности *W* (как на рис. 3а) к логиту (*Y*) на рис. 3б логистическая кривая превращается в прямую. Согласно визуальной оценке экспериментальные данные для логита (*Y*) в целом показывают линейный рост с ростом возраста. Статистическая проверка гипотезы о линейности связи логита (*Y*) с возрастом это подтверждает: согласно критерию согласия хи-квадрат Хосмера – Лемешова $\chi^2 = 0{,}16$ при 6 степенях свободы, при котором гипотеза о линейности не отвергается на обычном уровне значимости $\alpha = 0{,}05$, $p > 0{,}05$.

На примере данных Хосмера – Лемешова (рис. 3б) можно продемонстрировать неоднозначность результатов проверки гипотезы о линейности логита, которые получаются при различных способах стратификации данных. Хосмер и Лемешов разделили пациентов на 8 непересекающихся страт по возрастным категориям (20–29 лет, 30–34 г. и т. д.), каждая страта содержит от 8 до 17 пациентов, т. е. неравное число наблюдений в страте. Если обратиться к известным статистическим пакетам, то в них предиктор *X* обычно разделяется на страты с равным числом наблюдений [34]. В примере Хосмера – Лемешова общее число пациентов равно 100, так что в каждой из 8 страт должно быть 12 или 13 пациентов. Результаты разделения *X* на 8 страт с равным числом пациентов представлены на рис. 4а; в некоторых точках заметно явное отличие от рис. 3б. Мы провели расчеты критерия согласия Хосмера – Лемешова для 8 страт с равным числом пациентов в них и обнаружили резкое увеличение критерия $\chi^2 = 1{,}87$ (это значение критерия χ^2 в 11 раз больше, чем у Хосмера – Лемешова). При таком значении χ^2 нулевая гипотеза о линейной связи логита (*Y*) с возрастом также не отвергается (из-за недостаточного числа пациентов), но вероятность отвергнуть нулевую гипотезу значительно повышается.

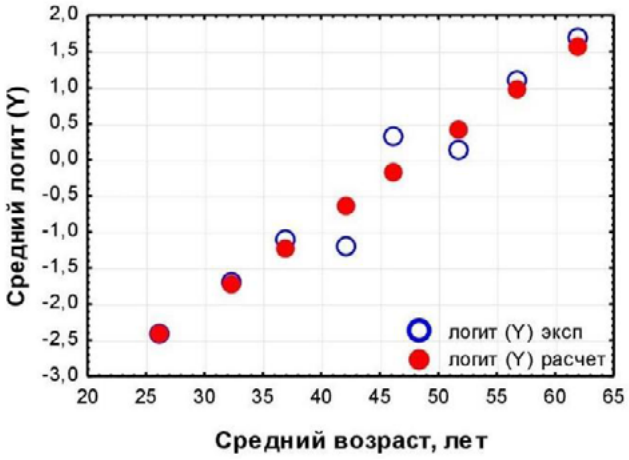

Рис. 4а. Стратификация 100 пациентов на 8 страт, равных по численности пациентов. Обозначения как на рис. 3б

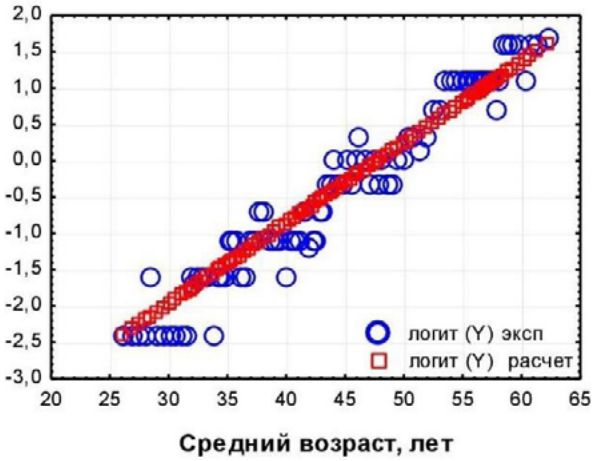

Рис. 4б. Средний логит вероятности иметь заболевание ССС по результатам скользящего среднего (окно $n_W = 12$, синие кружки). Красные квадраты – расчеты методом логистической регрессии

Поскольку экспериментальные данные прошли проверку на линейность связи логита с *X*, мы имеем право построить модель типа (6), описывающую связь логита вероятности *W* иметь заболевание ССС с возрастом:

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -5{,}309 + 0{,}1109 \cdot \text{возраст}. \quad (7)$$

Применение скользящего среднего к данным логистической регрессии

Применение скользящего среднего к данным логистической регрессии имеет особенность по сравнению с линейной регрессией – процедура скользящего среднего переводит дихотомическую переменную *Y* в количественную переменную. Рассмотрим несколько примеров применения скользящего среднего к реальным эпидемиологическим данным; каждый пример демонстрирует особенности данных и способы их выявления методами скользящего среднего.

Пример 1. Применим процедуру скользящего среднего к данным Хосмера – Лемешова. На рис. 4б показаны результаты скользящего среднего для отклика *Y* = заболевание ССС по предиктору «возраст» вместе с результатами логистической регрессии. Среди 89 точек на этом рисунке имеются 8 точек, которые есть на рис. 3б; кроме этого, здесь есть еще 81 точка для промежуточных значений возраста.

Данные рис. 4б можно использовать для построения модели связи логита вероятности заболевания ССС с возрастом. Из данных рис. 4б (синие точки скользящего среднего) получаем:

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -5{,}632 + 0{,}1179 \cdot \text{возраст}. \quad (8)$$

Очевидно, соотношение (8), полученное для данных скользящего среднего, практически не отличается от соотношения (7), полученного методом *логистической* регрессии. Этого следовало ожидать, поскольку график рис. 4б подтверждает справедливость соотношений типа (5), (6), которые являются условиями применимости логистической регрессии.

Пример 2. В предыдущем примере результаты стратификации в духе логистической регрессии хорошо согласуются с результатами скользящего среднего. На данном примере мы покажем отличия результатов скользящего среднего от результатов регрессии, притом что линейность логита в экспериментальных данных подтверждается критериями согласия.

В 2002 г. J. Peng с соавторами в журнале обучающей направленности опубликовали методическую статью [34] по применению логистической регрессии с подробным описанием тех процедур, которые необходимо выполнить при решении конкретной задачи, и тех показателей, которые необходимо отразить в сводке результатов. В частности, на примере конкретных эпидемиологических данных в этой работе была построена модель логистической регрессии для связи некоторого дихотомического отклика *Y* (*remedia*) с количественным предиктором *X* (*reading score*):

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -0{,}997 + 0{,}0281 \cdot \text{reading score}, \quad (9)$$

коэффициент регрессии при *X* оказался статистически значим (p = 0,0245 по t-критерию).

Были также рассчитаны критерии согласия для проверки линейности связи предиктора *X* с логитом (*Y*). Все критерии показали, что при разделении предиктора *X* на 10 групп (страт) с примерно одинаковым числом наблюдений гипотеза о линейности логита (*Y*) не отвергается.

В статье [34] не приведены графики связи средних значений предиктора *X* с логитом (*Y*), но они оказались достаточно показательными при их построении. На рис. 5а показаны результаты, полученные нами для этих данных после такой же стратификации, как и в оригинальной статье [34]. Видно, что логит (*Y*) при увеличении *X*, конечно, имеет тенденцию к уменьшению, но явной линейности здесь не наблюдается. Кроме того, расчетные данные во многих стратах весьма далеки от эксперимента (велик разброс экспериментальных значений около линии логистической регрессии).

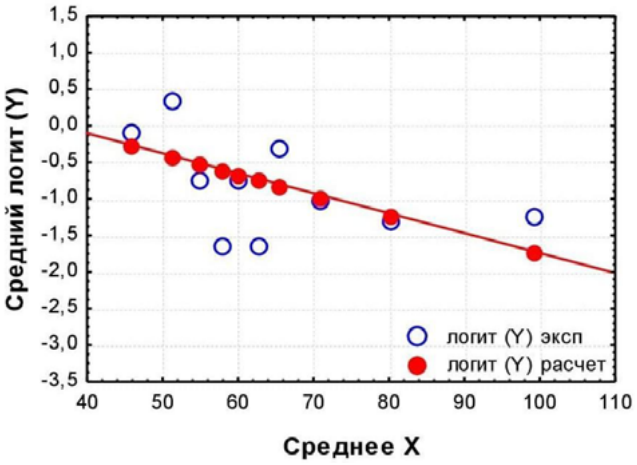

Рис. 5а. Красные кружки – расчетные данные логистической регрессии, синие кружки – результат стратификации на 10 страт [34]

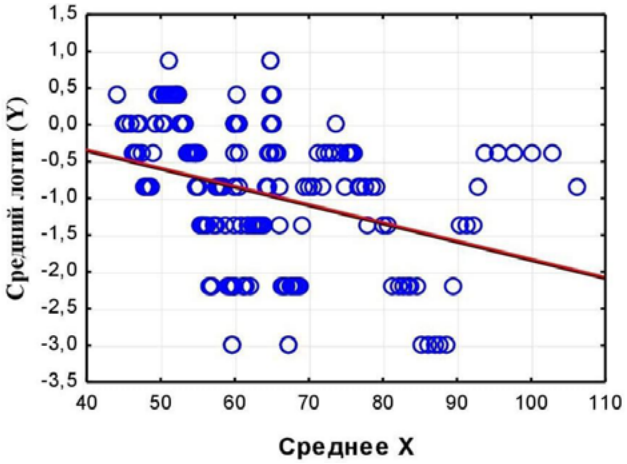

Рис. 5б. Кружки – результат скользящего среднего (эксперимент), сплошная линия – линейная интерполяция

Применим к данным [34] процедуру скользящего среднего, результат показан на рис. 5б. Этот график заставляет еще сильнее усомниться в строгой линейности связи между логитом (*Y*) и количественным предиктором *X*: на рис. 5б видны явные максимумы средних значений логита (*Y*) при *X*, равных 50, 62, 74 и 100, и минимумы логита при *X*, равных 60, 68 и 85–90. Особенно четко нелинейность логита (*Y*) видна для значений *X*, превышающих 65. Связь экспериментальных значений логита (*Y*) с *X* на рис. 5б в чем-то похожа на таковую на рис. 5а, но рис. 5б более подробный. В частности, размах значений логита (*Y*) на рис. 5б (от – 3 до + 1) гораздо больший, чем на рис. 5а (от – 1,7 до + 0,4).

Кажется, что данные логистической регрессии (сплошная линия на рис. 5а) кардинально отличаются от данных скользящего среднего (рис. 5б). Если, однако, аппроксимировать данные рис. 5б прямой линией, получается следующее соотношение:

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -0{,}631 + 0{,}0248 \cdot \text{reading score}, \qquad (10)$$

где значение коэффициента при *X*, равное $b_1 = -0{,}0248$, оказывается близко к соответствующему значению $b_1 = -0{,}0281$ в логистической регрессии для первичных данных в формуле (9). Таким образом, данные скользящего среднего (10) согласу-

ются с данными логистической регрессии (9), но скользящее среднее позволяет выявить более детально «структуру» связи логита (*Y*) с *X*.

Пример 3. В предыдущем примере показано, что графическое представление данных в логистической регрессии позволяет оценить возможную нелинейность связи логита (*Y*) с предиктором *X* даже в случае подтверждения линейности с помощью критериев согласия. В данном примере показано, что даже при выполнении всех условий применимости логистической регрессии результаты логистической регрессии могут *кардинально* отличаться от первичных данных. Это отличие удается установить только благодаря процедуре скользящего среднего. При отсутствии анализа скользящего среднего логистическая регрессия дает некорректный результат, который не согласуется с реальными данными.

Для построения модели использован эпидемиологический материал, включающий 100 женщин в возрасте старше 60 лет, для которых определялись антропометрические показатели и распространенность различных соматических патологий [35]. В данном примере описывается связь распространенности заболеваний щитовидной железы (ЩЖ) с индексом массы тела (ИМТ) женщин. Заболевания ЩЖ кодируются символом 0, если заболевания у данной женщины нет, и 1, если заболевание есть.

Нами был рассчитан критерий согласия Хосмера – Лемешова для проверки линейности связи предиктора *X* (ИМТ) с логитом вероятности иметь заболевание ЩЖ, который показал, что при разделении предиктора ИМТ на 10 страт с одинаковым числом наблюдений гипотеза о линейности логита не отвергается.

Применение метода логистической регрессии дает следующие результаты. Соотношение типа (6) для данной задачи имеет вид (*W* – вероятность иметь заболевание ЩЖ):

$$\log\left(\frac{W}{1-W}\right) = b_0 + b_1 \cdot X = -0{,}319 + 0{,}0111 \cdot \text{ИМТ}. \qquad (11)$$

Коэффициент регрессии b_1 статистически значимо не отличается от нуля, $p = 0{,}877$. Таким образом, согласно результатам логистической регрессии не существует значимой статистической связи между распространенностью заболеваний щитовидной железы с индексом массы тела женщин.

Сказанное позволяет сделать вывод: метод логистической регрессии использован при выполнении условий применимости, но значимой связи между *W*(ЩЖ) и ИМТ не обнаружено.

Проведем графический анализ связи *W*(ЩЖ) с ИМТ. На рис. 6а показаны результаты стратификации данных на 10 групп по 10 пациентов в каждой, а также расчетные данные методом логистической регрессии. Видно, что экспериментальные данные во многих случаях весьма далеки от расчетных данных.

Теперь применим к первичным данным процедуру скользящего среднего по ИМТ. На рис. 6б представлена связь логита вероятности *W*(ЩЖ) с ИМТ после выполнения операции скользящего среднего. Экспертная оценка данных рис. 6б показывает, что связь логита *W*(ЩЖ) иметь заболевание щитовидной железы с ИМТ у женщин существует, однако она не только не является линейной (как того требует логистическая регрессия), но даже не является монотонной: в интервале значений ИМТ от 22 до 26 кг/м^2 логит *W*(ЩЖ) в среднем уменьшается при увеличении ИМТ, а в интервале ИМТ от 26 до 30 кг/м^2 логит *W*(ЩЖ) возрастает.

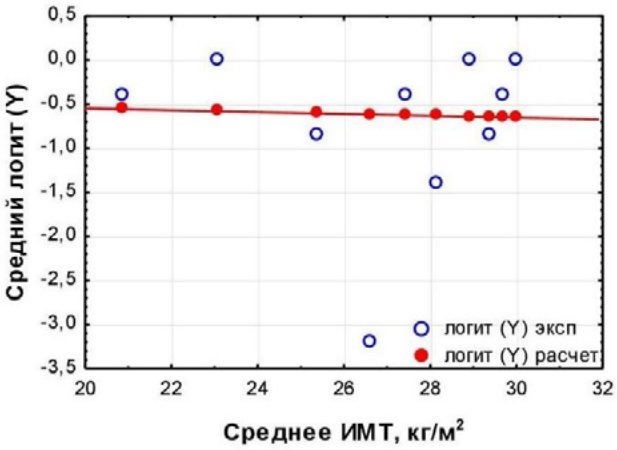

Рис. 6а. Связь логита W(ЩЖ) с ИМТ для данных 10 страт. Синие кружки (эксперимент) – средние значения логита W(ЩЖ), красные кружки – расчет методом логистической регрессии

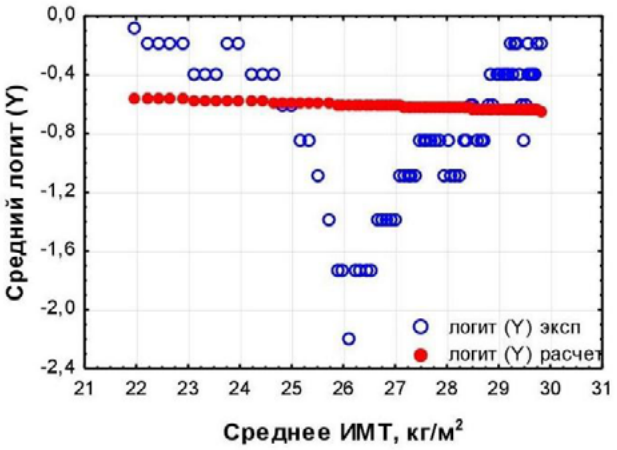

Рис. 6б. Связь логита W(ЩЖ) с ИМТ для данных 81 страты скользящего среднего. Синие кружки (эксперимент) – средние значения логита W(ЩЖ), красные кружки – расчет методом логистической регрессии

Проведем анализ статистической значимости различий *W*(ЩЖ) по данным скользящего среднего для различных значений ИМТ. Возьмем первую страту по ИМТ на рис. 6б. В этой страте среднее значение ИМТ = 21,90 кг/м^2, распространенность патологии ЩЖ в этой страте максимальна и равна $W = 0{,}476$. Возьмем 25-ю страту по ИМТ, среднее значение ИМТ = 26,10 кг/м^2, распространенность патологии ЩЖ в этой страте минимальна и равна $W = 0{,}143$. Страты № 1 и 25 содержат по 20 наблюдений и не пересекаются. По двустороннему критерию сравнения долей $W = 0{,}476$ и $W = 0{,}143$ различаются статистически значимо, $p = 0{,}0195$. В последней страте № 81 (среднее значение ИМТ = 29,83 кг/м^2) распространенность патологии ЩЖ равна $W = 0{,}429$. Различие со стратой № 25 статистически значимо, $p = 0{,}0404$.

Сравним выводы, получаемые для данного эпидемиологического материала методом логистической регрессии и методами скользящего среднего. Применение логистической регрессии показывает, что экспериментальные данные согласуются с предположением о линейности логита, но получаемый коэффициент регрессии не отличается от нуля статистически значимо. Это означает, что статистически значимой связи *W*(ЩЖ) и ИМТ нет. Применение методов скользящего среднего дает противоположный результат: логит *W*(ЩЖ) является не линейной, а V-образной функцией ИМТ; имеет место статистически и предметно значимая связь между распространенностью заболеваний щитовидной железы и ИМТ у пожилых женщин. Таким образом, основываясь только на результатах логистического регрессионного анализа, можно было «не заметить» явную связь между *W*(ЩЖ) и ИМТ.

4. Заключение

Линейная и нелинейная (логистическая) регрессия, описывающая статистическую связь отклика *Y* с предиктором *X*, должна использоваться в области своей применимости. Основное условие применимости данных регрессионных моделей – линейность связи *Y* с *X* в линейной регрессии и линейность связи логита (*Y*) с *X* в логистической регрессии. При обнаружении линейной связи между *Y* и *X* изменение отклика при изменении предиктора может быть описано одним коэффициентом (одним числом) для всей области изменения предиктора *X*. Если же связь нелинейная,

изменение отклика описывается разными значениями при различных значениях *X*. Поэтому так важно знать, является ли связь между *Y* и *X* линейной или нет. Если показатели линейной связи применяются к нелинейным данным, выводы статистического анализа будут некорректны.

Процедура проверки линейности зачастую является обособленным этапом анализа данных, видимо, поэтому во многих публикациях авторы вообще не сообщают о результатах проверки линейности связи *X* и *Y*, если таковая проверка вообще проводилась (например, работы [23, 25]).

В этой ситуации метод скользящего среднего может показать характер связи *Y* и *X* более наглядно и явно, чем обычная диаграмма рассеяния (сравнение рис. 1б с рис. 2а). Вместе с тем даже элементарный анализ диаграммы рассеяния может показать явную нелинейность связи *Y* и *X* (например, [26]).

Что касается методов проверки линейности связи для логистической регрессии, такая проверка заложена во многих статистических пакетах, для чего используются различные статистические критерии: критерий хи-квадрат Хосмера – Лемешова [30], индексы R^2 [32, 33]. При этом во многих публикациях не сообщается о проверке линейности в логистической регрессии (например, [36, 37]).

В данной работе показано, что ориентация только на статистические критерии может привести к искажению выводов логистической регрессии. В частности, критерии согласия могут не отвергать гипотезу о линейности связи предиктора *X* с логитом (*Y*), а графический анализ показывает явно нелинейную связь (пример 3).

Активными пропагандистами графического представления и анализа результатов статистического анализа были такие известные специалисты, как Дж. Тьюки и Э. Сигел. Американский статистик, профессор нескольких университетов Э. Сигел в монографии [18] практически каждый вывод иллюстрирует графиками с подробными комментариями. Один из создателей современной науки анализа данных Дж. Тьюки в монографии [38] писал: «Графики, подчеркивающие лишь то, что нам уже известно, нередко не стоят места, которое они занимают. Графики, которые надо рассматривать с лупой, заставляют нас тратить понапрасну время и мало полезны. График имеет наибольшую ценность тогда, когда он вынуждает нас заметить то, что мы совсем не ожидали увидеть». Приведенные выше рис. 2, 4, 5 и особенно 6 как раз и показывают то, что мы не ожидали увидеть.

5. Список литературы

1. *Osborne, J. W.* Multiple Regression Assumptions / J. W. Osborne, E. Waters. – ERIC Digest, 2002.
2. *Park, C. C.* A linearity test statistic in a simple linear regression / C. C. Park, K. E. Lee // J. of the Korean Data and Information Science Society. – 2014. – Vol. 25, No. 2. – P. 305–315. – DOI 10.7465/jkdi.2014.25.2.305.
3. *Nimon*, *K. F.* Statistical assumptions of substantive analyses across the general linear model: a mini-review / K. F. Nimon // Front. Psychology. – 2012. – Vol. 3, No. 322. – DOI 10.3389/fpsyg.2012.00322.
4. *Dhakal, C. P.* Regression invented as statistics / C. P. Dhakal // International J. of Interdisciplinary Research and Innovations. – 2018. – Vol. 6, No. 2. – P. 1–5.
5. *Stanton*, *J. M.* Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors / J. M. Stanton // J. of Statistics Education. – 2001. – Vol. 9, No. 3. – DOI 10.1080/10691898.2001.11910537.

6. *Aldrich, J.* Fisher and regression / J. Aldrich // Statistical Science. – 2005. – Vol. 20, No. 4. – P. 401–417.
7. *Draper, N. R.* Applied Regression Analysis, 3rd ed. / N. R. Draper, H. Smith. – Wiley-Interscience Publication, 1998. – 736 p. – ISBN 978-0-471-17082-2.
8. *Afifi, A. A.* Statistical Analysis: A Computer Oriented Approach / A. A. Afifi, S. P. Azen. – New York : Academic Press, 1972. – 366 p. – ISBN: 9780120444502.
9. *Afifi, A. A.* Computer-aided multivariate analysis, 4th ed. / A. A. Afifi, S. May, V. Clark. – Chapman&Hall/CRC, 2003. – 512 p. – ISBN 978-1584883081.
10. An Analysis of the Links between Smoking and BMI in Adolescents: A Moving Average Approach to Establishing the Statistical Relationship between Quantitative and Dichotomous Variables / A. N. Varaksin, E. D. Konstantinova, T. A. Maslakova [et al.] // Children (Basel). – 2022. – Vol. 9, No. 2. – DOI 10.3390/children9020220. – EDN GKBQJD.
11. *Владимиров, В. Г.* Радиозащитные рецептуры. Оптимизация состава и механизм действия / В. Г. Владимиров, Г. А. Поддубский, Г. И. Разоренов. – Л. : Воениздат, 1988. – 144 с.
12. *Налимов, В. В.* Теория эксперимента / В. В. Налимов. – М. : Наука, 1971. – 208 с.
13. Are in vivo and in vitro assessments of comparative and combined toxicity of the same metallic nanoparticles compatible, or contradictory, or both? A juxtaposition of data obtained in respective experiments with NiO and Mn_3O_4 nanoparticles / I. A. Minigalieva, T. V. Bushueva, E. Fröhlich [et al.] // Food and Chemical Toxicology. – 2017. – Vol. 109, No. 1. – P. 393–404. – DOI 10.1016/j.fct.2017.09.032. – EDN ZGWVNH.
14. *Fletcher, R. H.* Clinical epidemiology: The essentials / R. H. Fletcher, S. W. Fletcher. – 5th ed. – Lippincott Williams & Wilkins, 2012. – 272 p. – ISBN 978-1451144475.
15. *Расина, Л. Н.* Статистический анализ эколого-физиологических данных в интерпретации техногенных воздействий / Л. Н. Расина, Н. А. Орехова, А. Н. Вараксин // Экологические системы и приборы. – 2011. – № 12. – С. 50–53. – EDN SHXDFX.
16. Effects of Environmental Radioactive Pollution on the Cardiovascular Systems of Ural Region Residents: A Comparative Study / E. Konstantinova, Y. Shalaumova, T. Maslakova [et al.] // International J. of Medical Research & Health Sciences. – 2018. – Vol. 7, No. 3. – P. 1–7.
17. *Mosteller, F.* Data analysis and regression. A second course in statistics / F. Mosteller, J. W. Tukey. – Addison-Westly Publishing Company, 1978. – 317 p. – ISBN 978-0201048544.
18. *Siegel, A. F.* Practical Business Statistics, 7th ed. / A. F. Siegel. – Academic Press, 2016. – 642 p. – ISBN 978-0128042502.
19. *Гмурман, В. Е.* Теория вероятностей и математическая статистика : учеб. 12-е изд. / В. Е. Гмурман. – М. : Издательство Юрайт, 2020. – 479 с. – (Профессиональное образование). – ISBN 978-5-534-00859-3. – EDN DIXQIX.
20. *Sweet, S. A.* Data Analysis with SPSS. A First Course in Applied Statistics, 4th ed. / S. A. Sweet, K. Grace-Martin. – Pearson, 2010. – 288 p. – ISBN-13: 978-0205019670.
21. *Боровиков, В.* STATISTICA: искусство анализа данных на компьютере. Для профессионалов / В. Боровиков. – СПб.: Питер, 2001. – 656 с. – ISBN 5-318-00302-8.
22. *Norušis, M. J.* SPSS 16.0 Guide to Data Analysis / M. J. Norušis. – Prentice Hall, 2008. – 653 p.

23. *Nannan*, *N.* Estimating completeness of birth registration in South Africa, 1996–2011 / N. Nannan, R. Dorrington, D. Bradshaw // Bull World Health Organ. – 2019. – No. 97. – P. 468–476. – DOI 10.2471/BLT.18.222620.
24. *Hoekstra, R.* Are assumptions of well-known statistical techniques checked, and why (not)? / R. Hoekstra, H. A. L. Kiers, A. Johnson // Front. Psychology. – 2012. – Vol. 3, No. 137. – DOI 10.3389/fpsyg.2012.00137.
25. *Dye, C.* Tuberculosis decline in populations affected by HIV: a retrospective study of 12 countries in the WHO African Region / C. Dye, B. G. Williams // Bull World Health Organ. – 2019. – No. 97. – P. 405–414. – DOI: 10.2471/BLT.18.228577.
26. *Byass, P.* Correlation between noncommunicable disease mortality in people aged 30–69 years and those aged 70–89 years / P. Byass // Bull World Health Organ. – 2019. – No. 97. – P. 589–596. – DOI 10.2471/BLT.18.227132.
27. *Урбах, В. Ю.* Биометрические методы. Статистическая обработка опытных данных в биологии, сельском хозяйстве и медицине / В. Ю. Урбах. – М. : Наука, 1964. – 416 с.
28. Свидетельство о государственной регистрации программы для ЭВМ № 2019612871 Российская Федерация. Программа для расчета скользящего среднего и проверки условий применимости моделей регрессии : № 2019611507 : заявл. 18.02.2019 : опубл. 04.03.2019 / Ю. В. Шалаумова, А. Н. Вараксин, В. Г. Панов; заявитель Федеральное государственное бюджетное учреждение науки Институт промышленной экологии Уральского отделения Российской академии наук (ИПЭ УрО РАН). – EDN MVJCBS.
29. Применение методов скользящего среднего для построения регрессионных моделей в медико-экологических исследованиях / А. Н. Вараксин, Ю. В. Шалаумова, Т. А. Маслакова [и др.] // Экологические системы и приборы. – 2020. – № 6. – С. 12–21. – DOI 10.25791/esip.06.2020.1159. – EDN XTBFAV.
30. *Hosmer, D.* Applied Logistic Regression, 2nd ed. / D. Hosmer, S. Lemeshow. – New York: Wiley, 2000. – 397 p. – ISBN 0-471-35632-8.
31. *Shoukri, M. M.* Statistical methods for health sciences, 2nd ed. / M. M. Shoukri, C. A. Pause. – CRC Press, 1998. – 384 p. – ISBN 978-0849310959.
32. *Cox, D. R.* The analysis of binary data, 2nd ed. / D. R. Cox, E. J. Snell. – London: Chapman and Hall, 1989. – 240 p. – ISBN 9780412306204.
33. *Nagelkerke, N. J. D.* A note on a general definition of the coefficient of determination / N. J. D. Nagelkerke // Biometrica. – 1991. – Vol. 78, No. 3. – P. 691–692. – DOI 10.1093/biomet/78.3.691.
34. *P*eng*, J.* An Introduction to Logistic Regression Analysis and Reporting / J. Peng, K. L. Lee, G. M. Ingersoll // J. of Educational Research. – 2002. – Vol. 96, No. 1. – P. 3–14. – DOI 10.1080/00220670209598786.– EDN EGUCMX.
35. Эрозия слизистой оболочки влагалища у женщин с хирургической коррекцией пролапса гениталий в постменопаузе / А. А. Михельсон, М. В. Лазукина, А. Н. Вараксин [и др.] // Лечение и профилактика. – 2020. – Т. 10, № 4. – С. 55–64. – EDN ZCTUDM.
36. Fluid balance, intradialytic hypotension, and outcomes in critically ill patients undergoing renal replacement therapy: a cohort study / J. A. Silversides, R. Pinto, R. Kuint [et al.] // Critical Care. – 2014. – Vol. 18, No. 6. – DOI 10.1186/s13054-014-0624-8.
37. Hearing and vision screening for preschool children using mobile technology, South Africa / S. Eksteen, S. Launer, H. Kuper [et al.] // Bull World Health Organ. – 2019. – No. 97. – P. 672–680. – DOI 10.2471/BLT.18.227876.
38. *Тьюки, Дж.* Анализ результатов наблюдений / Дж. Тьюки. – М. : Мир, 1981. – 695 с.

Сведения об авторах:

Вараксин Анатолий Николаевич, д. ф.-м. н., профессор, главный научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, ул. С. Ковалевской, 20, Россия. Эл. почта: varaksin@ecko.uran.ru.

Шалаумова Юлия Валерьевна, канд. т. н., научный сотрудник.

Маслакова Татьяна Анатольевна, канд. ф.-м. н., научный сотрудник.

LINEAR AND NONLINEAR REGRESSION IN BIOLOGY AND MEDICINE: A MOVING AVERAGE APPROACH

A. N. Varaksin, Yu. V. Shalaumova, T. A. Maslakova

Institute of Industrial Ecology, Ural Branch of Russian Academy of Sciences, Ekaterinburg, Russia

Testing of assumptions of linear and non-linear (logistic) regressions are discussed in this method paper. We propose a moving average approach as one of the methods for checking the linearity of the relationship between predictor X and response Y in linear regression and the linearity of the relationship between predictor X and the logit of response Y in logistic regression. Specific examples show the advantages of a moving average over the common data stratification approach. The importance of graphical representation of moving average results in linear and (especially) logistic regression is emphasized.

Key words: moving average; linearity test; linear regression; logistic regression.

References

1. *Osborne, J. W.* Multiple Regression Assumptions / J. W. Osborne, E. Waters. – ERIC Digest, 2002.
2. *Park, C. C.* A linearity test statistic in a simple linear regression / C. C. Park, K. E. Lee // J. of the Korean Data and Information Science Society. – 2014. – Vol. 25, No. 2. – P. 305–315. – DOI 10.7465/jkdi.2014.25.2.305.
3. *Nimon*, *K. F.* Statistical assumptions of substantive analyses across the general linear model: a mini-review / K. F. Nimon // Front. Psychology. – 2012. – Vol. 3, No. 322. – DOI 10.3389/fpsyg.2012.00322.
4. *Dhakal, C. P.* Regression invented as statistics / C. P. Dhakal // International J. of Interdisciplinary Research and Innovations. – 2018. – Vol. 6, No. 2. – P. 1–5.
5. *Stanton*, *J. M.* Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors / J. M. Stanton // J. of Statistics Education. – 2001. – Vol. 9, No. 3. – DOI 10.1080/10691898.2001.11910537.
6. *Aldrich, J.* Fisher and regression / J. Aldrich // Statistical Science. – 2005. – Vol. 20, No. 4. – P. 401–417.
7. *Draper, N. R.* Applied Regression Analysis, 3rd ed. / N. R. Draper, H. Smith. – Wiley-Interscience Publication, 1998. – 736 p. – ISBN 978-0-471-17082-2.
8. *Afifi, A. A.* Statistical Analysis: A Computer Oriented Approach / A. A. Afifi, S. P. Azen. – New York : Academic Press, 1972. – 366 p. – ISBN: 9780120444502.
9. *Afifi, A. A.* Computer-aided multivariate analysis, 4th ed. / A. A. Afifi, S. May, V. Clark. – Chapman&Hall/CRC, 2003. – 512 p. – ISBN 978-1584883081.

10. An Analysis of the Links between Smoking and BMI in Adolescents: A Moving Average Approach to Establishing the Statistical Relationship between Quantitative and Dichotomous Variables / A. N. Varaksin, E. D. Konstantinova, T. A. Maslakova [et al.] // Children (Basel). – 2022. – Vol. 9, No. 2. – DOI 10.3390/children9020220. – EDN GKBQJD.
11. *Vladimirov, V. G.* Radioprotective formulations. Optimization of composition and mechanism of action / V. G. Vladimirov, G. A. Poddubsky, G. I. Razorenov. – Leningrad : Voenizdat, 1988. – 144 p. (In Russian).
12. *Nalimov, V. V.* Theory of experiment / V. V. Nalimov. – Moscow : Nauka, 1971. – 208 p. (In Russian).
13. Are in vivo and in vitro assessments of comparative and combined toxicity of the same metallic nanoparticles compatible, or contradictory, or both? A juxtaposition of data obtained in respective experiments with NiO and Mn_3O_4 nanoparticles / I. A. Minigalieva, T. V. Bushueva, E. Fröhlich [et al.] // Food and Chemical Toxicology. – 2017. – Vol. 109, No. 1. – P. 393–404. – DOI 10.1016/j.fct.2017.09.032. – EDN ZGWVNH.
14. *Fletcher, R. H.* Clinical epidemiology: The essentials, 5th ed. / R. H. Fletcher, S. W. Fletcher. – Lippincott Williams & Wilkins, 2012. – 272 p. – ISBN 978-1451144475.
15. *Rasina, L. N.* Statistical analysis of ecological and physiological data in the interpretation of technogenic impacts / L. N. Rasina, N. A. Orekhova, A. N. Varaksin // Ecological systems and devices. – 2011, No. 12. – P. 50–53. – EDN SHXDFX. (In Russian).
16. Effects of Environmental Radioactive Pollution on the Cardiovascular Systems of Ural Region Residents: A Comparative Study / E. Konstantinova, Y. Shalaumova, T. Maslakova [et al.] // International J. of Medical Research & Health Sciences. – 2018. – Vol. 7, No. 3. – P. 1–7.
17. *Mosteller, F.* Data analysis and regression. A second course in statistics / F. Mosteller, J. W. Tukey. – Addison-Westly Publishing Company, 1978. – 317 p. – ISBN 978-020-1048544.
18. *Siegel, A. F.* Practical Business Statistics, 7th ed. / A. F. Siegel. – Academic Press, 2016. – 642 p. – ISBN 978-0128042502.
19. *Gmurman*, *V. E.* Probability theory and mathematical statistics: textbook, 12th ed. / V. E. Gmurman. – Moscow : Yurayt Publishing House, 2020. – 479 p. – (Professional education). – ISBN 978-5-534-00859-3. – EDN DIXQIX. (In Russian).
20. *Sweet, S. A.* Data Analysis with SPSS. A First Course in Applied Statistics, 4th ed. / S. A. Sweet, K. Grace-Martin. – Pearson, 2010. – 288 p. – ISBN-13: 978-0205019670.
21. *Borovikov, V.* STATISTICA: the art of computer data analysis. For professionals / V. Borovikov. – St. Petersburg: Piter, 2001. – 656 p. – ISBN 5-318-00302-8.
22. *Norušis, M. J.* SPSS 16.0 Guide to Data Analysis / M. J. Norušis. – Prentice Hall, 2008. – 653 p.
23. *Nannan*, *N*. Estimating completeness of birth registration in South Africa, 1996–2011 / N. Nannan, R. Dorrington, D. Bradshaw // Bull World Health Organ. – 2019. – No. 97. – P. 468–476. – DOI 10.2471/BLT.18.222620.
24. *Hoekstra, R.* Are assumptions of well-known statistical techniques checked, and why (not)? / R. Hoekstra, H. A. L. Kiers, A. Johnson // Front. Psychology. – 2012. – Vol. 3, No. 137. – DOI 10.3389/fpsyg.2012.00137.
25. *Dye, C.* Tuberculosis decline in populations affected by HIV: a retrospective study of 12 countries in the WHO African Region / C. Dye, B. G. Williams // Bull World Health Organ. – 2019. – No. 97. – P. 405–414. – DOI: 10.2471/BLT.18.228577.

26. *Byass, P.* Correlation between noncommunicable disease mortality in people aged 30–69 years and those aged 70–89 years / P. Byass // Bull World Health Organ. – 2019. – No. 97. – P. 589–596. – DOI 10.2471/BLT.18.227132.
27. *Urbah, V. Yu.* Biometric methods. Statistical processing of experimental data in biology, agriculture and medicine. – Moscow : Nauka, 1964. – 416 p. (In Russian).
28. Certificate of state registration of the computer program No. 2019612871 Russian Federation. Program for calculating the moving average and checking the conditions for the applicability of regression models: No. 2019611507: Appl. 02/18/2019 : publ. 03/04/2019 / Yu. V. Shalaumova, A. N. Varaksin, V. G. Panov; applicant Federal State Budgetary Institution of Science Institute of Industrial Ecology of the Ural Branch of the Russian Academy of Sciences (IIE Ural Branch of the Russian Academy of Sciences). – EDN MVJCBS. (In Russian).
29. Application of moving average methods for constructing regression models in medical and environmental studies / A. N. Varaksin, Yu. V. Shalaumova, T. A. Maslakova [et al.] // Ecological systems and devices. – 2020, No. 6. – P. 12–21. – DOI 10.25791/esip.06. 2020.1159. – EDN XTBFAV. (In Russian).
30. *Hosmer, D.* Applied Logistic Regression, 2nd ed. / D. Hosmer, S. Lemeshow. – New York: Wiley, 2000. – 397 p. – ISBN 0-471-35632-8.
31. *Shoukri, M. M.* Statistical methods for health sciences, 2nd ed. / M. M. Shoukri, C. A. Pause. – CRC Press, 1998. – 384 p. – ISBN 978-0849310959.
32. *Cox, D. R.* The analysis of binary data, 2nd ed. / D. R. Cox, E. J. Snell. – London: Chapman and Hall, 1989. – 240 p. – ISBN 9780412306204.
33. *Nagelkerke, N. J. D.* A note on a general definition of the coefficient of determination / N. J. D. Nagelkerke // Biometrica. – 1991. – Vol. 78, No. 3. – P. 691–692. – DOI 10.1093/ biomet/78.3.691.
34. *Peng, J.* An Introduction to Logistic Regression Analysis and Reporting / J. Peng, K. L. Lee, G. M. Ingersoll // J. of Educational Research. – 2002. – Vol. 96, No. 1. – P. 3–14. – DOI 10.1080/00220670209598786.– EDN EGUCMX.
35. Erosion of the vaginal mucosa in postmenopausal women with surgical correction of genital prolapse / A. A. Mikhelson, M. V. Lazukina, A. N. Varaksin [et al.] // Treatment and prevention. – 2020. – Vol. 10, No. 4. – P. 55–64. – EDN ZCTUDM. (In Russian).
36. Fluid balance, intradialytic hypotension, aand outcomes in critically ill patients undergoing renal replacement therapy: a cohort study / J. A. Silversides, R. Pinto, R. Kuint [et al.] // Critical Care. – 2014. – Vol. 18, No. 6. – DOI 10.1186/s13054-014-0624-8.
37. Hearing and vision screening for preschool children using mobile technology, South Africa / S. Eksteen, S. Launer, H. Kuper [et al.] // Bull World Health Organ. – 2019. – No. 97. – P. 672–680. – DOI 10.2471/BLT.18.227876.
38. *Tukey, J.* Analysis of observational results / J. Tukey. – Moscow : Mir, 1981. – 695 p.