Korean J Fam Med. 2013;34:434-436 http://dx.doi.org/10.4082/kjfm.2013.34.6.434

Comments on Statistical Issues in November 2013

Commentary

Yong Gyu Park

Department of Biostatistics, The Catholic University of Korea College of Medicine, Seoul, Korea

In this section, we explain the actual number of observations used in a multivariate analysis when one or more explanatory variables have missing values, which appeared in the articles titled, "Postmarketing surveillance study of the efficacy and safety of Phentermine in patients with obesity," by Kim et al.[1] and "Relationships between dietary habits and allostatic load index in metabolic syndrome patients," by Kim[2] published in September 2013.

MISSING VALUES IN MULTIVARIATE ANALYSES

When there are some missing values in one or more variables, most researchers choose one of the following strategies for analysis: 1) delete all observations which have missing values or 2) use all observations regardless of missing values. The purpose of this section is to show how many observations are actually analyzed in multivariate analyses, such as multiple linear regression analysis or multiple logistic regression analysis, when there are different numbers of missing values in each explanatory variable. Let's perform a multiple linear regression using the following hypothetical data (Table 1).

Table 1. Hypothetical data

Response variable	Explanatory variables		
y	x1	x2	x3
1	1	1	1
2	2	2	2
3	1	1	2
4	2	3	3
5	3	2	3
6	2	4	3
7	.	3	4
8	.	4	4
9	.	.	4
10	.	.	4

In this data, explanatory variable x1 has four, x2 has two, and x3 has no missing values, respectively (denoted as a dot), and we will perform three analyzing processes using SPSS, 1) Pearson correlation analysis, 2) multiple linear regression analysis, and 3) stepwise multiple linear regression analysis.

PEARSON CORRELATION ANALYSIS

From the menus choose:

Analyze

Correlate

Bivariate...

Select all variables: y, x1, x2, x3

We obtain the following results: (Table 2).

X3 has the highest correlation with y, and explanatory variables, and x1, x2, and x3 are analyzed by using only their valid observations, 6, 8, and 10, respectively.

Table 2. Correlation coefficients

		y	x1	x2	x3
y	Pearson correlation coefficients	1	0.639	0.830	0.940
	P-value (two-sided)		0.172	0.011	0.000
	N	10	6	8	10

MULTIPLE LINEAR REGRESSION ANALYSIS

From the menus choose:

Analyze

Regression

Linear...

Choose dependent variable: y

Independent variables: x1, x2, x3

Options: statistics: descriptive statistics

We obtain the following results: (Tables 3-5).

X3 has the highest correlation with y, but all analyses (descriptive statistics, correlation analysis, and multiple regression analysis) are performed using only six observations which have no missing values for all dependent variables.

STEPWISE MULTIPLE LINEAR REGRESSION ANALYSIS

(Menus, variable selection, and options are the same as the above)

Variable selection methods: stepwise

We obtain the following results: (Tables 6-8).

Descriptive statistics (the same as above results)

Correlation coefficients (the same as above results)

Table 3. Descriptive statistics

	Mean	SD	N
y	3.50	1.871	6
x1	1.83	0.753	6
x2	2.17	1.169	6
x3	2.33	0.816	6

Table 4. Correlation coefficients

		y	x1	x2	x3
Pearson correlation	y	1.000	0.639	0.777	0.917
N	y	6	6	6	6

Table 5. Coefficients

Model		Unstandardized coefficient		Standardized coefficient	t	Significance
		Coefficient	SE	Beta		
1	Constant	−1.250	1.531		−0.816	0.500
	x1	−0.250	1.046	−0.101	−0.239	0.833
	x2	0.250	0.685	0.156	0.365	0.750
	x3	2.000	1.311	0.873	1.526	0.267

Table 6. Entered/removed variables

Model	Entered variable	Removed variable	Method
1	x3	.	Stepwise (criteria: enter = 0.050, remove = 0.100)

Table 7. Analysis of variance

Model		Sum of square	Degrees of freedom	Mean square	F-ratio	Significance
1	Regression	14.700	1	14.700	21.000	0.010
	Residual	2.800	4	0.700		
	Total	17.500	5			

Table 8. Coefficients

Model		Unstandardized coefficient		Standardized coefficient	t	Significance
		Coefficient	SE	Beta		
1	Constant	−1.400	1.122		−1.247	0.280
	x3	2.100	0.458	0.917	4.583	0.010

From the total degrees of freedom (df = 5) in the analysis of variance table, a stepwise multiple regression analysis is performed using only six observations which have no missing values for all explanatory variables, even though the results show that only one variable, x3, which has no missing values, remains in the final model.

As we can see from the above three results, the actual number of observations analyzed in a multivariate analysis is the minimum number of valid observations of all explanatory variables we had intended to include in the analysis, regardless of the variable selection methods.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Kim HO, Lee JA, Suh HW, Kim YS, Kim BS, Ahn ES, et al. Postmarketing surveillance study of the efficacy and safety of phentermine in patients with obesity. Korean J Fam Med 2013;34:298-306.
2. Kim JY. Relationships between dietary habits and allostatic load index in metabolic syndrome patients. Korean J Fam Med 2013;34:334-46.