Practical Multivariate Analysis

Sixth Edition

Abdelmonem Afifi
Susanne May
Robin A. Donatello
Virginia A. Clark

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Contents

Part I

Preparation for Analysis

Chapter 1

What is multivariate analysis?

1.1 Defining multivariate analysis

The expression **multivariate analysis** is used to describe analyses of data that are multivariate in the sense that numerous observations or variables are obtained for each individual or unit studied. In a typical survey 30 to 100 questions are asked of each respondent. In describing the financial status of a company, an investor may wish to examine five to ten measures of the company's performance. Commonly, the answers to some of these measures are interrelated. The challenge of disentangling complicated interrelationships among various measures on the same individual or unit and of interpreting these results is what makes multivariate analysis a rewarding activity for the investigator. Often results are obtained that could not be attained without multivariate analysis.

In the next section of this chapter several studies are described in which the use of multivariate analysis is essential to understanding the underlying problem. Section 1.3 provides a rationale for making a distinction between confirmatory and exploratory analyses. Section 1.4 gives a listing and a very brief description of the multivariate analysis techniques discussed in this book. Section 1.5 then outlines the organization of the book.

1.2 Examples of multivariate analyses

The studies described in the following subsections illustrate various multivariate analysis techniques. These are used later in the book as examples.

Depression study example

The data for the depression study have been obtained from a complex, random, multiethnic sample of 1000 adult residents of Los Angeles County. The study was a **panel** or **longitudinal** design where the same respondents were interviewed four times between May 1979 and July 1980. About three-fourths of the respondents were re-interviewed for all four interviews. The field work for the survey was conducted by professional interviewers from the Institute for Social Science Research at the University of California in Los Angeles.

This research is an epidemiological study of depression and help-seeking behavior among free-living (noninstitutionalized) adults. The major objectives are to provide estimates of the prevalence and incidence of depression and to identify causal factors and outcomes associated with this condition. The factors examined include demographic variables, life events stressors, physical health status, health care use, medication use, lifestyle, and social support networks. The major instrument used for classifying depression is the Depression Index (CESD) of the National Institute of Mental Health, Center of Epidemiological Studies. A discussion of this index and the resulting prevalence of depression in this sample is given in Frerichs et al. (1981).

The longitudinal design of the study offers advantages for assessing causal priorities since the time sequence allows us to rule out certain potential causal links. Nonexperimental data of this type cannot directly be used to establish causal relationships, but models based on an explicit theoretical

framework can be tested to determine if they are consistent with the data. An example of such model testing is given in Aneshensel and Frerichs (1982).

Data from the first time period of the depression study are described in Chapter 3. Only a subset of the factors measured on a subsample of the respondents is included in this book's web site in order to keep the data set easily comprehensible. These data are used several times in subsequent chapters to illustrate some of the multivariate techniques presented in this book.

Parental HIV study

The data from the parental HIV study have been obtained from a clinical trial to evaluate an intervention given to increase coping skills (Rotheram-Borus et al., 2001). The purpose of the intervention was to improve behavioral, social, and health outcomes for parents with HIV/AIDS and their children. Parents and their adolescent children were recruited from the New York City Division of Aids Services (DAS). Adolescents were eligible for the study if they were between the ages of 11 and 18 and if the parents and adolescents had given informed consent. Individual interviews were conducted every three months for the first two years and every six months thereafter. Information obtained in the interviews included background characteristics, sexual behavior, alcohol and drug use, medical and reproductive history, and a number of psychological scales.

A subset of the data from the study is available on this book's web site. To protect the identity of the participating adolescents we used the following procedures. We randomly chose one adolescent per family. In addition, we reduced the sample further by choosing a random subset of the original sample. Adolescent case numbers were assigned randomly without regard to the original order or any other numbers in the original data set.

Data from the baseline assessment will be used for problems as well as to illustrate various multivariate analysis techniques.

Northridge earthquake study

On the morning of January 17, 1994 a magnitude 6.7 earthquake centered in Northridge, CA awoke Los Angeles and Ventura County residents. Between August 1994 and May 1996, 1830 residents were interviewed about what happened to them in the earthquake. The study uses a telephone survey lasting approximately 48 minutes to assess the residents' experiences in and responses to the Northridge earthquake. Data from 506 residents are included in the data set posted on the book web site, and described in Appendix A.

Subjects were asked where they were, how they reacted, where they obtained information, whether their property was damaged or whether they experienced injury, and what agencies they were in contact with. The questionnaire included the Brief Symptom Inventory (BSI), a measure of psychological functioning used in community studies, and questions on emotional distress. Subjects were also asked about the impact of the damage to the transportation system as a result of the earthquake. Investigators not only wanted to learn about the experiences of the Southern California residents in the Northridge earthquake, but also wished to compare their findings to similar studies of the Los Angeles residents surveyed after the Whittier Narrows earthquake on October 1, 1987, and Bay Area residents interviewed after the Loma Prieta earthquake on October 17, 1989.

The Northridge earthquake data set is used in problems at the end of several chapters of the book to illustrate a number of multivariate techniques. Multivariate analyses of these data include, for example, exploring pre- and post-earthquake preparedness activities as well as taking into account several factors relating to the subject and the property (Nguyen et al., 2006).

Bank loan study

The managers of a bank need some way to improve their prediction of which borrowers will successfully pay back a type of bank loan. They have data from the past on the characteristics of persons

to whom the bank has lent money and the subsequent record of how well the person has repaid the loan. Loan payers can be classified into several types: those who met all of the terms of the loan, those who eventually repaid the loan but often did not meet deadlines, and those who simply defaulted. They also have information on age, sex, income, other indebtedness, length of residence, type of residence, family size, occupation, and the reason for the loan. The question is, can a simple rating system be devised that will help the bank personnel improve their prediction rate and lessen the time it takes to approve loans? The methods described in Chapter 12 and Chapter 13 can be used to answer this question.

Lung function study

The purpose of this lung function study of chronic respiratory disease is to determine the effects of various types of smog on lung function of children and adults in the Los Angeles area. Because they could not randomly assign people to live in areas that had different levels of pollutants, the investigators were very concerned about the interaction that might exist between the locations where persons chose to live and their values on various lung function tests. The investigators picked four areas of quite different types of air pollution and measured various demographic and other responses on all persons over seven years old who live there. These areas were chosen so that they are close to an air-monitoring station.

The researchers took measurements at two points in time and used the change in lung function over time as well as the levels at the two periods as outcome measures to assess the effects of air pollution. The investigators had to do the lung function tests by using a mobile unit in the field, and much effort went into problems of validating the accuracy of the field observations. A discussion of the particular lung function measurements used for one of the four areas can be found in Detels et al. (1975). In the analysis of the data, adjustments must be made for sex, age, height, and smoking status of each person.

Over 15,000 respondents have been examined and interviewed in this study. The data set is being used to answer numerous questions concerning effects of air pollution, smoking, occupation, etc. on different lung function measurements. For example, since the investigators obtained measurements on all family members seven years old and older, it is possible to assess the effects of having parents who smoke on the lung function of their children (Tashkin et al., 1984). Studies of this type require multivariate analyses so that investigators can arrive at plausible scientific conclusions that could explain the resulting lung function levels.

This data set is described in Appendix A. Lung function and associated data for nonsmoking families for the father, mother, and up to three children ages 7–17 are available from the book's web site.

School data set

The school data set is a publicly available data set that is provided by the National Center for Educational Statistics. The data come from the National Education Longitudinal Study of 1988 (called NELS:88). The study collected data beginning with 8th graders and conducted initial interviews and four follow-up interviews which were performed every other year. The data used here contain only initial interview data. They represent a random subsample of 23 schools with 519 students out of more than a thousand schools with almost twenty five thousand students. Extensive documentation of all aspects of the study is available at the following web site: `http://nces.ed.gov/surveys/NELS88/`. The longitudinal component of NELS:88 has been used to investigate change in students' lives and school-related and other outcomes. The focus on the initial interview data provides the opportunity to examine associations between school and student-related factors and students' academic performance in a cross-sectional manner. This type of analysis will be illustrated in Chapter 18.

1.3 Exploratory versus confirmatory analyses

A crucial component for most research studies and analyses is the testing of hypotheses. For some types of studies, hypotheses are specified in detail prior to study start (a priori) and then remain unchanged. This is typically the case, e.g., for clinical trials and other designed experiments. For other types of studies, some hypotheses might be specified in advance while others are generated only after study start and potentially after reviewing some or all of the study data. This is often the case for observational studies. In this section, we make a distinction between two conceptually different approaches to analysis and reporting based on whether the primary goal of a study is to *confirm* prespecified hypotheses or to *explore* hypotheses that have not been prespecified.

The following is a motivating example provided by Fleming (2010). He describes an experience where he walked into a **maternity ward** (when they still had such) while visiting a friend who had just given birth. He noticed that there were 22 babies, but only 2 of one gender while the other 20 were of the other gender. As a statistician, he dutifully calculated the p-value for the likelihood of seeing such (or worse) imbalance if in truth there are 50% of each. The two-sided p-value turns out to be 0.0001, indicating a very small likelihood (1 in 10,000) of such or more extreme imbalance being observed if in truth there are 50% of each. This is an example of where the hypothesis was generated after seeing the data. We will call such hypotheses *exploratory*.

Following Fleming, researchers might want to go out and test an exploratory hypothesis in another setting or with new data. In the example above, one might want to go to another maternity ward to collect further evidence of a strong imbalance in gender distribution at birth. Imagine that in a second (*confirmatory*) maternity ward there might be exactly equal numbers for each gender (e.g. 11 boys and 11 girls). Testing the same hypothesis in this setting will not yield any statistically significant difference from the presumed 50%. Nevertheless, one might be tempted to simply combine the two studies. A corresponding two-sided p-value remains statistically significant (p-value < 0.01).

The above example might appear silly, because few researchers will believe that the distribution of gender at birth (without human interference) is very different from 50%. Nevertheless, there are many published research articles which test and present the results for hypotheses that were generated by looking at data and noticing 'unusual' results. Without a clear distinction between whether hypotheses were specified a priori or not, it is difficult to interpret the p-values provided.

Results from confirmatory analyses provide much stronger evidence than results from exploratory analyses. Accordingly, interpretation of results from confirmatory analyses can be stated using much stronger language than interpretation of results from exploratory analyses. Furthermore, results from exploratory analyses should not be combined with results from confirmatory analyses (e.g. in meta analyses), because the **random high bias** (Fleming, 2010) will remain (albeit attenuated). To avoid random high bias when combining data or estimates from multiple studies only data/estimates from confirmatory analyses should be combined. However, this requires clear identification of whether confirmatory or exploratory analysis was performed for each individual study and/or analysis.

Many authors have pointed out that the medical literature is replete with studies that cannot be reproduced (Breslow, 1999; Munafò et al., 2017). As argued by Breslow (1999), **reproducibility** of studies, and in particular epidemiologic studies, can be improved if hypotheses are specified a priori and the nature of the study (exploratory versus confirmatory) is clearly specified.

Throughout this book, we distinguish between the two approaches to multivariate analyses and presentations of results and provide examples for each.

1.4 Multivariate analyses discussed in this book

In this section a brief description of the major multivariate techniques covered in this book is presented. To keep the statistical vocabulary to a minimum, we illustrate the descriptions by examples.

Simple linear regression

A nutritionist wishes to study the effects of early calcium intake on the bone density of postmenopausal women. She can measure the bone density of the arm (radial bone), in grams per square centimeter, by using a noninvasive device. Women who are at risk of hip fractures because of too low a bone density will tend to show low arm bone density also. The nutritionist intends to sample a group of elderly churchgoing women. For women over 65 years of age, she will plot calcium intake as a teenager (obtained by asking the women about their consumption of high-calcium foods during their teens) on the horizontal axis and arm bone density (measured) on the vertical axis. She expects the radial bone density to be lower in women who had a lower calcium intake. The nutritionist plans to fit a simple linear regression equation and test whether the slope of the regression line is zero. In this example a single outcome factor is being predicted by a single predictor factor.

Simple linear regression as used in this case would not be considered multivariate by some statisticians, but it is included in this book to introduce the topic of multiple regression.

Multiple linear regression

A manager is interested in determining which factors predict the dollar value of sales of the firm's personal computers. Aggregate data on population size, income, educational level, proportion of population living in metropolitan areas, etc. have been collected for 30 areas. As a first step, a multiple linear regression equation is computed, where dollar sales is the outcome variable and the other factors are considered as candidates for predictor variables. A linear combination of the predictors is used to predict the outcome or response variable.

Discriminant function analysis

A large sample of initially disease-free men over 50 years of age from a community has been followed to see who subsequently has a diagnosed heart attack. At the initial visit, blood was drawn from each man, and numerous other determinations were made, including body mass index, serum cholesterol, phospholipids, and blood glucose. The investigator would like to determine a linear function of these and possibly other measurements that would be useful in predicting who would and who would not get a heart attack within ten years. That is, the investigator wishes to derive a classification (discriminant) function that would help determine whether or not a middle-aged man is likely to have a heart attack.

Logistic regression

An online movie streaming service has classified movies into two distinct groups according to whether they have a high or low proportion of the viewing audience when shown. The company also records data on features such as the length of the movie, the genre, and the characteristics of the actors. An analyst would use logistic regression because some of the data do not meet the assumptions for statistical inference used in discriminant function analysis, but they do meet the assumptions for logistic regression. From logistic regression we derive an equation to estimate the probability of capturing a high proportion of the target audience.

Poisson regression

In a health survey, middle school students were asked how many visits they made to the dentist in the last year. The investigators are concerned that many students in this community are not receiving adequate dental care. They want to determine what characterizes how frequently students go to the dentist so that they can design a program to improve utilization of dental care. Visits per year are count data and Poisson regression analysis provides a good tool for analyzing this type of data. Poisson regression is covered in the logistic regression chapter.

Survival analysis

An administrator of a large health maintenance organization (HMO) has collected data for a number of years on length of employment in years for their physicians who are either family practitioners or internists. Some of the physicians are still employed, but many have left. For those still employed, the administrator can only know that their ultimate length of employment will be greater than their current length of employment. The administrator wishes to describe the distribution of length of employment for each type of physician, determine the possible effects of factors such as gender and location of work, and test whether or not the length of employment is the same for two specialties. Survival analysis, or event history analysis (as it is often called by behavioral scientists), can be used to analyze the distribution of time to an event such as quitting work, having a relapse of a disease, or dying of cancer.

Principal components analysis

An investigator has made a number of measurements of lung function on a sample of adult males who do not smoke. In these tests each man is told to inhale deeply and then blow out as fast and as much as possible into a spirometer, which makes a trace of the volume of air expired over time. The maximum or forced vital capacity (FVC) is measured as the difference between maximum inspiration and maximum expiration. Also, the amount of air expired in the first second (FEV1), the forced mid-expiratory flow rate (FEF 25–75), the maximal expiratory flow rate at 50% of forced vital capacity (V50), and other measures of lung function are calculated from this trace. Since all these measures are made from the same flow–volume curve for each man, they are highly interrelated. From past experience it is known that some of these measures are more interrelated than others and that they measure airway resistance in different sections of the airway.

The investigator performs a principal components analysis to determine whether a new set of measurements called principal components can be obtained. These principal components will be linear functions of the original lung function measurements and will be uncorrelated with each other. It is hoped that the first two or three principal components will explain most of the variation in the original lung function measurements among the men. Also, it is anticipated that some operational meaning can be attached to these linear functions that will aid in their interpretation. The investigator may decide to do future analyses on these uncorrelated principal components rather than on the original data. One advantage of this method is that often fewer principal components are needed than original variables. Also, since the principal components are uncorrelated, future computations and explanations can be simplified.

Factor analysis

An investigator has asked each respondent in a survey whether he or she strongly agrees, agrees, is undecided, disagrees, or strongly disagrees with 15 statements concerning attitudes toward inflation. As a first step, the investigator will do a factor analysis on the resulting data to determine which statements belong together in sets that are uncorrelated with other sets. The particular statements that form a single set will be examined to obtain a better understanding of attitudes toward inflation. Scores derived from each set or factor will be used in subsequent analyses to predict consumer spending.

Cluster analysis

Investigators have made numerous measurements on a sample of patients who have been classified as being depressed. They wish to determine, on the basis of their measurements, whether these patients can be classified by type of depression. That is, is it possible to determine distinct types of depressed patients by performing a cluster analysis on patient scores on various tests?

Unlike the investigator studying men who do or do not get heart attacks, these investigators do not possess a set of individuals whose type of depression can be known before the analysis is performed. Nevertheless, the investigators want to separate the patients into unique groups and to examine the resulting groups to see whether distinct types do exist and, if so, what their characteristics are.

Log-linear analysis

An epidemiologist in a medical study wishes to examine the interrelationships among the use of substances that are thought to be risk factors for disease. These include four risk factors where the answers have been summarized into categories. The risk factors are smoking tobacco (yes at present, former smoker, never smoked), drinking (yes, no), marijuana use (yes, no), and other illicit drug use (yes, no). Previous studies have shown that people who drink are more apt than nondrinkers to smoke cigarettes, but the investigator wants to study the associations among the use of these four substances simultaneously.

Correlated outcomes regression

A health services researcher is interested in determining the hospital-related costs of appendectomy, the surgical removal of the appendix. Data are available for a number of patients in each of several hospitals. Such a sample is called a **clustered sample** since patients are clustered within hospitals. For each operation, the information includes the costs as well as the patient's age, gender, health status and other characteristics. Information is also available on the hospital, such as its number of beds, location and staff size. A multiple linear regression equation is computed, where cost is the outcome variable and the other factors are considered as candidates for predictor variables. As in multiple linear regression, a linear combination of the predictors is used to predict the outcome or response variable. However, adjustments to the analysis must be made to account for the clustered nature of the sample, namely the possibility that patients within any one hospital may be more similar to each other than to patients in other hospitals. Since the outcomes within a given hospital are correlated, the researcher plans to use correlated outcomes regression to analyze the data.

1.5 Organization and content of the book

This book is organized into two major parts. Part One (Chapters 1–6) deals with data entry, preparation, visualization, screening, missing values, transformations, and decisions about likely choices for analysis. Part Two (Chapters 7–18) deals with regression analysis.

Chapters 2–6 are concerned with data preparation and the choice of what analysis to use. First, **variables** and how they are classified are discussed in Chapter 2. The next chapter concentrates on the practical problems of getting data into the computer, handling nonresponse, data management, getting rid of erroneous values, and preparing a useful codebook. Visualization techniques are discussed in Chapter 4. The next chapter deals with checking assumptions of normality and independence. The features of computer software packages used in this book are discussed. The choice of appropriate statistical analyses is discussed in Chapter 6.

Readers who are familiar with handling data sets on computers could skip some of these initial chapters and go directly to Chapter 7. However, formal course work in statistics often leaves an investigator unprepared for the complications and difficulties involved in real data sets. The material in Chapters 2–6 was deliberately included to fill this gap in preparing investigators for real world data problems.

For a course limited to multivariate analysis, Chapters 2–6 can be omitted if a carefully prepared data set is used for analysis. The depression data set, presented in Chapter 3, has been modified to make it directly usable for multivariate data analysis, but the user may wish to subtract one from the variables 2, 31, 33, and 34 to change the values to zeros and ones. Also, the lung function data, the