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

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

ЭКО•ВЕКТОР

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

BACKGROUND

Ayvazyan et al. [1, 2] proposed two applications of logistic regression models (LogR) in epidemiology: as a method for evaluating relationships (e.g., determining model coefficients, calculating odds ratios, and determining confidence intervals) and classifying data (e.g., constructing a classification matrix, calculating sensitivity and specificity, and performing a receiver operating characteristic [ROC] analysis). This paper only discusses LogR as a method for evaluating relationships. Using LogR for this task requires an assessment of model adequacy for the primary data, as well as a mandatory assessment of statistical significance.

Terminology

Primary data are epidemiological data collected for each study participant (i.e., each worker or patient). Primary data may include information such as the health status of each worker or patient (0, healthy; 1, ill), age, body mass index (BMI), and hemoglobin. *Stratification* is the division of primary data into intervals (strata). For example, age can be divided into the following strata: 20–24 years, 25–29 years, etc. In these age-based strata, the average values of all primary parameters of interest can be calculated. These characteristics include health status, mean BMI, and mean hemoglobin level. The *moving average* uses the same stratification, but with overlapping strata. For example, the age of 20–24 years for stratum 1, 21–25 years for stratum 2, and 22–26 years for stratum 3.

LogR models (and other statistical models) require two types of testing. The first type verifies the model's adequacy for the primary data. The second type verifies the statistical significance of the model if it has been deemed adequate.

Model adequacy

Afifi and Eizen [3] provided the clearest definition for adequacy criterion for statistical models: “By adequacy of the simple linear model, we mean that no other model significantly improves the prediction of *Y*.” This definition describes *linear* regression models that relate quantitative variables (predictor *X* and response *Y*). However, this criterion of model adequacy applies to any statistical model. The authors propose such a strict definition of model adequacy [3] that is impossible to achieve fully because of the large number of various models that can be constructed using specific primary data. Furthermore, criteria for improving *Y* predictions are often undefined [1].

From a practical standpoint, the proposal by Ayvazyan et al. seems more realistic (again: their approach involves linear regression, but all proposals are applicable to LogR). According to Ayvazyan et al. [1], *the adequacy criteria* cannot answer whether the hypothetical relationships being tested are the best or the most correct ones. They confirm or reject the consistency of the regression function type being tested based on the available primary data. In our paper, we further consider the adequacy of LogR models based on their consistency with the primary data. However, it is important to note that the strict model adequacy criterion proposed by Afifi and Eisen [3] can be used to understand the term *adequacy of a statistical model*.

Statistical significance of models

The statistical significance of a LogR model confirms that the relationship between *X* and *Y* is non-random at a level of significance α [1, 3, 4]. Adequacy and statistical significance of a model are two different concepts related to different aspects of constructing and analyzing statistical models.

Logistic regression

LogR is one of the most common nonlinear regression models. It is used to describe statistical relationships between a dichotomous response variable *Y* (*Y* takes two values: *Y*=0 or *Y*=1) and quantitative or rank predictor variables *X*. This type of data is commonly evaluated in epidemiological studies, where a dichotomous variable *Y* may indicate whether a patient has a disease or nor, whereas variable *X* may indicate whether a patient has a risk factor for a disease or not. It is generally accepted that *Y*=1 indicates the presence of a disease in a particular patient, whereas *Y*=0 indicates its absence.

In the LogR model, the statistical relationship between *Y* and a single predictor *X* is assumed to be as follows [5–7]:

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

where: $W(Y=1|X=x)$, probability of detecting $Y=1$ in the primary data given X=x.

Ratio (*1*) provides the following:

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

When condition (1) is met, there is a linear relationship between predictor X and complex ln(*W*/(1−*W*)), known as logit (W) [5–8].

Ratio (*2*) indicates the applicability of the LogR model. LogR typically has no specific limitations regarding the type of predictor *X* (quantitative or rank) [5]. Therefore, many authors believe it can be used wherever a dichotomous response *Y* exists. Examples of such publications are provided below. However, this is not true. First, in addition to LogR, there are other techniques of analyzing dichotomous responses *Y*, e.g., probit regression [9]. Secondly, for this type of specific epidemiological data, any relationship *W*(*Y*=1|*X*) between the dichotomous *Y* and predictor *X* is possible. Therefore, testing condition (*2*) is mandatory when using LogR to evaluate relationships. When condition (*2*) is met, the impact of predictor *X* on probability *W* is characterized by the odds ratio (OR). When *X* changes by one unit, the OR is calculated using the following formula: OR=exp(b1), where b_1 is the coefficient of model (*2*). The OR is a value that is the same for any *X*

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

only if condition (*2*) is met (a LogR model is characterized by a single value!). This is why OR is used in LogR rather than relative risk, which is common in many works on risk assessment [10]. If logit(W) is not a linear function of *X*, then for different *X*, OR will be different and the LogR model will no longer be characterized by a single OR value. Therefore, it is crucial to first confirm the linearity of the relationship between *X* and *Y* before using OR to characterize *X*'s impact on *Y*.

This paper discusses LogR models with one predictor (simple LogR models). Multiple regression models should be considered separately. Additionally, the paper considers only *quantitative* predictors *X* for which stratification and moving average calculation are possible in LogR.

Adequacy and statistical significance of logistic regression models

Adequacy of a LogR model should be tested by calculating goodness-of-fit tests for *stratified* primary data, such as χ^2 a Hosmer–Lemeshow test [5]:

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

where: the summation (index *i*) is performed by strata; n_i, $W_{obs,i}$, and $W_{calc,i}$ are the number of cases in a stratum, the actual and estimated probabilities, respectively; $W_{obs,i}$ probability is the mean dichotomous response *Y* in stratum *i* containing n_i observations, and $W_{calc,i}$ is the value calculated using the formula (*1*) for the mean *X* in stratum *i*. If the Hosmer–Lemeshow chi-squared (χ^2) test for the stratified primary data does not exceed the critical $\chi^2_{critical}(\alpha,\ \nu)$, the model is considered adequate for the primary data at a given significance level α (ν is the number of degrees of freedom). Adequacy testing is especially important when using the LogR model to evaluate relationships. Adequacy testing is the only way to guarantee the linearity of logit(W) for the LogR model. If the model appears to be adequate, it is logical to assess its statistical significance.

The statistical significance of the log R model is tested using either the t-test or the Wald test [5]. For a single predictor *X*, this is the way to determine the significance of the difference of coefficient b_1 from zero or the difference of the OR from 1 using formulas (1)–(2).

The study **aimed** to demonstrate how stratification techniques, the moving average, and the cumulative probability function can be used to construct and analyze LogR models for epidemiological tasks.

MATERIALS AND METHODS

The work illustrates the application of the above techniques using primary data from three sources: 1) Data from the monograph by Hosmer and Lemeshow [5]: 100 patients aged 20–69 years, some of them were diagnosed with cardiovascular disease (CVD); 2) Data from [11, 12], which present the results of preventive examinations of 820 male Ural enterprise workers aged 25–66 years who were diagnosed with CVD and had available body mass index (BMI); 3) Data from [13, 14], which present the results of an examination of 100 postmenopausal women aged 51–79 years, who were diagnosed with various concomitant diseases, including thyroid disease.

A statistical analysis of the data was performed using LogR models, stratification, a moving average, a cumulative probability function, a goodness-of-fit test, and proportion comparison techniques. Statistica 10.0 (StatSoft, USA) was used for calculations.

RESULTS

Specific epidemiological cases illustrate the analysis of LogR models using raw, stratified, or moving average data. Each case illustrates the specific application of stratification and the moving average methods for evaluating statistical relationships between dichotomous response *Y* and quantitative predictor *X* in various contexts.

Case 1. Rates of cardiovascular disease by age

Primary data. Fig. 1 shows the LogR estimates based on the primary data from the paper by Hosmer and Lemeshow [5]. *Y* represents the presence or absence of CVD in 100 patients: *Y*=0 indicates the absence of CVD and *Y*=1 indicates the presence of CVD in a patient of a given age.

As shown in Fig. 1, the data does not evaluate the extent to which ratios (1)–(2) are met. The presence of only two *Y* values (0 and 1) precludes any visual (expert) assessment of the relationship between *Y* and *X* [7], in contrast to linear regression [3, 4, 7]. Fig. 1 shows the LogR curve, which was plotted using a model based on primary data [5] for 100 patients:

$$\mathrm{W(CC3)}=\frac{\exp(-5{,}309+0{,}1109\times \mathrm{Age})}{1+\exp(-5{,}309+0{,}1109\times \mathrm{Age})}. \qquad (3)$$

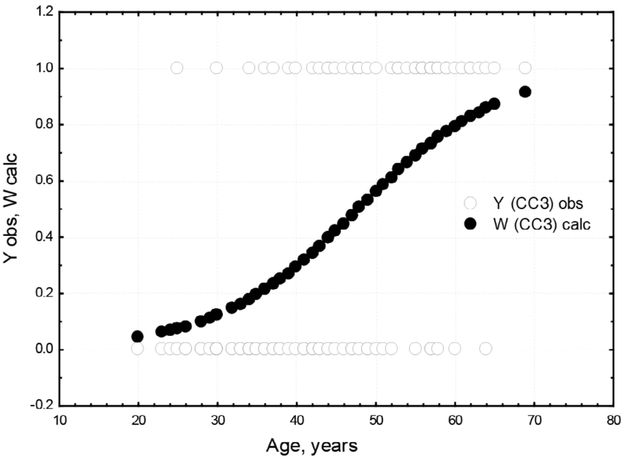

Fig. 1. Raw (obs) data for Y (Y=0 or Y=1 for each of 100 patients, open circles) and the probability of cardiovascular diseases $\mathrm{W(CVD)}_{calc}$, calculated based on logistic regression data (solid circles).

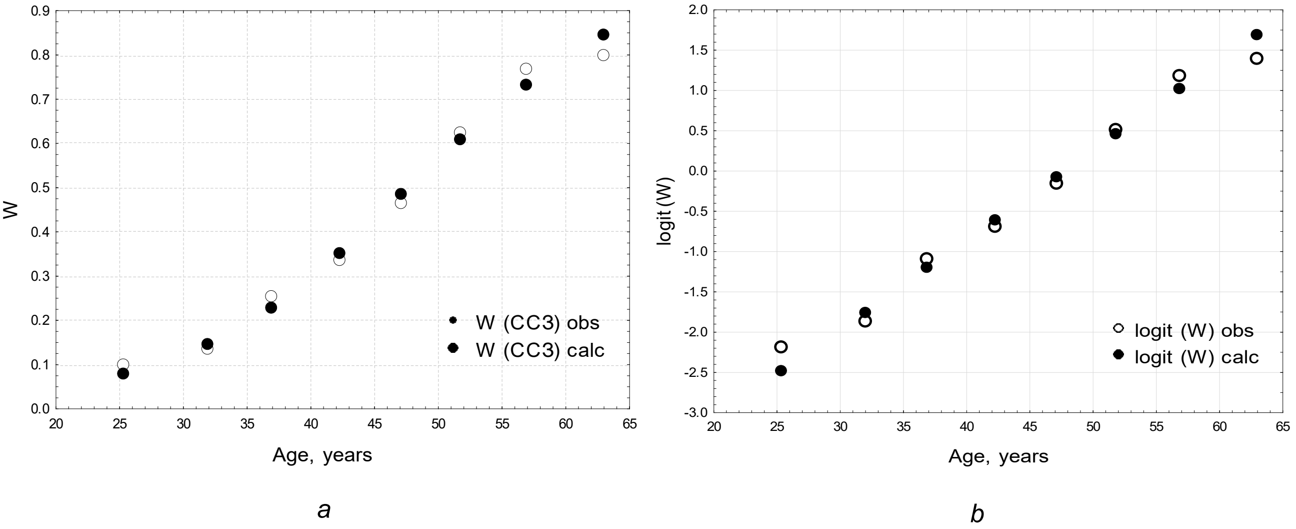

Fig. 2. Data on cardiovascular diseases in eight age strata: a, probability W; b, logit(W). The solid circles represent the logistic regression results (calc). The open circles represent the stratification results (obs).

OR=1.117; confidence interval (CI): 1.065–1.172; p <0.0001. However, relying solely on a visual analysis of values 0 and 1 in Fig. 1 does not guarantee that the probability of having CVD W(CVD) is accurately represented by this function.

Stratification. LogR has techniques for testing the linearity of the relationship between logit(W) and X type (*2*). These techniques evaluate model adequacy using stratified primary data. Many popular statistical packages, such as SAS and SPSS, include options for dividing predictor X into non-overlapping strata using a stratification procedure. The program then calculates statistical goodness-of-fit tests to evaluate the linearity hypothesis between logit(W) and X, for example by using the Hosmer–Lemeshow chi-squared test [5]. Although LogR users apply goodness-of-fit tests, they often only calculate the goodness-of-fit test and p-values without presenting the results graphically. As demonstrated below, graphical representation of stratification results is useful and sometimes provides unexpected findings.

Fig. 2*a* shows a comparison between the data calculated using the LogR model and primary data stratified by predictor X. In paper [5], predictor X (patient age) was divided into 8 strata based on age category. The mean age <Age> and probability W(CVD) were then calculated for each stratum. The results are shown in Fig. 2*a* and Fig. 2*b*. The transformation from probabilities W (as shown in Fig. 2*a*) to logit(W) (as shown in Fig. 2*b*) results in the conversion of the logistic curve into a straight line. A visual assessment of the stratified data for logit(W) shows a general linear increase with age. This is confirmed by statistical testing of the hypothesis about the linear relationship between logit(W) and age. When using Hosmer and Lemeshow goodness-of-fit test, the null hypothesis of linearity is not rejected at the 0.05 significance level, with χ^2=0.16 and 6 degrees of freedom (p=0.998).

Uncertainty of stratification. Age category is not the only possible basis for stratification. For example, Hosmer and Lemeshow [5] describe different techniques for stratification predictor X. All stratification techniques involve subjectivity (uncertainty) because the number of strata and their boundaries are chosen arbitrarily. Some statistical packages (SAS, SPSS, etc.) use stratification into 10 strata with an equal number of cases per stratum by default. This procedure is called *standard stratification*. The results of the stratification of the data published by Hosmer and Lemeshow [5] are shown in Fig. 3.

The LogR-calculated $W(CCZ)_{calc}$ values in Fig. 3 are the same as those in Fig. 2*a*. However, the actual values, which are obtained by averaging the primary data in the strata, differ. Clearly, the χ^2=2.43 for the data in Fig. 3 is smaller than the critical $\chi^2_{crit}(\alpha, \nu)$=15.51 for ν=8 degrees of freedom and a significance level α=0,05. Therefore, the model adequacy is confirmed (p=0.97 is significantly greater than 0.05), but χ^2=2.43 in Fig. 3 is 15 times higher than that in Fig. 2*a*. This case illustrates the uncertainty of stratification results, which can sometimes lead to inconsistent results.

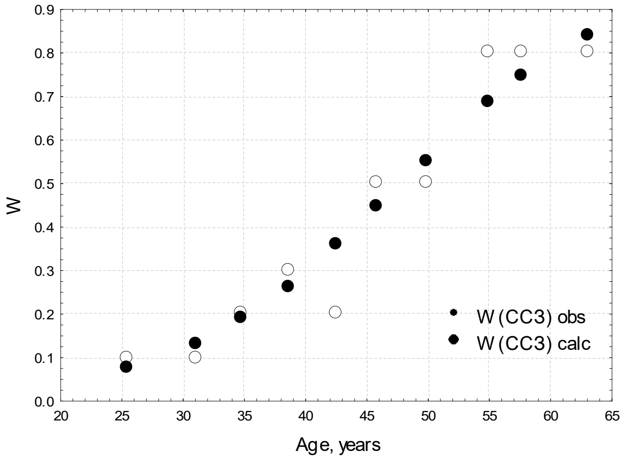

Fig. 3. Probability W of cardiovascular disease (CVD) depending on age: Stratification of 100 patients in the paper by Hosmer and Lemeshow [5] into 10 strata, each with an equal number of patients. For designations, see Fig. 2.

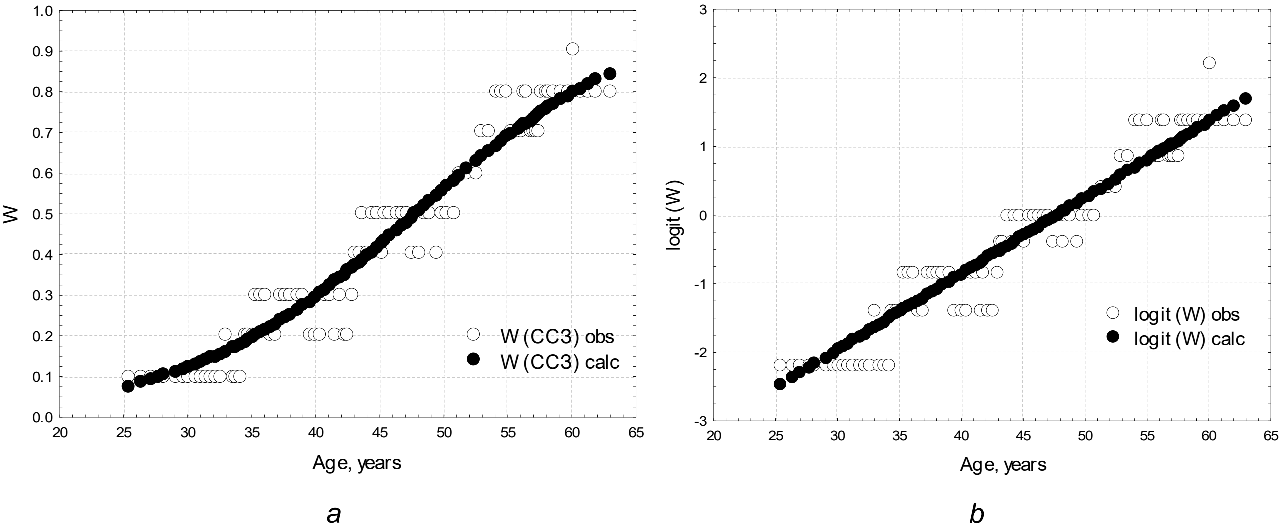

Fig. 4. Cardiovascular disease (CVD) data based on moving averages depending on the mean age in strata; averaging window n_W=10 (open circles): *a*, probability W; *b*, logit(W). The solid circles represent logistic regression results.

Moving average. If the stratification procedure is uncertain about the number of strata and their boundaries, the size of the averaging window is the only subjective parameter of the moving average. This parameter can be adjusted to achieve the best visual result [15]. Therefore, stratification uncertainty can be addressed using moving average techniques.

Fig. 4 shows the moving averages for the window of 10 cases, which coincides with the size of the strata in standard stratification procedure (Fig. 3). The moving average results cannot be used to perform any model adequacy tests, such as the Hosmer–Lemeshow test. These results are more likely to be perceived as an expert's visual assessment of goodness of fit between the theoretical model and the observed patterns. Theory predicts strict linearity of logit(W) depending on age. The findings either confirm or reject this linearity for the given data. As shown in Fig. 4*b*, the actual data confirms the theoretical linearity of logit(W). The following result confirms the linearity of logit(W). An attempt was made to add higher-order age terms to the logit(W) formula based on moving average data, but these terms were found to be statistically insignificant. Goodness of fit between the moving average results and the primary data is confirmed by goodness of fit between a logit(W) expression based on the moving average data (open circles in Fig. 4*b*) and a LogR expression based on the primary data (Fig. 1, equation *3*). For a moving average, logit(W)=–5.658+0.1183×Age. For the primary data (*3*), logit(W)=–5.309+0.1109×Age.

Conclusion for Case 1. The adequacy of the LogR model is tested using different stratification methods that produce different results (stratification uncertainty). Uncertainty can be reduced to some extent by using moving average techniques. The results of the moving averages show that, in this example, the W_{obs} probabilities of CVD detection are consistent with the LogR-based W_{calc} (Fig. 4*a*). Logit(W) is also a linear function of age (Fig. 4*b*).

Case 2. Incidence rates of cardiovascular disease by body mass index

Primary data. The case evaluates the statistical relationship between CVD rates and BMI in 820 male industrial workers aged 25–66 years in the Sverdlovsk region. The presence or absence of CVD was coded as 1 or 0 (a dichotomous variable in LogR). The quantitative predictor of BMI ranged 17.1–41.6 kg/m^2. The actual data were taken from [11, 12].

Testing the model adequacy by stratification. Using standard stratification, 820 participants were divided into 10 strata of 82 each. The Hosmer–Lemeshow chi-square goodness-of-fit test was 7.93 with 8 degrees of freedom, which is less than the critical $\chi^2_{crit.}$ of 15.51 for a significance level α=0.05. Therefore, the null hypothesis about the adequacy of the LogR model was not rejected (*p*=0.471). Therefore, the LogR model is considered adequate for the actual data and can be used to evaluate the relationship between CVD rates and BMI, including calculating OR.

The graphical results of the stratification are presented in Fig. 5*a* and Fig. 5*b*. As shown in Fig. 5*b*, the relationship between the observed logit(W) and BMI (open circles) appears to be linear. However, there are occasional deviations from the straight line representing the LogR estimates (solid circles).

The LogR equation, constructed using the primary data from 820 workers, is as follows: (*p* <0.0001 for all coefficients):

$$\text{logit(W)}=-4.6642+0.1554\times\text{BMI}; \quad \text{OR}=1.168 \text{ (CI: 1.126–1.212).} \qquad (4)$$

The regression plotted for the actual values of logit(W), which were obtained by stratifying the primary data into 10 strata (open circles in Fig. 5*b*), was as follows:

$$\text{logit(W)}=-4.6467+0.1545\times\text{BMI}. \qquad (5)$$

This is essentially the same as ratio (*4*). Attempts to include BMI nonlinear terms (quadratic and cubic) in equations (*4*) and (*5*) revealed their statistical insignificance.

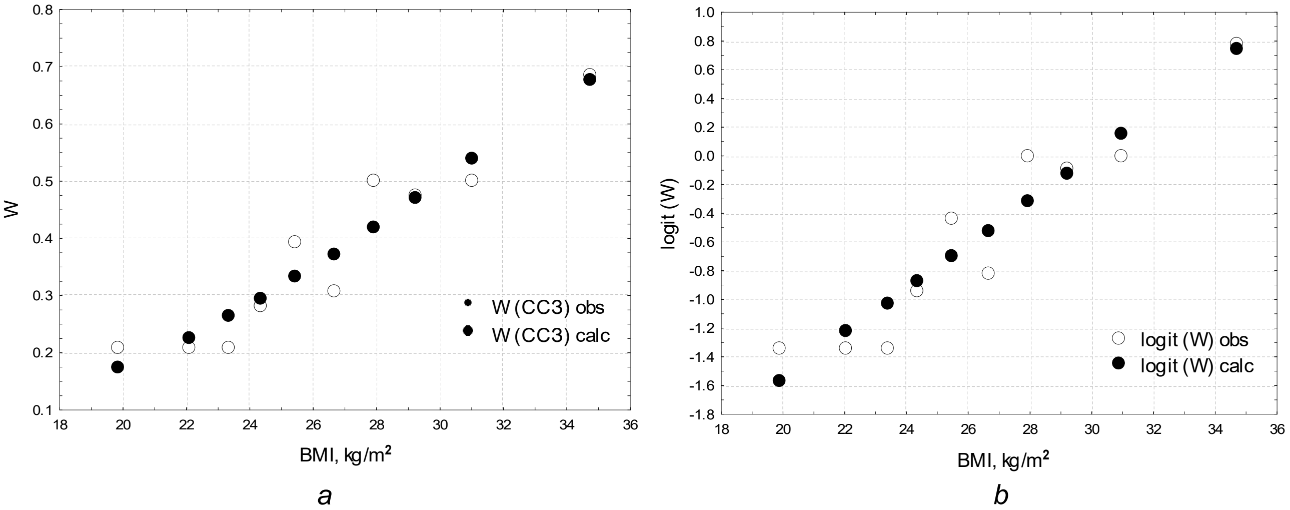

Fig. 5. Cardiovascular disease (CVD) data based on stratification of 820 workers into 10 strata with an equal number of workers per stratum using body mass index (BMI): *a*, probability W; *b*, logit(W). For designations, see Fig. 2.

As shown in Fig. 5*a*, the LogR estimates indicate that $W(CCC)_{calc}$ increases monotonically with an increase in BMI. The stratification results for low BMIs in the first three strata showed that the probability of W_{obs} did not increase with higher BMIs. The question is what range of BMIs corresponds to a low probability of CVD. This question arises from inconsistent stratification results. The stratification results cannot answer this question. However, another technique, the cumulative probability function, can be used to evaluate statistical relationships.

Cumulative probability function. The cumulative probability function $CUSUM_{nc}(X)$ for the response *Y* by predictor *X* is determined by the following ratio:

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

where: *nc*, the number of objects included in the function (this paper uses *CUSUM* abbreviation because Statistica for Windows has a built-in function with the same name and purpose). The cumulative probability function (*6*) is calculated as follows. First, the values of predictor *X* are ordered in ascending order. Then, the corresponding *Y* values are summed, as shown in equation (*6*). As a result, the first value of the $CUSUM_1$(*X* function is Y_1, corresponding to the minimum *X*. The second value of the $CUSUM_2(X)$ function is equal to half of the sum of the Y_1 and Y_2 values corresponding to the two minimum *X* values. The last point of the *CUSUM* function is the CVD rate in the full sample, which is calculated by summing all *Y* values and dividing by the number of objects in the sample. The *CUSUM* function has one feature. When the number of *nc* terms in sum (*6*) is small, i.e., for the initial region of the *CUSUM* function, a sharp conversion occurs when a new term is added. As *nc* increases, the *CUSUM* function becomes smoother. This smooth region can be used for a *CUSUM* analysis to draw conclusions.

Fig. 6 shows a *CUSUM* plot for CVD incidence in 820 workers with a BMI ranging 17.1–41.6 kg/m^2. In this graph, a point with a specific BMI* value indicates the mean CVD rate on the ordinate axis in participants with BMI ranging from the minimum to the specified BMI*. For example, when BMI is 24.0 kg/m^2 (with a range 17.1–24.0 kg/m^2, 254 participants), the *CUSUM* value is 0.202. Therefore, among 254 participants with BMI <24.0 kg/m^2, CVD rates were quite low (0.202). In participants with BMI >24.0 kg/m^2, CVD rates increased significantly and remained at a level of at least 0.2. The last *CUSUM* function value for maximum BMI is 0.376. This value was obtained by dividing the number of participants with CVD (308) by the total number of participants (820).

Based on this information, a new stratification procedure was performed. Some strata include BMI values less than 24 kg/m^2, whereas others include BMI values >24 kg/m^2. Table 1 and Fig. 7 show the results.

As shown in Table 1, 254 participants (quite a lot) had BMI <24 kg/m^2. Therefore, the BMI range was divided into 3 strata, each with relatively low CVD rates based on the actual W_{obs} data. The BMI range >24 kg/m^2 was divided into 6 strata, 4 of which (strata 4–7) had a range of 2 kg/m^2. Stratum 8 included the BMI range 32–34.5 kg/m^2. The cutoff BMI of 34.5 kg/m^2 was selected using the *CUSUM* procedure with a descending BMI order for *CUSUM* calculation. With this cutoff BMI in stratum 9 (BMI >34.5 kg/m^2), high CVD rates were recorded, equal to $W(CVD)_{obs}$=0.794. This rate was significantly higher than W(CVD)=0.690 in stratum 9 with standard stratification, as shown in Fig. 5.

Graphical representation. Fig. 7 shows that with BMI >24 kg/m^2, the estimated and actual CVD rates were concordant. This means that OR=1.168 (CI: 1.126–1.212), calculated by the LogR method using equation (*4*), was only true for BMI >24 kg/m^2. Using OR=1.168 for BMI <24 kg/m^2 would misinterpret the effect of BMI on CVD rates.

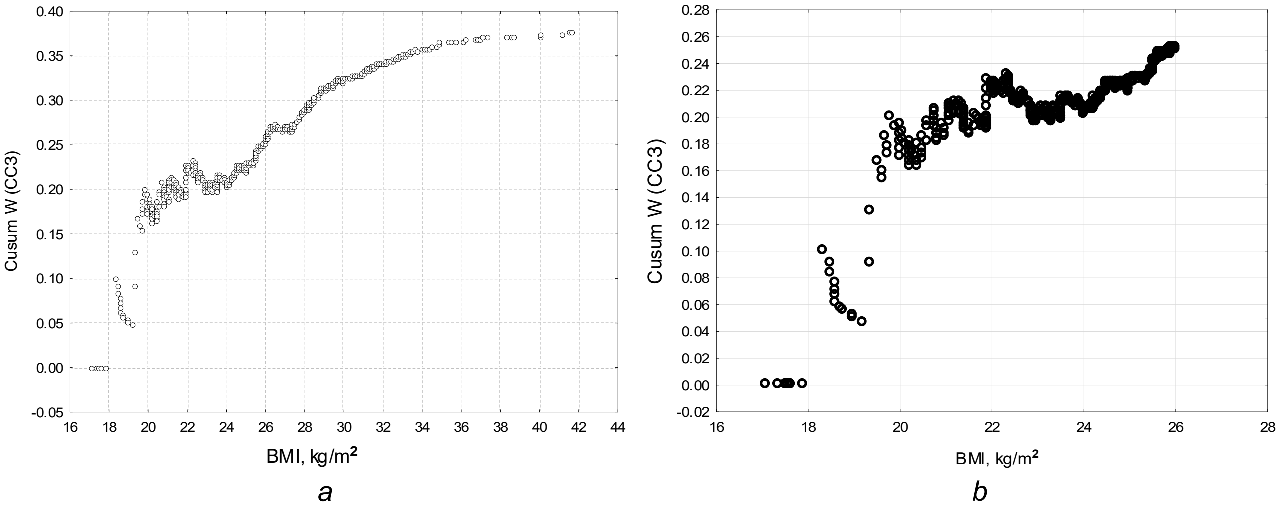

Fig. 6. Cumulative probability function for cardiovascular diseases (CVD) depending on body mass index (BMI): *a*, for 820 workers, *b*, initial region for BMI values < 26 kg/m2.

Table 1. *CUSUM* stratification scheme

No. of stratum	Range of BMI	Mean BMI	Number of participants per stratum	W_{obs} (cardiovascular disease), stratification	W (cardiovascular disease), logistic regression
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

Note. BMI, body mass index.

Conclusion for Case 2. Standard stratification into 10 strata, each containing an equal number of cases, shows that the LogR model adequately fits the primary data (the Hosmer–Lemeshow goodness-of-fit test yielded a value *significantly* lower than the critical value for a significance level of α=0.05). Graphical representation of the stratification results shows the possibility of differences between the stratification results and the LogR estimates in the initial strata. Due to the uncertainty of the stratification procedure, the use of a *CUSUM* cumulative probability function was proposed to confirm conclusions about the relationship between CVD and BMI. The *CUSUM* procedure with an ascending BMI order confirmed that there was no increase in CVD rates when BMI increased from the minimum to 24 kg/m^2. In this BMI range, CVD rates remained constantly low at W=0.20. The *CUSUM* procedure with a decreasing BMI order revealed a BMI range of 34.5 kg/m^2 with a high W(CVD) of 0.794, which the standard stratification did not show. Therefore, using the *CUSUM* function allows for more precise stratification to identify the relationship between W(CVD) and BMI. Such stratification revealed that the primary data are consistent only with the LogR estimates for BMI >24 kg/m^2.

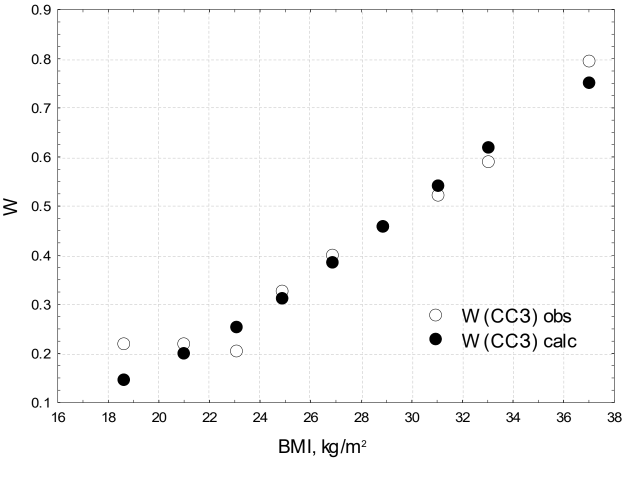

Fig. 7. Probability W of cardiovascular diseases (CVD) based on the stratification results shown in Table 1.

Case 3. Prevalence rates of thyroid disease and bmi

This case demonstrates that, even when the criteria for LogR applicability are met, its results can differ *significantly* from the primary data. The moving average procedure is the only way to identify this difference.

The LogR model was constructed using data from 100 postmenopausal women aged 51–79 years. Among other things, anthropometric parameters and prevalence rates of various medical conditions were determined for these women. The data were collected from the Research Institute for Maternal and Child Health in Yekaterinburg, Russia [13, 14]. The case evaluates the relationship between the prevalence of thyroid disease in women and their BMI. Thyroid diseases were coded as 0 if a woman did not have a disease, and as 1 if she did. BMI ranged 19.7–30.9 kg/m^2.

Stratification: adequacy testing of a LogR model. Fig. 8*a* shows the stratification results. These data were used to calculate the Hosmer–Lemeshow goodness-of-fit test in order to assess the linearity of the relationship between the logit(W) for thyroid disease and the BMI predictor. When the BMI predictor was divided into 10 strata with an equal number of cases, the hypothesis of logit(W) linearity was not rejected (chi-squared Hosmer–Lemeshow test $\chi^2 = 9.89$, which was significantly lower than the $\chi^2_{crit.}$ for 8 degrees of freedom at the significance level $\alpha = 0.05$. Therefore, we can use LogR to evaluate the relationship between thyroid gland prevalence rates and BMI.

Primary data. Using the LogR technique to assess the primary data shows that the regression coefficient $b_1 = 0.00995$ in the relationship like (*2*) is not statistically significantly different from zero ($p = 0.877$). Therefore, the use of statistical tests only concluded that the LogR technique used in this case, when the criteria for its applicability were *met*, *did not show* a statistically significant relationship between W(thyroid gland) and BMI. As for the graphical representation of the stratification results (Fig. 8*a*), the validity of such a conclusion no longer appears clear.

Moving average. Moving average techniques were used to more reliably assess the statistical relationship between BMI and the probability of thyroid disease. The moving average window, which was determined using the cumulative probability function, included 21 patients. The result shown in Fig. 8*b* differs from LogR estimates (W_{calc} in Fig. 8*a*) and from the standard stratification results (W_{obs} in Fig. 8*a*). Was there a statistically significant relationship between BMI and W(thyroid)? The statistical significance of differences in W(thyroid gland) for different BMI values was evaluated using the moving average data (Fig. 8*b*). The strata with the minimum and maximum W(thyroid gland) were compared. Stratum 1 in Fig. 8*b* (peak rate of W=0.476) included women with a BMI ranging 19.7–24.8 kg/m^2, with a mean BMI of 22.1 kg/m^2. In stratum 25 (range: 25.3–27.3 kg/m^2; mean BMI=26.5 kg/m^2), prevalence rates for thyroid disease were low (W=0.143). Strata 1 and 25 each contains 21 cases and did not overlap. When a two-sided test was used to compare proportions for independent samples, differences between W=0.476 and W=0.143 were statistically significant (*p*=0.0195). In stratum 80, which included women with a BMI ranging 29.6–30.9 kg/m^2 (mean BMI=29.8 kg/m^2), the prevalence rate of thyroid diseases was W=0.429. The difference with stratum 25 was also statistically significant (*p*=0.0404). Therefore, a statistically significant relationship was revealed between the prevalence rate of thyroid disease and BMI, which could not be described by a logistic function like (*1*).

Fig. 8*b* appears to be an outlier, resulting from a random combination of unusual factors and a rare exception to a typical pattern. However, this is not true. The analysis of the causes of the unusual representation of Fig. 8*b* revealed the following: Let $\langle x_0 \rangle$ and $\langle x_1 \rangle$ be the mean values of predictor *X* in groups *Y*=0 and *Y*=1, respectively, and let σ_0^2 and

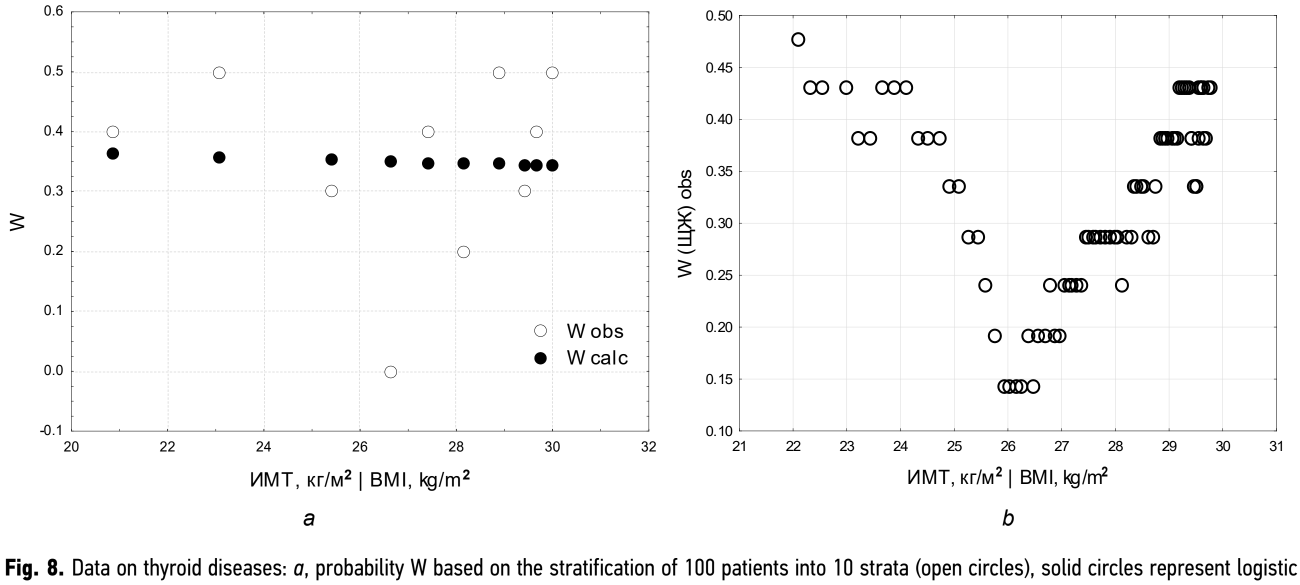

Fig. 8. Data on thyroid diseases: *a*, probability W based on the stratification of 100 patients into 10 strata (open circles), solid circles represent logistic regression results; *b*, relationship of probability W with body mass index (BMI) for moving averages (moving average window + 21, total of 80 strata).

σ_1^2 be variances *X* in the same groups. Then, if $<x_0>=<x_1>$ and $_0^2\neq\sigma_1^2$, diagrams like Fig. 8*b* are logical. Moreover, Fig. 8*b* with a decline in the center is obtained when $\sigma_1^2 > \sigma_0^2$. If $\sigma_1^2 < \sigma_0^2$, Fig. 8*b* demonstrates a peak in the center. Note that diagrams like Fig. 2 (linear logit) could be obtained when $<x_0>\neq<x_1>$ and $\sigma_0^2=\sigma_1^2$.

Conclusion for Case 3. Even if the LogR model meets the goodness-of-fit test, the actual data can be radically inconsistent with the model. *This is especially true when the model proves to be statistically insignificant.* The statistical insignificance of the model may be caused by the logit(W) deviation from linearity. In some cases, the model can be considered adequate for the actual data. The above case shows that such situations are possible.

DISCUSSION

The criteria for logit linearity should be met to correctly use the LogR for evaluation of the statistical relationship between the dichotomous response *Y* and the quantitative predictor *X* (i.e., the adequacy of the LogR model should be tested). A review of publications on the use of LogR in epidemiological studies reveals that most of them did not test the adequacy of the model [16, 17]. There are a few rare exceptions. For example, Konyrtaeva et al. [18] tested both the adequacy (using the Hosmer–Lemeshow test) and statistical significance of the LogR model.

The absence of tests for the adequacy of logR models in many publications is difficult to explain because such tests are mandatory in all fields of mathematical statistics to ensure concordance between theory and actual data. For example, before presenting data as a mean and standard deviation, it is necessary to test whether the actual data are adequate for a normal distribution [19]. In addition, when using the t-test to compare the mean values of *X* in two independent samples, the normality of *X* data in the samples should be first tested, and only then the statistical significance of the difference in means should be assessed. However, many authors do not consider it mandatory to test the logR model for adequacy. Nevertheless, testing the LogR model for adequacy does more than confirm or reject the model. It also evaluates the statistical relationship between the dichotomous *Y* and the quantitative *X* based on the available data.

As mentioned above, the adequacy and statistical significance of the LogR model are two distinct concepts. Some authors confuse adequacy with significance. For example, in their *educational* paper, Peng et al. [8] proposed that "rejecting such a null hypothesis [$H_0 : b_1=0$] implies that a linear relationship exists between *X* and the logit of Y.[1]"

This statement is incorrect. In fact, rejecting the null hypothesis $H_0 : b_1=0$ means that, an equation like (*2*) for logit(W) contains a linear *term* for *X*, though it may also contain nonlinear terms for *X*. If logit(W) contains nonlinear terms, it is impossible to use OR as a parameter for the relationship between dichotomous *Y* and quantitative *X*. Linearity of Logit(W) can only be guaranteed by testing the model for adequacy, not by statistical significance of b_1.

It should be noted that adequacy testing is critical when using LogR to *evaluate relationships*. This is necessary to ensure that the estimated OR accurately reflects the actual situation. If a LogR model is used for *classification* (prediction) and yields high sensitivity and specificity results, then it is not mandatory to test it for adequacy. The most important thing is that the model can accurately predict outcomes, regardless of how the prediction rule is obtained. In some works [17], the LogR model is used in both capacities simultaneously. In these cases, adequacy testing is important for correctly interpreting the OR.

The benefits of diagrams. This paper demonstrates that relying solely on statistical criteria can lead to incorrect LogR conclusions. For example, goodness-of-fit tests may not reject the hypothesis of a linear relationship between predictor *X* and logit(W), whereas graphical analysis clearly shows a nonlinear relationship (see Case 3). Prominent experts such as Tukey and Siegel actively promoted graphical representation for statistical results. In his paper [4], American statistician Andru Siegel illustrates almost every conclusion with graphs and detailed comments. John Tukey [20], one of the founders of modern data analysis, wrote, "Pictures that emphasize what we already know ... are frequently not worth the space they take. Pictures that have to be gone over with a reading glass to see the main point are wasteful of time and inadequate of effect. **The greatest value of a picture** is when it *forces* us to notice **what we never expected to see**" (the text was emphasized by the author). Figures 3, 7, and especially 8 show exactly what we did not expect to see.

Let us discuss benefits of diagrams again. The null hypothesis of logit(W) linearity was considered true, if the probability of rejecting it increased with an increase in the number of cases. For example, if there were 100 patients in case 1, then the null hypothesis would not be rejected with a high probability. If the number of patients increased from 100 to 650, then the null hypothesis of logit(W) linearity (which was based on the standard stratification results presented in Fig. 3) would be rejected at a significance level α=0.05, maintaining all relationships between age and the probability of CVD. This is the so-called *oversampling effect* [1], which results in the rejection of any null hypothesis if enough cases are available [1, 21]. This is another reason to use the LogR diagrams for expert assessment of logit(W) linearity, especially if the number of cases is really large.

CONCLUSION

In all of the above cases where the statistical relationship between the dichotomous response *Y* and quantitative

[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].

predictor *X* was evaluated, the hypothesis that the LogR model adequately describes the actual data was not rejected. Therefore, the use of the LogR model is justified. However, the LogR results were completely different in three cases. In case 1, the artificial database model demonstrated complete goodness of fit between the model and the data. In case 2, some regions of the predictor values deviated from the LogR model, though the model and the actual data were generally concordant. In case 3, the actual data demonstrated a non-linear and even non-monotonic relationship between *Y* and *X*. However, the Hosmer–Lemeshow goodness-of-fit test identified the logR model (linear for *X*) as an adequate for the actual data. Any differences between the LogR models and the actual data can be identified using stratification, moving averages, and cumulative probability functions, as well as graphical representations and analyses of the results.

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.

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.

REFERENCES | СПИСОК ЛИТЕРАТУРЫ

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.

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

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

* Corresponding author / Автор, ответственный за переписку

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