•Biostatistics in psychiatry (38)•

Inconsistency between univariate and multiple logistic regressions

Hongyue WANG[1,] Jing PENG[1], Bokai WANG[1], Xiang LU[1], Julia Z. ZHENG[3], Kejia WANG[1], Xin M. TU[4], Changyong FENG[1,2,*]

Summary: Logistic regression is a popular statistical method in studying the effects of covariates on binary outcomes. It has been widely used in both clinical trials and observational studies. However, the results from the univariate regression and from the multiple logistic regression tend to be conflicting. A covariate may show very strong effect on the outcome in the multiple regression but not in the univariate regression, and vice versa. These facts have not been well appreciated in biomedical research. Misuse of logistic regression is very prevalent in medical publications. In this paper, we study the inconsistency between the univariate and multiple logistic regressions and give advice in the model section in multiple logistic regression analysis.

Key words: Conditional expectation; model selection; logistic regression

[*Shanghai Arch Psychiatry.* 2017; **29**(2): 124-128. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.217031]

1. Introduction

Many medical studies have binary primary outcomes. For example, to study the treatment effect of a new intervention on patients with severe anxiety disorders, patients are randomized to the new intervention or treatment as usual (control) groups. The outcome is significant clinical improvement (yes or no) within a period such as 12 months. For this kind of outcome, we use 1 (0) to denote the occurrence or success (no occurrence or failure) of the outcome of interest such as significant (no significant) clinical improvements. The treatment effects can be measured by the difference or ratio of success rates in the two groups. Pearson’s chi-square test (or Fisher’s exact test) can be easily used if the treatment effect of the treatment method is better than the current method.

It is not uncommon that treatment effect is confounded by differences between treatment groups such as age, medication use and comorbid conditions. If such confounding covariates are categorical, such as gender and smoking status, contingency table methods can be easily used to study treatment differences. For continuous covariates such as age, although still possible to apply such methods by categorizing them into categorical variables, results depend on how continuous variables are categorized such as the number of end cut-points for categories.

The multiple logistic regression[1] provides a more objective approach for studying effects of covariates on the binary outcome. It addresses both categorical and continuous covariates, without imposing any subjective element to categorize a continuous covariate. Coefficients of continuous as well as non-continuous covariates, which are readily obtained using well-established estimation procedures such as the maximum likelihood, have clear interpretation. Also, its ability to model relationships for case-control studies has made logistic regression one of the favorite statistical models in epidemiologic studies.[2]

[1]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

[2]Department of Anesthesiology, University of Rochester, Rochester, NY, USA

[3]Department of Microbiology and Immunology, McGill University, Montreal, QC, Canada

[4]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA, USA

*correspondence: Dr. Changyong Feng. Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Ave., Box 630, Rochester, NY, 14642, USA. E-mail: Changyong_feng@urmc.rochester.edu

Model selection offers advantages of increasing power for detecting as well as improving interpretation of effects of covariates on the binary outcome, especially when there are numerous covariates to consider. Here is how model selection was carried out in multiple logistic regression in a paper recently published in JAMA surgery[3]:

'Associations between preoperative factors and adenocarcinoma or HGD were determined with univariate binary logistic regression analysis. Variables with statistically significant association on univariate analysis were included in a multivariable binary logistic regression model.'

Such a univariate analysis screening (UAS) method to select covariates for multiple logistic regression has been widely used in research studies published in top medical journals[4-6] since it seems very intuitive, reasonable, and easy to understand. In this paper we take a closer look at this popular approach and show that the UAS is quite flawed, as it may miss important covariates in the multiple logistic regression and lead to extremely biased estimates and wrong conclusions. The paper is organized as follows. In Section 2 we give a brief overview of the logistic regression model. In Section 3 we study the relationship between the univariate regression analysis, the basis for selecting covariates for further consideration in multiple logistic regression, and the multiple logistic regression model. In Section 4 we use the theoretical findings derived, along with simulation studies, to show the flaws of the UAS. In Section 5, we give our concluding remarks.

2. Logistic regression model

We use *Y*=1 or 0 to denote 'success' or 'failure' of the outcome. Here 'success' and 'failure' only indicate two opposite statuses and should not be interpreted literally. For example, if we are interested in the relation between the exposure of high density of radiation and cancer, we can use *Y*=1 to denote that the subject develops cancer after the exposure. Aside from the outcome, we also observe some factors (covariates) which may have significant effects on the outcome, denoting them by X_1, X_2, ..., X_p. The relation between the outcome and the covariates is characterized by the conditional probability distribution of *Y* given X_1, X_2, ... X_p. In multiple logistic regression, the conditional distribution is assumed to be of the following form

$$\Pr\{Y = 1|X_1, \dots, X_p\} = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}, \qquad (1)$$

where $\beta_1\beta_2 \dots \beta_p \neq 0$. This is the model on which our following discussions will be based. The covariates may include both continuous and categorial variables. A more familiar equivalent form of (1) is

$$\log\frac{\Pr\{Y = 1|X_1, \dots, X_p\}}{\Pr\{Y = 0|X_1, \dots, X_p\}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where the left hand side is called the conditional log-odds.

Given a random sample, the parameters (β_0, β_1, … ,β_p) in (1) can be easily estimated by maximum likelihood estimation (MLE) method, see for example.[7,8]

3. Univariate regression model

Suppose we are interested in the marginal relation between the outcome and a single factor X_1, i.e. we need to find $\Pr\{Y = 1|X_1\}$.

From the property of conditional expectation[9] we know that

$$\Pr\{Y = 1|X_1\} = \mathrm{E}\{\Pr\{Y = 1|X_1, \dots, X_p\}|X_1\}$$

$$= \mathrm{E}\left\{\frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}|X_1\right\}. \qquad (2)$$

If the joint distribution of X_1, X_2,…,X_p is unknown, generally it is impossible to find the analytical form of (2). In this section we consider the univariate regression model with following some specific distributions.

3.1 Univariate regression with categorical covariate

First assume X_1 is a 0-1 valued covariate. For example, in the randomized clinical trial, we can use X_1 as the group indicator (=1 for the treatment group and for the control group). It is easy to prove that there exist unique constants α_0 and α_1 such that

$$\Pr\{Y = 1|X_1\} = \frac{\exp(\alpha_0 + \alpha_1 X_1)}{1 + \exp(\alpha_0 + \alpha_1 X_1)}, \qquad (3)$$

where both α_0 and α_1 are functions of β_0, β_1, … ,β_p. Usually the form of these functions are complex as they depend on the joint distribution of X_1, X_2,…,$X_{p.}$ There is no obvious qualitative relation between α_1 in (3) and β_1 in (1).

Equation (3) indicates the marginal relation between *Y* and X_1 still satisfies the logistic regression model, and

$$\log\frac{\Pr\{Y = 1|X_1 = 1\}\Pr\{Y = 0|X_1 = 0\}}{\Pr\{Y = 1|X_1 = 0\}\Pr\{Y = 0|X_1 = 1\}} = \alpha_1,$$

which means that α_1 in (2) is the log odds ratio. Furthermore, if X_1 is independent of (X_1, X_2, ..., X_p), we can prove that (i) $\alpha_1>0$ if and only if $\beta_1>0$, (ii) $\alpha_1<0$ if and only if $\beta_1<0$, and (iii) $\alpha_1=0$ if and only if $\beta_1=0$. The independent assumption is true for completely randomized clinical trials. However, it seldom holds in practice, especially in observational studies.

Now assume X_1 is a covariate with k-categories, denoted by 1, ... k. Let $Z_j=1\{X_1=j\}$. We can also prove that there exist constant α_0, α_1, … ,$\alpha_{(k-1)}$ such that

$$\Pr\{Y = 1|X_1\} = \frac{\exp(\alpha_0 + \sum_{j=1}^{k-1} \alpha_j Z_j)}{1 + \exp(\alpha_0 + \sum_{j=1}^{k-1} \alpha_j Z_j)}.$$

All those parameters have similar interpretation as in the binary case.

This section shows that for categorical covariate, the univariate regression still has the form of logistic regression. However, the interpretation of the parameter is different from that in multiple logistic regression.

3.2 Univariate regression with continuous covariate

Assume X_1 is a continuous covariate, for example, the age of the patient. We want to know if Pr{Y=1|X_1} can be written in the (3) if (1) is the true multiple logistic regression model. Before answering this question, let's us take a look a the following example.

Example 1. Suppose there are only two covariates in the multiple logistic regression model (1), where X_1 is a continuous covariate with range R, X_2 is 0-1 valued random variable with Pr{X_2=1}=1/2, and X_1 and X_2 are independent. We further assume $\beta_0=\beta_1=\beta_2=1$ in (1). Then

$$\Pr\{Y=1|X_1\}=\frac{1}{2}\left\{\frac{\exp(2+X_1)}{1+\exp(2+X_1)}+\frac{\exp(1+X_1)}{1+\exp(1+X_1)}\right\}.$$

If (3) is true, then we should have

$$\frac{1+\exp(\alpha_0+\alpha_1X_1)}{1+\exp(2+X_1)}+\frac{1+\exp(\alpha_0+\alpha_1X_1)}{1+\exp(1+X_1)}=2. \qquad (4)$$

Let $X_1 \to \infty$ in (4) we have α_1=1. Let X_1=0 in (4) we have

$$\alpha_0=1+\log\frac{1+e+2e^2}{2+e+e^2}.$$

However, if X_1=1 in (4), then

$$\alpha_0=1+\log\frac{1+e+2e^3}{2+e^2+e^3}.$$

Since these two solutions of α_0 do not match, model (3) does not hold.

This example shows that, for continuous covariate X_1, the regression of Y on X_1 does not in general satisfy the univariate logistic regression model even if X_1 is an essential component in the multiple logistic regression. Hence, the univariate logistic regression model should not be used to estimate the marginal relation between the outcome and a continuous covariate.

4. Inconsistency between univariate and multiple logistic regressions

In Section 3 we show that in multiple logistic regression, the univariate regression of the outcome on each individual covariate may not satisfy the logistic regression any more. This fact has serious implications for model selection and interpretation of results in data analysis. In this section, we demonstrate this important issue using simulation studies.

4.1 Significant effect in multiple but not in univariate logistic regression

In this section we show an example where a continuous covariate is a necessary part in the multiple logistic, but the univariate regression indicates that the covariate has no effects in the univariate regression. The following preliminary result will be used in our discussion.

Lemma 1. *Suppose c is a positive constant and the random variable X has standard normal distribution. Then E[X/(1+cexp(ϑX))]=0 if and only if ϑ=0.*

The proof of this result is available from authors upon request.

Example 2. Let X_2 and X_3 be independent random variables with standard normal distributions, and $X_1=X_2+2X_3$. Consider the following multiple logistic regression model

$$\Pr(Y=1|X_1,X_2)=\frac{\exp(\alpha_0+\alpha_1X_1+\alpha_2X_2)}{1+\exp(\alpha_0+\alpha_1X_1+\alpha_2X_2)}, \qquad (5)$$

where $\alpha_1=-\alpha_2/5, \alpha_2\neq 0$. Using the result in Lemma 1 we can prove that if

$$\Pr\{Y=1|X_1\}=\frac{\exp(\theta_0+\theta_1X_1)}{1+\exp(\theta_0+\theta_1X_1)}, \qquad (6)$$

then θ_1=0.

What does this result mean within the current context? Although X_1 and X_2 both are in the multiple logistic regression, if their coefficients satisfy the condition (5), the regression of Y on X_1 is no longer a univariate logistic regression. Moreover, if (Y_{i1}, X_{i1}, X_{i2}),i=1,…,n is a random sample from (5), X_1 and X_2 will become increasingly significant in the multiple logistic regression, but X_1 will remain nonsignificant regardless of sample size, as illustrated by the following simulation results.

The data was generated according (5) with $\alpha_0=1,\alpha_1=-3/5,\alpha_2=3$. Shown in Table 1 are the estimates and standard deviations of the coefficient of X_1 in both univariate and multiple logistic regression after 10,000 Monte Carlo (MC) replicates. The parameters were estimated by MLE. For a wide range of sample sizes, the maximum likelihood estimator of the coefficient of X_1 in the multiple logistic regression was very close to the true value, and the standard errors decreased with the sample size, as expected. However, the estimated coefficient in the univariate analysis was consistently close to 0 in all cases.

Table 1 also reports the chance that p-value is >0.2 (or >0.1) in the univariate logistic regression. It shows that although X_1 is a necessary part of the multiple logistic regression, X_1 will most likely be excluded from the multiple logistic regression, if the cutoff of the p-value is set at 0.2 (or 0.1).

Reported in Table 2 are the estimates of the coefficient of X_2 in the logistic regression if X_1 is

Table 1. Estimate of regression coefficient of X_1 in Example 2

	Univariate regression				Multiple regression	
n	Estimate	SD	p-value>0.2	p-value>0.1	Estimate	SD
100	-0.0042	0.0983	0.7963	0.8991	-0.6533	0.2103
200	-0.0015	0.0674	0.7952	0.898	-0.6173	0.1308
500	-0.0004	0.0429	0.791	0.889	-0.6085	0.0828
1,000	-0.0009	0.0284	0.801	0.907	-0.6056	0.0566
1,500	-0.0002	0.0239	0.799	0.902	-0.6046	0.0465
2,000	-0.0004	0.0205	0.809	0.905	-0.6027	0.0392

Table 2. Estimates of coefficients of X_2 in logistic regression with X_1 being removed in Example 2

	Coefficient of X_2 (α_2=3)	
n	Estimate	SD
100	2.0243	0.4268
200	1.9843	0.2903
500	1.9579	0.1789
1,000	1.9556	0.1231
1,500	1.9498	0.1040
2,000	1.9495	0.0857

mistakenly excluded due to UAS method. The true coefficient of X_3 is 3 in the multiple logistic regression, but the estimated coefficient of X_2 became extremely biased if X_1 was excluded.

Taken together, the results show that the UAS not only most likely misses some important covariates in the multiple logistic regression, but also leads to severely biased estimates of effects of other covariates on the response.

4.2 Significant effect in univariate but not in multiple regression

In this section we show a case where a continuous covariate has significant effect in the univariate regression, but is not significant if it is included in the multiple regression.

Example 3. Suppose X_1, X_2, X_4 and ε are independent standard normal random variables, and X_3=X_1+X_4. Consider the following multiple logistic regression model

$$\Pr(Y=1|X_1,X_2)=\frac{\exp(\alpha_0+\alpha_1X_1+\alpha_2X_2)}{1+\exp(\alpha_0+\alpha_1X_1+\alpha_2X_2)}, \qquad (7)$$

where α_1 α_2≠0.

In the simulation study, the data was generated according model (7) with α_0=0,α_1=2,α_2=1. Shown in Table 3 are the estimates of the coefficient of X_3 in both univariate and multiple linear regression (with X_1,X_2 and X_3 as covariates) after 10000 replicates. For all sample sizes, X_3 shows very significant effect on Y in the univariate regression, but no significant effect in the multiple logistic regression.

Table 3. Estimate of the regression coefficient of X_3

	Univariate regression		Multiple regression	
n	Estimate	SD	Estimate	SD
100	0.7120	0.2012	0.0130	0.3079
200	0.6907	0.1320	-0.0021	0.1953
500	0.6787	0.0800	-0.0039	0.1221
1,000	0.6777	0.0588	0.0005	0.0865
1,500	0.6772	0.0463	-0.0012	0.0681
2,000	0.6771	0.0400	0.0000	0.0602

5. Discussion

Although the logistic regression is a very powerful analytical method for binary outcome, the results from the univariate and multiple logistic regressions tend to be conflicting. A covariate may show very significant effect in the univariate analysis but has no role in the multiple logistic regression model. On the other hand, a covariate may be an essential part of the multiple logistic regression but shows no significant effect on the outcome in the univariate regression. The UAS method uses the univariate analysis as an initial step to select covariates for further consideration in the multiple regression. This method may mistakenly exclude important covariates in the multiple logistic regression and lead to extremely biased estimates of the effects of other covariates in the multiple model. Hence the UAS is not a valid method in model selection. It should be removed from the tool kits of biomedical researchers and even some PhD statisticians. Formal model selection methods based on solid theory, such as Akaike's information criterion (AIC) and Schwarz' Bayesian information criterion (BIC) discussed in [10] should be implemented in all regression analyses.

Funding

None

Conflict of interest statement

The authors report no conflict of interest related to this manuscript.

Authors' contribution

Hongyue Wang, Bokai Wang, Xiang Lu, Xin M. Tu and Changyong Feng: theoretical derivation and revision

Julia Zheng, Jing Peng, and Kejia Wang: Simulation studies and manuscript drafting

单因素与多因素逻辑回归的不一致性

WANG H, PENG J, WANG B, LU X, ZHENG J, WANG K, TU X, FENG C

概述：逻辑回归是研究协变量对二元结果影响的一种常用的统计方法。它已被广泛应用于临床试验和观察性研究。然而，单因素回归得到的结果和多元逻辑回归得到的结果往往是相互矛盾的。在多元回归中可能对结果会显示出非常强烈的影响的一个协变量在单因素回归中可能不会，反之亦然。这些事实在生物医学研究中并没有引起足够的重视。误用逻辑回归在医学出版物中非常普遍。在本文中，我们研究了单因素和多因素逻辑回归分析的不一致性，并在多元逻辑回归分析的模型部分中给出建议。

关键词：条件期望；模型选择；逻辑回归

References

1. Cox DR. The regression analysis of binary sequences (with discussion). *J Roy Stat Soc B*. 1958; **20**: 215–242
2. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrica*. 1979; **63**: 403–411. doi: http://dx.doi.org/10.1093/biomet/66.3.403
3. Postlewait LM, Ethun CG, McInnis MR, Merchant N, Parikh A, Idrees K, et al. Association of preoperative risk factors with malignancy in pancreatic mucinous cystic neoplasms: A multicenter study. *JAMA Surg*. 2017; **152(**1): 19-25. doi: http://dx.doi.org/10.1001/jamasurg.2016.3598
4. Karcutskie CA, Meizoso JP, Ray JJ, Horkan D, Ruiz XD, Schulman CI, et al. Association of mechanism of injury with risk for venous thromboembolism after trauma. *JAMA Surg*. 2017; **152**(1): 35-40. doi: http://dx.doi.org/10.1001/jamasurg.2016.3116
5. Templin C, Ghadri JR, Diekmann J, Napp LC, Bataiosu DR, Jaguszewski M, et al. Clinical features and outcomes of takotsubo (stress) cardiomyopathy. *N Engl J Med*. 2015; **373**(10): 929-38. doi: http://dx.doi.org/10.1056/NEJMoa1406761
6. Nor AM, Davis J, Sen B, Shipsey D, Louw SJ, Dyker AG. The Recognition of Stroke in the Emergency Room (ROSIER) scale: Development and validation of a stroke recognition instrument. *Lancet Neurol*. 2005; **4**(11): 727-734. doi: http://dx.doi.org/10.1016/S1474-4422(05)70201-5
7. McCullagh P, Nelder JA. *Generalized Linear Models (2nd ed)*. New Yrok: Chapman & Hall; 1989
8. Agresti A. *Categorical Data Analysis (3rd ed)*. Hoboken, NJ: Wiley; 2010
9. Durrett R. *Probability: Theory and Examples* (4th ed). New York: Cambridge University Press; 2010
10. Claeskens G, Hjort NL. *Model Selection and Model Averaging*. New York: Cambridge University Press; 2008

Hongyue Wang obtained her BS in Scientific English from the University of Science and Technology of China (USTC) in 1995, and PhD in Statistics from the University of Rochester in 2007. She is a Research Associate Professor in the Department of Biostatistics and Computational Biology at the University of Rochester Medical Center. Her research interests include longitudinal data analysis, missing data, survival data analysis, and design and analysis of clinical trials. She has extensive and successful collaboration with investigators from various areas, including Infectious Disease, Nephrology, Neonatology, Cardiology, Neurodevelopmental and Behavioral Science, Radiation Oncology, Pediatric Surgery, and Dentistry. She has published more than 80 statistical methodology and collaborative research papers in peer-reviewed journals.