УДК 519.233.5, 51.76

АДЕКВАТНОСТЬ РЕГРЕССИОННЫХ МОДЕЛЕЙ В ЭКОЛОГИИ И МЕДИЦИНЕ: ПЕРВИЧНЫЕ ДАННЫЕ, СТРАТИФИКАЦИЯ И СКОЛЬЗЯЩЕЕ СРЕДНЕЕ

А. Н. Вараксин [1], Ю. В. Шалаумова [2], Е. Д. Константинова [1], Т. А. Маслакова [1]

[1] Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия
[2] Институт экологии растений и животных УрО РАН, г. Екатеринбург, Россия

Рассмотрены подходы к установлению адекватности моделей линейной и логистической регрессии, применяемых в экологии и медицине. Традиционные методы установления адекватности моделей линейной регрессии, основанные на корреляционном отношении, практически не используются в реальных исследованиях. Недостаточное распространение в работах, опубликованных в периодической печати, получили методы визуальной (экспертной) оценки линейности связи отклика с предиктором. В данной работе в качестве метода установления адекватности моделей линейной регрессии предлагается использовать методы скользящего среднего. На конкретных примерах показана работоспособность данного подхода. Для моделей логистической регрессии, в дополнение к традиционным методам стратификации установления адекватности, предлагаются методы скользящего среднего и кумулятивной функции вероятности в сочетании с графическим представлением результатов. Эффективность данного подхода показана на примерах.

Ключевые слова: линейная регрессия; логистическая регрессия; адекватность моделей; стратификация; скользящее среднее; кумулятивная функция вероятности.

1. Введение

Линейные регрессионные модели – модели типа (1), линейные по коэффициентам b_i:

$$Y(x_1, x_2, x_3, \ldots, x_k) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_kx_k, \quad (1)$$

где Y – отклик, x_i (i=1,2,…, k) – набор предикторов [1–5].

Нелинейные регрессионные модели – нелинейные по коэффициентам b_i, например, модели типа (2):

$$Y(x_1, x_2, x_3, \ldots, x_k) = exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_kx_k), \quad (2)$$

где коэффициенты b_i стоят под знаком экспоненты [3, 4].

В данной работе речь идет об анализе линейных и нелинейных *однофакторных* регрессионных моделей, т. е. моделей с одним предиктором. Некоторые результаты для моделей *множественной* регрессии, полученные нами при анализе на основе первичных данных, стратификации и скользящего среднего, можно найти в [6, 7].

2. Терминология

Первичные данные – это данные, полученные в результате обследования (эксперимента) для каждого объекта исследования. Примеры: возраст, пол, индекс массы тела (ИМТ), уровень гемоглобина и т. д. каждого пациента.

Стратификация – разделение первичных данных на интервалы (страты). Например, возраст разделяют на страты 20–25 лет, 26–30 лет и т. д. В этих стратах можно рассчитать усредненные характеристики всех остальных показателей, например, соотношение мужчин и женщин (пол), средние значения ИМТ, гемоглобина и т. д. Параметры процедуры стратификации – число страт и их границы.

Скользящее среднее — это та же стратификация, только с перекрывающимися стратами. Например, возраст в первой страте от 20 до 25 лет; возраст во второй страте от 21 до 26 лет и т. д. В этих стратах скользящего среднего также можно рассчитать усредненные характеристики всех остальных показателей. Параметры процедуры скользящего среднего – «окно n_W и шаг n_S скользящего среднего». Окно n_W показывает, сколько наблюдений входит в страту, а шаг n_S – это сдвиг одной страты относительно другой. В приведенном выше примере со скользящим по возрасту шаг n_S равен одному году.

Фактические (наблюдаемые) данные (observed data). Термин – обобщение терминов «Первичные данные», «Данные стратификации», «Данные скользящего среднего». Мы будем пользоваться этим термином, когда необходимо отличать результаты, полученные с использованием первичных или преобразованных данных без применения статистических моделей, от результатов, полученных расчетным путем с использованием различных статистических моделей *(calculated data).*

3. Адекватность модели

Статистический анализ адекватности модели (верификация модели) – это различные процедуры сопоставления модельных заключений, оценок, следствий и выводов с реально наблюдаемой действительностью [8]. Это общий взгляд на проблему адекватности любых статистических моделей.

Одну из конкретных формулировок понятия адекватности для регрессионных моделей дают А. Афифи и С. Эйзен: «Адекватность модели означает, что никакая другая модель не дает значимого улучшения предсказания отклика» [5]. Понятие адекватности модели в таком виде четко указывает на необходимость сопоставления модели и первичных (наблюдаемых) данных. Поскольку «других» моделей может быть очень много, проверка адекватности конкретной модели может оказаться затруднительной. Более реалистичной с точки зрения практического применения статистических моделей выглядит позиция С. А. Айвазяна и соавторов. Говоря о критериях адекватности, авторы пишут [3]: «…они (критерии адекватности*. – Вставка наша*) не могут ответить на вопрос: является ли проверяемый гипотетический вид зависимости наилучшим, единственно верным? Они лишь либо подтверждают факт непротиворечивости проверяемого вида функции регрессии имеющимся у исследователя исходным данным, либо отвергают его». Именно с этих позиций (непротиворечивость модели первичным данным) будет в дальнейшем рассматриваться адекватность регрессионных моделей. Подчеркнем, однако, что «предельная» позиция А. Афифи и С. Эйзена [5] относительно понятия адекватности модели очень полезна для общего понимания термина «адекватность статистической модели».

4. Линейная регрессия

Рассмотрим вопросы адекватности моделей линейной регрессии. В работе речь идет о простой линейной регрессии, в которой отклик *Y* связывается с *одним* предиктором *X линейной* связью, причем *Y* и *X* – количественные переменные [1, 2]. В первоначальной постановке задачи регрессионного анализа предполагалось, что предиктор *X* принимает ряд фиксированных (неслучайных) значений, для каждого из которых имеется некоторое множество значений отклика *Y* (см. рис. 1а); это так называемый Fixed-*X* case [4,5]. Именно для такого случая регрессионная модель формулируется в виде:

$$E(Y|X) = b_0 + b_1 X, \qquad (3)$$

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

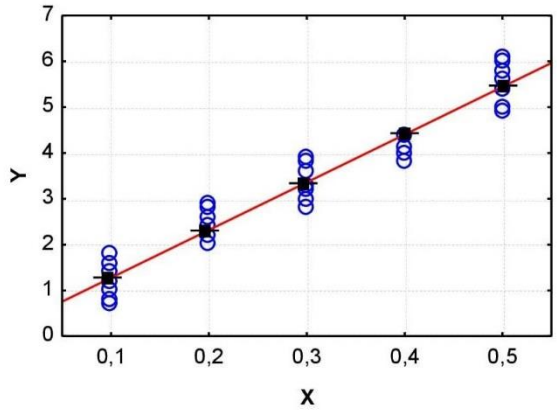

Рис. 1а. Иллюстративный рисунок. Открытые кружки – отдельные наблюдения, заполненные квадраты – средние значения *Y* для данного *X*

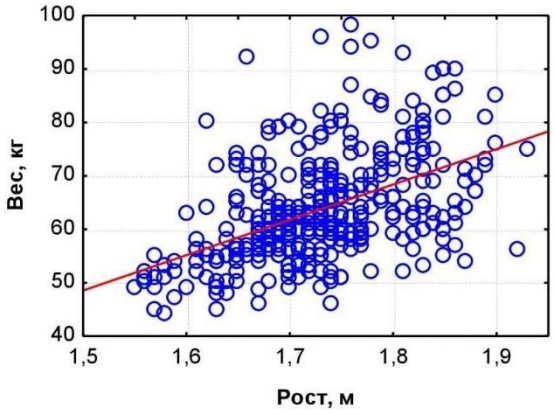

Рис. 1б. Реальные данные. Диаграмма рассеяния, показывающая соотношение веса и роста 375 подростков Челябинской области [9]

Ситуация, изображенная на рис. 1а, чаще соответствует случаю так называемых активных (планируемых) экспериментов, когда исследователь имеет возможность задавать конкретные фиксированные значения предиктора *X* исходя из поставленной задачи. Это область биологических экспериментов или клинических испытаний (см., например, [10, 11].

Совсем другая ситуация имеет место при проведении так называемых обсервационных исследований (пассивные эксперименты), когда исследователь использует существующие данные, не имея возможности их изменять или на них повлиять [8, 12]. К этой категории относятся натурные исследования в биологии или эпидемиологические исследования в медицине.

При анализе результатов обсервационных исследований предиктор *X* принимает ряд случайных значений, и для каждого значения *X* имеется (как правило) одно значение отклика *Y* – это так называемый Variable-*X* case. Данные для линейного регрессионного анализа при обсервационных исследованиях имеют характерный вид, изображенный на рис. 1б.

Прежде чем применять регрессионные методы для анализа данных, необходимо провести проверку *адекватности модели* реальным данным. Линейная регрессия типа (3) может быть использована только при выполнении ряда условий:

А) средние значения отклика *Y* линейно связаны со значениями *X* как в соотношении (3);

Б) распределение (*X*, *Y*) должно соответствовать двумерному нормальному распределению [1].

Условие А) относится к применимости модели (3): в линейной регрессии *Y* должен быть линейно связан с *X*. Условие Б) относится к статистическим свойствам коэффициентов b_0 и b_1 модели (3). В настоящей работе речь идет только о проверке выполнения условия А), т. е. об адекватности модели простой линейной регрессии в случае Variable-*X* case.

Математические методы построения линейных регрессионных моделей типа (3) в обоих случаях, изображенных на рис. 1а и рис. 1б, одинаковы (метод наименьших квадратов, метод максимального правдоподобия). Для конкретных экспериментальных данных, состоящих из пар значений (x_i, y_i), модель линейной регрессии записывается в виде [1, 2, 5]:

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \qquad (4)$$

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

модели. Применяя метод наименьших квадратов для получения значений коэффициентов, имеем:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i, \quad (5)$$

где $\hat{b}_0$ и $\hat{b}_1$ (обычно изображаются символами с крыжом [3]) – значения коэффициентов модели (4), полученные методом наименьших квадратов. При выполнении условия (5) *изменение Y* при изменении *X* описывается *одним* параметром b_1 для любого X. В этом случае приращение *Y* при изменении *X* на единицу равно b_1 для любого значения *X*. При любых других формах связи *Y* и *X* (например, связь *Y* с *X* описывается квадратичной формой, включающей *X* и X^2) *изменение Y* при изменении *X* оказывается разным для разных *X* [1–3]. Таким образом, для того чтобы иметь право описывать наблюдаемые данные одним коэффициентом b_1 модели (5), необходимо быть уверенным в выполнении соотношения (5). Об этом пишут многие известные статистики [1–5], однако в реальности это требование соблюдается не всегда.

В настоящей работе для проверки условия А) линейности связи *Y* и *X* для данных типа рис. 1б предлагается использовать методы стратификации и скользящего среднего. Эти методы широко используются для сглаживания временных рядов [2, 3, 13]; мы предлагаем использовать методы скользящего среднего для независимых наблюдений (не временных рядов). Как будет показано, для данных регрессии типа рис. 1б скользящее среднее позволяет выявить «структуру» средних значений *Y* при изменении *X* с применением минимального количества субъективных факторов. Надо отметить, что скользящее среднее предлагает лишь экспертный (визуальный) метод проверки линейности связи *Y* и *X*. Тем не менее даже такая визуальная проверка линейности – это шаг вперед по сравнению с полным отсутствием такой проверки, и ее (визуальную проверку) рекомендуют многие авторы [2, 3, 5, 14].

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), а статистическая значимость отклонения от линейности может быть проверена с помощью корреляционного отношения [3, 15]. Ситуация, изображенная на рис. 1б, в большинстве случаев не позволяет визуально оценить линейность связи *Y* и *X* в той мере, в какой это возможно для ситуации рис. 1а. Для проверки статистической значимости отклонения от линейности в принципе и здесь возможно применение корреляционного отношения; отметим лишь, что идеология корреляционного отношения предполагает разделение значений предиктора *X* на *страты,* в каждой 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 – результат субъективного выбора исследователя. 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 [2–5]. Кроме того, процедура разделения на страты и проверка линейности связи *Y* с *X* для данных типа рис. 1б не предусмотрена в известных статистических пакетах (Statistica, SPSS, SAS и др.). Видимо, поэтому при анализе опубликованных в научной литературе последних десятилетий результатов мы не обнаружили ни одной статьи, посвященной линейной регрессии, в которой авторы сообщали бы о проверке линейности связи *Y* и *X* помощью корреляционного отношения. Более того, во многих публикациях не проводится даже визуальная (экспертная) оценка линейности связи *Y* и *X* с помощью диаграмм типа рис. 1б [16–20]; при этом в данных публикациях приводятся и анализируются значения коэффициентов b_1 линейной регрессионной модели (или коэффициентов корреляции Пирсона между *Y* и *X*), которые по определению имеют смысл только при наличии линейной связи типа (3).

Применение стратификации и скользящего среднего к данным линейной регрессии

Пример 1

В качестве примера применения скользящего среднего рассмотрим данные рис. 1б (статистическая связь веса и роста). На рис. 2а и рис. 2б показаны данные стратификации и скользящего среднего для показателя «Вес» по предиктору «Рост». Для расчета скользящих средних в среде Statistica for Windows использована компьютерная программа [21]. На рис. 2 показаны результаты применения к данным рис. 1б методом стратификации и скользящего среднего.

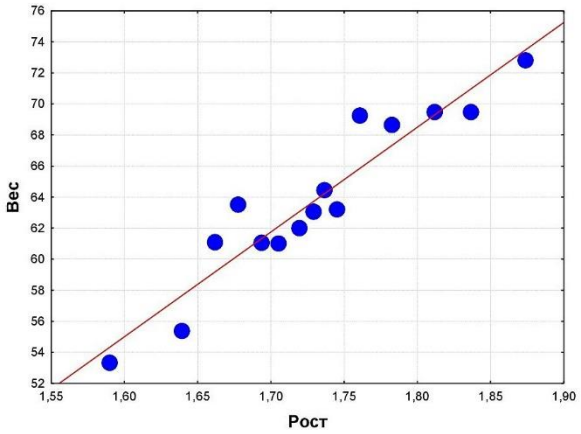

Рис. 2а. Соотношение средних значений веса и роста 375 подростков, стратифицированных по росту с размером страты 25 наблюдений; всего 31 страта, страты не пересекаются

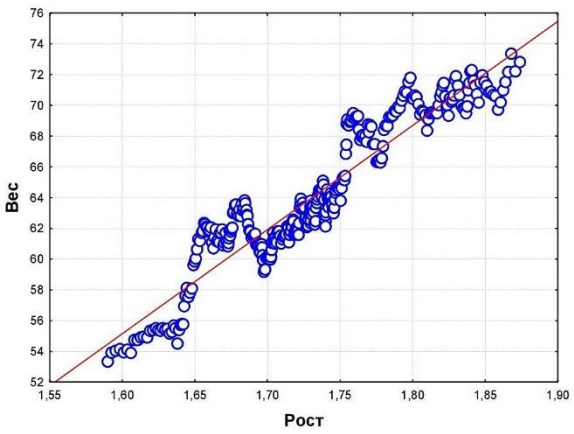

Рис. 2б. Соотношение средних значений веса и роста 375 подростков после выполнения скользящего среднего по росту, окно n_W =25, шаг n_S =1 (всего 351 пересекающаяся страта)

На рис. 2а значения предиктора *X* (рост юношей) разделены на непересекающиеся страты, в каждой из которых рассчитаны средние значения роста и веса. Рисунок наглядно показывает, что точной линейной связи средних значений веса и роста нет; более того, 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. Рис. 2б показывает более подробную (чем рис. 2а) картину связи средних значений роста и веса. Среди 351 точки на рис. 2б есть, очевидно, 31 точка, которая показана на рис. 2а.

Для наглядности еще раз сопоставим диаграммы рассеяния, основанные на первичных данных и данных скользящего среднего (рис. 3).

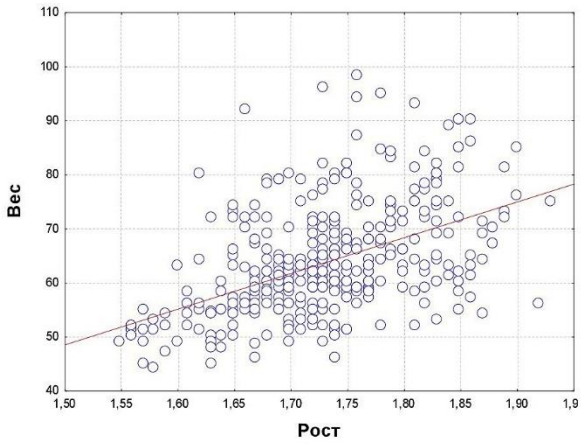

Рис. 3а. Диаграмма рассеяния для первичных данных, показывающая соотношение веса и роста 375 подростков (повторение рис. 1б.)
$Вес(кг) = -50{,}6 + 0{,}661 \cdot Рост(см)$

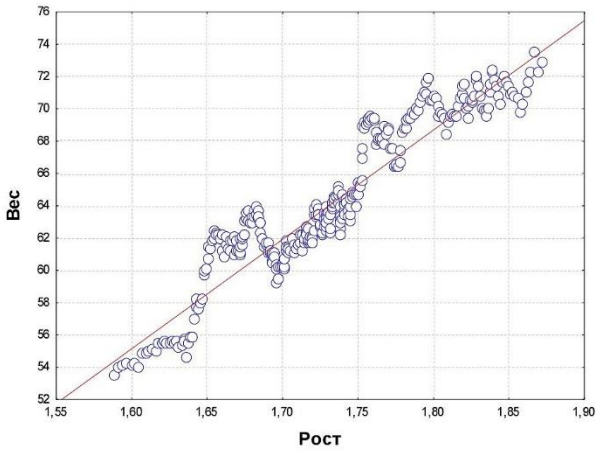

Рис. 3б. Диаграмма рассеяния для данных скользящего среднего, показывающая соотношение веса и роста 375 подростков (повторение рис. 2б.)
$Вес(кг) = -53{,}0 + 0{,}675 \cdot Рост(см)$

Какие выводы можно сделать при сравнении двух рисунков, представляющих первичные данные и результаты скользящего среднего?

1 – линии регрессии на двух рисунках очень близки (см. формулы в нижней части рисунков), наклон линий кажется разным из-за различий в масштабах по оси Вес.

2 – разброс значений около линии регрессии значительно меньше на рис. 3б для скользящего среднего.

3 – согласно экспертной (визуальной) оценке данных рис. 3б, связь «Веса» и «Роста» ***в целом***, несомненно, линейная (не квадратичная или другая);

4 – на рисунке скользящего среднего на фоне прямой линии имеются явные взлеты и провалы, откуда они? Очевидно, это результат воздействия на показатель «Вес» не только роста, но и других факторов, например, потребления продуктов питания, наследственных факторов, индивидуальных и т. д. В моделях множественной регрессии можно наблюдать, как при включении в модель новых предикторов взлеты и провалы относительно линии регрессии постепенно исчезают [6].

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*, описанный ниже в примере 5. Подробности можно найти в работах [6, 7]). Шаг усреднения n_S можно в любом случае выбирать равным единице, как на рис. 2б (получаем максимум информации о средних значениях *Y* и *X*).

Пример 2

В примере рассмотрим возможные отклонения от линейности связи Y с X и способы обнаружения этих отклонений. Анализируется статистическая связь распространенности туберкулеза (ТБ) в Свердловской области с возможными факторами, влияющими на распространенность ТБ (данные Роспотребнадзора по Свердловской области в 2002 г. [22]). Заболеваемость ТБ в 53 городах и сельских районах Свердловской области представлена как число болеющих жителей в расчете на 10 000 человек. Факторы риска повышения распространенности туберкулеза в Свердловской области в основном те же, что и в России в целом, и за рубежом: низкий уровень дохода, недостаток врачей, малая жилая площадь, неблагоприятные

гигиенические условия проживания (например, отсутствие централизованного водоснабжения и канализации) и др. [23–25].

Рис. 4а (первичные данные для 53 городов и районов) показывает статистически значимое уменьшение распространенности ТБ с ростом доходов населения (средние значения доходов в 53 городах и районах в тысячах рублей в месяц). Коэффициент корреляции между доходами и ТБ равен r = – 0,52 ($p < 0,0001$). Вопрос, который интересует нас в рамках данной статьи: является ли связь дохода с ТБ линейной? Рис. 4а не дает однозначного ответа. Чтобы ответить на данный вопрос более обоснованно, построим диаграмму рассеяния «Доход – ТБ» с использованием скользящего среднего. Результат представлен на рис. 4б.

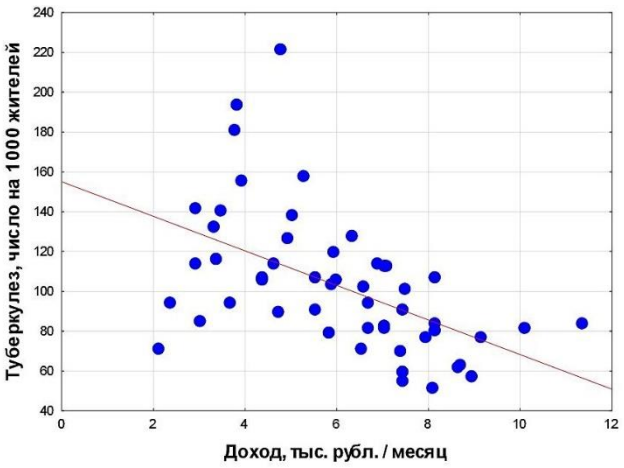

Рис. 4а. Первичные данные: статистическая связь между распространенностью ТБ и доходами населения: первичные данные для 53 городов и районов Свердловской области

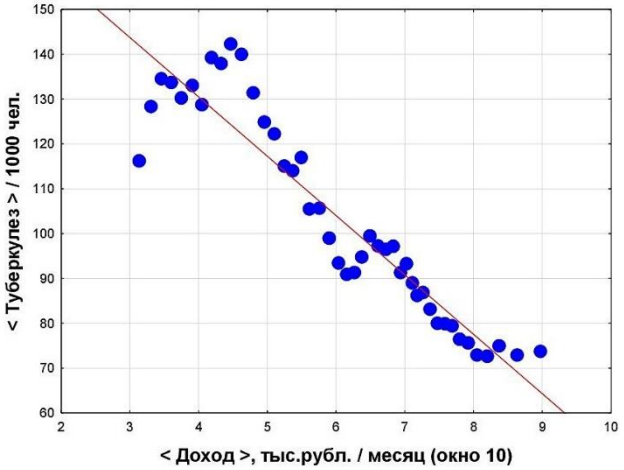

Рис. 4б. Данные скользящего среднего: связь между распространенностью ТБ и доходами населения, усредненными по 10 наблюдениям (рис. 4а)

Каковы причины увеличения распространенности ТБ при увеличении дохода при малых его значениях? Согласно существующим представлениям распространенность ТБ должна уменьшаться при увеличении доходов. Это происходит из-за наличия корреляций между показателем «Доход» и другими факторами, влияющими на распространенность ТБ, такими как размер жилой площади на человека, обеспеченность централизованным водоснабжением и др. На рис. 5 представлена связь между показателями «Доход» и «Обеспеченность жильем», полученная с использованием скользящего среднего.

Можно было ожидать, что увеличение дохода приведет к увеличению обеспеченности жильем, но при малых доходах (от минимума до примерно 4–5 тыс. руб. в месяц) видим снижение обеспеченности жильем. Это происходит в силу каких-то случайных (или неизвестных нам) причин, но такова реальная обстановка в Свердловской области. Далее, увеличение обеспеченности жильем означает улучшение комфортности жилья, что должно приводить к снижению распространенности ТБ (наши данные это подтверждают). Поэтому снижение обеспеченности жильем при увеличении дохода на начальном участке может служить причиной увеличения распространенности ТБ.

Отметим, что на рис. 4б для данных скользящего среднего коэффициент регрессии для связи ТБ с доходом равен b_1= –13,2. Это гораздо больше, чем для первичных данных рис. 4а. (b_1= –8,68). Если же на рис. 4б учесть только «спадающий» участок диаграммы (доход выше 4,5 тыс.), коэффициент регрессии будет еще больше

b_1= – 18,02. Также коэффициент корреляции на рис. 4б (r = – 0,94) гораздо выше соответствующего коэффициента на рис. 4а (r = – 0,52).

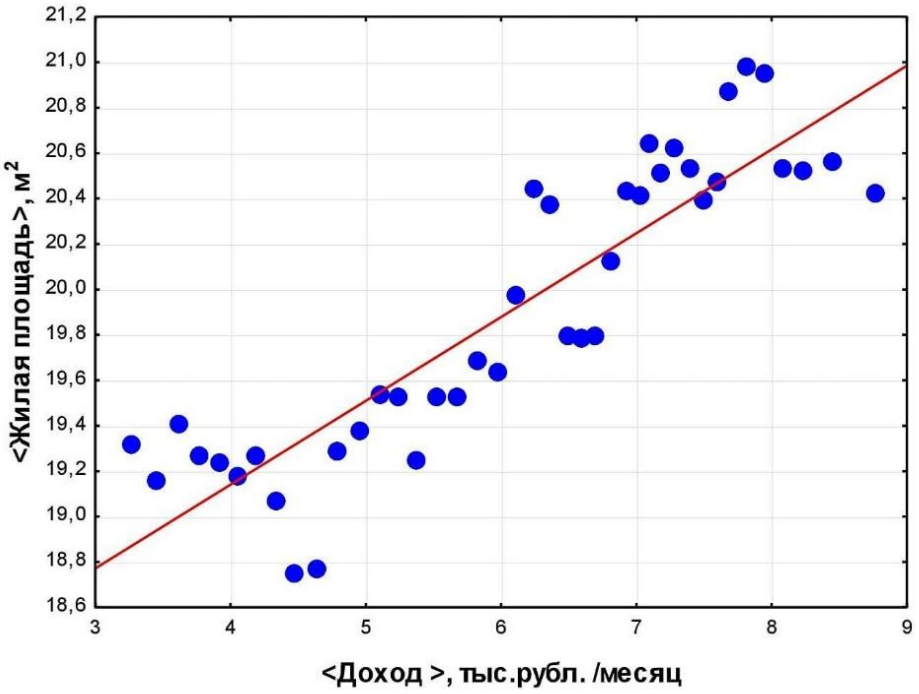

Рис. 5. Данные скользящего среднего (усреднение по 10 наблюдениям, показанным рис. 4а): связь между размером жилой площади на одного жителя и доходами населения

5. Логистическая регрессия

Одна из наиболее распространенных моделей нелинейной регрессии – логистическая регрессия (ЛогР) [26]. В этом разделе рассмотрим методы оценки адекватности моделей ЛогР с использованием стратификации и скользящего среднего. Логистическая регрессия применяется для описания статистической связи между дихотомическом откликом Y (Y принимает два значения: Y=0 и Y=1) и предиктором X. Каких-либо ограничений на тип предиктора X обычно не предъявляется [26], но «истинное» назначение ЛогР – это работа с количественным предиктором X; для категоризованных предикторов вместо ЛогР можно использовать таблицы сопряженности признаков. Данные такого рода (дихотомический Y и количественный X) часто встречаются в эпидемиологических исследованиях, когда Y кодирует отсутствие (Y=0) или наличие (Y=1) некоторого заболевания, а X – фактор риска возникновения заболевания (например, возраст пациента).

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

$$W(Y = 1|X = x) = \frac{exp(b_0+b_1x)}{1+exp(b_0+b_1x)}, \quad (6)$$

где $W(Y=1|X=x)$ – вероятность обнаружения в экспериментальных данных значения Y=1 для заданного значения X=x [26]. При использовании соотношения (6) имеем

$$ln(\frac{W}{1-W}) = logit\ (W) = b_0 + b_1 \cdot x. \quad (7)$$

Таким образом, как и в случае линейной регрессии, имеем линейную связь, но не между X и Y, а между предиктором X и комплексом $ln(\frac{W}{1-W})$, который называют логитом (W). Соотношение (7) – это условие применимости логистической регрессии: если условие (7) выполняется, значит, модель логистической регрессии (6), (7)

адекватна фактическим данным, в противном случае для связи *Y* с *X* необходимо искать другую форму.

Примечание. Shoukri and Pause [27] используют обозначение logit (*W*), подчеркивая, что комплекс в (7) определяется через *W*. Практически аналогичное обозначение используют Peng et al. [28], именно logit (*Y*). Afifi, May and Clark [4] используют обозначение logit (*X*), имея в виду, что комплекс ln(*W*/(1-*W*)) в (7) является функцией *X*; Hosmer and Lemeshow [26] пишут просто logit или используют обозначение *g*(*X*). В данной статье мы будем использовать обозначение logit (*W*).

Аналогично случаю линейной регрессии (3) при выполнении условия линейности логита (7) изменение логита (*W*) при изменении *X* описывается одним параметром b_1, одинаковым для любого *X*. Для характеристики связи *Y* и *X* вместо коэффициента b_1 в ЛогР традиционно используют параметр «отношение шансов *OR»*. Отношение шансов определяется следующим образом. Вместо вероятности *W* вводится показатель ($\frac{W}{1-W}$), который в эпидемиологии называется «шанс». Отношение шансов – это отношение шанса получить значение *Y*=1 для значения (*x*+Δ*x*) к шансу получить значение *Y*=1 для значения *x*. Таким образом,

$$OR = \frac{W(x+\Delta x)/(1-W(x+\Delta x))}{W(x)/(1-W(x))}.$$

Если рассчитывать *OR* при Δ*x*=1, получаем *OR*=*exp*(b_1), где b_1 – коэффициент модели (7).

Что известно об условиях выполнения линейного соотношения (7)? Известно достаточное (но не обязательно необходимое) условие, при котором выполняется соотношение (7): распределение *X* в группах *Y*=0 и *Y*=1 подчиняется нормальному закону с равными дисперсиями [27]. В остальных случаях линейность связи *X* с логитом (*W*) требует специальной проверки.

Как выполнить эту проверку? В отличие от линейной регрессии, в логистической регрессии уделяется внимание проверке линейности связи логита (*W*) с *X*, т. е. соотношения (7). Многие известные статистические пакеты (SAS, SPSS) включают опции, аналогичные процедуре корреляционного отношения в линейной регрессии [29]. 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 (*W*) с *X*, например, критерий согласия Hosmer – Lemeshow [26].

Таким образом, адекватность модели ЛогР проверяется путем расчета критериев согласия для *стратифицированных* первичных данных. Критерий χ^2 Hosmer – Lemeshow [26] имеет вид

$$\chi^2(H-L) = \sum_i n_i \frac{(W_{obs,i}-W_{calc,i})^2}{W_{calc,i}(1-W_{calc,i})},$$

где суммирование (индекс *i*) идет по стратам, n_i, $W_{obs,i}$, $W_{calc,i}$ – соответственно число наблюдений в страте, наблюдаемые и расчетные вероятности; вероятность $W_{obs,i}$ – это среднее значение дихотомического отклика *Y* в *i*-страте, содержащей n_i наблюдений, а $W_{calc,i}$ – результат расчета по формуле (6) для среднего в *i*-страте значения *X*. Если значение χ^2 Hosmer – Lemeshow, рассчитанное по первичным данным в результате их стратификации, не превосходит критического значения $\chi^2_{крит.}(\alpha, \nu)$, модель признается адекватной первичным данным на заданном уровне значимости α (ν – число степеней свободы). Проверка адекватности особенно важна, когда ЛогР используется как метод исследования зависимостей; для модели ЛогР гарантировать линейность logit(*W*) может только проверка адекватности модели. Если модель оказывается адекватной, тогда есть смысл проверить ее статистическую значимость, в противном случае необходимо искать другую модель.

Пользователи логистической регрессии активно (но не всегда) применяют критерии согласия; однако даже в случае применения критерия ограничиваются лишь

определением значений критерия согласия и уровня значимости *p* без графического анализа полученных результатов. Иногда данные стратификации представляют в виде таблицы средних значений *X* и логита (*W*) в сравнении с результатами логистической регрессии, но также без графического представления. Между тем графическое представление результатов стратификации показывает часто неожиданные результаты. Ниже приведены несколько примеров применения логистической регрессии к эпидемиологическим и клиническим данным.

Применение стратификации и скользящего среднего к данным логистической регрессии

Пример 3

На рис. 6а представлены первичные данные для логистической регрессии (количественный *X* и дихотомический *Y*), взятые из монографии [26]. Здесь отклик *Y* представляет сердечно-сосудистые заболевания (ССЗ) у пациентов данного возраста: *Y*=0 означает отсутствие заболевания, *Y*=1 – наличие заболевания у данного пациента, *X* – предиктор ССЗ (возраст пациента). Очевидно, что данные типа рис. 6а не позволяют оценить выполнение соотношения (7) даже в той минимальной мере, в какой это возможно для линейной регрессии, как это показано на рис. 1б и рис. 4а. 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 проведена согласно модели (6), однако нет никакой уверенности, что вероятность иметь ССЗ описывается именно такой кривой, если опираться на визуальный анализ нулей и единиц.

На рис. 6б показано сравнение данных логистической регрессии с данными, полученными путем стратификации предиктора *X*. Предиктор *X* (возраст пациента) в работе [26] был разделен на 8 страт, в каждой страте рассчитаны средние значения возраста и логита (*W*), которые являются результатом усреднения первичных данных. На рис. 6б приведены также значения логита (*W*), вычисленные в этих же стратах методом логистической регрессии. В результате перехода от вероятности *W* (как на рис. 6а) к логиту (*W*) (рис. 6б) логистическая кривая превращается в прямую линию. После выполнения процедуры стратификации можно провести визуальную оценку согласия между наблюдаемыми и расчетными данными; для логита (*W*) в целом видим линейный рост с увеличением возраста. Статистическая проверка гипотезы о линейности связи логита (*W*) с возрастом это подтверждает: согласно критерию согласия хи-квадрат Hosmer – Lemeshow χ^2=0,16 при 6 степенях свободы, при котором гипотеза о линейности не отвергается на обычном уровне значимости, α = 0,05, *p* >> 0,05.

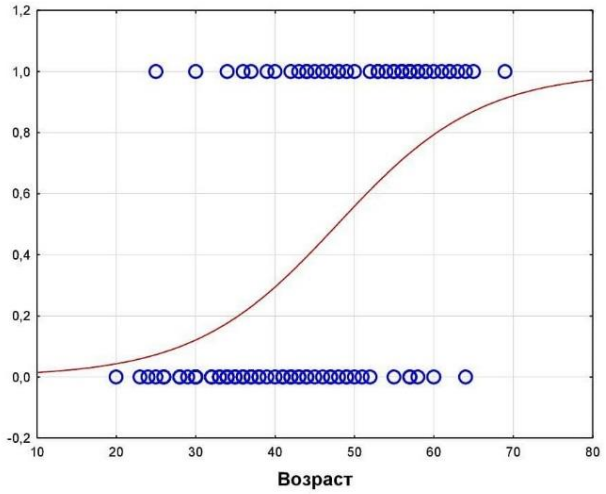

Рис. 6а. Кружки – первичные данные Hosmer – Lemeshow. Количественный предиктор *X* (возраст) и дихотомический отклик *Y* (*Y*=1, если данный пациент имеет заболевание ССЗ; *Y*=0 – заболевания нет). Сплошная линия – расчетная кривая логистической регрессии для вероятности *W*(ССЗ)

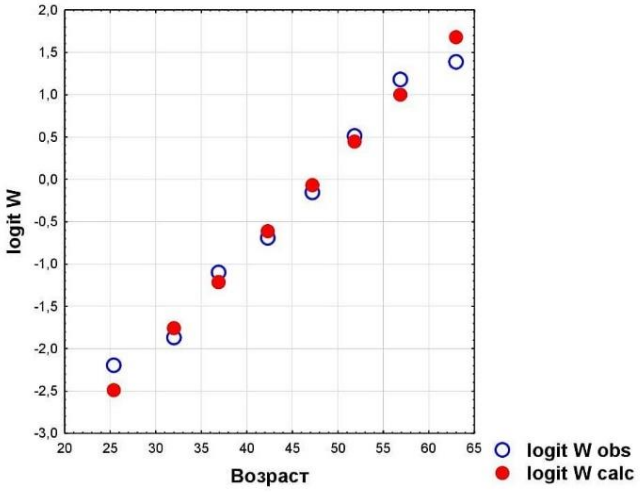

Рис. 6б. Расчетные (calc – полные кружки) и наблюдаемые (obs – открытые кружки) значения логита в результате стратификации Hosmer – Lemeshow на 8 страт по возрасту. Значения logit *W* calc согласно (7) лежат на прямой линии

Неоднозначность стратификации. На примере данных Hosmer – Lemeshow (рис. 6б) можно продемонстрировать неоднозначность результатов проверки гипотезы о линейности логита, которые получаются при различных способах стратификации данных. Разделение на страты по возрастным категориям, как на рис. 6б, не является единственно возможным, например, в монографии [26] описаны несколько различных способов стратификации предиктора *X*. Во всех способах стратификации есть элементы субъективности (неоднозначности), которые обусловлены произвольным выбором числа страт и их границ.

Hosmer и Lemeshow разделили пациентов на 8 непересекающихся страт по возрастным категориям (20–29 лет, 30–34 года и т. д.), каждая страта содержит от 8 до 17 пациентов, т. е. неравное число наблюдений в страте. Если обратиться к известным статистическим пакетам, в них предиктор *X* обычно разделяется на страты с равным числом наблюдений [28]. В примере Hosmer – Lemeshow общее число пациентов равно 100, так что численность каждой из 8 страт должна быть 12 или 13 пациентов. Результаты разделения *X* на 8 страт с равным числом пациентов представлены на рис. 7б. Для удобства сравнения на рис. 7а показан повтор рис. 6б. В некоторых точках заметно явное отличие результатов рис. 7б от рис. 7а.

Мы провели расчеты критерия согласия Hosmer – Lemeshow для 8 страт с равным числом пациентов в стратах и обнаружили резкое увеличение критерия χ^2 = 1,87 по сравнению с результатом Hosmer – Lemeshow χ^2 = 0,16. При таких значениях χ^2 гипотеза о линейной связи логита (*W*) с возрастом, конечно, не отвергается (из-за недостаточного числа пациентов), но вероятность отвергнуть нулевую гипотезу об адекватности модели ЛогР значительно повышается.

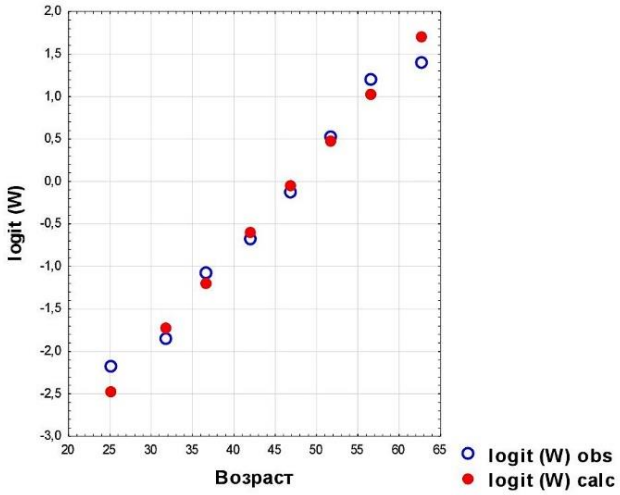

Рис. 7а. Расчетные и наблюдаемые значения логита в результате стратификации Hosmer – Lemeshow на 8 страт по возрасту (повтор рис. 6б). χ^2 =0,16 при 6 степенях свободы, p >> 0,05

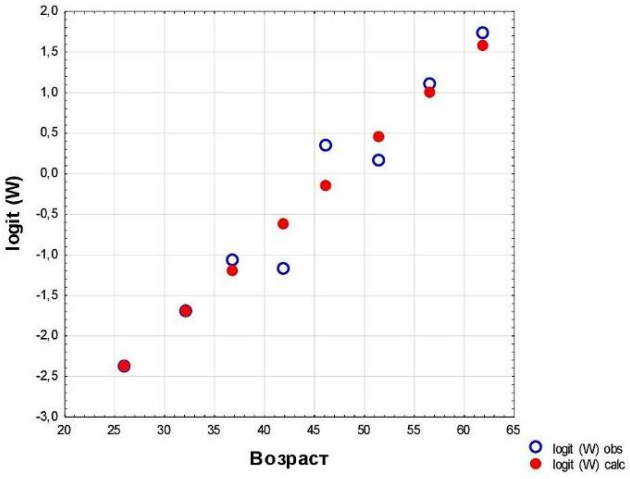

Рис. 7б. Логит (*W*) в зависимости от возраста: стратификация 100 пациентов Hosmer – Lemeshow на 8 страт с равным числом пациентов в страте, обозначения как на рис. 7*а*. χ^2 = 1,87, p > 0,05

В некоторых статистических пакетах (SAS, SPSS и др.) по умолчанию применяется стратификация на 10 страт с равным числом наблюдений в страте (такую процедуру можно называть *стандартной стратификацией* [26, 28]), такая процедура будет использована нами в следующих примерах.

Скользящее среднее. Если при стратификации имеем неопределенность в числе страт и их границах, то в скользящем среднем имеется только один субъективный показатель – размер окна усреднения, который можно варьировать для получения наиболее наглядного результата. Поэтому неоднозначность стратификации в определенной степени может быть преодолена использованием методов скользящего среднего.

На рис. 8б показаны результаты скользящего среднего с окном усреднения, равным 10 наблюдениям (что совпадает с размером страт в процедуре стандартной стратификации на рис. 7а). Результаты скользящего среднего не позволяют рассчитать какой-либо критерий адекватности модели, типа критерия Hosmer – Lemeshow; скорее, эти результаты можно воспринимать как визуальную (экспертную) оценку согласия теории и наблюдаемых закономерностей: теория предсказывает строгую линейность logit(*W*) в зависимости от возраста, а наблюдения либо подтверждают, либо опровергают эту линейность для данного фактического материала. По данным рис. 8б, фактический материал скорее подтверждает теоретическую линейность logit(*W*). Подтверждением линейности logit(*W*) можно также считать следующий результат: в формулу для logit(*W*) по данным скользящего среднего была попытка добавить члены более высокого порядка по возрасту, чем линейный член, они оказались статистически незначимы. Подтверждение согласованности результатов скользящего среднего и первичных данных – также согласие выражения для logit(*W*), полученного по данным скользящего среднего (открытые кружки на рис. 8б) с выражением, полученным по первичным данным методом ЛогР (рис. 6а, формула типа (7)). В случае скользящего среднего logit (*W*) = – 5,658 + 0,1183 · Возраст, а по первичным данным logit (*W*) = – 5,309 + + 0,1109 · Возраст. Очевидно, эти соотношения практически не отличаются друг от друга. Этого следовало ожидать, поскольку график на рис. 8б подтверждает

справедливость соотношения (7), которое является условием применимости логистической регрессии.

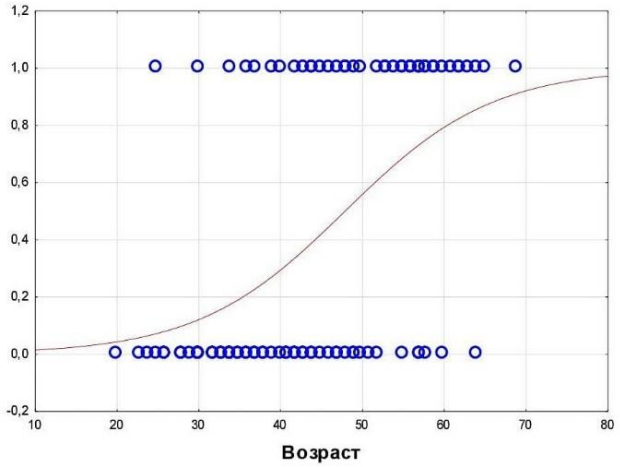

Рис. 8а. Повтор рис. 6а для первичных данных. Сплошная линия – логистическая кривая
$logit\ (W) = b_0 + b_1 \cdot X =$
$= -5{,}309 + 0{,}1109 \cdot Age$

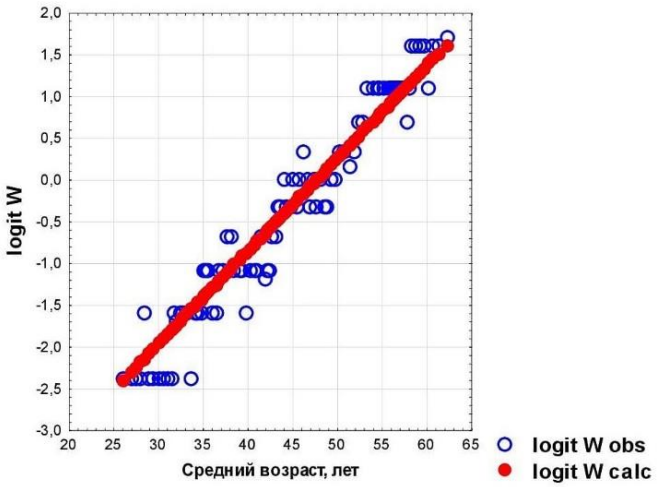

Рис. 8б. Данные скользящего среднего (окно n_W=12, открытые кружки). Сплошные кружки – расчеты методом логистической регрессии
$logit\ (W) = b_0 + b_1 \cdot X =$
$= -5{,}632 + 0{,}1179 \cdot Age$

Пример 4

В работе C. Peng [28], опубликованной в журнале обучающей направленности, описаны основы логистического анализа и даются рекомендации по представлению результатов в журнальных статьях. В статье есть все, кроме графического представления результатов, которое, как будет показано ниже, описывает особенности связи отклика с предиктором.

Авторы [28] анализируют связь отклика *Y* (remedial reading recommended – рекомендация по дополнительным занятиям по чтению: *Y*=0 – не рекомендуется, *Y*=1 – рекомендованы дополнительные занятия) с предиктором X (Reading Score – отметка по чтению у школьников). Предполагается, что детям с низкими отметками по чтению будет чаще рекомендовано посещать дополнительные занятия.

Для проверки данного предположения была построена модель логистической регрессии:

$$logit\ (W) = b_0 + b_1 \cdot X = 0{,}9968 - 0{,}02811 \cdot ReadingScore, \quad (8)$$

в которой коэффициент у предиктора оказался статистически значим (*p*=0,0245), следовательно, существует статистическая связь между оценками и рекомендациями по дополнительным занятиям. Авторы [28] проводили проверку адекватности модели (8) фактическим данным с использованием методов стратификации: весь массив наблюдений (189 детей) был разделен на 10 групп по 19 наблюдений в каждой (последняя группа содержала 18 наблюдений, т. е. проведена стандартная стратификация по предиктору *X*). Такое разделение *X* на 10 групп по умолчанию выполняется во многих современных статистических пакетах, в том числе в SPSS [29], который использован в работе [28]. В каждой группе были определены средние значения *W* и логита (*W*), полученные в результате стратификации (наблюдаемые значения – *obs*); также были рассчитаны W и логит (W) с использованием логистической регрессии (расчетные значения – *calc*). Результаты проверки линейности логита с использованием критерия хи-квадрат, представленные в работе [28], таковы: проверка по критерию хи-квадрат Hosmer – Lemeshow показала адекватность модели фактическим данным (для 8 степеней свободы в [28] получено

χ^2 = 10,76). Критическое значение χ^2 = 15,51, p = 0,31>0,05, т. е. модель признается адекватной фактическим данным. Таким образом, авторы [28] имели право использовать модель (8) для заключения о необходимости дополнительных занятий у определенной группы школьников.

Что может дать исследователю графическое представление результатов логистического анализа, которое отсутствует в работе [28]? Первоначально мы построили график связи отклика *Y* (Remedial reading recommended) с количественным предиктором *X* (Reading Score) при разделении 189 наблюдений по *X* на 10 равных по численности страт (повтор процедуры [28]), результат показан на рис. 9а. Если пользоваться результатами логистической регрессии (полные кружки на рис. 9а), видим ожидаемую линейность расчетного логита *W* в зависимости от *X*, т. е. шанс появления значения *Y*=1 монотонно уменьшается с ростом *X* [28]. Совсем другую картину мы видим для наблюдаемых значений логита (*W*): открытые кружки на рис. 9а не показывают монотонность изменения логита при увеличении предиктора. Например, если взять две точки с минимальными и две точки с максимальными значениями *X* и соединить их прямыми линиями, наклон линий оказывается положительным (шанс появления значения *Y*=1 увеличивается с ростом *X*, и *OR* оказывается больше единицы, что противоречит выводам логистической регрессии). Таким образом, отсутствие графика типа рис. 9а в работе [28] не позволило авторам в полной мере оценить характер связи между *Y* and *X*.

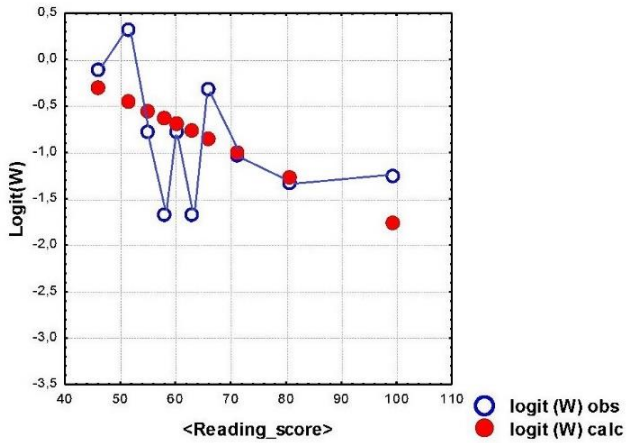

Рис. 9а. Связь расчетных и наблюдаемых значений логита *W* с предиктором *X*

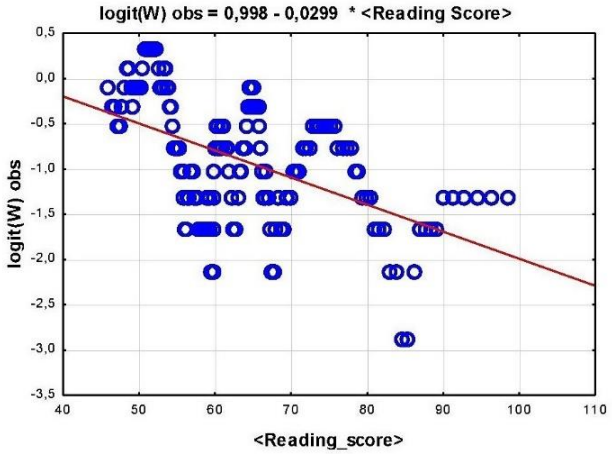

Рис. 9б. Диаграмма рассеяния для скользящего среднего логита отклика *W* и предиктора *X* (окно=19 наблюдений, всего 171 страта)

Для более детального (чем при стратификации) анализа связи *Y* с *X* использованы методы скользящего среднего. Мы построили скользящее среднее по количественному предиктору *X* (Reading Score) для дихотомического отклика *Y* (Remedial reading recommended) с окном усреднения n_W =19 наблюдений (как при стратификации) и с шагом скользящего среднего, равного единице. В результате применения процедуры скользящего среднего получили 171 страту вместо 10 страт у авторов [28]. Для каждой страты в скользящем среднем рассчитаны средние значения *Y* и *X*. На рис. 9б показана диаграмма рассеяния для наблюдаемых значений логита *W*. График на рис. 9б показывает, что логит *Y* ***в среднем*** действительно уменьшается при увеличении X (сплошная линия – линия регрессии). Это совпадает с выводами логистической регрессии (8), причем коэффициент b_1 для логита (*W*) на рис. 9б,

равный – 0,02989, близок к значению b_1 = – 0,02811 (8), получаемому в логистической регрессии [28].

Однако график на рис. 9б (данные скользящего среднего) заставляет усомниться в строгой линейности связи между логитом W и количественным предиктором *X* во всем диапазоне значений *X*: на рис.9б видны явные максимумы среднего logit (*W*) в районе значений *X*, равных 52, 65 и 75, и минимумы в областях *X*, равных 60, 68 и 85. Связь наблюдаемых значений логита с *X* на рис. 9б в чем-то похожа на таковую на рис. 9а, но рис. 9б более подробный. В частности, «провал» наблюдаемых значений логита в районе *X*=85 на рис. 9б отсутствует на рис. 9а; при этом на рис. 9б среди 179 точек есть те 10 точек, которые представлены на рис. 9а.

Пример 5

В примере анализируется статистическая связь между смертностью пациентов с почечной недостаточностью (*Y*=0 – пациент жив, *Y*=1 – пациент умер) и такими предикторами смертности *X*, как баланс жидкости на момент начала диализа (Total CFB on day of RRT в литрах) и разность баланса жидкости на момент начала диализа и после диализа (Delta Volume в литрах) [30]. Для анализа связи между *Y* и *X* использована логистическая регрессия. Для предиктора «Баланс жидкости на момент начала диализа» (сокр. «Total CFB») получено уравнение ЛогР:

$$logit\,(W) = -0{,}435 + 0{,}0794 \cdot TotalCFB. \qquad (9)$$

Коэффициенты модели (9) статистически высоко значимы (p < 0,0001).

Проведем проверку адекватности модели (9) с использованием процедуры стандартной стратификации (разделение пациентов на 10 страт по «Total CFB» с равным числом пациентов во всех стратах). По критерию хи-квадрат Hosmer – Lemeshow нулевая гипотеза о линейности связи смертности с предиктором «Total CFB» не отклоняется (логит может считаться линейной функцией *X*). Результаты стандартной стратификации для вероятности и для логита графически представлены на рис. 10.

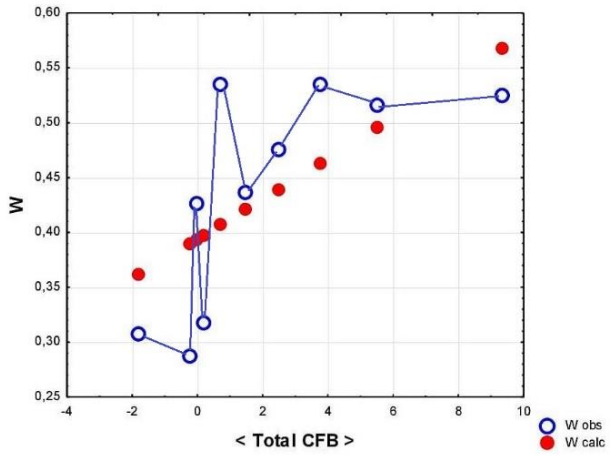

Рис. 10а. Результаты стандартной стратификации (вероятность)

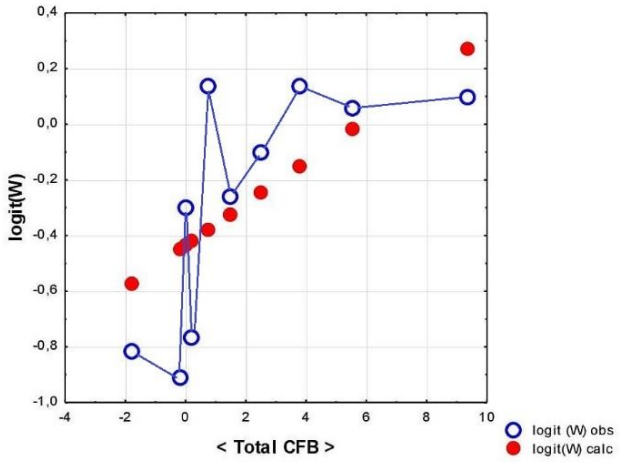

Рис. 10б. Результаты стандартной стратификации (логит)

Данные рис. 10б показывают, что наблюдаемые значения логита не укладываются точно на прямую линию, описываемую соотношением (9): при малых (в основном отрицательных) значениях предиктора «Total CFB», наблюдаемые значения меньше расчетных, а при «Total CFB» выше нуля, наблюдаемые значения логита много больше расчетных. Использованная здесь стандартная стратификация разделяет пациентов на страты автоматически без обращения к деталям связи смертности с «Total CFB». Эти детали можно выявить, если провести «ручную»

стратификацию с использованием кумулятивной функции вероятности умереть при заданных значениях «Total CFB».

Кумулятивная функция вероятности

Как осознанно провести стратификацию? Один из подходов – кумулятивная функция для вероятности иметь *Y*=1. Кумулятивная функция $Y_{NC}(X)$ для отклика *Y* по предиктору *X* определяется соотношением

$$Y_{N_C}(X) = \frac{1}{N_C}\sum_{i=1}^{N_C} y_i,$$

где N_C – количество объектов, включенных в кумулятивную функцию (N_C = 1,2,3…*n*).

На рис. 11 показана кумулятивная функция для вероятности смерти в зависимости от «Total CFB» (точнее, от ранга «Total CFB»), при котором на графике проявляются особенности связи смертности с «Total CFB». Видно, что минимум кумулятивной функции приходится на значение ранга «Total CFB», равного 194. Это означает, что в группе из 194 пациентов с минимальными значениями «Total CFB» средняя по группе вероятность умереть минимальна и равна *W*(death) =0,289. Для этой группы среднее значение «Total CFB» равно 1,04 (примерно соответствует тому, что видно на рис. 10а при стандартной стратификации).

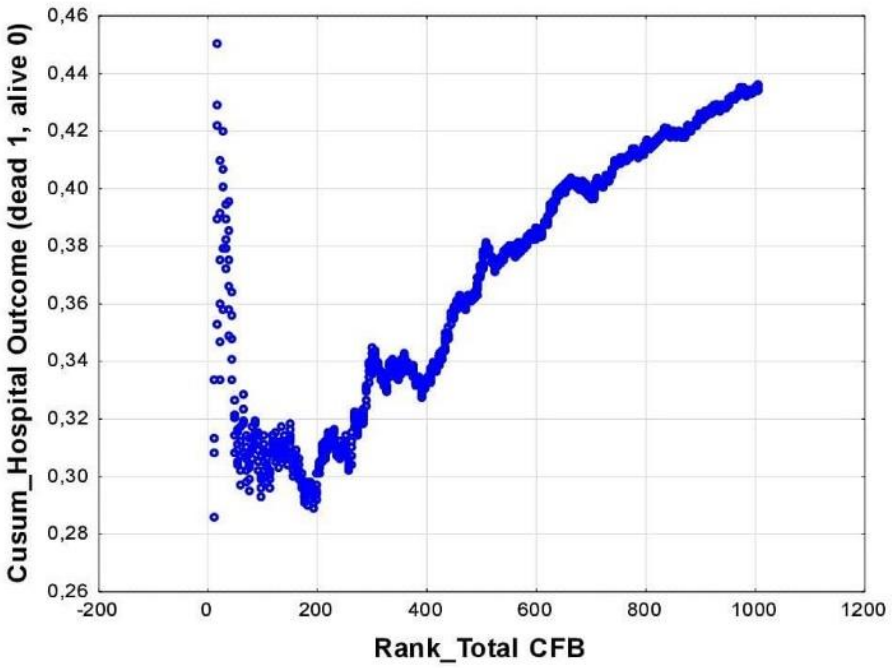

Рис. 11. Кумулятивная функция вероятности

Для нахождения второй группы пациентов с использованием кумулятивной функции удалим из базы данных 194 пациента с минимальными значениями «Total CFB» и снова построим график типа рис. 11. Анализируя особенности этого графика, находим вторую группу пациентов с минимальным средним значением вероятности умереть *W*(death)=0,371, для которой среднее значение «Total CFB» равно +0,084. Продолжая аналогичным образом, находим 5 страт, каждая из которых соответствует особенностям типа рис. 11. Вероятности *W*(death) для этих пяти страт, вместе с результатами логистической регрессии, показаны на рис. 12.

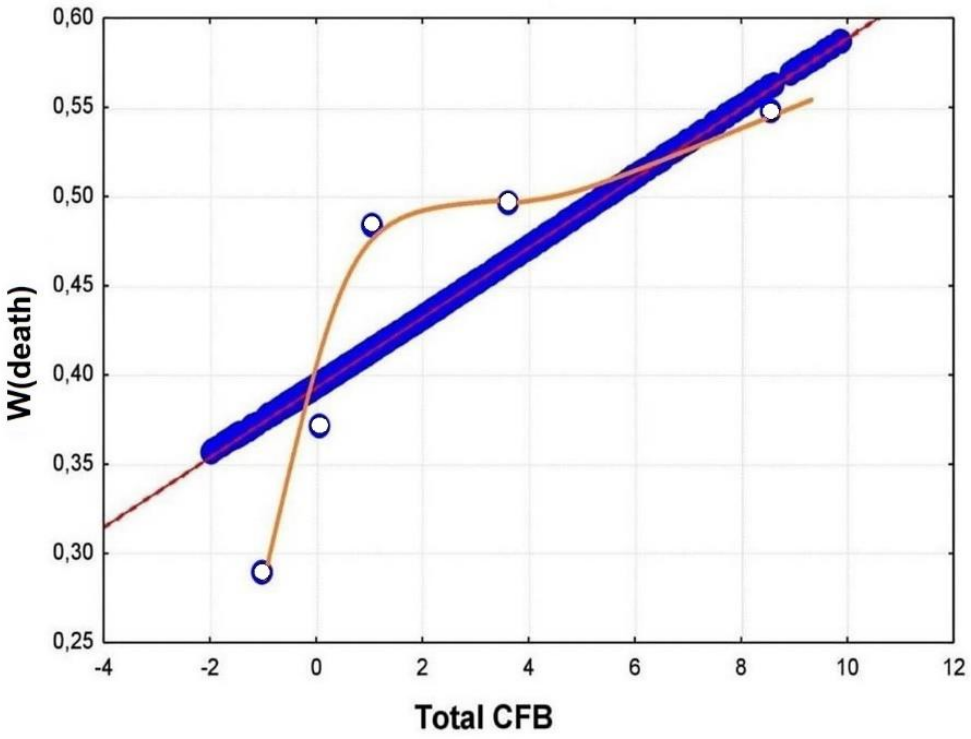

Рис. 12. Вероятность умереть *W*(death) в зависимости от «Total CFB». Открытые кружки – наблюдаемые значения вероятности по результатам стратификации, сплошная линия – расчет методом ЛогР

Теперь выполним расчеты, аналогичные представленным выше, для второго предиктора смертности – показателя «разность баланса жидкости на момент начала диализа и после диализа» (сокр. «Delta Volume»).

Для анализа связи между *Y* и *X* построим модель логистической регрессии

$$logit\,(W) = -0{,}386 - 0{,}0322 \cdot DeltaVolume. \qquad (10)$$

Коэффициент модели (10) у предиктора «Delta Volume» оказался статистически незначим (p=0,33), следовательно, статистическая связь смертности с «Delta Volume» не обнаружена.

Проведем проверку адекватности модели (10) с использованием процедуры стандартной стратификации (разделение пациентов на 10 страт по «Delta Volume» с равным числом пациентов во всех стратах). По критерию хи-квадрат Hosmer – Lemeshow нулевая гипотеза о линейности связи смертности с предиктором «Delta Volume» отклоняется, т. е. логит не может считаться линейной функцией *X*. Результаты стандартной стратификации графически представлены на рис. 13. Если воспользоваться представлением рис. 13а, можно отметить, что отклонение гипотезы о линейности логита по критерию Hosmer – Lemeshow произошло в основном за счет двух точек: это точка с минимальным (*W* порядка 0,2) и точка с максимальным (*W* порядка 0,7) значениями вероятности.

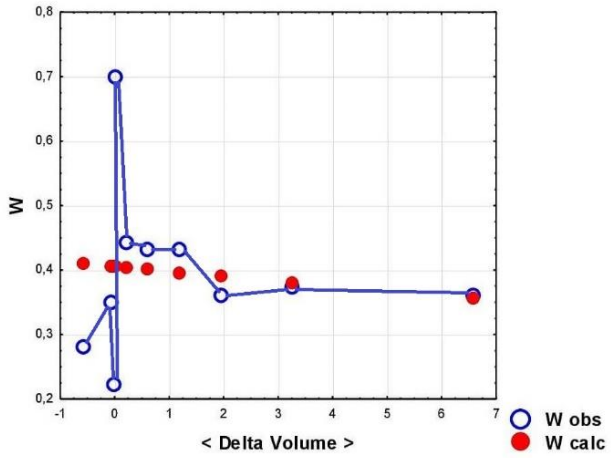

Рис. 13а. Результаты стандартной стратификации (вероятность)

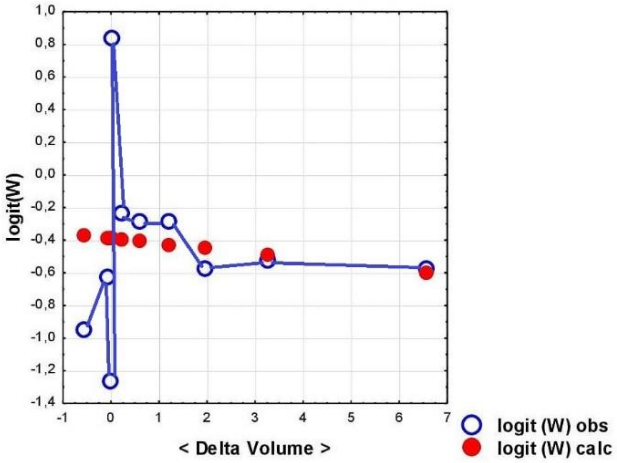

Рис. 13б. Результаты стандартной стратификации (логит)

Данные рис. 13б так же, как рис. 10б, показывают, что наблюдаемые значения логита не укладываются точно на прямую линию, описываемую соотношением (10): при малых значениях предиктора «Delta Volume» наблюдаемые значения логита (логит меньше –0,6) оказываются меньше расчетных, а при более высоких «Delta Volume», наблюдаемые значения логита лежат в районе расчетных; исключение составляет одна точка, в которой логит выше 0,8. Как и ранее, детали связи смертности с «Delta Volume» можно выявить, если использовать кумулятивную функцию вероятности. Результат стратификации, выполненной с использованием кумулятивной функции вероятности, показан в табл. 1 и на рис. 14.

Таблица 1. Результаты стратификации с использованием кумулятивной функции

Номер страты	Число пациентов в страте	<Delta Volume>	W obs	*p*-значение
1	168	– 0,320	0,321	(1 и 2) = 0,0178
2	134	0	0,455	(1 и 3) = 0,0265
3	224	0,373	0,431	(1 и 4) = 0,089
4	221	1,99	0,405	(2 и 5) = 0,081
5	113	6,01	0,336	
Всего	860	1,33	0,394	

Табл. 1 показывает, что имеются статистически значимые (p <0,05) различия вероятностей умереть в стратах 1 и 2, а также в стратах 1 и 3; значимость различий вероятностей в стратах 1 и 4, а также 2 и 5 оценивается как p <0,10.

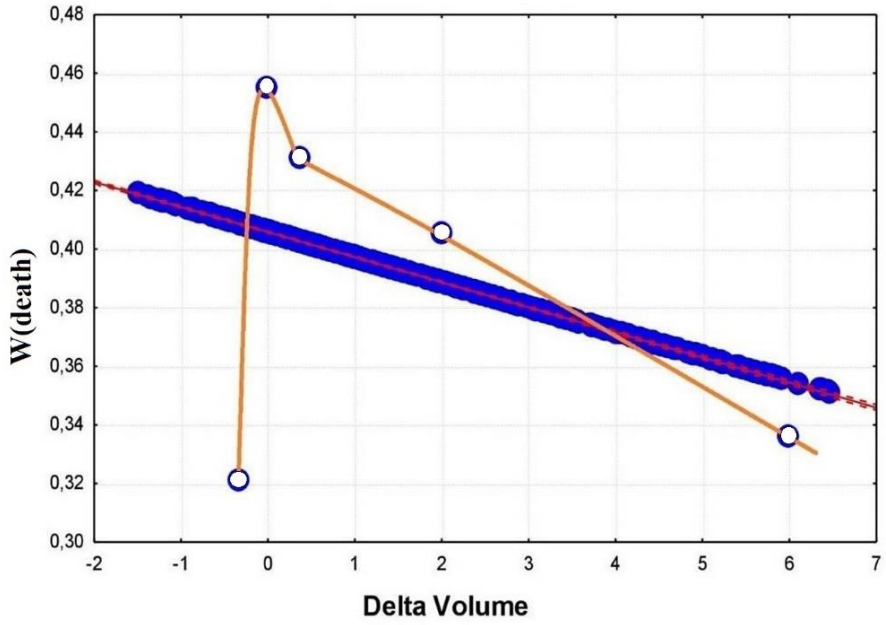

Рис. 14. Вероятность умереть *W*(death) в зависимости от «Delta Volume». Открытые кружки – наблюдаемые значения вероятности по результатам стратификации табл. 1, сплошная линия – расчет методом ЛогР

Таким образом, причина неадекватности и статистической незначимости модели ЛогР (10) заключается, как показано на рис. 14, в немонотонности наблюдаемого изменения доли умерших при увеличении «Delta Volume»: на начальном участке предиктора «Delta Volume» вероятность умереть возрастает, а затем падает; логистическая регрессия предполагает монотонное изменение вероятности или логита при увеличении предиктора *X* – см. формулы (6) и (7). Согласно данным табл. 1, «Delta Volume» – это значимый предиктор смерти, если ориентироваться на статистическую значимость различий долей умерших в различных стратах.

Пример 6

В предыдущем примере показано, что графическое представление данных стратификации и скользящего среднего позволяет оценить возможную нелинейность связи логита (*W*) с предиктором *X*. В данном примере показано, что даже при выполнении всех условий применимости логистической регрессии (адекватность модели фактическим данным по критериям согласия) результаты логистической регрессии могут *кардинально* отличаться от фактических данных. Это отличие удается установить только благодаря процедуре скользящего среднего.

Для построения модели использован эпидемиологический материал, включающий 100 женщин в возрасте старше 60 лет, для которых определялись антропометрические показатели и распространенность различных соматических патологий – первичные данные НИИ Охраны материнства и младенчества, Екатеринбург [31, 32]. В данном примере описывается связь распространенности заболеваний щитовидной железы (ЩЖ) с индексом массы тела (ИМТ). Заболевания ЩЖ кодируются символом 0, если заболевания у данной женщины нет, и 1, если заболевание есть.

Применение метода логистической регрессии дает следующие результаты. Соотношение типа (7) для данной задачи имеет вид (*W* – вероятность иметь заболевание ЩЖ):

$$logit\,(W) = b_0 + b_1 \cdot X = -0{,}319 - 0{,}0111 \cdot \text{ИМТ}. \qquad (11)$$

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.

Проверка адекватности модели (11). Нами был рассчитан критерий согласия Hosmer – Lemeshow для проверки линейности связи предиктора *X* (ИМТ) с логитом вероятности иметь заболевание ЩЖ, который показал, что при использовании процедуры стандартной стратификации гипотеза о линейности логита не отвергается (модель (11) адекватна фактическим данным). Сказанное позволяют сделать вывод: метод логистической регрессии в данном примере использован при выполнении условий применимости, но статистически значимой связи между *W*(ЩЖ) и ИМТ не обнаружено.

Проведем графический анализ связи *W*(ЩЖ) с ИМТ. На рис. 15а показаны результаты стратификации ИМТ на 10 групп по 10 пациентов в каждой, а также расчетные данные, полученные методом логистической регрессии. Видно, что наблюдаемые данные (открытые кружки) во многих случаях весьма далеки от расчетных данных (полные кружки).

Теперь применим к первичным данным процедуру скользящего среднего по ИМТ. Как и в примере 5, для выбора окна усреднения n_W использовали кумулятивную функцию вероятности, которая позволила обоснованно выбрать n_W =21 (см. ниже). На рис. 15б представлена связь вероятности *W*(ЩЖ) с ИМТ после выполнения операции скользящего среднего. Экспертная оценка данных рис. 15б показывает, что связь W(ЩЖ) иметь заболевание щитовидной железы с ИМТ у женщин существует, однако она совсем не похожа на результаты логистической регрессии (полные кружки на рис. 15а). Эта связь даже не является монотонной: в интервале значений ИМТ от 22 до 26 кг/м^2 вероятность *W*(ЩЖ) в среднем уменьшается при увеличении ИМТ, а в интервале ИМТ от 26 до 30 кг/м^2 *W*(ЖКТ) возрастает.

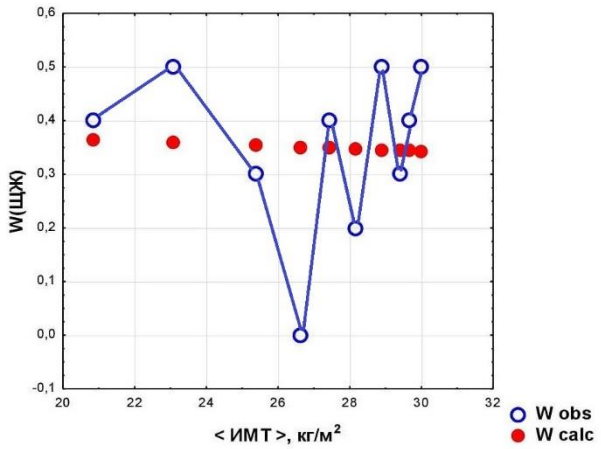

Рис. 15а. Вероятности *W*(ЩЖ) в зависимости от ИМТ по данным стратификации на 10 страт. Открытые кружки – результат стратификации (наблюдаемые значения), полные кружки – расчет методом логистической регрессии

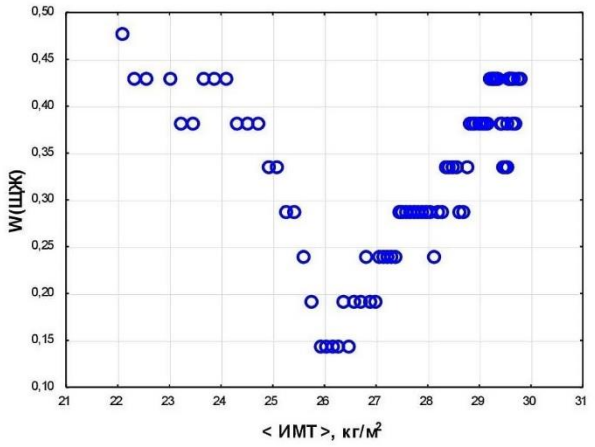

Рис. 15б. Связь вероятности *W*(ЩЖ) с ИМТ для данных скользящего среднего, 81 страта

Проведем анализ статистической значимости различий *W*(ЩЖ) для различных значений ИМТ по данным скользящего среднего. В первой страте по ИМТ на рис. 15б (21 пациентка с минимальными значениями ИМТ) 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 21 наблюдению 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.

Для данного эпидемиологического материала сравним выводы, получаемые методом логистической регрессии и методами скользящего среднего. Проверка адекватности модели логистической регрессии показывает, что наблюдаемые данные согласуются с предположением о линейности логита, но получаемый коэффициент регрессии не отличается от нуля статистически значимо. Это означает, что статистически значимой связи между *W*(ЩЖ) и ИМТ нет. Применение методов скользящего среднего дает противоположный результат: логит *W* (график логита похож на график вероятности на рис. 15б) является не линейной, а V-образной функцией ИМТ; при этом для данных скользящего среднего имеет место статистически и предметно значимая связь между распространенностью заболеваний щитовидной железы и ИМТ у женщин старшего возраста. Таким образом, основываясь только на результатах логистического регрессионного анализа, можно было «не заметить» явную связь между *W*(ЩЖ) и ИМТ.

О выборе окна усреднения в скользящем среднем.

Для выбора окна усреднения в скользящем среднем (рис. 15б) была построена кумулятивная функция вероятности иметь заболевание ЩЖ в зависимости от ИМТ. График кумулятивной функции вероятности показан на рис. 16. Видно, что после колебаний на начальном участке вследствие малого числа слагаемых в кумулятивной функции эта функция имеет максимум на ранге ИМТ, равном 21. Это означает, что средняя распространенность заболевания ЩЖ среди 21 женщины с наименьшими значениями ИМТ оказывается максимальной и равной W=0,476. По результатам анализа данных рис. 16 было принято решение выбрать в качестве окна скользящего среднего n_W =21.

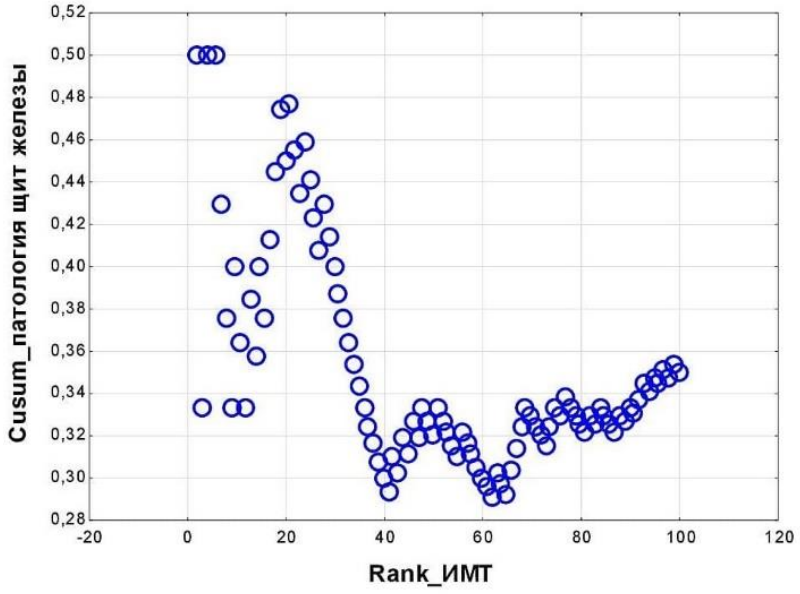

Рис. 16. Кумулятивная функция для связи вероятности иметь заболевание ЩЖ с ИМТ

Таким образом, в работе показано, что ориентация только на статистические критерии может привести к искажению выводов логистической регрессии. Например, показано, что критерии согласия могут не отвергать гипотезу о линейности связи предиктора *X* с логитом (*W*), а графический анализ показывает явно нелинейную связь

(пример выше). Активными пропагандистами графического представления результатов статистического анализа были такие известные специалисты, как Дж. Тьюки и Э. Сигел. Американский статистик Эндрю Сигел (Andrew F. Siegel) в монографии [2] практически каждый вывод иллюстрирует графиками с подробными комментариями. Один из создателей современной науки анализа данных Джон Тьюки (John W. Tukey) в монографии [14, с.9] 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 малополезны. График имеет наибольшую ценность тогда, когда он ***вынуждает*** нас заметить то, что мы совсем не ожидали увидеть». Приведенные выше рис. 2б, 4б, 7б, 9б и особенно рис. 14 и 15б как раз и показывают то, что мы не ожидали увидеть.

6. Предметная значимость статистической значимости при проверке статистических гипотез

Статистическая значимость связей, различий, эффектов в математической статистике проверяется с помощью статистических гипотез. Для этого формируется некий критерий, который обладает известным распределением и с его помощью можно проверить, какова вероятность того, что изучаемая связь, различие или эффект являются НЕ-случайными. Например, при проверке различий математических ожиданий некоторого параметра X в двух выборках формируется известный критерий Стьюдента t, который содержит такие показатели выборок, как выборочные средние и дисперсии, а также число наблюдений в выборках. Критерий t имеет известное распределение (распределение Стьюдента), с помощью которого можно рассчитать вероятность того, что математические ожидания в выборках не различаются (точнее, что обе выборки принадлежат одной генеральной совокупности).

При выполнении расчетов по конкретным данным, вероятность того, что эффект является неслучайным, выражается так называемым p-значением. Даже те пользователи статистических процедур, которые совсем не знают математики, тем не менее знают, что p <0,05 – это хорошо (вероятность того, что эффект случаен, меньше 5 %, и можно говорить о статистической значимости).

Хорошо известно, что с ростом числа наблюдений при сохранении величины эффекта p-значение уменьшается; значит, увеличением числа наблюдений мы можем снизить p-значение до любого, сколь угодно низкого уровня. Ф. Мостеллер по этому поводу писал: «… при достаточно большом числе наблюдений можно установить все что угодно» (цит. по [33]).

Применительно к теме данной статьи: если мы проверяем адекватность модели с помощью статистического критерия, при увеличении числа наблюдений мы можем ***отклонить*** гипотезу об адекватности модели даже в том случае, когда модель безусловно ***адекватна***! Это известный «эффект» слишком большого объема выборки (oversampling) [8, с. 315]. Специалистам, которые интересуются вопросами слишком большого и слишком малого объема выборки, рекомендуем раздел 9.4 монографии [8]. Таким образом, *статистическая значимость при проверке статистических гипотез имеет низкую предметную значимость вследствие зависимости статистической значимости от объема выборки.*

В этой ситуации особое значение приобретают графические методы проверки адекватности моделей, типа моделей регрессии. В какой-то мере здесь могут быть использованы методы стратификации, но методы скользящего среднего (в сочетании с анализом кумулятивной функции вероятности) в некоторых случаях явно работают лучше.

Исследование выполнено за счет субсидий Минобрнауки РФ на выполнение научной темы FUMN-2024-0002.

7. Список литературы

1. *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.
2. *Сигел, Э*. Практическая бизнес-статистика / Э. Сигел. – М. : ООО «И. Д. Вильямс», 2017. – 1056 с. – ISBN 978-5-8459-1367-8.
3. *Айвазян, С. А*. Прикладная статистика. Исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М. : Финансы и статистика, 1985. – 487 с.
4. *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.
5. *Афифи, А.* Статистический анализ: подход с использованием ЭВМ / А. Афифи, С. Эйзен. – М. : Мир, 1982. – 488 с.
6. 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.
7. *Вараксин, А. Н*. Регрессионные модели в задачах медико-экологического мониторинга: алгоритм учета корреляций между предикторами / А. Н. Вараксин // Экологические системы и приборы. – 2025. – № 1. – С. 3–14. – DOI 10.25791/esip.1.2025.1495. – EDN NVRCWH.
8. *Айвазян, С. А.* Прикладная статистика. Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М. : Финансы и статистика. 1983. – 471 с.
9. 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.
10. *Владимиров, В. Г*. Радиозащитные рецептуры. Оптимизация состава и механизм действия / В. Г. Владимиров, Г. А. Поддубский, Г. И. Разоренов. – Л. : Воениздат, 1988. – 144 с.
11. *Налимов, В. В.* Теория эксперимента / В. В. Налимов. – М. : Наука, 1971. – 206 с.
12. *Fletcher, R. H.* Clinical epidemiology: The essentials / R. H. Fletcher, S. W. Fletcher. – 5th ed. – Lippincott Williams & Wilkins, 2012. – 272 p. – ISBN 978-1451144475.
13. *Мостеллер, Ф.* Анализ данных и регрессия / Ф. Мостелер, Дж. Тьюки. – М. : Финансы и статистика, 1982. – 378 с.
14. *Тьюки, Дж.* Анализ результатов наблюдений. Разведочный анализ / Дж. Тьюки. – М. : Мир, 1981. – 400 с.
15. *Гмурман, В. Е.* Теория вероятностей и математическая статистика : учеб / В. Е. Гмурман. – 12-е изд. – М. : Издательство Юрайт, 2020. – 479 с. – (Профессиональное образование). – ISBN 978-5-534-00859-3. – EDN DIXQIX.
16. *Гайсенок, О. В.* Анализ прогнозирования выявления каротидного атеросклероза в зависимости от градаций уровня липопротеина (а) / О. В. Гайсенок // Анализ риска здоровью. – 2022. – № 3. – С. 133–142. – DOI 10.21668/health.risk/2022.3.13. – EDN PXLAPE.
17. *Попов, В. В.* Факторы профессионального выгорания врачей-терапевтов поликлиник / В. В. Попов, Ю. А. Дьякова, И. А. Новикова // Гигиена и санитария. – 2017. – Т. 96, № 3. – С. 265–269. – DOI 10.18821/0016-9900-2017-96-3-265-269. – EDN YHSWPD.
18. Стереотипы пищевого поведения и состояние сердечно-сосудистой системы населения / С. А. Максимов, М. В. Табакаев, Я. В. Данильченко [и др.] // Гигиена и санитария. – 2017. – Т. 96, № 6. – С. 585–589. – DOI 10.18821/0016-9900-2017-96-6-585-589. – EDN ZAPEFZ.
19. Кардиоваскулярный риск у работников нефтехимических производств / З. Ф. Гимаева, И. В. Бухтияров, А. Б. Бакиров [и др.] // Гигиена и санитария. – 2020. – Т. 99, № 5. – С. 498–503. – DOI 10.33029/0016-9900-2020-99-5-498-503. – EDN HQARKF.
20. *Марцев, А. А.* Роль факторов окружающей среды в реализации эпидемического процесса по аскаридозу / А. А. Марцев // Гигиена и санитария. – 2021. – Т. 100, № 3. – С. 218–222. – DOI 10.47470/0016-9900-2021-100-3-218-222. – EDN UMYRAH.
21. 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.

22. Использование регрессионных моделей в системе поддержки принятия решений по управлению риском для здоровья населения в результате воздействия социально-экономических факторов / В. Б. Гурвич, Б. И. Никонов, О. Л. Малых [и др.] // Уральский медицинский журнал. – 2008. – № 8(48). – С. 26–33. – EDN JXCVXL.
23. Global Tuberculosis Report 2019 / World Health Organization. – Geneva: WHO Press, 2019. – 297 p. – ISBN 978-92-4-156571-4.
24. Туберкулез у взрослых: клинические рекомендации / И. А. Васильева, А. В. Абрамченко, С. А. Андронов [и др.] – М. : Минздрав РФ, 2024. – 167 с.
25. Глобальное бремя туберкулеза в России и в мире как проблема общественного здоровья (историко-аналитический обзор) / З. М. Загдын, Н. В. Кобесов, Л. И. Русакова [и др.] // Туберкулез и болезни легких. – 2023. – Т. 101, № 5. – С. 78–88. – DOI 10.58838/2075-1230-2023-101-5-78-88. – EDN RNIUPD.
26. *Hosmer, D*. Applied Logistic Regression / D. Hosmer, S. Lemeshow. – 2nd ed.– New York: Wiley, 2000. – 397 p. – ISBN 0-471-35632-8.
27. *Shoukri, M. M*. Statistical methods for health sciences. Second Edition / M. M. Shoukri, C. A. Pause. – CRC Press, 1999. – 390 p. – ISBN: 0-8493-1095-4.
28. *Peng, C. Y. J.* An Introduction to Logistic Regression Analysis and Reporting / C. Y. J. Peng // Journal of Educational Research. – 2002. – Vol. 96, No. 1. – P. 3–14. – EDN EGUCMX.
29. *Sweet, S. A.* Data Analysis with SPSS. A First Course in Applied Statistics / S. A. Sweet, K. Grace-Martin. – 4th ed. – Pearson, 2010. – 288 p. – ISBN-13: 978-0205019670.
30. Association between fluid accumulation at initiation of acute renal replacement therapy and mortality in critically ill patients with acute kidney injury / I. Skorniakov, A. Varaksin, D. Vikulova [et al.] // Critical Care. – 2017. – Vol. 21, No. 1. – P. 515. – DOI 10.1186/s13054-017-1629-x. – EDN ESOSQD.
31. 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. – EDNZCTUDM.
32. Effects of Preoperative Preparation on the Vaginal Mucosa in Women with Genital Prolapse Associated with Genitourinary Menopausal Syndrome / A. A. Mikhelson, M. V. Lazukina, A. N. Varaksin [et al.] // Acta Scientific Women's Health. –2023. – Vol. 5, No. 4. – P. 83–97. – DOI: 10.31080/ASWH.2023.05.0494. – EDN UWBRGC.
33. *Аптон Г*. Анализ таблиц сопряженности / Г. Аптон. – М. : Финансы и статистика, 1982. – 143 с.

Сведения об авторах:

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

Шалаумова Юлия Валерьевна, к. т. н., старший научный сотрудник Института экологии растений и животных УрО РАН, г. Екатеринбург, Россия. Эл. Почта: yulyash@gmail.com.

Константинова Екатерина Даниловна, к. ф.-м. н., старший научный сотрудник, и. о. заведующего лабораторией биостатистики Института промышленной экологии УрО РАН, г. Екатеринбург, Россия. Эл. почта: K_Konst@ecko.uran.ru.

Маслакова Татьяна Анатольевна, к. ф.-м. н., научный сотрудник Института промышленной экологии УрО РАН, г. Екатеринбург, Россия. Эл. почта: masta@ecko.uran.ru.

ADEQUACY OF REGRESSION MODELS IN ECOLOGY AND MEDICINE: PRIMARY DATA, STRATIFICATION AND MOVING AVERAGE

A. N. Varaksin [1], Yu. V. Shalaumova [2], E. D. Konstantinova [1], T. A. Maslakova [1]

[1] Institute of Industrial Ecology, Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russia

[2] Institute of Plant and Animal Ecology, Ural Branch of the Russian Academy of Sciences, 620144, Ekaterinburg, Russia

Approaches to investigating the adequacy of linear and logistic regression models used in the fields of ecology and medicine are considered. Conventional approaches to testing the adequacy of linear regression models based on correlation relations are practically not used in real research. The methods of visual (expert) assessment of the linearity of the response-predictor relationship are not widely used in scientific publications. In this paper, we propose the utilisation of moving average methods as a method of investigating the adequacy of linear regression models. For logistic regression models, moving average and cumulative likelihood function methods, combined with graphical presentation of results, are proposed in addition to traditional stratification methods for testing the adequacy. The effectiveness of these approaches is demonstrated through examples.

Keywords: linear regression; logistic regression; model adequacy; stratification; moving average; cumulative probability function.

References

1. *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.
2. *Siegel, A.* Practical Business Statistics / A. Siegel. – M. : OOO «I. D. Vil'jams», 2017. – 1056 p. – ISBN 978-5-8459-1367-8. (In Russian).
3. *Aivazyan, S. A*. Applied Statistics. Study of Relationships / S. A. Aivazyan, I. S .Yenyukov, L. D. Meshalkin. – M.: Finansy i statistika, 1985. – 487 p. (In Russian).
4. *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.
5. *Afifi, A*. Statistical Analysis: A Computer Oriented Approach / A. A. Afifi, S. P. Azen. – M. : Mir, 1982. – 488 p. (In Russian).
6. 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).
7. *Varaksin, A. N.* Regression models in medico-ecological monitoring: an algorithm for taking into account correlations between predictors / A. N. Varaksin // Ecological systems and devices. – 2025. – No. 1. – P. 3–14. – DOI 10.25791/esip.1.2025.1495. – EDN NVRCWH. (In Russian).
8. *Aivazyan, S. A*. Applied Statistics. Bases of Modelling and Initial Data Processing / S. A. Aivazyan, I. S. Yenyukov, L. D. Meshalkin – M. : Finansy i statistika. 1983. – 471 p. (In Russian).
9. 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.
10. *Vladimirov, V. G.* Radioprotective formulations. Optimization of composition and mechanism of action / V. G. Vladimirov, G. A. Poddubsky, G. I. Razorenov. – L. : Voen izdat, 1988. – 144 p. (In Russian).
11. *Nalimov, V. V*. Theory of experiment / V. V. Nalimov. – M. : Nauka, 1971. – 206 p. (In Russian).
12. *Fletcher, R. H.* Clinical epidemiology: The essentials / R. H. Fletcher, S. W. Fletcher. – 5th ed. – Lippincott Williams & Wilkins, 2012. – 272 p. – ISBN 978-1451144475.
13. *Mosteller, F.* Data analysis and regression / F. Mosteller, J. W. Tukey. – M. : Finansy i statistika, 1982. – 378 p. (In Russian).

14. *Tukey, J. W.* Analysis of the observation results. Exploratory analysis / J. W. *Tukey* – M. : Mir,. 1981. – 400 p. (In Russian).
15. *Gmurman, V. E*. Probability theory and mathematical statistics: textbook. / V. E. Gmurman 12th ed. – M. : Yurayt Publishing House, 2020. – 479 p. – (Professional education). – ISBN 978-5-534-00859-3. – EDN DIXQIX. (In Russian).
16. *Gaisenok, O. V*. Predicting the detection of carotid atherosclerosis depending on the gradations of lipoprotein (a) levels / O. V. Gaisenok // Health risk analysis. – 2022. – No. 3. – P. 133-142. – DOI 10.21668/health.risk/2022.3.13. – EDN PXLAPE. (In Russian).
17. *Popov, V. V*. Professional burnout factors in physicians of polyclinics / V. V. Popov, Yu. A. Dyakova, I. A. Novikova // Hygiene and Sanitation. – 2017. – Vol. 96, No. 3. – P. 265–269. – DOI 10.18821/0016-9900-2017-96-3-265-269. – EDN YHSWPD.(In Russian).
18. Dietary patterns and cardiovascular health of the population / S. A. Maksimov, M. V. Tabakaev, Y. V. Danilchenko [et al.] // Hygiene and Sanitation. – 2017. – Vol. 96, № 6. – С. 585–589. – DOI 10.18821/0016-9900-2017-96-6-585-589. – EDN ZAPEFZ.(In Russian).
19. Cardiovascular risk in petrochemical workers / Z. F. Gimaeva, I. V. Bukhtiyarov, A. B. Kaptsov [et al.] // Hygiene and Sanitation. – 2020. – Vol. 99, No. 5. – P. 498–503. – DOI 10.33029/0016-9900-2020-99-5-498-503. – EDN HQARKF.(In Russian).
20. *Martsev, A. A.* The role of environmental factors in realizing the epidemic process on ascariasis / A. A. Martsev // Hygiene and Sanitation. – 2021. – Vol. 100, No. 3. – P. 218–222. – DOI 10.47470/0016-9900-2021-100-3-218-222. – EDN UMYRAH.(In Russian).
21. 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: publ. 03.04.2019 / Yu. V. Shalaumova, A. N. Varaksin, V. G. Panov; The applicant is the Federal State Budgetary Institution of Science Institute of Industrial Ecology of the Ural Branch of the Russian Academy of Sciences (IPE Ural Branch of the Russian Academy of Sciences). – EDN MVJCBS.(In Russian).
22. The use of regression models in the decision support system for managing the risk to public health as a result of the impact of socio-economic factors / V. B. Gurvich, B. I. Nikonov, O. L. Malyh [et al.] // Ural Medical Journal – 2008. – No. 8(48). – P. 26–33. – EDN JXCVXL. (In Russian).
23. Global Tuberculosis Report 2019 / World Health Organization. – Geneva: WHO Press, 2019. – 297 p. – ISBN 978-92-4-156571-4.
24. Tuberculosis in adults: clinical recommendations / I. A. Vasil'eva, A. V. Abramchenko, S. A. Andronov [et al.] – M. : Minzdrav RF, 2024. – 167 p. (In Russian).
25. Global tuberculosis burden in Russia and the world as a public health problem (historical and analytical review) / Z. M. Zagdyn, N. V. Kobesov, L. I .Rusakova [et al.] // Tuberculosis and Lung Diseases. – 2023. – Vol. 101, No. 5. – P. 78–88. – DOI 10.58838/2075-1230-2023-101-5-78-88. – EDN RNIUPD. (In Russian).
26. *Hosmer, D*. Applied Logistic Regression / D. Hosmer, S. Lemeshow. – 2nd ed.– New York: Wiley, 2000. – 397 p. – ISBN 0-471-35632-8.
27. *Shoukri, M. M*. Statistical methods for health science. Second Edition / M. M. Shoukri, C. A. Pause. – CRC Press, 1999. – 390 p. – ISBN: 0-8493-1095-4.
28. *Peng, C. Y. J.* An Introduction to Logistic Regression Analysis and Reporting / C. Y. J. Peng // Journal of Educational Research. – 2002. – Vol. 96, No. 1. – P. 3–14. – EDN EGUCMX.
29. *Sweet, S. A.* Data Analysis with SPSS. A First Course in Applied Statistics / S. A. Sweet, K. Grace-Martin. – 4th ed. – Pearson, 2010. – 288 p. – ISBN-13: 978-0205019670.
30. Association between fluid accumulation at initiation of acute renal replacement therapy and mortality in critically ill patients with acute kidney injury / I. Skorniakov, A. Varaksin, D. Vikulova [et al.] // Critical Care. – 2017. – Vol. 21, No. 1. – P. 515. – DOI 10.1186/s13054-017-1629-x. – EDN ESOSQD.
31. 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).
32. Effects of Preoperative Preparation on the Vaginal Mucosa in Women with Genital Prolapse Associated with Genitourinary Menopausal Syndrome / A. A. Mikhelson, M. V. Lazukina, A. N. Varaksin [et al.] // Acta Scientific Women's Health. –2023. – Vol. 5, No. 4. – P. 83-97. – DOI: 10.31080/ASWH.2023.05.0494. – EDN UWBRGC.
33. *Upton, G.* The Analysis of Cross-tabulated Data / G. Upton – M. : Finansy i statistika, 1982. – 143 p. (In Russian).