DOI: https://doi.org/10.17816/humeco642576 EDN: XXYJJP

Применение логистической регрессии в эпидемиологии: первичные данные, стратификация и скользящее среднее

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

[1] Институт промышленной экологии Уральского отделения Российской академии наук, Екатеринбург, Россия;
[2] Институт экологии растений и животных Уральского отделения Российской академии наук, Екатеринбург, Россия

АННОТАЦИЯ

Обоснование. Методы логистической регрессии являются наиболее используемыми для установления статистических связей между количественными предикторами *X* и дихотомическим откликом *Y* (*Y*=0 или *Y*=1). Именно поэтому разработка новых подходов к анализу связей между *X* и *Y* такого типа является актуальной.

Цель. Показать особенности применения методов стратификации, скользящего среднего и функции кумулятивной вероятности при построении и анализе моделей логистической регрессии в задачах оценки риска здоровью.

Материалы и методы. Для анализа моделей логистической регрессии используются методы стратификации, скользящего среднего, функции кумулятивной вероятности, а также критерии согласия и методы сравнения долей.

Результаты. Показано, что стандартные методы стратификации недостаточны для оценки характера связей между дихотомическим *Y* и количественным *X*. Дополнительные методы (скользящее среднее и функция кумулятивной вероятности) позволяют выявить особенности этих связей. Показана роль графического представления результатов логистической регрессии для понимания статистических связей между переменными *X* и *Y*. Результаты применения методов стратификации, скользящего среднего и функции кумулятивной вероятности иллюстрируются примерами из области эпидемиологии.

Заключение. Методы скользящего среднего и функции кумулятивной вероятности в сочетании со стратификацией позволяют надёжно идентифицировать характер связи между дихотомическим *Y* и количественным *X* и выявить возможные отклонения от условий применимости моделей логистической регрессии.

Ключевые слова: модели логистической регрессии; адекватность модели; статистическая значимость; стратификация; скользящее среднее; функция кумулятивной вероятности; сердечно-сосудистые заболевания; заболевания щитовидной железы.

Как цитировать:
Вараксин А.Н., Шалаумова Ю.В., Маслакова Т.А. Применение логистической регрессии в эпидемиологии: первичные данные, стратификация и скользящее среднее // Экология человека. 2024. Т. 31, № 9. С. 678–691. DOI: 10.17816/humeco642576 EDN: XXYJJP

Рукопись поступила: 05.12.2024 **Рукопись одобрена:** 18.02.2025 **Опубликована online:** 23.03.2025

DOI: https://doi.org/10.17816/humeco642576 EDN: XXYJJP

Application of logistic regression in epidemiology: primary data, stratification and moving average

Anatoly N. Varaksin[1], Yulia V. Shalaumova[2], Tatiana 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, Ekaterinburg, Russia

ABSTRACT

BACKGROUND: Logistic regression is the most commonly used method for establishing statistical relationships between quantitative predictors X and a dichotomous response Y (Y=0 or Y=1). Therefore, it is relevant to develop new approaches to the analysis of relationships between X and Y of this type.

AIM: To demonstrate the specific characteristics of the application of stratification, moving average and cumulative probability function methods in the construction and analysis of logistic regression models in the context of health risk assessment.

MATERIALS AND METHODS: The analysis of logistic regression models employs a range of statistical methods, including the stratification, moving average, cumulative probability function, goodness-of-fit tests, and proportion comparison tests.

RESULTS: It is shown that the standard stratification methods are not sufficient for exploring the nature of the relationships between dichotomous Y and quantitative X. Additional methods, including moving average and cumulative likelihood function, facilitate the identification of features characterizing these relationships. The utility of graphical representations of logistic regression results in elucidating the statistical relationships between variables X and Y is demonstrated. The efficacy of the stratification, moving average and cumulative probability function methods is illustrated by examples from the field of epidemiology.

CONCLUSION: The combination of moving average and cumulative probability function methods with stratification enables the reliable identification of the nature of the relationship between dichotomous Y and quantitative X, as well as the potential for deviations from the conditions of applicability of logistic regression models.

Keywords: logistic models; model adequacy; statistical significance; stratification; moving average; cumulative probability function; cardiovascular diseases; thyroid diseases.

To cite this article:
Varaksin AN, Shalaumova YuV, Maslakova TA. Application of logistic regression in epidemiology: primary data, stratification and moving average. *Ekologiya cheloveka (Human Ecology).* 2024;31(9):678–691. DOI: 10.17816/humeco642576 EDN: XXYJJP

Submitted: 05.12.2024 **Accepted:** 18.02.2025 **Published online:** 23.03.2024

DOI: https://doi.org/10.17816/humeco642576 EDN: XXYJJP

流行病学中的逻辑回归应用：原始数据、分层和移动平均数

Anatoly N. Varaksin[1], Yulia V. Shalaumova[2], Tatiana 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, Ekaterinburg, Russia

摘要

论证。逻辑回归法是建立定量预测因子X与二元响应变量Y（Y=0或Y=1）之间统计关系的最常用方法。这就是开发新的方法来分析X和Y之间的关系变得如此迫切的原因。

目的。说明在健康风险评估任务中构建和分析逻辑回归模型时应用分层、移动平均数和累积概率函数方法的特殊性。

材料和方法。使用分层、移动平均数、累积概率函数，以及拟合优度准则和份额比较方法来分析逻辑回归模型。

结果。结果表明，标准的分层方法不足以评估二元变量Y与定量X之间关系的性质。其他方法（移动平均数和累积概率函数）可以确定这些关系的特性。逻辑回归结果的图形表示法在理解变量X和 Y之间的统计关系方面的作用显而易见。以流行病学领域的实例说明了分层法、移动平均数和累积概率函数法的应用结果。

结论。移动平均数和累积概率函数法与分层相结合，能够可靠地确定二元变量Y与定量X之间关系的性质，并确定逻辑回归模型适用条件的可能偏差。

关键词：逻辑回归模型；模型充分性；统计意义；分层；移动平均数；累积概率函数；心血管疾病；甲状腺疾病。

引用本文:
Varaksin AN, Shalaumova YuV, Maslakova TA. 流行病学中的逻辑回归应用：原始数据、分层和移动平均数. *Ekologiya cheloveka (Human Ecology).* 2024;31(9):678–691. DOI: 10.17816/humeco642576 EDN: XXYJJP

收到: 05.12.2024 **接受:** 18.02.2025 **发布日期:** 23.03.2024

ОБОСНОВАНИЕ

Модели логистической регрессии (ЛогР) в эпидемиологии, по терминологии С.А. Айвазяна и соавт. [1, 2], могут быть использованы в двух качествах: как метод исследования зависимостей (определение коэффициентов модели, расчёт отношения шансов и доверительных интервалов) и как метод классификации (построение классификационной матрицы, расчёт чувствительности и специфичности, ROC-анализ). В данной работе мы обсуждаем ЛогР только как метод исследования зависимостей. При таком использовании ЛогР, кроме стандартной оценки статистической значимости, обязательна проверка адекватности модели первичным данным.

Терминология

Первичные данные (primary data) — эпидемиологические данные для каждого объекта исследования (каждого работника, пациента). Примеры первичных данных: статус здоровья каждого работника или пациента (0 — здоров, 1 — болен), возраст, индекс массы тела (ИМТ), уровень гемоглобина и т.д. Стратификация (stratification) — разделение первичных данных на интервалы (страты). Например, возраст можно разделить на страты 20–24 года, 25–29 лет и т.д. В этих стратах, организованных в данном случае по показателю «возраст», можно рассчитать усреднённые характеристики всех интересующих исследователя первичных показателей, таких как статус здоровья (доля больных), средние значения ИМТ и уровня гемоглобина и т.п. Скользящее среднее (moving average) — это та же стратификация, только с перекрывающимися стратами. Например, возраст в первой страте от 20 до 24 лет, возраст во второй страте — от 21 до 25 лет, в третьей — от 22 до 26 лет и т.д.

При построении моделей ЛогР (и других статистических моделей) необходимо выполнить две проверки: проверка модели на адекватность первичным данным, в случае адекватности модели — проверка статистической значимости модели.

Адекватность модели

Наиболее однозначно смысл понятия адекватности статистических моделей сформулировали А. Афифи и С. Эйзен [3]: «Под адекватностью модели простой линейной регрессии подразумевается, что никакая другая модель не даёт значимого улучшения в предсказании *Y*». В данном тексте речь идёт конкретно о моделях *линейной* регрессии, связывающих количественные переменные — предиктор *X* и отклик *Y*; такое же понятие адекватности модели, очевидно, применимо к любой статистической модели. Предельная «экстремальность» высказывания авторов [3] о понятии адекватности наталкивается на невозможность его реализации в полном объёме просто в силу большого числа различных моделей, которые в принципе могут быть построены на основе конкретных первичных данных; также зачастую встаёт вопрос о критериях улучшения предсказания *Y* [1].

Более реалистичной с точки зрения практического применения статистических моделей выглядит позиция С.А. Айвазяна и соавт. (речь опять идёт о линейной регрессии, но все предложения в полной мере применимы для ЛогР). Говоря о критериях адекватности, С.А. Айвазян и соавт. пишут [1] следующее: «...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 моделей ЛогР. Подчеркнём, однако, что «предельная» позиция А. Афифи и С. Эйзена [3] относительно понятия адекватности модели является полезной для общего понимания термина «адекватность статистической модели».

Статистическая значимость модели

Статистическая значимость модели ЛогР — это установление факта, что связь между *X* и *Y* является неслучайной на некотором уровне значимости α [1, 3, 4]. Понятия адекватности и статистической значимости модели — это два различных понятия, относящиеся к различным сторонам построения и анализа статистических моделей.

Логистическая регрессия

Одной из наиболее распространённых моделей нелинейной регрессии является ЛогР, которая применяется для описания статистических связей между дихотомическим откликом *Y* (*Y* принимает два значения: *Y*=0 и *Y*=1) и предикторами *X*, которые могут быть количественными или ранговыми. Данные такого рода часто встречаются в эпидемиологических исследованиях, когда дихотомический *Y* кодирует, например, 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 ЛогР статистическая связь между *Y* и одним предиктором *X* предполагается в следующем виде [5–7]:

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

где $W(Y=1|X=x)$ — вероятность обнаружения в первичных данных значения *Y*=1 для заданного значения *X*=x.

При использовании соотношения (*1*) имеем следующее:

$$\ln\left(\frac{W}{1-W}\right)=b_0+b_1x. \quad (2)$$

DOI: https://doi.org/10.17816/humeco642576

Таким образом, при выполнении (*1*) имеем линейную связь между предиктором *X* и комплексом ln(*W*/(1–*W*)), который называют логитом — logit (*W*) [5–8].

Соотношение (*2*) — это условие применимости модели ЛогР. Поскольку особых ограничений на тип предиктора *X* (количественный, ранговый) в ЛогР обычно не предъявляется [5], многие авторы считают, что ЛогР можно использовать везде, где есть дихотомический отклик *Y* (примеры таких публикаций будут приведены ниже). Это, однако, не так. Во-первых, кроме ЛогР, существуют другие методы анализа для дихотомического отклика *Y*, например, пробит-регрессия [9]. Во-вторых, для конкретных эпидемиологических данных вида $W(Y=1|X)$ связь дихотомического *Y* с предиктором *X* может оказаться любой. Поэтому проверка условия (*2*) при использовании ЛогР в качестве метода исследования зависимостей является обязательной. При выполнении условия (*2*) воздействие предиктора *X* на вероятность *W* характеризуется отношением шансов (OR), которое при изменении *X* на единицу рассчитывается по формуле: OR=exp(b1), где b_1 — коэффициент модели (*2*). Только при выполнении условия (*2*) OR — это число, одинаковое для любого значения *X* (модель ЛогР характеризуется одним числом!). Именно поэтому в ЛогР используют OR, а не относительный риск, как принято во многих работах по оценке риска [10]. Если же logit(W) не является линейной функцией *X*, тогда для различных *X* OR будет разным и модель ЛогР уже не будет характеризоваться одним значением OR. Поэтому очень важно сначала подтвердить линейность связи *X* и *Y* и только потом использовать показатель OR для характеристики воздействия *X* на *Y*.

В данной работе мы рассматриваем модели ЛогР с одним предиктором (модели простой ЛогР), модели множественной регрессии заслуживают отдельного рассмотрения. Рассмотрены также только *количественные* предикторы *X*, для которых в ЛогР возможна не только стратификация, но и расчёты скользящего среднего.

Адекватность и статистическая значимость модели логистической регрессии

Адекватность модели ЛогР проверяется путём расчёта критериев согласия для *стратифицированных* первичных данных, например, критерия χ^2 D. Hosmer и S. Lemeshow [5]:

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

где суммирование (индекс *i*) идёт по стратам, n_i, $W_{obs,i}$, $W_{calc,i}$ — соответственно число наблюдений в страте, наблюдаемые и расчётные вероятности; вероятность $W_{obs,i}$ — среднее значение дихотомического отклика *Y* в *i*-страте, содержащей n_i наблюдений, а $W_{calc,i}$ — результат расчёта по формуле (*1*) для среднего в *i*-страте значения *X*. Если значение χ^2 D. Hosmer и S. Lemeshow, рассчитанное по первичным данным в результате их стратификации, не превосходит критического значения $\chi^2_{крит.}(\alpha, \nu)$, модель признаётся 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, есть смысл проверить её статистическую значимость.

Статистическая значимость модели ЛогР проверяется по критерию Стьюдента или по критерию Вальда [5]. В нашем случае одного предиктора *X* — это проверка значимости отличия от нуля коэффициента b_1 или отличия от единицы OR в формулах (*1*)–(*2*).

Цель исследования. Показать особенности применения методов стратификации, скользящего среднего и функции кумулятивной вероятности при построении и анализе моделей ЛогР в задачах эпидемиологии.

МАТЕРИАЛЫ И МЕТОДЫ

Для иллюстрации применения названных методов в работе использованы первичные данные из трёх источников: 1) данные из монографии D. Hosmer и S. Lemeshow [5] — 100 пациентов в возрасте от 20 до 69 лет, у некоторых из них диагностированы сердечно-сосудистые заболевания (ССЗ); 2) данные из работ [11, 12], в которых представлены результаты профилактических осмотров 820 мужчин 25–66 лет — работников Уральских предприятий, у которых диагностированы ССЗ и есть сведения об ИМТ; 3) данные из работ [13, 14], в которых представлены результаты обследования 100 женщин 51–79 лет в состоянии постменопаузы, у которых диагностированы различные сопутствующие заболевания, в том числе щитовидной железы (ЩЖ).

Статистический анализ данных проведён с использованием методов ЛогР, стратификации, скользящего среднего, функции кумулятивной вероятности, критерия согласия, методов сравнения долей. Расчёты выполнены в пакете Statistica 10.0 (StatSoft, USA).

Этическая экспертиза

Проведение исследования одобрено локальным этическим комитетом ИПЭ УрО РАН (протокол №3 от 05.06.2023).

РЕЗУЛЬТАТЫ

Анализ моделей ЛогР, использующих первичные данные либо данные стратификации или скользящего среднего, показан на конкретных эпидемиологических примерах. Каждый пример демонстрирует определённые особенности применения методов стратификации и скользящего среднего для анализа статистической связи между дихотомическим откликом *Y* и количественным предиктором *X* в различных ситуациях.

Пример 1. Заболеваемость ССЗ в зависимости от возраста

Первичные данные. На рис. 1 представлены результаты расчётов методом ЛогР для первичных данных, взятых из монографии D. Hosmer и S. Lemeshow [5]. Здесь отклик *Y* представляет ССЗ у 100 пациентов: *Y*=0 означает отсутствие заболевания, *Y*=1 — наличие заболевания у пациента данного возраста.

Очевидно, что данные рис. 1 не позволяют оценить степень выполнения соотношений (1)–(2). Действительно, наличие у отклика *Y* всего двух значений (нуль и единица) исключает всякую возможность визуальной (экспертной) оценки формы связи между *Y* и *X* [7], как это возможно, например, в линейной регрессии [3, 4, 7]. Кривая ЛогР на рис. 1 проведена согласно модели, которая построена по первичным данным [5] для 100 пациентов:

$$W(ССЗ)=\frac{\exp(-5{,}309+0{,}1109\times Возраст)}{1+\exp(-5{,}309+0{,}1109\times Возраст)}. \qquad (3)$$

OR=1,117; доверительный интервал (ДИ): 1,065–1,172; p <0,0001. Нет, однако, никакой уверенности, что вероятность W(ССЗ) иметь заболевание ССЗ описывается именно такой функцией, если опираться только на визуальный анализ нулей и единиц на рис. 1.

Стратификация. В ЛогР известны методы проверки линейности связи logit(W) с *X* типа (*2*), то есть проверки адекватности модели методом стратификации первичных данных. Многие известные статистические пакеты (например, SAS, SPSS) включают опции, которые предполагают разделение предиктора *X* на непересекающиеся страты (процедура стратификации) с последующим расчётом статистических критериев согласия для проверки гипотезы о линейности связи logit(W) с *X,* например, критерий χ^2

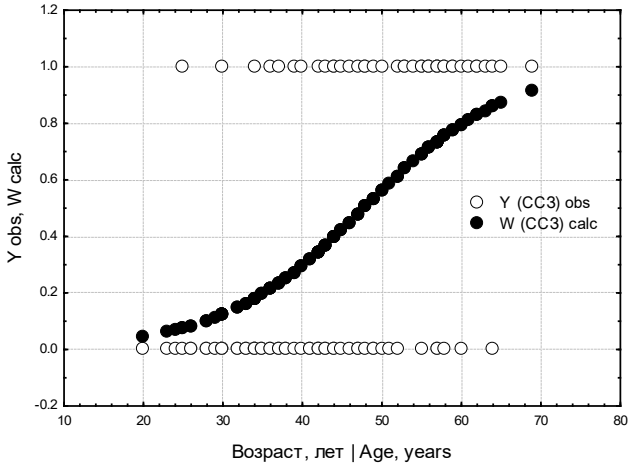

Рис. 1. Первичные (obs) данные для *Y* (*Y*=0 либо *Y*=1 для каждого из 100 пациентов, открытые кружки) и вероятности иметь сердечно-сосудистые заболевания $W(ССЗ)_{calc}$ согласно расчётам методом логистической регрессии (сплошные кружки).

Fig. 1. Primary (obs) data for *Y* (*Y*=0 or *Y*=1 for 100 patients, open circles) and probabilities $W(cardiovascular\ disease)_{calc}$ according to logistic regression calculations (full circles).

D. Hosmer и S. Lemeshow [5]. Пользователи статистических пакетов в ЛогР применяют критерии согласия, однако часто ограничиваются лишь определением значений критерия согласия и уровня значимости *p* без графического представления полученных результатов. Как будет показано ниже, графическое представление результатов стратификации оказывается полезным, а в некоторых случаях даёт неожиданные результаты.

На рис. 2*a* показано сравнение данных расчёта методом ЛогР с результатами стратификации первичных данных по предиктору *X*. Предиктор *X* (возраст пациента) в работе [5] был разделён на 8 страт по возрастным категориям, в каждой страте рассчитаны средние значения возраста <Возраст> и вероятности W(ССЗ). Результаты показаны на рис. 2*a* и рис. 2*b*. В результате перехода от вероятностей W (как на рис. 2*a*) к logit(W) на рис. 2*b* логистическая кривая превращается в прямую линию. Согласно визуальной оценке, данные стратификации для logit(W) в целом также показывают линейный рост с увеличением возраста. Статистическая проверка гипотезы о линейности связи logit(W) с возрастом это подтверждает: по критерию согласия χ^2 D. Hosmer и S. Lemeshow χ^2=0,16 при шести степенях свободы (*p*=0,998), при котором гипотеза о линейности logit(W) не отвергается на уровне значимости α=0,05.

Неоднозначность стратификации. Разделение на страты по возрастным категориям не является единственно возможным, например, в монографии D. Hosmer и S. Lemeshow [5] описаны несколько различных способов стратификации предиктора *X*. Во всех способах стратификации есть элементы субъективности (неоднозначности), которые обусловлены произвольным выбором числа страт и их границ. 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 «*стандартная стратификация*»). Если для данных D. Hosmer и S. Lemeshow [5] использовать такой способ стратификации, получим результаты, изображённые на рис. 3.

Значения $W(ССЗ)_{calc}$ на рис. 3, рассчитанные методом ЛогР, остаются, очевидно, теми же самыми, что и на рис. 2*a*, но наблюдаемые значения, полученные усреднением первичных данных в стратах, получаются другими. Значение критерия χ^2=2,43 для данных на рис. 3, конечно, меньше критического $\chi^2_{крит.}(\alpha, \nu)$=15,51 для ν=8 степеней свободы и при уровне значимости α=0,05. Таким образом, адекватность модели подтверждается (*p*=0,97 много больше 0,05), но значение χ^2=2,43 на рис. 3 в 15 раз выше, чем на рис. 2*a*. Этот пример демонстрирует неоднозначность результатов стратификации, которая в определённых ситуациях может привести к неоднозначным результатам.

Скользящее среднее. Если при стратификации имеем неопределённость 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

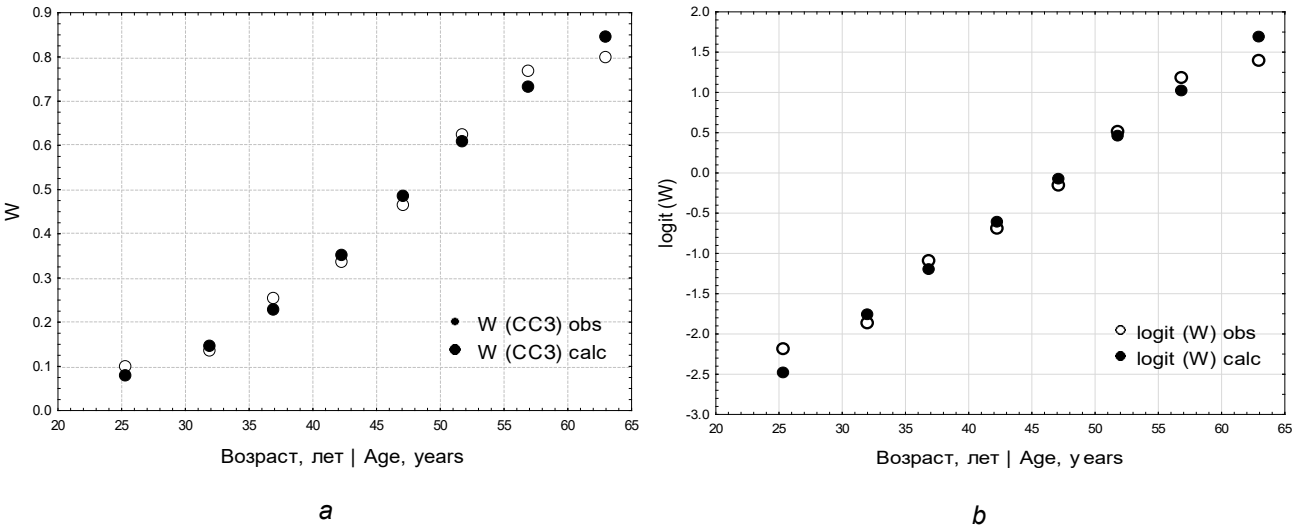

Рис. 2. Данные о сердечно-сосудистых заболеваниях (ССЗ) в восьми стратах по возрасту: *a* — вероятность W, *b* — значения logit(W). Сплошные кружки — расчёт методом логистической регрессии (calc), открытые кружки (obs) — результат стратификации.
Fig. 2. Cardiovascular disease in 8 age strata: *a*. probabilities W, *b*. logit(W). Open circles (obs) represent results from stratification; full circles show results from logistic regression calculations (calc).

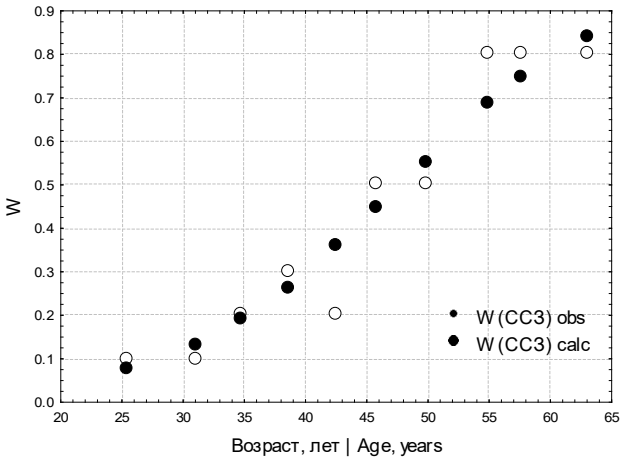

Рис. 3. Вероятности W иметь сердечно-сосудистые заболевания (ССЗ) в зависимости от возраста: стратификация 100 пациентов D. Hosmer, S. Lemeshow [5] на 10 страт с равным числом пациентов в страте; обозначения см. на рис. 2.
Fig. 3. Probabilities W (cardiovascular disease) in 10 strata with equal number of patients in each stratum as a function of age, designations as in Fig. 2.

результата [15]. Именно поэтому неоднозначность стратификации в определённой степени может быть преодолена использованием методов скользящего среднего.

На рис. 4 показаны результаты скользящего среднего с окном усреднения, равным 10 наблюдениям (что совпадает с размером страт в процедуре стандартной стратификации на рис. 3). Результаты скользящего среднего не позволяют рассчитать какой-либо критерий адекватности модели, типа критерия D. Hosmer и S. 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. 4*b* 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. Подтверждением согласованности результатов скользящего среднего и первичных данных является согласие выражения для logit(W), полученного по данным скользящего среднего (открытые кружки на рис. 4*b*) с выражением, полученным по первичным данным методом ЛогР (рис. 1, формула *3*). В случае скользящего среднего logit(W)=–5,658+0,1183×Возраст, а по первичным данным (*3*) logit(W)=–5,309+0,1109×Возраст.

Уроки примера 1. Стратификация, применяемая для проверки адекватности модели ЛогР, показывает различные результаты при различных способах стратификации (неоднозначность стратификации). Неоднозначность может быть в известной степени преодолена путём использования методов скользящего среднего. Результаты скользящего среднего показывают, что в данном примере вероятности W_{obs} обнаружения ССЗ хорошо согласуются с расчётами методом ЛогР W_{calc} (рис. 4*a*), а logit(W) является линейной функцией возраста (рис. 4*b*).

Пример 2. Заболеваемость ССЗ в зависимости от ИМТ

Первичные данные. В примере анализируется статистическая связь между распространённостью ССЗ и ИМТ у 820 мужчин 25–66 лет, являющихся работниками промышленных предприятий Свердловской области. Наличие/отсутствие ССЗ кодируется цифрами 1 и 0

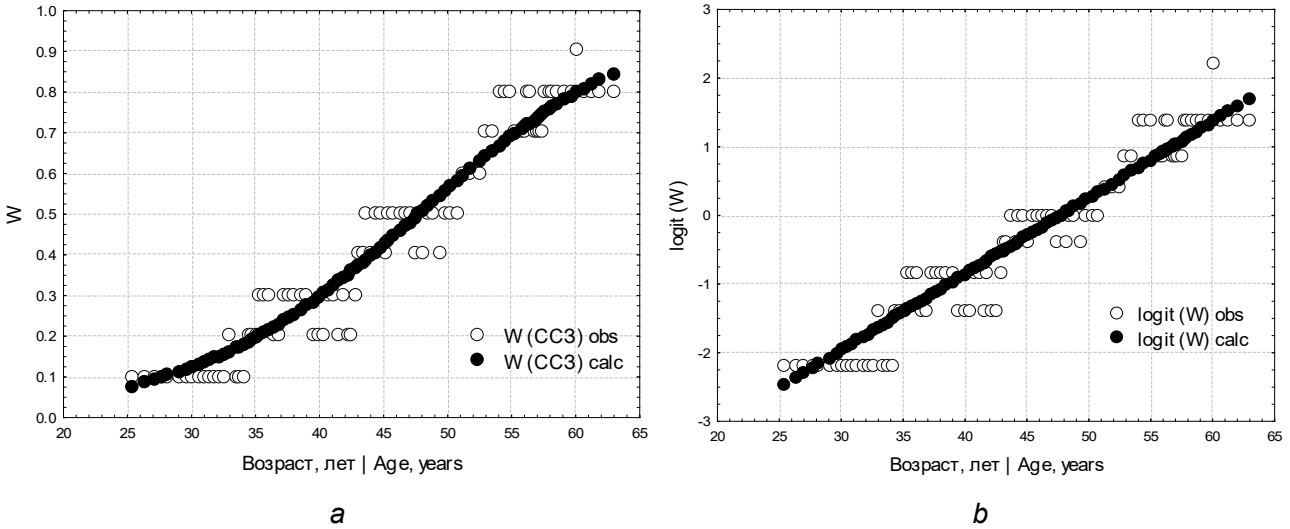

Рис. 4. Данные о сердечно-сосудистых заболеваниях (ССЗ) по результатам расчёта скользящего среднего в зависимости от среднего возраста в стратах, окно усреднения n_W=10 (открытые кружки): *a* — вероятность W, *b* — значения logit(W). Сплошные кружки — расчёты методом логистической регрессии.

Fig. 4. Cardiovascular disease from moving average data as a function of mean age in strata, averaging window is n_W=10 (open circles): *a*. probabilities W, *b*. logit(W). Full circles represent the results of logistic regression calculations (calc).

(дихотомическая переменная в ЛогР), количественный предиктор ИМТ принимает значения от 17,1 до 41,6 кг/м^2. Фактические данные взяты из работ [11, 12].

Проверка адекватности модели путём стратификации. При использовании процедуры «стандартной стратификации» 820 работников разделены на 10 страт по 82 работника в страте. Значение критерия согласия хи-квадрат D. Hosmer и S. Lemeshow, равное 7,93, при 8 степенях свободы оказалось меньше критического значения $\chi^2_{крит.}$=15,51 для уровня значимости α=0,05; следовательно, гипотеза об адекватности модели ЛогР не отклоняется (*p*=0,471). Таким образом, модель ЛогР признается адекватной фактическим данным и может использоваться для анализа связей между распространённостью ССЗ и ИМТ, в том числе для расчёта OR.

В графическом виде результаты стратификации представлены на рис. 5*a* и 5*b*. Согласно рис. 5*b*, связь наблюдаемого logit(W) с ИМТ (открытые кружки) выглядит вполне линейной, хотя и со случайными отклонениями от прямой линии, представляющей результаты расчёта методом ЛогР (сплошные кружки).

Уравнение ЛогР, построенное по первичным данным для 820 работников, имеет вид (*p*-значения для всех коэффициентов меньше 0,0001):

$$\text{logit (W)}=-4{,}6642+0{,}1554\times\text{ИМТ};$$
$$\text{OR}=1{,}168\ (\text{ДИ: } 1{,}126–1{,}212). \qquad (4)$$

Регрессия, построенная по наблюдаемым значениям logit(W), полученным в результате стратификации первичных данных (10 страт, открытые кружки на рис. 5*b*), имеет вид:

$$\text{logit(W)}=-4{,}6467+0{,}1545\times\text{ИМТ}. \qquad (5)$$

Это практически не отличается от соотношения (*4*). Попытка включить в выражение (*4*) или (*5*) нелинейные по ИМТ члены (квадратичные и кубические) показала их статистическую незначимость.

Рис. 5*a* показывает, что согласно расчётам ЛогР вероятность $W(CCC)_{calc}$ монотонно возрастает с ростом ИМТ. Согласно же результатам стратификации в первых трёх стратах для малых значений ИМТ вероятность W_{obs} не возрастает с ростом ИМТ, поэтому возникает вопрос: до какого значения ИМТ вероятность ССЗ остаётся низкой? Этот вопрос возникает потому, что результаты стратификации являются неоднозначными. Результаты стратификации не могут дать ответа на этот вопрос, однако в нашем распоряжении есть ещё один метод исследования статистических связей — функция кумулятивной вероятности, которая может ответить на поставленный вопрос.

Функция кумулятивной вероятности. Функция кумулятивной вероятности $CUSUM_{nc}(X)$ для отклика *Y* по предиктору *X* определяется соотношением:

$$\text{CUSUM}_{nc}(X)=\frac{1}{nc}\sum_{i=1}^{nc} y_i \qquad (6)$$

где *nc* — количество объектов, включённых в функцию (мы используем аббревиатуру *CUSUM*, поскольку встроенная функция с таким названием и такого же назначения имеется в пакете Statistica for Windows). Для расчёта функции кумулятивной вероятности (*6*) сначала проводится упорядочение значений предиктора *X* по возрастанию, а затем — суммирование соответствующих значений *Y*, как показано в формуле (*6*). В результате первым значением функции $CUSUM_1(X)$ будет значение Y_1, соответствующее минимальному *X*. Вторым значением функции $CUSUM_2(X)$ будет половина суммы значений Y_1 и Y_2, соответствующих двум минимальным *X*. Последним значением функции *CUSUM* будет распространённость ССЗ во всей

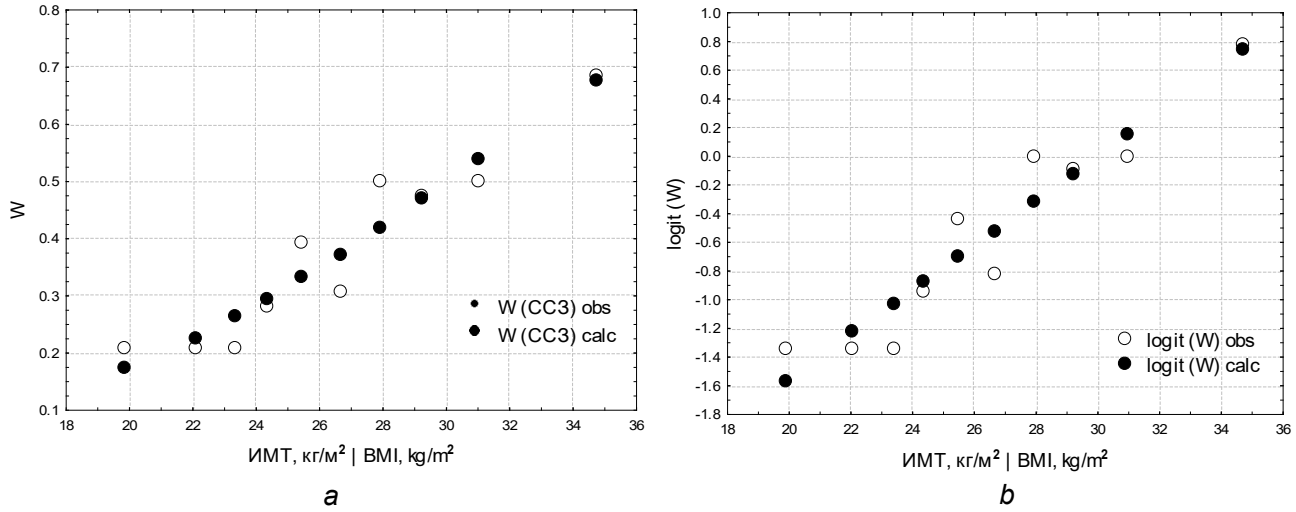

Рис. 5. Данные о сердечно-сосудистых заболеваниях (ССЗ) по результатам стратификации 820 работников на 10 страт с равным числом работников в страте в зависимости от индекса массы тела (ИМТ): *a* — вероятность W, *b* — значения logit(W), обозначения см. на рис. 2.
Fig. 5. Cardiovascular disease from data obtained by stratifying 820 workers into 10 strata with the same number of workers in each stratum according to body mass index (BMI): *a*. probabilities W, *b*. logit(W), designations as in Fig. 2.

выборке (сумма всех *Y*, делённая на число объектов в выборке). Отметим особенность функции *CUSUM*: при малом числе *nc* членов в сумме (*6*), то есть для начального участка функции *CUSUM*, характерно резкое изменение функции при добавлении нового слагаемого, а при увеличении *nc* функция *CUSUM* становится более гладкой. Именно на этом гладком участке мы можем проводить анализ *CUSUM* для получения выводов.

На рис. 6 приведён график *CUSUM* для ССЗ у 820 работников, имеющих ИМТ от 17,1 до 41,6 кг/м^2. На этом графике точка с некоторым значением ИМТ* показывает на оси ординат среднюю распространённость ССЗ для работников, имеющих ИМТ от минимального значения ИМТ до ИМТ*. Например, при значении ИМТ*=24,0 кг/м^2 (интервал ИМТ от минимального 17,1 до 24,0 кг/м^2 включает 254 работника) имеем значение *CUSUM*=0,202. Это означает, что среди всех 254 работников, у которых ИМТ меньше 24,0 кг/м^2, распространённость ССЗ достаточно низкая и равна 0,202. При увеличении ИМТ выше 24,0 кг/м^2 начинается явный рост распространённости ССЗ, она уже никогда не опускается до уровня порядка 0,2. Последнее значение функции *CUSUM* для максимального значения ИМТ равно 0,376; оно получается путём деления числа работников, имеющих ССЗ (таких работников 308), на общее число работников — 820.

Используя эту информацию, проведём новую стратификацию таким образом, чтобы одни страты включали значения ИМТ меньше 24 кг/м^2, а другие — больше 24 кг/м^2. Результаты показаны в табл. 1 и на рис. 7.

Согласно табл. 1, значения ИМТ ниже 24 кг/м^2 имеют 254 работника (достаточно много), поэтому этот диапазон ИМТ мы раздели на 3 страты. В каждой из них фактические данные W_{obs} показывают достаточно низкие значения распространённости ССЗ. Значения ИМТ выше 24 кг/м^2 были разделены на 6 страт, из которых 4 (страты 4–7) имеют диапазон 2 кг/м^2. Страта 8 включает диапазон ИМТ от 32 до 34,5 кг/м^2. Граничное значение ИМТ= 34,5 кг/м^2 выбрано также с использованием процедуры *CUSUM,* но «в обратном направлении» (значения ИМТ для расчёта *CUSUM* упорядочиваются по убыванию). При таком выборе граничного значения ИМТ в последней страте 9 со значениями ИМТ выше 34,5 кг/м^2 наблюдаем высокое значение распространённости ССЗ, равное $W(CC3)_{obs}$=0,794, что значительно выше, чем W(ССЗ)=0,690 в последней страте при стандартной стратификации на рис. 5.

Графика. Как показывает рис. 7, при ИМТ больше 24 кг/м^2 расчётные и наблюдаемые значения распространённости ССЗ хорошо согласуются друг с другом. Отсюда следует, что OR=1,168 (ДИ: 1,126–1,212), рассчитанное методом ЛогР по формуле (*4*), соответствует действительности только при ИМТ выше 24 кг/м^2. Если же использовать OR=1,168 при значениях ИМТ, меньших 24 кг/м^2, мы получим неверное представление о влиянии ИМТ на распространённость ССЗ.

Уроки примера 2. Стандартная стратификация на 10 страт с равным числом наблюдений в стратах показывает адекватность модели ЛогР первичным данным (значение критерия согласия D. Hosmer и S. Lemeshow *значительно* меньше критического для уровня значимости α=0,05). Графическое представление результатов стратификации показывает, что между данными стратификации и расчётами методом ЛогР могут наблюдаться различия в начальных стратах. Поскольку процедура стратификации содержит элементы неоднозначности, для подтверждения выводов о характере связи ССЗ и ИМТ предлагается использовать функцию кумулятивной вероятности *CUSUM*, которая не содержит неоднозначностей. Использование процедуры *CUSUM* «в прямом направлении» (то есть при увеличении ИМТ)

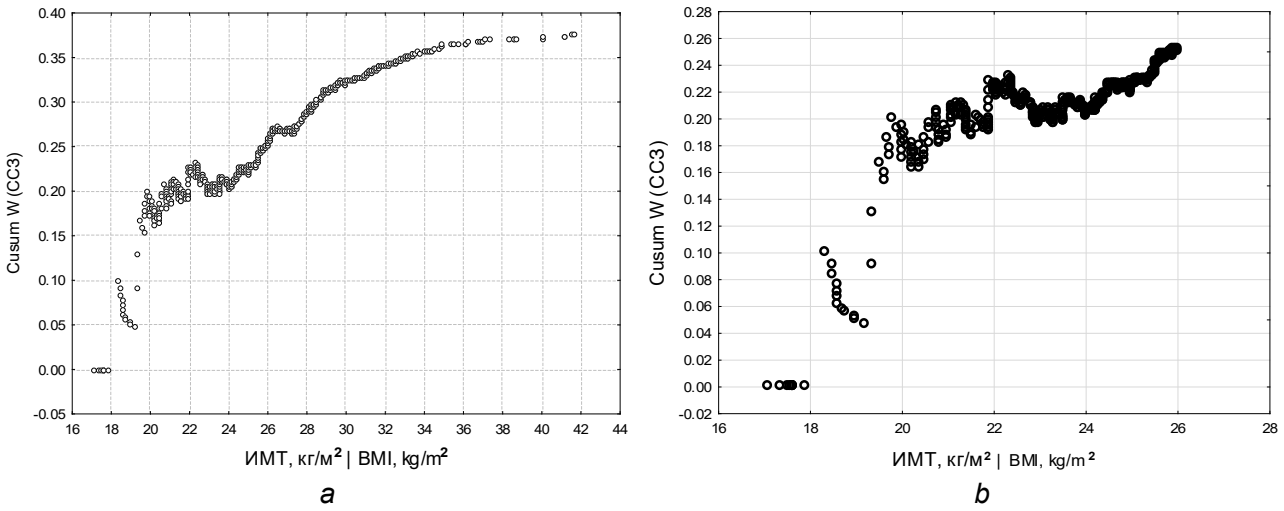

Рис. 6. Функция кумулятивной вероятности для сердечно-сосудистых заболеваний (ССЗ) в зависимости от индекса массы тела (ИМТ): *a* — для 820 работников, *b* — начальный участок для значений ИМТ, меньше 26 кг/м^2.
Fig. 6. Cumulative probability function for cardiovascular disease as a function of body mass index (BMI): *a*. for 820 workers. *b*. its initial section for BMI values less than 26 kg/m^2.

подтвердило отсутствие увеличения распространённости ССЗ при увеличении ИМТ от минимума до ИМТ=24 кг/м^2 (в этом диапазоне ИМТ распространённость ССЗ остаётся постоянной и низкой — на уровне W=0,20). Процедура *CUSUM* «в обратном направлении» (при уменьшении ИМТ) показала наличие диапазона от максимума до ИМТ=34,5 кг/м^2 с высоким значением W(ССЗ)=0,794, который не выявлялся при стандартной стратификации. Таким образом, использование функции *CUSUM* позволяет провести стратификацию более обоснованно с точки зрения выявления связи W(ССЗ) с ИМТ. В результате такой стратификации показано, что первичные данные хорошо согласуются с расчётами методом ЛогР только для ИМТ выше 24 кг/м^2.

Пример 3. Распространённость заболеваний щитовидной железы и ИМТ

В данном примере показано, что даже при выполнении условий применимости ЛогР её результаты могут *кардинально* отличаться от первичных данных, и это отличие удаётся установить только благодаря процедуре скользящего среднего.

Для построения модели ЛогР использован материал, включающий 100 женщин в возрасте от 51 до 79 лет в состоянии постменопаузы, для которых определяли в том числе антропометрические показатели и распространённость различных соматических патологий — данные НИИ охраны материнства и младенчества, Екатеринбург [13,

Таблица 1. Схема стратификации согласно рекомендациям *CUSUM*
Table 1. Stratification scheme according to *CUSUM*

Номер страты Stratum number	Диапазон ИМТ BMI interval	Среднее значение ИМТ Mean BMI	Число работников в страте Number of workers in stratum	W (сердечно-сосудистые заболевания)$_{obs}$, стратификация W_{obs} (cardiovascular disease), stratification	W (сердечно-сосудистые заболевания), метод логистической регрессии W (cardiovascular disease), logistic regression calculations
1	17,1–19,99	18,70	32	0,219	0,146
2	20,0–21,99	21,03	83	0,217	0,198
3	22,0–23,99	23,07	139	0,201	0,253
4	24,0–25,99	24,95	150	0,327	0,313
5	26,0–27,99	26,89	123	0,398	0,381
6	28,0–29,99	28,88	133	0,459	0,456
7	30,0–31,99	31,04	75	0,520	0,541
8	32,0–34,49	33,06	51	0,588	0,617
9	34,5+	37,03	34	0,794	0,751

Примечание. ИМТ — индекс массы тела.
Note. BMI — Body mass index.

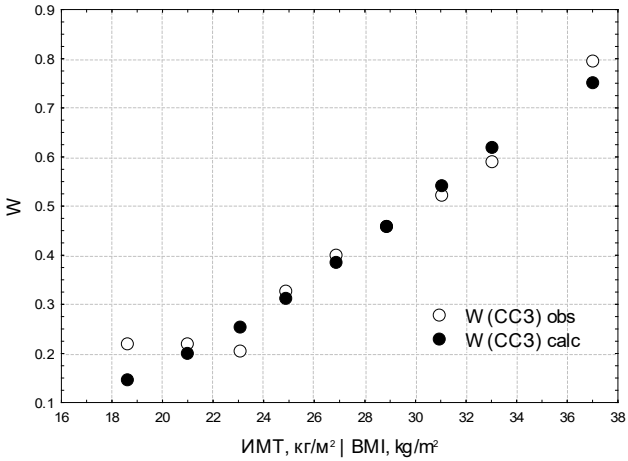

Рис. 7. Вероятности W сердечно-сосудистых заболеваний (ССЗ) по результатам стратификации, показанной в табл. 1.
Fig. 7. Probabilities W(cardiovascular disease) based on the stratification results in table 1.

14]. В примере анализируется связь распространённости заболеваний ЩЖ у женщин с ИМТ. Заболевания ЩЖ кодируются символом 0, если заболевания у данной женщины нет, и 1, если заболевание есть, ИМТ изменяется от 19,7 до 30,9 кг/м^2.

Стратификация (проверка адекватности модели ЛогР). Результаты стратификации показаны на рис. 8*a*. На их основе был рассчитан критерий согласия D. Hosmer и S. Lemeshow для проверки линейности связи logit(W) вероятности иметь заболевание ЩЖ с предиктором ИМТ. Оказалось, что при разделении предиктора ИМТ на 10 страт с равным числом наблюдений в стратах, гипотеза о линейности logit(W) не отвергается (хи-квадрат D. Hosmer и S. Lemeshow χ^2=9,89 гораздо меньше критического значения $\chi^2_{крит.}$ для восьми степеней свободы на уровне значимости α=0,05), мы имеем право использовать ЛогР для анализа связи между распространённостью ЩЖ и ИМТ.

Первичные данные. Применение метода ЛогР к первичным данным показывает, что коэффициент регрессии b_1=0,00995 в соотношении типа (*2*) статистически значимо не отличается от нуля (p=0,877). Таким образом, если ограничиться только статистическими критериями, приходим к выводу, что метод ЛогР, использованный в данном примере при *выполнении* условий его применимости, *не показал* статистически значимой связи между W(ЩЖ) и ИМТ. Если же обратиться к графическому представлению результатов стратификации (рис. 8*a*), тогда обоснованность такого вывода уже не выглядит столь однозначной.

Скользящее среднее. Для более надёжного установления характера статистической связи между ИМТ и вероятностью W иметь заболевание ЩЖ используем методы скользящего среднего (окно скользящего среднего, равное 21 пациент, выбрано с использованием функции кумулятивной вероятности). Результат, показанный на рис. 8*b*, отличается от расчётов методом ЛогР (W_{calc} на рис. 8*a*) и от результатов стандартной стратификации (W_{obs} на рис. 8*a*). Возникает вопрос: имеется ли статистически значимая связь между ИМТ и W(ЩЖ)? Проведём анализ статистической значимости различий вероятностей W(ЩЖ) для различных значений ИМТ по данным скользящего среднего (рис. 8*b*), для чего сравним страты с минимальной и максимальной W(ЩЖ). Первая страта на рис. 8*b* (максимум распространённости W=0,476) включает женщин с ИМТ от 19,7 до 24,8 кг/м^2; среднее значение ИМТ=22,1 кг/м^2. В 25-й страте по ИМТ (ИМТ от 25,3 до 27,3; среднее ИМТ=26,5 кг/м^2) наблюдаем минимальную распространённость патологии ЩЖ (W=0,143). Страты 1 и 25 содержат по 21 наблюдению и не пересекаются. По двустороннему критерию сравнения долей для независимых наблюдений W=0,476 и W=0,143 различаются статистически значимо, p=0,0195. В страте 80 (включает женщин с ИМТ от 29,6 до 30,9 кг/м^2; среднее значение ИМТ=29,8 кг/м^2) распространённость патологии ЩЖ равна W=0,429. Различие со стратой 25 также статистически значимо, p=0,0404. Таким образом, имеем статистически значимую связь распространённости заболевания ЩЖ с ИМТ, которая не описывается логистической функцией типа (*1*).

Глядя на «экзотический» рис. 8*b* можно предположить, что он появился в результате случайного стечения каких-то необычных причин и является редким исключением из общего правила. Это, однако, не так. Анализ ситуации, которая привела к рис. 8*b*, показал следующее. Пусть $\langle x_0 \rangle$ и $\langle x_1 \rangle$ — средние значения предиктора X в группах Y=0 и Y=1, а σ_0^2 и σ_1^2 — дисперсии X в этих же группах. Тогда диаграммы типа рис. 8*b* неизбежно возникают при условии $\langle x_0 \rangle = \langle x_1 \rangle$ и $\sigma_0^2 \neq \sigma_1^2$, причем рис. 8*b* с «провалом» в центре получается, когда $\sigma_1^2 > \sigma_0^2$; если же $\sigma_1^2 < \sigma_0^2$, тогда в центре рис. типа 8*b* увидим пик. Отметим также, что рисунки типа рис. 2 (линейный логит) получаются при условии $\langle x_0 \rangle \neq \langle x_1 \rangle$ и $\sigma_0^2 = \sigma_1^2$.

Уроки примера 3. Даже в случае адекватности модели ЛогР по критериям согласия возможны ситуации, когда фактический материал радикально не согласуется с моделью. *Особенно это касается случаев, когда модель оказывается статистически незначимой.* Статистическая незначимость модели может происходить именно из-за отклонения logit(W) от линейности. При этом в некоторых случаях модель признаётся адекватной фактическому материалу. Приведённый пример показывает, что такие ситуации возможны.

ОБСУЖДЕНИЕ

Для корректного использования методов ЛогР при исследовании статистической связи между дихотомическим откликом Y и количественным предиктором X необходимо выполнение условия линейности логита (проверка адекватности модели ЛогР). Анализ публикаций

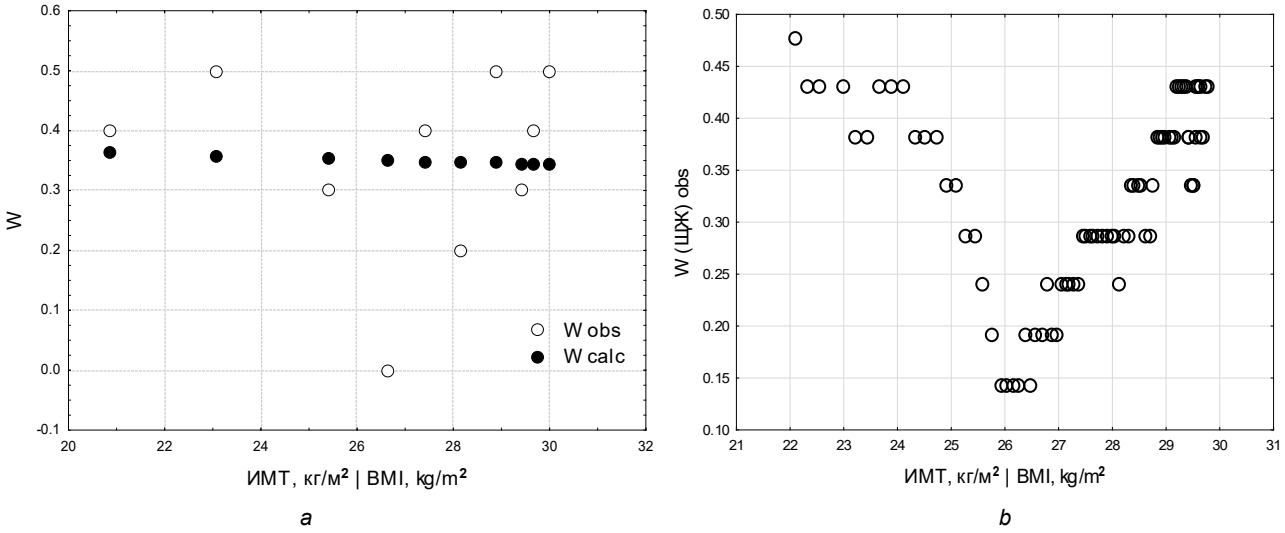

Рис. 8. Данные о заболеваниях щитовидной железы (ЩЖ): *a* — вероятность W по результатам стратификации 100 пациенток на 10 страт (открытые кружки), сплошные кружки – расчёт методом логистической регрессии, *b* — связь вероятности W с индексом массы тела (ИМТ) для данных скользящего среднего (окно скользящего среднего — 21, всего 80 страт).

Fig. 8. Thyroid disease: *a*. probabilities W from data obtained by stratifying 100 patients into 10 strata (open circles) as a function of body mass index (BMI), full circles represent results of logistic regression calculations (calc), *b*. probabilities W_{obs} for moving average data as a function of BMI (moving average window is 21 patients, total 80 strata).

по применению ЛогР в задачах эпидемиологии показывает, что в большинстве из них адекватность моделей не проверяется [16, 17]. Редкие исключения — работы типа Н.Н. Коныртаевой и соавт. [18], где выполнены обе проверки: адекватности (критерий D. Hosmer и S. Lemeshow) и статистической значимости модели ЛогР.

Отсутствие во многих публикациях проверки адекватности моделей ЛогР выглядит необъяснимым, поскольку проверка адекватности (согласия теории фактическому материалу) является обязательной во всех разделах математической статистики. Например, прежде чем представить данные в виде среднего значения и стандартного отклонения, обязательно проверяется адекватность фактических данных форме нормального распределения [19]. Также при использовании критерия Стьюдента для сравнения средних значений *X* в двух независимых выборках сначала проверяется нормальность распределения *X* в выборках и только потом — статистическая значимость различия средних. При этом проверка адекватности именно модели ЛогР многими авторами не считается обязательной. Между тем проверка адекватности модели ЛогР описанными выше способами не только отвечает на вопрос об адекватности (варианты ответа: да — нет), но и позволяет установить реальную форму статистической связи между дихотомическим *Y* и количественным *X* в имеющихся фактических данных.

Выше подчёркивалось, что адекватность и статистическая значимость модели ЛогР — это два различных понятия. Есть авторы, которые путают адекватность со значимостью. Например, C.-Y.J. Peng и соавт. [8] в статье *обучающей* направленности пишут: «отклонение нулевой гипотезы H_0: b_1=0 означает, что между *X* и logit(W) имеется линейная связь»[1].

Здесь авторы, очевидно, ошибаются. На самом деле, отклонение нулевой гипотезы H_0: b_1=0 означает, что в формуле типа (*2*) для logit(W) имеется линейное по *X слагаемое*, но не обязательно оно одно, возможно, в logit(W) присутствуют другие члены, нелинейные по *X*. Если же в logit(W) есть нелинейные члены, тогда использование OR как показателя связи между дихотомическим *Y* и количественным *X* невозможно, а гарантировать линейность logit(W) может только проверка адекватности модели, а вовсе не статистическая значимость коэффициента b_1.

Отметим, что проверка адекватности критически важна именно при использовании ЛогР как метода *исследования зависимостей*, когда необходимо быть уверенным, что расчётное значение OR действительно описывает реальную ситуацию. Если же модель ЛогР используется как *метод классификации* (прогнозирования) и даёт высокие значения чувствительности и специфичности, тогда проверка адекватности модели ЛогР не так важна; главное, чтобы модель работала (давала правильный прогноз) независимо от того, каким способом получено прогностическое правило. В некоторых работах [17] модель ЛогР используется одновременно в обоих качествах, тогда проверка адекватности модели важна именно для корректной интерпретации OR.

В данной работе показано, что ориентация только на статистические критерии может привести к искажению

[1] «Within the framework of inferential statistics, the null hypothesis states that b1 equals zero, or there is no linear relationship in the population. Rejecting such a null hypothesis implies that a linear relationship exists between X and the logit of Y» [8].

выводов, получаемых методом ЛогР. Например, показано, что критерии согласия могут не отвергать гипотезу о линейности связи предиктора *X* с logit(W), а графический анализ показывает явно нелинейную связь (см. пример 3). Активными пропагандистами графического представления и анализа результатов статистического анализа были такие известные специалисты, как Дж. Тьюки и A.F. Siegel. Американский статистик Andrew Siegel в монографии [4] практически каждый вывод иллюстрирует графиками с подробными комментариями. Один из создателей современной науки анализа данных Джон Тьюки [20] писал: «Графики, подчёркивающие лишь то, что нам уже известно, нередко не стоят места, которое они занимают. Графики, которые надо рассматривать с лупой, чтобы увидеть в них главное, 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**» (выделения в тексте принадлежат автору). Приведённые выше рис. 3, 7 и особенно 8 как раз и показывают то, что мы не ожидали увидеть.

При увеличении числа наблюдений имеет место увеличение вероятности отклонить нулевую гипотезу о линейности logit(W) в случае, если она верна. Например, в примере 1 при наличии 100 пациентов нулевая гипотеза не отклоняется с высокой вероятностью. Если же увеличить число пациентов со 100 до 650 при сохранении всех соотношений между возрастом и вероятностью ССЗ, тогда нулевая гипотеза о линейности logit(W) по результатам стандартной стратификации рис. 3 будет отклонена на уровне значимости α=0,05. Это так называемый эффект *слишком большого объёма выборки* [1], в результате которого любая нулевая гипотеза будет отклонена, если число наблюдений достаточно велико [1, 21]. Это ещё одно основание использовать графики в ЛогР для экспертной оценки линейности logit(W), особенно если число наблюдений действительно велико.

ЗАКЛЮЧЕНИЕ

Во всех приведённых выше примерах, в которых исследовалась статистическая связь между дихотомическим откликом *Y* и количественным предиктором *X*, гипотеза об адекватности модели ЛогР фактическим данным не отклоняется, следовательно, использование ЛогР является обоснованным. В то же время результаты применения ЛогР в трёх примерах совсем разные. В первом примере модельная (искусственно созданная) база данных демонстрирует полное согласие модели и данных. Во втором примере имеются отклонения от модели ЛогР в некоторых областях значений предиктора, хотя в целом модель и фактические данные находятся в согласии. В третьем примере фактические данные показывают нелинейную и даже немонотонную связь между *Y* и *X*, в то время как по критерию согласия D. Hosmer и S. Lemeshow модель ЛогР (линейная по *X*) признаётся адекватной фактическим данным. Все возможные разногласия между моделями ЛогР и фактическими данными могут быть выявлены при совместном использовании методов стратификации, скользящего среднего и функции кумулятивной вероятности в сочетании с графическим представлением и анализом результатов.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

Вклад авторов. А.Н. Вараксин — идея и концепция исследования, анализ данных, написание текста, Ю.В. Шалаумова — создание компьютерных программ для расчёта скользящего среднего и функции кумулятивной вероятности, анализ данных, написание текста, Т.А. Маслакова — анализ данных, работа с литературными источниками, написание текста. Все авторы подтверждают соответствие своего авторства международным критериям ICMJE (все авторы внесли существенный вклад в разработку концепции, проведение исследования и подготовку статьи, прочли и одобрили финальную версию перед публикацией).

Этическая экспертиза. Проведение исследования одобрено локальным этическим комитетом ИПЭ УрО РАН (протокол №3 от 05.06.2023).

Источники финансирования. Исследование выполнено в соответствии с Государственным заданием и планом НИР ИПЭ УрО РАН за счёт субсидий Минобрнауки Российской Федерации на выполнение научной темы FUMN-2024-0002 и Государственным заданием и планом НИР ИЭРиЖ УрО РАН за счёт субсидий Минобрнауки Российской Федерации на выполнение научной темы АААА-А19-119111990097-4.

Раскрытие интересов. Авторы заявляют об отсутствии отношений, деятельности и интересов за последние три года, связанных с третьими лицами (коммерческими и некоммерческими), интересы которых могут быть затронуты содержанием статьи.

Оригинальность. При создании настоящей работы авторы не использовали ранее опубликованные сведения (текст, иллюстрации, данные).

Доступ к данным. Редакционная политика в отношении совместного использования данных к настоящей работе не применима, новые данные не собирали и не создавали.

Генеративный искусственный интеллект. При создании настоящей статьи технологии генеративного искусственного интеллекта не использовали.

Рассмотрение и рецензирование. Настоящая работа подана в журнал в инициативном порядке и рассмотрена по обычной процедуре. В рецензировании участвовали два внешних рецензента, член редакционной коллегии и научный редактор издания.

ADDITIONAL INFORMATION

Author contributions. A.N. Varaksin — idea, data analysis, drafting of the text and general scientific supervision. Yu.V. Shalaumova — development of computer programs for calculation of moving average and cumulative probability function, data analysis, drafting of the text. T.A. Maslakova — data analysis, work with literary sources, drafting of the text. All authors confirm that their authorship is in accordance with the international ICMJE criteria (all authors contributed substantially to the design, conduct of the study and writing of the article, and read and approved the final version before publication).

Ethical expertise. The study was approved by the local ethics committee of the Institute of Industrial Ecology, Ural Branch of the Russian Academy of Sciences (protocol No. 3 of 06/05/2023)

Funding source. The study was supported by Ministry of Science and Higher Education of the Russian Federation, project FUMN-2024-0002 in accordance with the state assignment and the research plan of the Institute of Industrial Ecology, Ural Branch of the Russian Academy of Sciences and project AAAA-A19-119111990097-4 in accordance with the state assignment and the research plan of the Institute of Plant and Animal Ecology, Ural Branch of the Russian Academy of Sciences.

Disclosure of interests. The authors have no relationships, activities or interests for the last three years related with for-profit or not-for-profit third parties whose interests may be affected by the content of the article.
Statement of originality. In creating this work, the authors did not use previously published information (text, illustrations, data).
Data availability statement. The editorial policy regarding data sharing does not apply to this work, and no new data was collected or created.
Generative AI. Generative AI technologies were not used for this article creation.
Provenance and peer-review. This paper was submitted to the journal on an unsolicited basis and reviewed according to the usual procedure. Two external reviewers, a member of the editorial board, and the scientific editor of the publication participated in the review.

СПИСОК ЛИТЕРАТУРЫ | REFERENCES

1. Ayvazyan SA, Yenyukov IS, Meshalkin LD. *Applied statistics. Addiction research*. Moscow: Finansy i statistika; 1985. 487 p. (In Russ.)
2. Ayvazyan SA, Buchstaber VM, Yenyukov IS, Meshalkin LD. *Applied statistics. Classification and reduction of dimensionality*. Moscow: Finansy i statistika; 1989. 606 p. (In Russ.)
3. Afifi AA, Azen SP. *Statistical analysis. A computer oriented approach*. Moscow: Mir; 1982. 488 p. (In Russ.)
4. Siegel AF. *Practical business statistics*. Irwin: McGraw-Hill; 1999. 800 p.
5. Hosmer D, Lemeshow S. *Applied logistic regression*. New York: Wiley & Sons; 2000. 373 p.
6. Shoukri MM, Pause CA. *Statistical methods for health sciences*. Boca Raton: CRC Press; 1999. 384 p.
7. Afifi AA, May S, Clark V. *Computer-aided multivariate analysis*. Boca Raton: Chapman&Hall/CRC; 2003. 512 p.
8. Peng C-YJ, Lee KL, Ingersoll GM. An introduction to logistic regression. Analysis and reporting. *J. Educational Research*. 2002;96(1):3–14. doi: 10.1080/00220670209598786
9. Wooldridge JM. *Introductory econometrics: a modern approach*. Mason: South-Western; 2009. 865 p.
10. Schmidt CO, Kohlmann T. When to use the odds ratio or the relative risk? *Int J. Public Health*. 2008;53(3):165–167. doi: 10.1007/s00038-008-7068-3
11. Bakhtereva EV, Shirokov VA, Varaksin AN, Panov VG. Assessing the risk of carpal tunnel syndrome exposure occupational factors. *Ural Medical Journal*. 2015;(10):9–13. EDN: VLMSTX
12. Varaksin AN, Bakhtereva EV, Panov VG, et al. Risk factors for neurological diseases development in workers of Urals industrial plants: prognostic models based on discriminant analysis. *Ecological Systems and Devices*. 2016;(5):27–33. EDN: WMATKB
13. Mikhelson AA, Lazukina MV, Varaksin AN, et al. Erosion of the vaginal mucosa in postmenopausal women with surgical correction of genital prolapse. *Treatment and prevention*. 2020;10(4);55–64. EDN: ZCTUDM
14. Mikhelson AA, Lazukina MV, Varaksin AN, et al. Effects of preoperative preparation on the vaginal mucosa in women with genital prolapse associated with genitourinary menopausal syndrome. *Acta Scientific Women's Health*. 2023;5(4):83–97. doi: 10.31080/ASWH.2023.05.0494 EDN: UWBRGC
15. Varaksin AN, Shalaumova YuV, Maslakova TA, et al. Application of moving average methods for the construction of regression models in medical and environmental research. *Ecological Systems and Devices*. 2020;(6):12–21. doi: 10.25791/esip.06.2020.1159 EDN: XTBFAV
16. Maksimov DM, Maksimova ZV. Prevalence of smoking and hazardous drinking among industrial workers in the Sverdlovsk region. *Ekologiya cheloveka (Human Ecology)*. 2021;28(3):34–41. doi: 10.33396/1728-0869-2021-3-34-41 EDN: ICGEPK
17. Kretova IG, Vedyasova OA, Komarova MV, Shiryaeva OI. Analysis and forecasting of reserve capabilities of the organism of students according to indices of heart rate variability. *Hygiene and Sanitation*. 2017;96(6):556–561. doi: 10.18821/0016-9900-2017-96-6-556-561 EDN: ZAPEEB
18. Konyrtaeva NN, Ivanov SV, Kausova GK, et al. Leech therapy in kazakhstan: patients' characteristics and compliance with treatment. *Ekologiya cheloveka (Human Ecology)*. 2016;23(2):42–48. doi: 10.33396/1728-0869-2016-2-42-48 EDN: VQGTMZ
19. Kharkova OA, Grjibovski AM. Analysis of one and two independent samples using STATA software: parametric tests. *Ekologiya cheloveka (Human Ecology)*. 2014;21(3):57–61. EDN: RYIEZL
20. Tukey JW. *Exploratory data analysis*. Moscow: Mir; 1981. 693 p. (In Russ.)
21. Glantz S. *Primer of biostatistics*. New York: McGraw-Hill; 1992. 440 p.

ОБ АВТОРАХ

*** Вараксин Анатолий Николаевич,** д-р физ.-мат. наук, профессор;
адрес: Россия, 620990, Екатеринбург, ул. С. Ковалевской, д. 20;
ORCID: 0000-0003-2689-3006;
eLibrary SPIN: 9910-2326;
e-mail: varaksin@ecko.uran.ru

Шалаумова Юлия Валерьевна, канд. техн. наук;
ORCID: 0000-0002-0173-6293;
eLibrary SPIN: 3163-6856;
e-mail: jvshalaumova@gmail.com

Маслакова Татьяна Анатольевна, канд. физ.-мат. наук;
ORCID: 0000-0001-6642-9027;
eLibrary SPIN: 3233-7652;
e-mail: t9126141139@gmail.com

AUTHORS' INFO

*** Anatoly N. Varaksin,** Dr. Sci. (Physics and Mathematics), Professor;
address: 20 S. Kovalevskoy st, Ekaterinburg, Russia, 620990;
ORCID: 0000-0003-2689-3006;
eLibrary SPIN: 9910-2326;
e-mail: varaksin@ecko.uran.ru

Yulia V. Shalaumova, Cand. Sci. (Engineering);
ORCID: 0000-0002-0173-6293;
eLibrary SPIN: 3163-6856;
e-mail: jvshalaumova@gmail.com

Tatiana A. Maslakova, Cand. Sci. (Physics and Mathematics);
ORCID: 0000-0001-6642-9027;
eLibrary SPIN: 3233-7652;
e-mail: t9126141139@gmail.com

* Автор, ответственный за переписку / Corresponding author