AMERICAN ACADEMY OF OTOLARYNGOLOGY–HEAD AND NECK SURGERY
FOUNDATION

Otolaryngology–
Head and Neck Surgery
148(2) 185–190
© American Academy of
Otolaryngology—Head and Neck
Surgery Foundation 2013
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0194599812462685
http://otojournal.org
SAGE

Practical Guide to Understanding Multivariable Analyses: Part A

J. Gail Neely, MD[1], Randal C. Paniello, MD[1],
Judith E. Cho Lieu, MD, MSPH[1], Courtney C. J. Voelker, MD, DPhil[1],
David J. Grindler, MD[1], Sunitha M. Sequeira, MD[1], and
Brian Nussenbaum, MD[1]

No sponsorships or competing interests have been disclosed for this article.

Abstract

Multivariable analyses are complex statistical methods to evaluate the impact of multiple variables on outcomes of interest. Books have been written on each of these methods detailing the mathematical and statistical objectives and processes. However, we have found very little in the way of brief reports that help the nonstatistically trained physician obtain a basic understanding of multivariable analyses in order to have some understanding of the increasing literature using these methods. This work is organized in 2 parts. This article, Part A, addresses the "big 4" algebraic methods of multivariable analysis. The primary focus of Part A is to present a brief "primer" to help the reader understand the methods and uses; it expressly avoids the many details of statistical assumptions, calculations, and myriad branching alternatives. Part B will concentrate on conjunctive consolidation and will focus on enough information to allow the interested reader to actually perform the analysis. For the statistical scholar, we have included references to several voluminous serious works.

Keywords

statistics, multivariable analysis, linear regression, logistic regression, Cox proportional hazard analysis, discriminant function analysis

Received August 3, 2012; accepted September 6, 2012.

We are familiar with concentrating on one experimental independent, predictor variable (eg, treatment) relating to one dependent, outcome variable (eg, outcome). However, many nonexperimental baseline variables inherent in test subjects (eg, age, race, sex, socioeconomic level, comorbidities) or in the application of the treatment (eg, quality of performance, time of performance, duration, dosage) may influence the outcome. Multivariable analyses are used to explore and determine which other independent (predictor) variables might play a role in confounding or modifying the effect on one or more dependent (outcome) variables; these nonexperimental variables that might influence the outcomes are called covariates.[1] Multivariable analytic methods are complex statistical techniques. The objective of this guide is to help readers of all backgrounds understand these methods conceptually; as such, detailed explanations of some of the deeper mathematical derivations are excluded.

Multivariable analyses may be used to (1) identify baseline variables that have significant effects on the outcome of interest in addition to the intervention being tested; (2) improve hypothesis testing by controlling for important covariates; (3) identify possible etiologic factors leading to disease; (4) establish "weights," which reflect the relative importance, of variables used in diagnostic or prognostic scoring systems; and (5) develop new rating scales.[2,3]

The many multivariable techniques are too numerous to mention. However, there are 2 general methods to approach multivariable analysis: (1) fit data into mathematical models and/or (2) arrange data into clusters. Feinstein[2] characterized the mathematical models into the "big 4": multiple linear regression, multiple logistic regression, proportional hazard (Cox) analysis, and discriminant function analysis. Arranging data into clusters is often performed using conjunctive consolidation. In this article, we present the essence of the mathematical models in a way, it is hoped, that will be understandable and useful in reading articles using multivariable analyses. The objective is not to present the enormous details that fill books on each method but to accurately present enough information that will allow to reader to better understand the "what" and "why" these methods have use. Conjunctive consolidation will be discussed in a subsequent article.

The multivariable analytic algebraic method of choice depends on the data scale in which the dependent variable is reported (**Table 1**).[2-4] Feinstein[2] emphasized, "The choice of

[1] Washington University School of Medicine, St Louis, Missouri, USA

Corresponding Author:
J. (John) Gail Neely, MD, Professor and Director, Otology/Neurotology/Base of Skull Surgery, Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine, 660 S Euclid Avenue, Box 8115, St Louis, MO 63110, USA
Email: neelyg@ent.wustl.edu

Table 1. Selection of Multivariable Methods by Scales of Dependent Variable[2]

Dependent Variable Scale	Method
Continuous (sometimes ordinal or binary)	Multiple linear regression
Ordinal	None—however, can be performed using linear or logistic techniques
Dichotomous (binary), nominal (sometimes ordinal)	Multiple logistic regression
"Moving" binary (survival curves)	Cox proportional hazard
Nominal (sometimes binary)	Discriminant function analysis

Categorical variables: ordinal, dichotomous (binary), and nominal.

Table 2. Characteristics of Multivariable Methods

Method	Characteristic Function	Formula
Multiple linear regression	Predicts the value of *Y*, given *X*s	$Y = b_0 + b_1X_1 + b_2X_2 + \text{etc}$
Multiple logistic regression	Predicts the probability (or odds) of *Y* occurring, given *X*s	Probability: $P(Y) = 1/1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \text{ etc})}$ Odds: *Log (odds) = $b_0 + b_1X_1 \ldots b_nX_n$
Cox proportional hazards regression	Predicts the probability of event occurring at time *t*, for an individual	$Y_{i,t} = S(t)^{eG}$ *Log$(h_it/h_0t) = b_0 + b_1(x_1 - \bar{x}_1) \ldots + b_n(x_n - \bar{x}_n)$
Discriminant function analysis	Attempts to discriminate between nominal groups using mathematical models	$L = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \ldots$

Note: *G* (dependent variable for the method) changes as the method changes; for example, *G* in multiple linear regression is *Y*, and in multiple logistic regression, *G* is *P*(*Y*). *G* in Cox proportional hazards regression is used as a double exponent, and in discriminant function analysis, *G* is *L*.[2]
*Personal communication with Kenneth Schechtman, PhD.

analytic methods is seldom affected by the type of data contained in the *independent variables*" (emphasis added).

The general configuration of the multivariable mathematical models is in the form of regression. The usual linear regression model for bivariate analysis (one dependent variable and one independent variable) is

$$Y = a + bX,$$

where Y is the dependent variable; a is the intercept, also known as the constant (the value of Y when X is 0); b is the slope of the line, also known as the regression coefficient, which shows the impact of X on Y; and X is the independent variable.[5]

The generic algebraic model for all 4 multivariable analyses is[2]

$$G = b_0 + b_1X_1 + b_2X_2 + b_3X3 \ldots,$$

where G is the dependent variable relative to analysis; b_0 is the intercept (value of G when X is 0); b_1, b_2, b_3, and so on are the regression coefficients showing the impact of X on G; and X_1, X_2, X_3, and so on are the independent variables (**Table 2**).

In building the candidate independent (predictor) variables to be included in the multivariable model, several factors are considered for each: (1) Does it make biological sense? (2) Have previous articles or pilot projects suggested it is important? (3) Does bivariate analysis show it is statistically significant? Candidate variables should (1) significantly correlate with the outcome variable, and (2) they should not highly correlate with each other (known as collinearity). The ideal multivariable model should have the smallest number of variables predicting the largest amount of variation.[6] A loose rule of thumb suggests that for every variable included in the model, at least 10 subjects who develop the outcome of interest should be available.[7]

The first step in multivariable analysis is to determine which variables are statistically significant in bivariate analysis relative to the outcome of interest. **Table 3** shows the results of bivariate analysis of an arbitrary sample data set.

Multiple Linear Regression

The objective of linear regression is to find the line that best fits the data. To begin to understand the computational process, the data of 2 arbitrary variables are displayed in a scatter plot in **Figure 1**. The target dependent variable (outcome) is displayed on the y-axis, and the independent (predictor) variable is on the x-axis. If this is done once, as in this figure, it is simple linear regression. If this is done for more than one independent variable to see how they affect the outcome, it is multiple linear regression. Note, however, that multiple linear regression is not just a series of simple linear regressions; it is a regression that combines multiple independent variables that affect a single outcome into one equation, as shown earlier in the "generic equation."

Computer programs generating a regression line from the actual sample data attempt to find a line that best fits the data. The fit of the data is measured with the residual value, which

Table 3. Bivariate Analyses: 5-Year Survival Fixed-Point Analysis (Comparison Groups: Alive and Dead at 5 Years)

	Variable *A* (Continuous Scale)	Variable *B* (Ordinal)	Variable *C* (Ordinal)	Variable *D* (Ordinal)	Variable *E* (Continuous Scale)	Variable *F* (Continuous Scale)
P value	<.001	.477	.344	.342	<.001	.977
Test	*t* test	Mann-Whitney	Mann-Whitney	Mann-Whitney	*t* test	*t* test

Variables *A* and *E* are statistically significant using bivariate analysis.

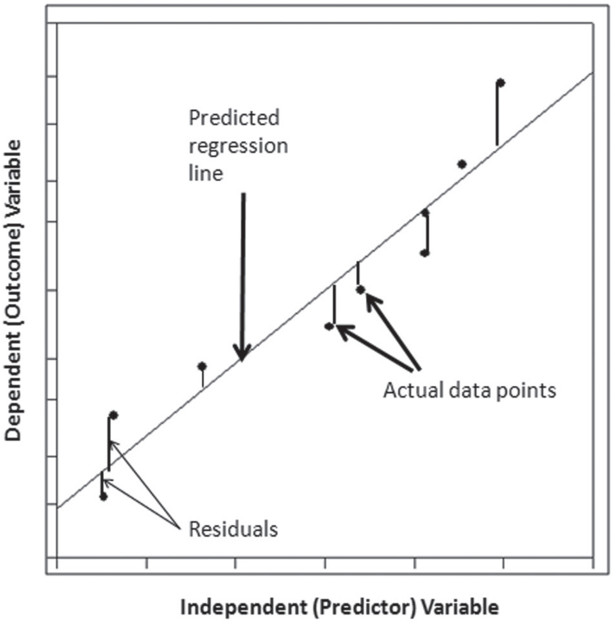

Figure 1. Illustration of scatter plot of arbitrary data showing individual data points, regression line fitting the data, and residuals. By convention, the independent (predictor) variable is on the x-axis and the dependent (outcome) variable is on the y-axis.

is the value of the amount of deviation of the observed data from the predicted regression line. The residual value can be positive or negative. The goal of regression analysis is to minimize the residual values or, in other words, derive a linear relationship that has the smallest values of residuals. The best fit typically uses "least squares"; this means squaring each actual value deviation from the predicted regression line (giving all the residuals a positive sign) and adding all these squared values. To use this technique of minimization of residual values, a number of very specific statistical assumptions must be met. The details of these assumptions are beyond the scope of this article; however, it does emphasize the need for a well-versed statistician to be in consultation.

A multiple linear regression was performed on the arbitrary sample data set using SPSS version 20 (SPSS, Inc, an IBM Company, Chicago, Illinois); the results are seen in **Table 4**. In this example, the overall model is significant ($P < .05$) in predicting the outcome. In the coefficients portion of the table, the only variable in the model that was statistically significant was variable *A*. In multiple linear regression, *B* or beta regression coefficients show the predicted change in *Y* with each unit change in *X* (relative to the specific independent variable), with all other independent variables held constant.[6] The sign of the beta coefficient tells the direction of change in *Y* caused by *X*. The unstandardized *B* coefficient is calculated in the units specific for that variable. However, the standardized beta regression coefficients reflect the comparative impacts between the independent variables on the dependent variable because standardization removes the various units used to measure these variables.

For the example in **Table 4**, the *B* regression coefficients seem to all have close to the same degree of influence on outcome. However, they suggest that a unit change in variables *A* and *C* tends to increase the dependent target outcome variable *E*, and a unit change in variables *B* and *D* tends to decrease the dependent variable *E* (based on the signs of the coefficients). However, only variable *A* has a statistically significant effect on outcome. When comparing the gradient impacts of the independent variables by looking at the standardized beta regression coefficients, variable *A* has a much greater impact on the dependent variable (outcome) than do the others when the units in which those variables are measured are removed.

Multiple Logistic Regression

Multiple logistic regression predicts the *probability* (or better, the odds) of *Y* occurring, given *X*,[8] unlike multiple linear regression, which predicts the *value* of *Y*, given one or more *X*s (**Table 2**). The name *logistic* refers to the process of using logarithms (in this case, the natural log, base *e*). When the outcome variable is dichotomous (binary), it cannot behave in a linear fashion and therefore fails to meet the fundamental assumption for linear regression. However, by using logarithmic transformation, the relationship between the variables behaves closer to a linear relationship. In this case, the logarithmic transformation of a linear regression equation is called a *logit*.[8] Note that in **Table 2**, the exponent of *e* has the same form as the independent variable side of the multiple linear regression equation ($b_0 + b_1x_1 + b_2x_2 + \ldots$).

Next, we must review probability and odds. Probabilities are proportions (ratios, fractions) ranging between 0 (impossible) and 1 (certain). Fundamentally, a probability is a ratio of a frequency count of occurrences of events divided by all possible events.[9] Knowing one probability, we can rapidly obtain the converse probability as follows. If the probability of event is $P = 0.25$, then the probability of not getting the event is $1 - P$, also known as Q, $= 0.75$. If the 2 ratios are divided, we get

Table 4. Multiple Linear Regression

Model 1	Sum of Squares	Mean Square	*F*	*P* Value
Regression	1390.102	347.526	11.243	< .001
Residual	463.655	30.910		
Total	1853.757			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
Model 1	*B*	Standard Error	β	*t*	*P* Value
(Constant)	4.489	5.106		.879	.393
Variable *A* (continuous scale)	.572	.085	.880	6.698	< .001
Variable *B* (ordinal)	–.756	1.261	–.079	–.599	.558
Variable *C* (ordinal)	.733	.603	.160	1.216	.243
Variable *D* (ordinal)	–.673	1.298	–.068	–.519	.612

Dependent variable: variable *E*.

Table 5. Logistic Regression (Dependent Variable Fixed-Point Analysis[a]: Survival at 5 Years)

		Variables in the Equation							
								95% CI for Exp(*B*)	
		B	SE	Wald	*df*	Significance	Exp(*B*)	Lower	Upper
Step 1									
	Variable *E*	.279	.121	5.309	1	.021	1.321	1.042	1.674
	Constant	–5.290	2.196	5.805	1	.016	.005		

Abbreviation: CI, confidence interval.

[a] In this example, fixed-point analysis at 5 years rather than time-to-event data using Kaplan-Meier curves as for Cox regression.

an odds ratio such as $P/1 - P = P/Q = 0.25/0.75 = 0.333$ or 1/3, stated as there is a 1 to 3 odds of an event or conversely a 3 to 1 odds for not having an event. Note that when we talk of a probability, it is a single fraction. When we speak of odds, it is a ratio of 2 fractions, and the result is spoken of as though it were one thing "odds."

Getting back to logistic regression, which predicts the probability of Y occurring, given X, if $\hat{Y}$ is the estimated (predicted) occurrence, then the odds (odds ratio) of occurrence is $\hat{Y}/(1 - \hat{Y})$. If this is logarithmically transformed, it becomes a logit $\ln(P/Q)$ or $\ln[\hat{Y}/(1 - \hat{Y})]$.[2] *ln* stands for natural logarithm. The "natural log" base is written as *e;* the value of *e* is 2.718281828. For example, the $\text{Log}_e\ 100 = 4.60517$.

Logistic regression for the arbitrary sample data is seen in **Table 5**. The binary dependent, target variable (eg, survival status at 5 years) is coded 0 (dead) or 1 (alive), and the independent variables are in differing scales. The resulting coefficients (Exp(*B*)) may be interpreted as odds ratios.[2,8] When all the sensible variables were included in the logistic model, only variable *E* was significantly predictive of being alive at 5 years. Using the Exp(*B*), the odds of surviving 5 years is increased by 1.321 for every unit increase in variable *E*.

Cox Proportional Hazards Regression

Classic survival curves, such as the Kaplan-Meier curve,[10] can be modified for more detailed analysis by multiple baseline variables. If these prognostic baseline variables are favorable, the curves may be better than if not favorable. The purpose of the Cox proportional hazard regression is to estimate the probability of an event for a single person at a specific time.[1,2] It is not used to assess the entire curve; the log-rank test is used to compare the differences between whole Kaplan-Meier curves.[1]

In a survival curve (such as a Kaplan-Meier curve) for an entire group, $S(t)$ is the proportion of survivors at time t. When $S(t)$ is raised to the double exponent power e^G, the result is an estimation of the survival of a person (i) at time (t) (written $\hat{Y}_{i,t}$). The "hat" ^ above a letter indicates that it is an estimate; Y without a "hat" would be the actual observed value. G contains a modification of the predictor variables in the general formula $G = b_0 + b_1X_1 + b_2X_2$ . . . and so on, as seen in the previous formulas (**Table 2**).

The value of the Cox proportional hazards regression is to determine the impact of baseline variables on the survival curve for an individual and is reported as a hazard ratio (much like one would see with an odds ratio).[11] It is a method of looking at multiple variables as they might influence a moving target (eg, time to event, such as tumor recurrence). It should be emphasized that the Cox proportional hazard model is quite complex and often requires consultation with a statistician for both proper analysis and interpretation.[1]

Discriminant Function Analysis

Uniquely different from the other multivariable analytic methods, which focus on one dependent variable, this method seeks to derive a mathematical model for many dependent unranked nominal (named) variables as targets, such as diagnostic categories (trauma, metabolic, infectious, congenital, rheumatological disease). Ultimately, the effort is to derive discriminant function lines (L) that incorporate multiple independent variables (X_1, X_2 . . ., etc), such as serological tests, electrical tests, and imaging, using the standard linear regression formula $L = b_0 + b_1X_1 + b_2X_2 \ldots$ and so on to derive the probability of each diagnostic category and then use these lines to separate the dependent nominal variables into groups that are different from each other.[2]

This technique is complex and somewhat arbitrary. It is infrequently used today in biomedical research. It is mentioned here only to identify that such analyses can potentially be done but are fraught with difficulty. Linear and logistic regression techniques are currently more useful.

Summary

Multivariable analyses are used to explore and determine which independent (predictor) variables, in addition to experimental interventions, play a role in predicting one or more dependent (outcome) variables. These nonexperimental variables that might influence the outcomes are called covariates.

Goals of multivariable analysis are to (1) identify baseline variables that have significant effects on the outcome of interest in addition to the intervention being tested, (2) improve hypothesis testing by controlling for important covariates, (3) identify possible etiologic factors leading to disease, (4) establish weights of variables used in diagnostic or prognostic scoring systems, and (5) develop new rating scales.

There are two methods to approach multivariable analysis: (1) fit data into mathematical models and/or (2) arrange data into clusters. The "big 4" of the mathematical models are multiple linear regression, multiple logistic regression, proportional hazard (Cox) analysis, and discriminant function analysis. Arranging data into clusters, such as in conjunctive consolidation, is the second category of multivariable analysis, which will be discussed in detail in Part B.

In this article, we present the essence of the mathematical models in a way, it is hoped, that will be understandable and useful in reading articles using multivariable analyses. The objective was not to present the enormous details that fill books on each method but to accurately present enough information that will allow the reader to better understand the "what" and "why" these methods have use.

Acknowledgments

Our thanks for critical review of the manuscript go to Jay F. Piccirillo, MD, Vice-Chairman for Research, Director of Clinical Outcomes Research Office, Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine, St Louis, Missouri; Kenneth B. Schechtman, PhD, Associate Professor of Biostatistics, Director of Consulting Activities in Division of Biostatistics, Washington University School of Medicine, St Louis, Missouri; and Dorina Kallogieri, MD, MPH, senior data control coordinator, Clinical Outcomes Research Office, Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine, St Louis, Missouri.

Author Contributions

J. Gail Neely, substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data, drafting the article and revising it critically for important intellectual content, and final approval of the version to be published; **Randal C. Paniello**, substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data, drafting the article and revising it critically for important intellectual content, and final approval of the version to be published; **Judith E. Cho Lieu**, substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data, drafting the article and revising it critically for important intellectual content, and final approval of the version to be published; **Courtney C. J. Voelker**, substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data, drafting the article and revising it critically for important intellectual content, and final approval of the version to be published; **David J. Grindler**, substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data, drafting the article and revising it critically for important intellectual content, and final approval of the version to be published; **Sunitha M. Sequeira**, substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data, drafting the article and revising it critically for important intellectual content, and final approval of the version to be published; **Brian Nussenbaum**, substantial contributions to conception and design, acquisition of data, and analysis and interpretation of data, drafting the article and revising it critically for important intellectual content, and final approval of the version to be published.

Disclosures

Competing interests: None.

Sponsorships: None.

Funding source: None.

References

1. Altman DG. *Practical Statistics for Medical Research*. New York: Chapman & Hall/CRC; 1991:80, 336-364, 387-388.
2. Feinstein AR. *Multivariable Analysis: An Introduction*. New Haven, CT: Yale University Press; 1996:2-3, 68-70, 74-77, 264-294, 297-369, 370-430, 431-474, 512-524.
3. Jekel JF, Katz DL, Elmore JG. *Epidemiology, Biostatistics, and Preventive Medicine*. 2nd ed. Philadelphia, PA: W. B. Saunders; 2001:209-218.
4. Riegelman RK. *Studying a Study and Testing a Test: How to Read the Medical Evidence*. Philadelphia, PA: Lippincott Williams and Wilkins; 2005:350-368.
5. Feinstein AR. *Clinical Epidemiology: The Architecture of Clinical Research*. Philadelphia, PA: W. B. Saunders; 1985:124, 174-178.

6. Peat J, Barton B. *Medical Statistics: A Guide to Data Analysis and Critical Appraisal*. Malden, MA: Blackwell; 2005:162-201.
7. Lieu JEC. Development of staging and stratification systems. *ORL*. 2004;66:173-179.
8. Field A. *Discovering Statistics Using SPSS*. 3rd ed. Thousand Oaks, CA: Sage; 2009:197-263, 264-315.
9. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 2nd ed. Upper Saddle River, NJ: Prentice Hall Health; 2000:388, 509-535.
10. Rich J, Neely J, Paniello R, et al. A practical guide to understanding Kaplan-Meier curves. *Otolaryngol Head Neck Surg*. 2010;143:331-336.
11. Flick RP, Katusic SK, Colligan RC, et al. Cognitive and behavioral outcomes after early exposure to anesthesia and surgery. *Pediatrics*. 2011;128:1053-1061.