Original Article

Multivariate logistic regression analysis using multilevel model

Ahmad Vakili Basir[1], Mohammad Gholami Fesharaki[2]*, Mohsen Rowzati[3]

[1] Biostatistics Department, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
[2] Assistant Professor of Biostatistics, Biostatistics Department, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
[3] Worksite Follow-Up Unit, Occupational Health Center, Mobarakeh Steel Company, Isfahan, Iran

ARTICLE INFO

ABSTRACT

Background & Aim: One of the basic assumptions in simple linear regression models is the statistical independence of observations. Sometimes this assumption is not true for study subject and consequently the use of general regression models may not be appropriate. In this case, one of the leading methods is the use of multilevel models. The present study utilizes multivariate logistic regression model using a multilevel model to exhibit the chance of having elbow, wrist and knee disorders over the past year based on elbow, wrist and disorders during the past week.

Methods & Materials: This study is a cross-sectional study that was carried out from April 2015 to May 2016 in Mobarakeh Steel Company, Isfahan. The study population includes 300 male employees of Mobarakeh Steel Company, with a mean age of 41.40±8.17 years and an average working experience of 16.0±7.66 years. Data were analyzed using SPSS (version 24) and MLwiN software.

Results: Based on this study, results obtained from single variable and multivariable regression were different.

Conclusion: Based on this study, it can be suggested that multivariable regression cause a better and more accurate deduction compared to single variable method.

Introduction

The aim of statistics is the development and application of methods for extraction of valuable results from data sources and experiences. This science is divided into two main statistics groups including descriptive and inferential. The aim of inferential statistics is to discuss the effects of descriptive variables on response variables through modeling. Different statistical modeling can be described in generalized linear model format. A generalized linear model is composed from three components, including random, systematic and link function. Random component can be defined as distribution of variables' response which is continuous or discrete and the mean of random component can be considered as a function of covariates. The systematic component is a relationship between the mean of dependent variable with covariates $(x\_1, x\_2, ..., x\_p)$ and link function, correlation between the mean of random and systematic components. In most of medical applications and studies the desired answer variable is defined as zero and one. For example, if the individual age is supposed as an independent variable, whether he/she has cardiovascular disease or not, the dependent (response) variable can be zero (with heart failure) or one (without heart failure). If n number of individuals are examined, data of $(x\_1, y\_1), (x\_2, y\_2), ..., (x\_n, y\_n)$ will be obtained in which x_i is considered as age and y_i is zero or one. Logistic regression is, therefore, used in this condition. Odds ratio, which is defined as a ratio of occurrence probability to lack of occurrence probability, is designed in logistic regression. However, general logistic regression cannot be used in multivariate response. For instance, it has been supposed that a researcher

* Corresponding Author: Mohammad Gholami Fesharaki, Postal Address: Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
Email: mohammad.gholami@modares.ac.ir

Please cite this article in press as: Vakili Basir A, Gholami Fesharaki M, Rowzati M. Multivariate logistic regression analysis using multilevel model. *J Biostat Epidemiol.* 2018; 4(3): 119-123

http://jbe.tums.ac.ir

wants to evaluate the effect of shift working, work experience and education on musculoskeletal disorders incidence (1–3). Musculoskeletal disorders depend on work patterns. It is not specific to the industry or specialty and may be seen in all industries and professions. Symptoms of this disease are mainly observed in areas of the lower back, neck and upper extremities of the body (4, 5). According to NIOSH (National Institute for Occupational Safety and Health), which is done just for occupational patients, Work-related Musculoskeletal Disorders are ranked second after respiratory illnesses (6). Symptoms of the disease are reported for back, neck and upper limbs. If the author wants to consider correlations, a simple logistic regression for back problems, additional simple logistic regression for neck and one another for upper limbs is needed. Indeed, three logistic regression models are used for one hypothesis. Use of three models instead of one model increases first type error. Furthermore, consideration of relationships between these three variables in three separate regressions is impossible. For this purpose, see the Figure 1.

Multilevel model
One of the basic assumptions in simple linear

regression models is the statistical

independence of observations (7). Sometimes this assumption is not true for study subject and consequently, the use of general regression models may not be appropriate. This status occurred for multilevel of cluster data that contain intergroup relation structures (8, 9). Appropriate model for the analysis of these data is multilevel models (10, 11). Since multilevel models consider correlation between observations, they provide more accurate results than compared to simple linear regression analysis (10, 12). In follow-up studies, there is the possibility of data loss which can lead to loss of information and missed values. The presence of missed values may decline accuracy of calculated statistical indicators and subsequently increase the complexity of model and statistical methods. Since results from incomplete data can result in bias, these data analysis should be done in appropriate way (8, 9, 13). Multilevel model is a suitable method because it does not require data with balanced structure (11, 12). Multilevel models are a kind of analytical models that include variables measured at hierarchy different levels. Indeed, they include a class of models that use hierarchical linear regression. These models consider hierarchical structure of data. Multilevel models are also

A : Logistic regression model for effects of covariates on back

B : Logistic regression model for effects of covariates on neck

C : Logistic regression model for effects of covariates on back
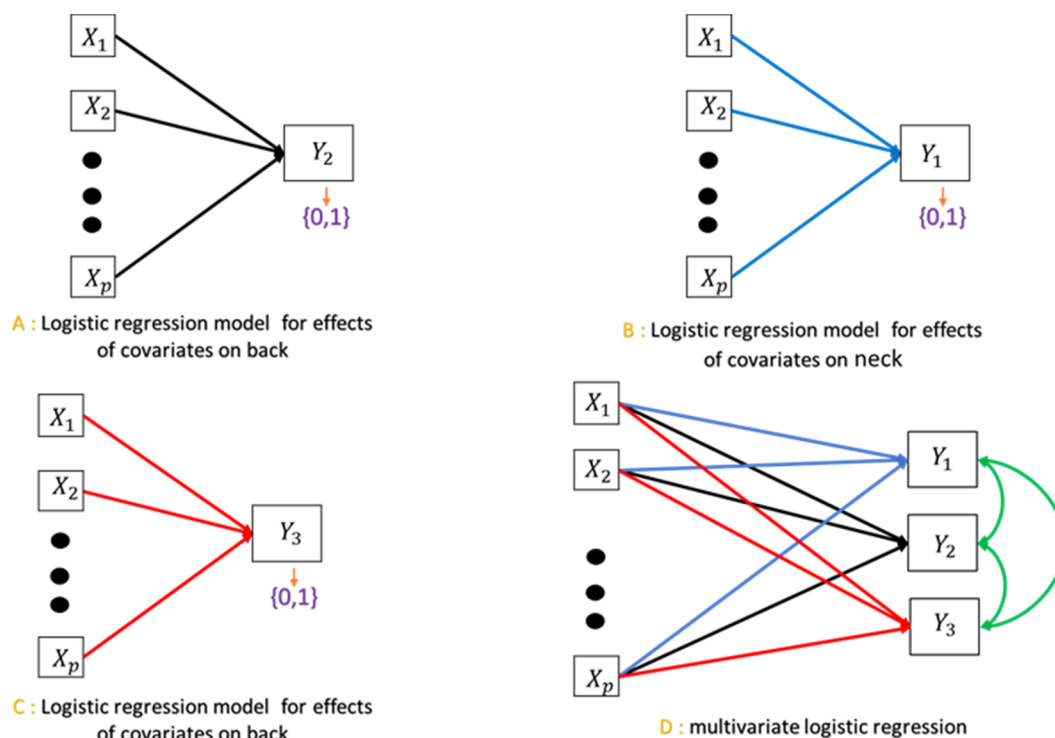
D : multivariate logistic regression

Figure 1. Comparison of the relationships between three variables in three separate regressions (A, B, C) and multivariate logistic regression (D)

known as random coefficients model, variants component model, and hierarchical linear model (14, 15). For example, it has been supposed that there are binary variables (y_1, y_2, y_2) with a predictor variable (x_1). For the fitness of the three variables logistic regression model, the following Equation can be used:

$$\begin{cases} logit(y_{i1}) = \beta_{01} + \beta_1 x_{1i} \\ logit(y_{i2}) = \beta_{02} + \beta_2 x_{1i} \\ logit(y_{i3}) = \beta_{03} + \beta_3 x_{1i} \end{cases} \qquad (1)$$

Data structure in this form, which has known as $Wide$ structure, is shown in Table 1.

It has been also supposed that if the correlation matrix $Y = [y_1, y_2, y_3]$ is summarized as follow:

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

Data structure can be converted to that in Table (2), which is known as $Long$ structure. Considering the long structure, Equation fitness form of Equation (1) can be converted to that in Equation (2):

$$\begin{aligned} logit(y_{ij}) = & \ \beta_{01} + \beta_1 x_{1i} \qquad (2) \\ & + (\beta_{02} + \beta_2 x_{1i})I_2 \\ & + (\beta_{03} + \beta_3 x_{1i})I_3 \\ & + u \end{aligned}$$

where $I_2 = \begin{cases} 1 & j = 2 \\ 0 & o.w \end{cases}$ and $I_3 = \begin{cases} 1 & j = 3 \\ 0 & o.w \end{cases}$ also assume that with a conception of $u \sim N(0, \sigma_u^2)$. With this assumption the value of ρ can be calculated with following formula:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\pi^2}{3}}$$

Methods

This is a cross-sectional study which was carried out from April 2015 to May 2016 in Mobarakeh Steel Company, Isfahan. Study

Table 1. Wide structure

y_3	y_2	y_1	x_1
1	0	1	2
0	1	1	6
1	1	0	3
0	0	1	4
1	0	0	8
1	1	1	7
1	0	0	9
0	1	1	7
1	0	0	5
0	1	1	9

Table 2. Long structure

$x_1 * I_3$	$x_1 * I_2$	I_3	I_2	Y	$index$	x_1
0	0	0	0	1	1	2
0	2	0	1	0	2	2
2	0	1	0	1	3	2
0	0	0	0	0	1	6
0	6	0	1	1	2	6
6	0	1	0	1	3	6
0	0	0	0	1	1	3
0	3	0	1	1	2	3
3	0	1	0	0	3	3
0	0	0	0	0	1	4
0	4	0	1	0	2	4
4	0	1	0	1	3	4
0	0	0	0	1	1	8
0	8	0	1	0	2	8
8	0	1	0	0	3	8
0	0	0	0	1	1	7
0	7	0	1	1	2	7
7	0	1	0	1	3	7
0	0	0	0	1	1	9
0	9	0	1	0	2	9
9	0	1	0	0	3	9
0	0	0	0	0	1	7
0	7	0	1	1	2	7
7	0	1	0	1	3	7
0	0	0	0	1	1	5
0	5	0	1	0	2	5
5	0	1	0	0	3	5
0	0	0	0	1	1	9
0	9	0	1	1	2	9
9	0	1	0	1	3	9

population contained 300 male employers in Mobarakeh Steel Company, Isfahan. The inclusion criteria were as follow:

✓ Being official employment or Contract

✓ At least one year of services in the factory

✓ Lack of absence within the last 6 months

✓ At least 20 h activity each week

✓ History of limb fractures during the last year

✓ Severe damage

✓ History of physiotherapy and surgery during the last 6 months.

Individuals who had fever, chronic infection, acute inflammatory reactions, and limb fractures during the study, as well as unwilling participants were excluded from the study. In this study, three questionnaire tools, examination and figures were used for data collection. Nordic questionnaire was applied to evaluate musculoskeletal disorders in different organs of employers. The validity and reliability of this questionnaire, which has been confirmed in different studies, were performed by the researcher (16–18). The physical status of employers including Gait status, Posture in standing and working modes, and muscular-

Table 3. Comparison of estimated coefficients of multivariate logistic regression and ordinary logistic regression

	Univariate				Multivariate			
	$\hat{\beta}_i's$	P-value	SE	OR	$\hat{\beta}_i's$	P − value	SE	OR
Intercept (Elbows)	0.511	0.484	0.73	1.667	-3.306	0.011	1.3	0.037
Intercept(Wrists)	-1.70	<0.001	0.44	0.182	-5.52	<0.001	1.16	0.004
Intercept (Knees)	3.738	<0.001	1.01	42.014	3.793	<0.001	1.07	44.389
Elbows	2.74	0.001	0.84	15.487	1.76	0.292	1.67	5.812
Wrists	4.303	<0.001	0.56	73.921	3.33	0.032	1.56	27.938
Knees	1.09	0.444	1.43	2.974	1	0.491	1.45	2.718

skeletal system evaluation through observation, touch, range of joints movements, and also assessment of strength and flexibility of muscles. The abdominal muscle and vertebrae strength were evaluated using Kendall method, while the strength of other muscles was evaluated by standard muscle measurement technique (MMT). Laboratory and paraclinical examinations were also provided in case there were pathological findings during clinical assessment of each person. QEC method was applied to evaluate the level of muscular-skeletal disorders (19, 20). Data were analyzed through IBM's SPSS (Version 24) and *MLwiN* Software.

Results

In this study, 300 workers from Mobarakeh Steel Company, Isfahan with a mean age of 41.40±8.17 years and a work experience of 16.0±7.66 years were included. In this research, a variable of chronic problem in each week was used as an independent variable for prediction of the presence of problem chance in one year. Therefore, the presence of chronic problem over the last year in elbow, wrist, and knee was a functional problem within the last one week. In the both single and multivariable models, effect of knee failure in last week on knee failure in last year was rejected. However, a chance of 15.4 and 73.9 times were found to be significant for elbow and wrist problems in single variable (Table 3). These results for multivariable showed a chance of 28 for the wrist problem. Furthermore, in multivariate condition, effect of elbow problem in last week on the last year was rejected.

Discussion and Conclusion

Since results obtained from single variable and multivariable regression were different, it can be suggested that multivariable regression cause a better and more accurate deduction compared to single variable method.

Ethics approval and consent to participate

Medical Ethics Committee of Tarbiat Modares University Medical School approved the study (code number: IR.TMU.REC. 1395.398).

Conflicts of interests

The authors declare that there is no conflict of interest regarding the publication of this article.

References

1. Kheiri H, Moshfegh H, Hatani H, Ranjbarian M. Investigation of relationship between musculoskeletal disorders and working conditions among workers at a pharmaceutical industry in Iran (2011-2012). Iran J Heal Saf Environ. 2014;1(3):145–50.
2. Chiu YJ, Liao WC, Wang TH, Shih YC, Ma H, Lin CH, et al. A retrospective study: Multivariate logistic regression analysis of the outcomes after pressure sores reconstruction with fasciocutaneous, myocutaneous, and perforator flaps. J Plast Reconstr Aesthetic Surg. 2017;70 (8):1038–43.
3. Mesquita CC. Musculoskeletal Disorders in Workers-risk factors: What Can We Do? Occup Med Heal Aff. 2013; 27 Apr.
4. Mortazavi SB, Khavanin A, Motamedzade M, Hajizade E, Mohseni M. The effect of postures on musculoskeletal disorders in work places. J Kermanshah Univ Med Sci. 2012;16(5):367–74.
5. Punnett L, Wegman DH. Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. J Electromyogr Kinesiol. 2004;14(1):13–23.
6. Choobineh AR, Daneshmandi H, Deilami F, Khoshnami S. Ergonomic workplace assessment and survey of musculoskeletal injuries in a generator manufacturing company. J Heal Syst Res. 2013;9:20–30.

7. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied linear statistical models. Vol. 4. Irwin Chicago; 1996.

8. Harvey G. Multilevel Models in Educational and Social Research. John Wiley & Sons, 2011.

9. McCulloch CE, Neuhaus JM. Generalized linear mixed models. Wiley Online Library; 2001.

10. Sullivan LM, Dukes KA, Losina E. Tutorial in biostatistics. An introduction to hierarchical linear modelling. Stat Med. 1999;18(7):855–88.

11. Buxton R. Statistics: Multilevel modeling. Math Learn Support Centre–Technical report. 2018. Online available at: www.statstutor.ac.uk/ resources/uploaded/multilevel modelling.pdf.

12. Haji Ahmadi M, Ayatollahi MT, Behboudian J. The effect of missing data in growth curves. J Babol Uni Med Sci. 2004;6(4):23-29.

13. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. Vol. 998. John Wiley & Sons; 2012.

14. Guo G, Zhao H. Multilevel modeling for binary data. Annu Rev Sociol. 2000;26(1):441–62.

15. Song X, Lee S. Model comparison of generalized linear mixed models. Stat Med. 2006;25(10): 1685–98.

16. Kuorinka I, Jonsson B, Kilbom A, Vinterberg H, Biering-Sørensen F, Andersson G, et al. Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. Appl Ergon. 1987;18(3):233–7.

17. Crawford JO. The Nordic Musculoskeletal Questionnaire. Occup Med (Chic Ill) [Internet]. 2007;57(4):300–1. Available from: http://dx.doi. org/10.1093/occmed/kqm036

18. Dickinson CE, Campion K, Foster AF, Newman SJ, O'rourke AMT, Thomas PG. Questionnaire development: an examination of the Nordic Musculoskeletal questionnaire. Appl Ergon. 1992;23(3):197–201.

19. David G, Woods V, Li G, Buckle P. The development of the Quick Exposure Check (QEC) for assessing exposure to risk factors for work-related musculoskeletal disorders. Appl Ergon. 2008;39(1):57–69.

20. Li G, Buckle P. A practical method for the assessment of work-related musculoskeletal risks-Quick Exposure Check (QEC). In: Proceedings of the human factors and ergonomics society annual meeting. SAGE Publications Sage CA: Los Angeles, CA; 1998. p. 1351–5.