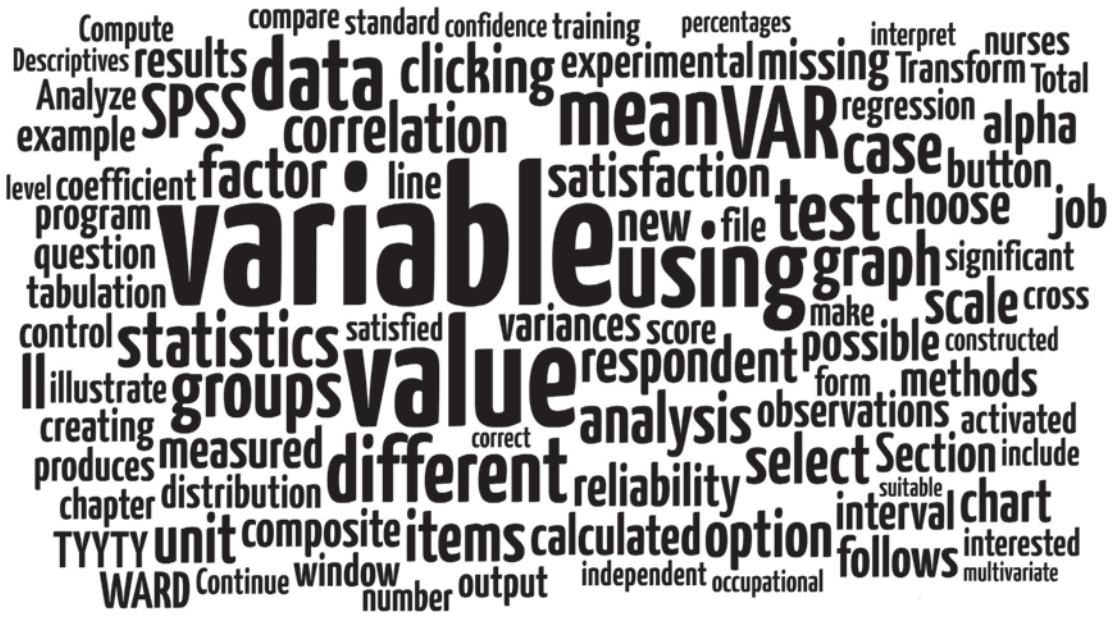
SECTION I

Basics of SPSS Software



CHAPTER 1

Introduction to SPSS

Goals of the Chapter

1. To orientate toward analyzing the data, using a statistical software package.
2. To obtain an overview of the literature discussing the SPSS software package.

The IBM Statistical Package for Social Sciences (SPSS) has been specifically designed to analyze quantitative data. Therefore, the aim of this Chapter is to equip the novice researcher with the necessary tools for using SPSS in analyzing quantitative data. Versions 17 and 18 of the software package were called Predictive Analysis Software (PASW) and from Version 19 onward, this software package is called IBM SPSS.[1] In what follows, the SPSS abbreviation has been used in the text. While I am writing this book, SPSS 23 is in use though the screen shots are from SPSS 22.

Advanced topics will not be discussed in this Section; instead, the focus will be on basic data editing and the performing simple analyses. As an example of data analysis, the same dataset introduced in Section V of Volume 1 is used; a simple visual analogue scale (VAS) type of a measurement instrument was used to examine job satisfaction. In Section II of this volume, the methods for the multivariate analysis are introduced in the SPSS environment. In Section I of Volume 3, more advanced methods in the SPSS environment are introduced. The experimental research is introduced in Section II, the multilevel modeling in Section III, the structural equating modeling in Section IV, and the survival analysis in Section V.

There is plenty of reference material on SPSS because it is a very commonly used statistical analysis software. On the side of this book at hand and the excellent manuals by Marija J. Norušis (2010; 2011a; 2011b; 2011c; see also ftp://public.dhe.ibm.com/software/

[1] SPSS Inc. was acquired by IBM in October 2009.

There are lots of literature to help …

analytics/spss/documentation/statistics/20.0/en/client/Manuals for free manuals), some materials are compared in the appendix on the basis of their published list of contents. The textbooks vary from very *elementary basics* or *short* books (e.g., Cronk, 2008; Griffith, 2010; Kirkpatrick & Feeney, 2011) to very *extensive and comprehensive books* (e.g., Field, 2009; Gray & Kinnear, 2011; Kremelberg, 2010; Weinberg & Abramowitz, 2008). Except Norušis' manuals and this book, surprisingly few SPSS books contain any piece of writing on estimating the reliability of the test scores (see, however, e.g., Berkman & Reise, 2012; Ho, 2006) nor Multilevel Modeling or Survival Analysis. It is good to also remember Internet materials like Garson's material for an advanced audience.[2]

…as well as Internet materials

The use of this basic guide to SPSS presupposes some familiarity with the basics of descriptive statistics.[3] Therefore, technical details (e.g., formulas for calculations) will not be addressed. The reader is also assumed to be familiar with different kinds of scales and the techniques of constructing an instrument of measurement.[4] The text begins with a situation of one having acquired the data: "What to do with the data?"

This section was originally written to correspond to SPSS Version 18 when its name was, for a short while, PASW. Most of the screenshots are taken from Version 22. Now, while writing this book, Version 23 of SPSS was already available, but the differences between these two versions are not particular up to the extent where the mention of a guide is required. We won't go through all the niceties of the software; instead, the reader is given pointers as how to perform simple analyses easily. Even an advanced user of SPSS may benefit from reading this section, for example, when estimating reliability. In Section II, some more demanding procedures of SPSS are handled.

Other parts of the book deepen the knowledge

It must be noted that a great deal of statistical computation (like basic descriptive statistics, computing probabilities, or performing tests on the Sampling distribution) can be performed in a spreadsheet software package, for example, in Microsoft Excel or in OpenOffice Calc. While these software packages do not include built-in functions in creating more complex statistics, these can be created quite easily by entering the relevant formulas into the spreadsheet software. On the other hand, it has often been easier to content oneself with figures that are produced by a software package like SPSS and be fairly confident that the computations are carried out correctly.

Many statistical computations can be performed in simple spreadsheet programs

[2] http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm (accessed on May 6, 2016).

[3] Sections I, V, and VI in Volume 1.

[4] Section II of Volume 1.

CHAPTER 2

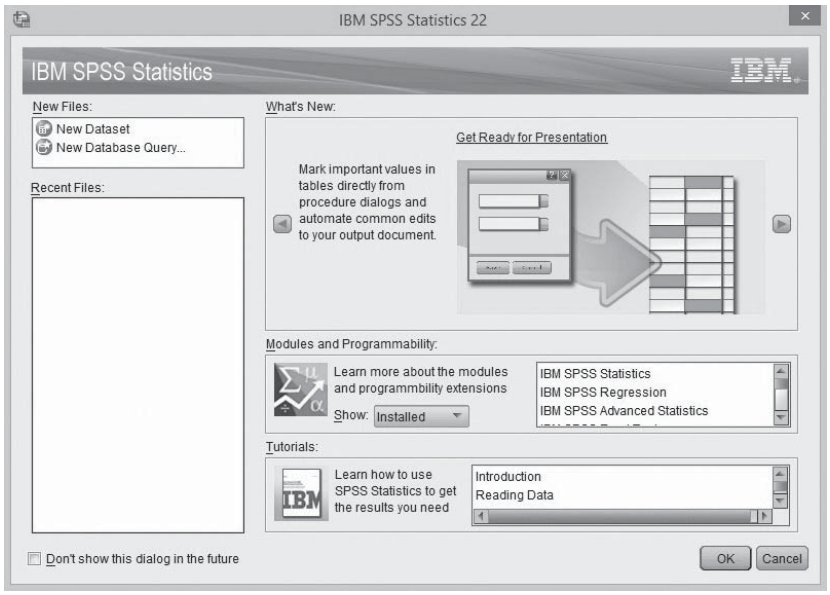
Basic Matters and Data Entry

Goals of the Chapter

1. To be capable of inputting data using the SPSS' own data editor and spreadsheet software.
2. To be able to define variables in SPSS.
3. To be able to check the data and replace missing values when necessary.
4. To get familiar with the concepts of "multicollinearity" and "outlier."

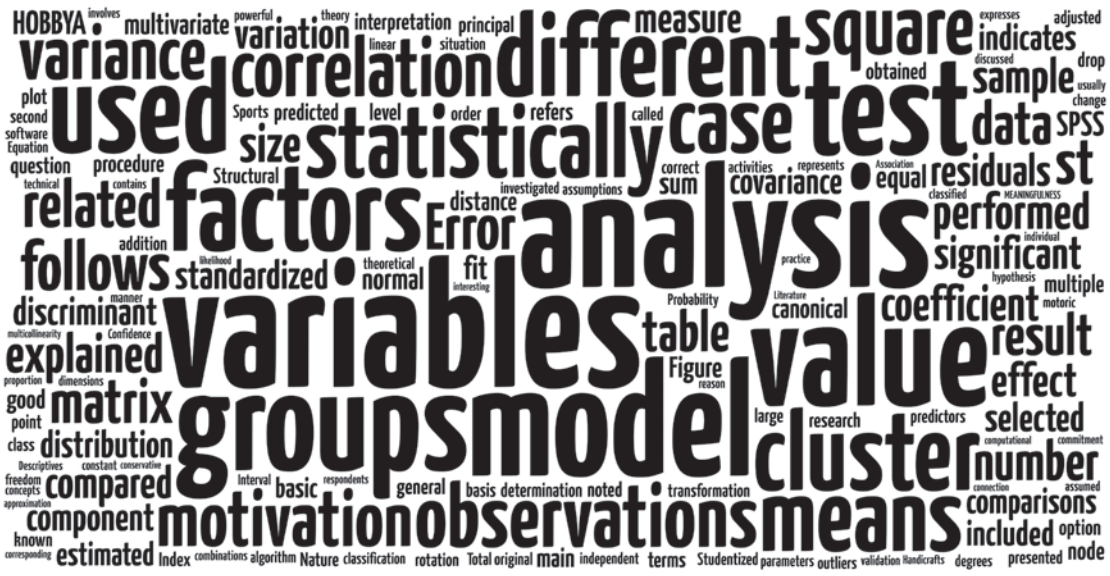
2.1 The Basic View

When the SPSS software package is opened, the following screen is displayed (Versions 20-22 look the same):



SECTION II

Basics of Multivariate Analysis



A Glossary for a Beginner Researcher

Determinant	Describes the general variance of the matrix. If the matrix is singular, its determinant is zero. See Singularity.
Estimation	An approximate computation of the values of statistics and parameters. See Parameter.
Heteroscedasticity	An undesirable characteristic of the residuals' variance faced in many multivariate analysis methods. The basic assumption is homoscedasticity, that is, that the residuals are evenly distributed. If the residuals are heteroscedastic, they may be small with the small values of a variable but large with the large values of a variable. In other words, the difference between the predicted and the observed value depends on the original variable values. This is not tolerated. See Homoscedasticity.
Homoscedasticity	A desirable characteristic of the residuals' variance faced in many multivariate analysis methods. The basic assumption is homoscedasticity, that is, that the residuals are evenly distributed. When the residuals are homoscedastic, small or large variable values can produce small or large residual values. The difference between the predicted and observed value does not depend on the original variable values. See Heteroscedasticity.
Latent variable	An unobserved, hypothetic, or hidden variable that is often measured through the observed variables. A fundamental construct in the test theory and factor analysis.
Linear combination	An expression constructed from a set of terms by multiplying each term by a constant and adding the results.
Linear relation	A positive correlation between two variables; means that when the value of the one variable increases so does the value of the other variable. A nonlinear relation is a curvilinear relation.
Matrix extraction	The matrix is divided into smaller matrices for certain computations, for example, to find the eigenvalues. There are many methods for matrix extraction.
Matrix operations	In multivariate analysis methods, the data is usually presented in a form of a matrix. Statistical software manipulates matrices with an aid of a matrix interpreter: the matrices are divided, multiplied, rotated, and saved. These all are matrix operations. See also Appendix 2.
Multicollinearity	Undesirable character of the data where the dependent variables are highly intercorrelated. This might happen if one triangulates the same thing in several, slightly different, ways or use a slightly different measure with two or more variables. Only one of the multicollinear variables is informative, the others reproduce the same information. See also Singularity.
Multi-Normality	All the variables included in the analysis and their linear combinations, are distributed Normally.

Multivariate	Involving more than one variable at the same time. For example, a multivariate outlier is a deviant case with several variables.
Multivariate methods	In multivariate analysis, several variables are handled at the same time.
Nonlinear relation	A curvilinear relation. See linear relation.
Notation	A way of presenting the concepts as symbols. In LISREL-analysis, for example, there is an exceptional way to symbolize the variables and connection between them.
Outlier	A deviant observation, "to lie out of the other cases" as the name suggests.
Parameter	From Ancient Greek: "side measure." A characteristic, a feature, a measurable factor that can help in defining a particular system in statistical analyses. Parameters are the moving particles of the model, that is, values depending on the data, the amount of variables, model, and selected variables.
Residual	The unexplained variation in the model. In technical terms, a residual is either the difference or the square between the predicted and the observed value.
Robust	Stable. A method is stable if it produces a correct result even though its assumptions could be violated. For example, there might be an assumption of Normality for all the variables. A robust method can still compute the correct (or very close) result, even though some of the variables would not be Normally distributed.
Singularity	One or more of the variables in a matrix is a combination of several other variables. For example, a compound (summed) variable induces singularity. The lack of variance in a variable also induces singularity; all the observations have the same value in some variable.
Syntax	A coding language. Software executes commands created by syntaxes. In some software, such as SPSS, the syntax is hidden behind the icons, but in reality the software only obeys the commands that are written in a certain way. In some cases, such as computing a canonical correlation in SPSS, the user has to write the syntax directly.
Transformation of a variable	Transforming a variable into another by computing. For example, the square root of a variable can be taken—this is called square root transformation—or a constant can be added to it. The basic idea behind the variable transformation is that all the individual values of the variable will be changed. Transformation of a variable is performed, for example, when the variable is not Normally distributed.

CHAPTER 7

Multivariate Analysis in Human Sciences

Goals of the Chapter

1. To orientate analyzing the data with multivariate methods.
2. To obtain an overview of the characteristics of multivariate analysis.
3. To comprehend the importance of a preliminary examination of the data for analysis.
4. To understand the meaning of correlation coefficient and its anomalies in the multivariate methods.
5. To deepen the understanding about the concepts of "outlier," "multicollinearity," and "singularity," and to acknowledge their effects on the results of multivariate analysis.

7.1 A Brief History of Multivariate Analysis

Multivariate analysis has a long tradition in the statistical science. Sir Francis Galton (1822–1911) created the concepts of *standard deviation* (Galton, 1886) and *correlation* with the same meaning as we know it today (Galton, 1888, 1889) and developed the basics of *Regression Analysis* (*Regression toward Mediocrity,* Galton, 1885, 1886) on the basis of the discoveries of Adrien-Marie Legendre (1805) and Carolo Friderico Gauss (1809, 1821). Galton himself was not a mathematician and hence he left the refinement of his techniques to talented mathematicians like Karl Pearson.

• before 1900: Galton

A British mathematician Karl Pearson (1857–1936) introduced the *Chi-squared test* and *p-value* as the basis for statistical inference (Pearson, 1900a) and *Phi-coefficient* for cross-tabulations (Pearson, 1900b), developed the concepts of correlation, or longer, *Pearson's* Product-moment *correlation* coefficient (Pearson, 1896), and developed the *Regression Modeling* following Galton and demonstrated the connection between the correlation and regression (Pearson, 1896). He also initiated the *principal component analysis* (Pearson, 1901).

• 1900: Pearson

• 1910:
Spearman

At the beginning of the 20th century, another pioneer, an English psychologist Charles Spearman (1863–1945) developed *Spearman Rank correlation* (Spearman, 1904) and *correction for attenuation* (Spearman, 1904, 1907) and introduced the first ideas of *factor analysis* (Spearman, 1904), coined the term, and applied the method extensively in his studies on cognitive performance. Later, during the 1930s the method was further developed—and forgotten, see Mulaik (1987)—in the USA by Louis Leon Thurstone (1937, 1947) and his students. Louis Guttman (1954, 1955) raised the topic of factor analysis again and from 1969 the method has become one of the most used methods in human sciences (see Figure 8.1).

• 1920s: Fisher

From the 1920s to 1935, a British statistician Sir Ronald Aylmer Fisher was well-known, for example, for *F-distribution*, *Analysis of Variance*, *Fisher information*, and *Fisher's exact test*. He is "*the architect of multivariate analysis*" as expressed by Rao (1964), and contributed, among other things, to the *maximum likelihood estimation* by giving its name and promoting it.[1] Later on, it has been used, for example, in Structural Equation Modeling.

• 1930s:
Fisher
Wilks
Hotelling
Mahalanobis

In the 1930s, Harold Hotelling,[2] inspired by the works of Fisher, generalized Student's *t*-distribution to multivariate settings and thus created *Hotelling's T^2-distribution* (Hotelling, 1931; 1951) which is used in *discriminant analysis*. He contributed to the development of *principal component analysis* (1933), and created

[1] A Danish historian Anders Hald (1998, 738) describes Fisher as "a genius who almost single-handedly created the foundation for modern statistical science without detailed study of his predecessors." Rao (1992) calls him—not only "*the architect of multivariate analysis*" (1964), but also "*the founder of modern statics.*" It is worth noting that Fisher popularized statistical methods in his early book *Statistical Methods for Research Workers* (Fisher, 1925a) which has influenced several generations of researchers and went through many editions. His later book *Design of Experiment* (Fisher, 1935a) made the similar contribution in the area of experimental studies. Hence, not only that Fisher developed substantially the modern statistical thinking, he also educated several generations through his books. Thus it may not be overrating to say that Sir Fisher is the father of modern science. An interested reader finds many interesting readings of Fisher, for example, at http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/rafframe.htm (accessed on May 7, 2016).

[2] From the statistical history viewpoint, Hotelling's role is remarkable not only due to his own innovations, but also because he educated and supported dozens of future statisticians, such as Abrahan Wald (renowed from *Wald test* statistic), Henry Mann (*Mann–Whithey U-test*), W. Allen Wallis (*Kruskal–Wallis test*), Jacob Wolfowitz (*Wald–Wolfowitz Runs test*) in Columbia University and S. N. Roy (*Roy's Largest Root Criteria*), M. Friedman (*Friedman test*), W. H. Kruskal (*Kruskal–Wallis test*), D. N. Nanda (*Bartlett–Nanda–Pillai trace*), and I. Olkin (*Kaiser–Maier–Olkin test*) at University of North Carolina (see details in Neyman (1960)).

CHAPTER 8

Family of Factor Analysis

Goals of the Chapter

1. To know the research questions when PCA, EFA, and CFA are relevant.
2. To understand the main differences between the methods.
3. To be able to use PCA, EFA, and CFA when relevant.
4. To deepen the skills and knowledge of SPSS in the area of factor analysis.

8.1 Introduction

Historical overview of the family of Factor Analysis.

Factor analysis is one of the oldest multivariate methods and it is still widely used. Charles Spearman is considered to be the father of the method. He played a crucial role in the development of the method at the beginning of the 20th century and presented the one-factor solution in 1904. However, prior to this, in 1901, Karl Pearson published an article on the principal axis method. A major development came in the 1930s when Thurstone developed the multiple factor model. Until the 1960s, it was remarkably difficult to perform the computations related to factor analysis—even the computation of the correlation matrix was a time-consuming process.

In the last ten years, there has been a phenomenal increase in interest in factor analysis family, especially in educational research. Figure 8.1 shows that in ERIC database, which contains more than a million scientific papers, the hits of the keyword "factor analysis" were decreasing from 1973 to 2003, but since that threshold year, the writings on *exploratory factor analysis* (EFA) as well as on such newcomers as *confirmatory factor analysis* (CFA) and *Structural Equation Modeling* (SEM) have been expanded. It could be that the new methods opened new perspectives for the old ones.

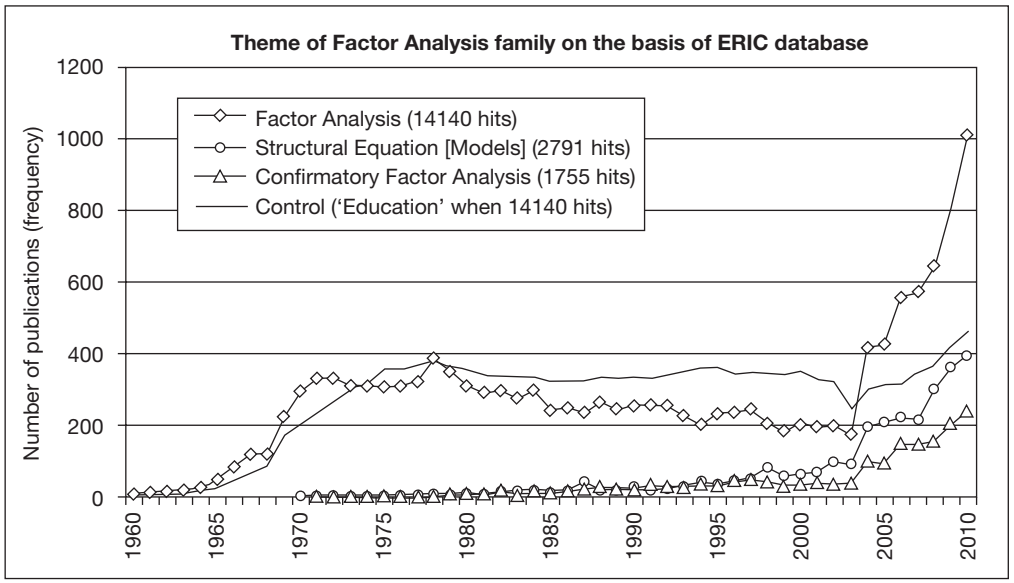


Figure 8.1 Historical Trends of Factor Analysis Family in ERIC Database

Some relevant literature on factor analysis

Even though the method is old there is relevant recent literature. In particular, a notable body of publications has appeared on SEM analysis, which has significantly developed since the 1970s. This is mostly due to Karl G. Jöreskog, a Swedish researcher who published a set of seminal articles on maximum likelihood estimation in CFA at the end of the 1960s (see Chapter 8). The classic sources on SEM analysis include Jöreskog and Sörbom (1988; 1993), Bentler (1985; 1995), and Bollen (1989a). Older classics, Harman (1976) and Gorsuch (1983), are also considered the basic reading on factor analysis. Practical information on all the methods discussed here can be found in the book by Tabachnick and Fidell (2006). Some other recent publications in the area are Bartholomew, Knot and Moustaki (2011), Brown (2006), Child (2006), Kline (2000), Loehling (2003), Mulaik (2009), Thompson (2004), and Walkey and Welch (2010).

8.1.1 Variations

Where to use FA:
1. To sum up the variation in a set of variables as a few central principal components or factors.

The family of factor analysis has three main areas of application. First, it is frequently used to locate in a large set the variables that strongly correlate with each another and form a coherent entity. Several coherent entities can be detected in a large set of variables. principal component analysis (PCA) and EFA are used to sum up the variation in many variables in a few central principal components or factors. In the analysis, there can be dozens of variables that have to be measured on, at least, a good ordinal scale (such as the

Likert scale).[1] Although PCA and EFA are mathematically and philosophically different analysis methods, based on their purpose, they belong to the same family of analysis methods. Both methods are based on the premise of finding, in a large set of variables, common variations, and forming interpretable subsets out of the variables to reduce the number of variables. Both analysis methods can be considered "black box experiments"; variables are entered into a black box (the computer), and then the output is interpreted.

The purpose is to reduce the number of variables.

Second, factor analysis can be used to confirm the fit of the given theoretical structure to the real life data. Therefore, in addition to PCA and EFA, CFA is included in factor analysis family. CFA is sometimes identified with SEM[2]; the designations are flexible. The difference between EFA and CFA is that EFA is used to *search* a model among the combinations of variables, that is, the analysis is performed in an exploratory mode to investigate the structure of the correlation matrix. In CFA, an existing model is investigated to *confirm* that the data is compatible with the model.

2. In examination of an existing model and to answer the question: "Is the model compatible with the data at hand?"

Third, both PCA and EFA are well suited for constructing measurement instruments and EFA for estimating the measurement error, or the reliability, of the measurement instrument. While it is not always viable to compute a direct sum of a large set of variables, it is possible to locate among the variables three or four reliable "subinstruments" to measure the phenomenon wider than by using only one summed score. This characteristic of factor analysis is not addressed in this section. One can find further discussion in Section II of Volume 1 and Chapter 14 therein, and a practical example in Section II of this volume and Chapter 3 therein.

3. To create a measurement instrument.

8.1.2 Factor Analysis and Software Packages

Factor analysis is rarely performed manually because of the complexity of computation. Hence, one needs a statistical software package for computations. All main statistical packages include algorithms for solving the computational challenges of EFA and PCA. Within this Section, SPSS software is used in examples. It can be slightly confusing that when you perform factor analysis in SPSS, the default option carries out PCA. There is nothing wrong with this since the background assumptions of PCA are not as strict as those of EFA. However, sometimes a student in his/her

In SPSS, one must pay attention to the choice of the analysis method

[1] Of the type of scales, see Section I of Volume 1 and Chapter 7 therein, and Section II of Volume 1 and Chapter 11 therein.

[2] These designations seems to be flexible. Some consider SEM to be a subcategory of CFA, while others consider SEM to be the parent category that includes both CFA and path models. In this section, SEM analysis is discussed in terms of LISREL software package. In Section IV of Volume 3, the topic is expanded and addressed in terms of AMOS software package.

doctoral dissertation or even a professional researcher mistakenly states that she/he performed factor analysis. Then in their manuscripts, the wrong terms such as "factor" or "loadings on factors" are used instead of "principal component" and "loadings on principal components."

When it comes to CFA, the basic module of SPSS software does not enable SEM analysis. One needs to get specific software for SEM analysis. The first was LISREL software.[3] Later, such software packages as EQS (Bentler, 1995), MPLUS[4] (Muthén & Muthén, 1998–2007), and AMOS[5] (Arbuckle, 2011) were created to analyze the models. In addition, SAS software has its own procedure (SAS CALIS) of modeling. In this Section, LISREL software is used. One can easily download a free student version online.[6] In Section IV of Volume 3, AMOS software is introduced.

Software for SEM/CFA

8.1.3 Common Elements and Discrepancies of PCA and EFA

Stages of PCA and EFA analysis:
1. Correlation matrix
2. Estimating the loadings
3. Rotation
4. Calculating the factor or principal component scores
5. Naming of the dimensions

Both PCA and EFA are carried out in four stages. First, the correlation or covariance matrix of the original variables is computed. Second, the matrix thus generated is used in estimating the principal component or factor loadings. Third, the loadings are rotated; in rotation, the original axes (see Figure 8.2) are transformed in order to make the loadings easier to interpret. Fourth, and optionally, factor or component scores can be computed for each observation. What is common to both PCA and EFA is that both analysis methods involve forming linear combinations based on the original variables; these linear combinations are then called either principal components or factors. Finally, the principal components and

[3] The current version is LISREL 8.8 (Jöreskog & Sörbom, 1999, 2002a, 2002b; Jöreskog et al., 2003).

[4] Petri Nokelainen and Markus Mattsson suggest Bengt Muthén's MPLUS software for the analysis of variables in nominal and ordinal scales. This is valuable in the context of human sciences for variables measured with a Likert scale. Note that a Likert type of scale is an *ordinal* scale while many analyzes require at least an *interval* type of scale for the measurement (see Section I of Volume 1 and Chapter 7 therein). The current version is MPLUS 6.2. At http://www.statmodel.com/references.shtml (accessed on May 7, 2016), one can find a large amount of relevant literature on the software and SEM.

[5] In literature, the abbreviation Amos (*analysis of moment structures*) is used for the software. However, here the capital letters are used in abbreviation (AMOS) for uniformity with other software in the book. The current (Summer 2016) version is SPSS AMOS 23—the numbering system follows SPSS system from AMOS 7 to AMOS 16 (Arbuckle, 2011, 8).

[6] http://www.ssicentral.com/lisrel/down-loads.html (accessed on May 7, 2016).

factors are usually named on the basis of the content of the variables having high loadings in the component or factor.

The fundamental differences between PCA and EFA can be summarized as follows: First, in PCA the whole intervariable variance is analyzed, whereas in EFA, only a part of the variance is analyzed and the rest is considered error. Thus, EFA is suitable for estimating reliability whereas PCA is not. Second, PCA mechanically computes the relationships between the variables; the variables produce the components. EFA, on its behalf, is based on the idea of an underlying existing theory: the theory influences the relationships between the variables and the (underlying) factors produce the variables. Third, the practical difference is that in PCA, the components are "unambiguous," irrespective of how many components are formed, the first components and their loadings are equal. In EFA, the factor loadings are dependent on the number of factors included in the analysis. It has been suggested (Tabachnick & Fidell, 2006) that PCA is best suited for finding the structure of a correlation matrix; EFA is best when the researcher has an inkling of the underlining theory.

The fundamental differences between PCA and EFA

PCA can be used in investigating composition of the data

EFA is used when a possible composition is known

Because PCA and EFA are technically similar methods, the same text is included in the Chapters that deal with them so that the reader does not have to go back and forth between the chapters. Tabachnick and Fidell, for example, solved the issue by reviewing the theoretical assumptions of PCA and EFA together.

8.2 Principal Component Analysis (PCA)

When the aim is to categorize a large number of variables into few sets to reduce the heterogeneity of the phenomenon under investigation, principal component analysis as well as the traditional exploratory factor analysis (see section 8.3) suits well. The aim is to find something in common among many variables—something that connects variables in a manner congruent with both theory and practice.

8.2.1 Suitability

PCA can be applied in various situations. The objectives of the analysis can be (a) to summarize the correlations between the variables; (b) to represent a large set of variables with a few principal components; (c) to operationalize the process behind the observed variables; and (d) to test a theory of the nature of the underlying process. PCA is especially well suited to a situation when the researcher does *not* have an idea on the kind of theory that might connect the variables under investigation (cl. EFA in section 8.3.2).

PCA can be used when the objective is to
- summarize information
- reduce the number of variables
- frame a process as a mathematical theory
- test the nature of a process

CHAPTER 9

Family of Regression Analysis

Goals of the Chapter

1. To know the research questions when RA, LRA, and CCA are relevant.
2. To understand the main differences between the methods.
3. To be able to use RA, LRA, and CCA when relevant.
4. To get familiar with the commonality analysis.
5. To deepen the skills and knowledge of SPSS in the area of regression analysis.

9.1 Introduction to Regression Analysis

A historical note

Regression analysis (RA), together with its different variations, must be one of the most central and most popular analysis methods of modeling. Like factor analysis, RA is one of the oldest multivariate methods that are still popular. Although the method had been used at the beginning of the 19th century by Legendre (1805) and Gauss (1809) in astronomy,[1] Sir Francis Galton is credited with the "christening" of the method. In the late 19th century, he studied the heights of father–son pairs and observed that tall fathers' sons did not, on average, end up quite as tall as their fathers. Respectively, he noted that the sons of shorter-than-average fathers ended up slightly taller than their fathers. When the sons' height was explained by the fathers' height, it was noted that the sons' average height tended to be closer to the population mean than to either of the extremes. Galton originally dubbed the phenomenon *regression*

[1] Originally, Legendre published *Sur la Méthode des moindres quarrés* as an appendix of his book of astronomical studies (Legendre, 1805). Gauss (1809) also used the method to calculate the orbits of the planets.

toward mediocrity but later called it *regression toward mean*—thus, the name of the method—"regression analysis." The theoretical foundation of logistic RA had been also established in the 19th century by Alphonse Quetelet and his student Pierre François Verhults (see the history in Cramer, 2002). Canonical correlation analysis (CCA) was introduced by Harold Hotelling (1936).

Even though its mathematical foundation remains the same, the method is presented in every advanced and intermediate textbook on econometrics, biometrics, or psychometrics—this book is no exception. For example, Tabachnick, and Fidell (2006) spend roughly 170 pages on RA and its cognates. Hair et al. (2009) also dedicate more than 100 pages for RA. Specific recent publications in RA are, for example, Hill and Gelman (2006), Kleinbaum et al. (2007), Mendenhall and Sincich (2011), Montgomery, Peck, and Vining (2007), and Price and Chatterjee (2006). The relevant materials of Garson (2012b) are available online.[2]

9.1.1 Variations

RA is used for finding the variables that can explain variation in a variable of interest

- RA: When one or more variable(s) are used to explain variation in one continuous variable
- LRA: When the dependent variable is categorical
- CCA: When there are multiple independent and dependent variables

By now, the reader is familiar with the concept of the correlation coefficient: it expresses the strength of a relationship between two variables.[3] The square of the correlation coefficient, the coefficient of determination r^2, expresses the proportion of variance in one variable that is explained by the other variable. In RA, there can be multiple predictor variables. In that case, R^2 (squared multiple correlation, SMC) expresses the proportion of variation in the explained variable that the group of variables explains. The explained variable is called a *dependent variable* (*DV*), a *response variable* or a *regressand;* and the explanatory variables are called *independent variables* (*IVs*), *predictor variables*, *controlled variables*, *factors*, or *regressors*. A common research question is: "Which independent variables predict (and in what way) the observed variation in a given dependent variable?" Usually, the dependent variable is a continuous variable measured on an interval scale. If several variables are used to explain the variation in a continuous dependent variable, the method of choice is multivariate RA. When the dependent variable is dichotomously categorical (e.g., content/non-content individuals or dropouts/non-dropouts), LRA is used. If there are several dependent and independent variables, the procedure of choice is called CCA. These three classical members of the "family" are introduced in this chapter. *General linear modeling* (GLM) is one variation of RA, and it is introduced along with the

[2] http://www.statisticalassociates.com/regression.pdf (accessed on May 7, 2016).

[3] Correlation coefficients are discussed in detail in Section VI of Volume 1and Chapter 33 therein.

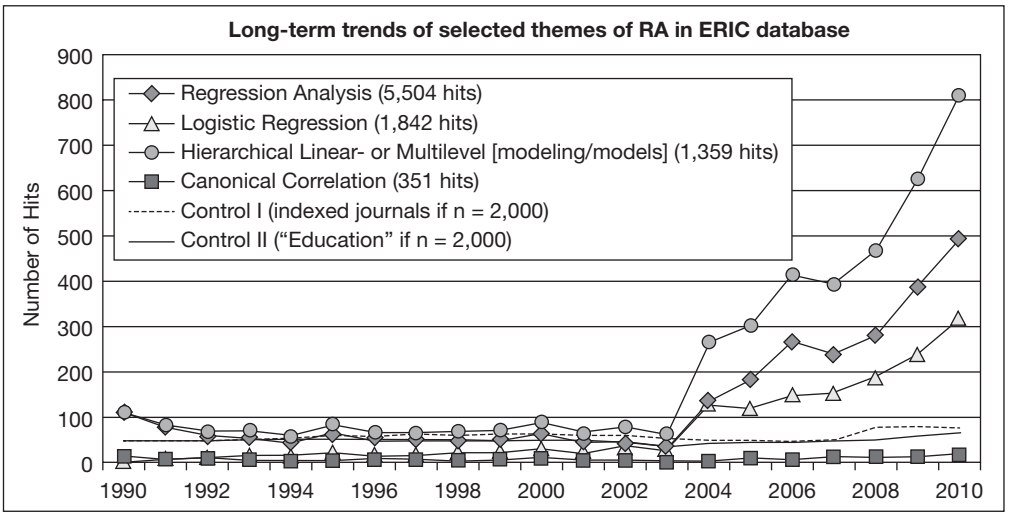


Figure 9.1 Historical Trends of RA Family in ERIC Database

analysis of variance in Chapter 10. Another is *multilevel modeling* (MLM) or *hierarchical linear modeling* (HLM), used to analyze clustered datasets such as datasets collected in schools; it is discussed in detail in Section III of Volume 3. The RA family can include *path analysis* (also known as *dependence analysis*) that combines several regression analyses. However, in its modern form, path analysis is not considered to be in RA family; the procedure is now a part of structural equation modeling (SEM). SEM analysis is introduced as a part factor analysis family (see section 8.4). Path analysis is mainly discussed in Section IV of Volume 3. Yet another variation is *Cox Regression*, introduced in Section V of Volume 3, when the researcher knows that some of the cases were dropped from the data. It may be good for the reader, at least, to be aware what family members are *not* included here; the family of RA is quite large. This book does not include such variation of RA as curve estimation, partial least-squares estimation, multinomial LRA, ordinal RA, probit analysis, nonlinear regression, weight estimation, or two-stage least-squares estimation, all of which are available in SPSS. One may know from the Introduction in Volume 1 and Figure I.3 therein that the historical trend of RA applications was in decline from the 1970s to 2003, when its popularity had rocketed. The same phenomenon is observed with logistic regression and hierarchical linear modeling (Figure 9.1). CCA differs from the others; it has deep roots and low but steady numbers of publications over the years.

Other family members within this book:
- GLM
- Multilevel modeling when the data is clustered
- Path modeling when willing to combine several regression analyses
- Cox Regression when having missing values and willing to utilize the information

Other variations:
- Curve estimation
- Partial least squares
- Multinomial LRA
- Ordinal RA
- Probit analysis
- Non-Linear RA
- Weight estimation,
- 2-stage least squares

9.1.2 Basic Procedure in RA

RA is a basic method in many disciplines

Stages in RA:
1. Choose variables
2. Calculate coefficients
3. Perform regression diagnostics

In an *exploratory* (i.e., investigative) RA, multiple variables are included as predictors with the intent of determining which variables are essential from the point of view of the phenomenon. However, it would be more sophisticated to include only theoretically *relevant* variables in the analysis; this could be called *confirmatory* RA.[4] As a method, RA consists of three stages. At the first stage, the variables to analyze are chosen; one or more as the explained variable (*Dependent variable, DV*) and the others as predictors (*Independent variable*, *IV*). At the second stage, RA itself is performed. At the third stage, the values of the so-called regression diagnostics are computed for the obtained model. Regression modeling involves several assumptions that are discussed in detail in this chapter. The central assumption is that the part of the model that remains unexplained, that is, the *residuals* related to individual observations, are Normally distributed and their variance is constant, that is, *homoscedastic* (see the figures in section 9.2.4.6). At the same time, it is customary to check that the dataset contains no *outliers.* Anomalous observations tend to have a large effect on the value of the correlation coefficient. RA is also sensitive to too high correlations among the predictors, that is, *multicollinearity* and to *singularity* (for details, see section 7.2.4).The fundamental starting point for the analysis is that the predictor variables correlate with the factor but not necessarily with one another. The higher the common variation (i.e., correlation) is among the predictors, the larger bias is introduced to the model if an incorrect method is chosen (multicollinearity is addressed later in this chapter; see also section 7.2.4). Hence, *commonality analysis* is introduced in sections 9.2.4.6 and 9.3.4.6, to advance the process; the common explanatory power is partitioned so that the real effect of the individual predictors can be observed.

9.2 Traditional Linear Regression Analysis (RA)

The traditional RA is a basic analysis method in scientific research. It involves explaining the variation in one—usually continuous—variable by several predictor variables. RA is suitable for *exploring* for variables that are relevant to the phenomenon, or when *confirming* the effects of theoretically interesting variables. A regression model can be used to model reality in a mathematical manner and, thus, to control the phenomenon through the use of mathematical

[4] However, I have not seen this concept used.

tools. Therefore, RA can be used for modeling a phenomenon. On the other hand, regression modeling is also suitable for predicting the values of new observations that are not a part of the original data. Consequently, if, say, admission test scores have been successfully used in modeling one's study performance, they can be used in predicting study performance in the age classes to come.

RA is suitable for
- modeling a phenomenon
- confirming existing model
- predicting the observations

9.2.1 Suitability

As mentioned above, RA is, in principle, suitable for two situations. On the one hand, it can be used for finding, among a large group of variables, the factors that are able to explain the variation in a continuous variable. For instance, admission test scores, motivational levels, studying techniques, or teaching styles could be used in accounting for the variation in study performance; or, to take another example, customer or work satisfaction can be explained, say, by personnel's attitude toward customer service, their degree of expertise, etc. When used in this manner, RA is being used in an *exploratory* mode, as a kind of "black box experiment": after pouring variables into the analysis engine, we start up the machine and wait for the results. On the other hand, a more sophisticated way of running RA involves investigating the relative importance of various predictor variables that are *apriori* known to be important from the point of view of the phenomenon investigated. Such factors might include explaining life expectancy by certain factors that are *a priori* known to be reliable predictors, such as women's literacy, child mortality, or, on the other hand, by the degree of cancer mutations, the years of smoking. When used in this manner, RA is being used in an *confirmatory* mode. In addition to these two basic modes of operation, RA can also be used in *testing whether certain variables are better predictors than others*. The analysis can be further refined through regressing a variable in a group of variables with simultaneously eliminating the effect of (a) certain other variable(s).[5] Several regression models can also be combined; this happens in path modeling which is discussed in Section IV of Volume 3.

[5] This idea is implemented in an analysis method called the ANCOVA; the method is discussed in more detail in the context of ANOVA Section 10.3. During the years, the methods used in calculating ANOVA have become more similar to the basic idea of RA: *General Linear Modeling, GLM*, is a procedure that is used in the next chapter to calculate MANOVA, ANCOVA, and MANCOVA.

9.2.2 Limitations and Assumptions

Assumptions in RA:

- Variables are relevant predictors
- Reasonably high amount of oservations for each variable

An interesting problem the researcher faces when carrying out RA is that there may be no theoretical grounds for choosing a particular set of variables as predictors over another set. In an RA—and other multivariate methods—the software package has no other information at its disposal than what the user inputs. In case irrelevant variables are included in the analysis, the results will be more or less uncertain or unstable. In RA, as with other multivariate methods, the number of observations should be reasonably high in relation to the number of variables included in the analysis. In case too few observations in relation to the number of variables are included, the model's coefficient of determination is inflated due to technical reasons. Intuitively, this is because each observation comes to have its own predictor, and the phenomenon is over fitted. Green (1991) has suggested that if predictors are otherwise exemplary and correlate moderately with the dependent variable, an adequate sample size might be 50 + 8 times the number of predictors when estimating the value of the multiple correlation coefficient and 104+ the number of predictors when testing the predictors. Thus, if there were five predictors, the minimum sample size would be 50 + 40 = 90 or 104 + 5 = 110. Of these two, whichever is greater will be chosen. When using traditional statistical regression procedures, which are covered in section 9.2.3, the sample size needs to be greater than these stated sizes. In this case, Tabachnick and Fidell (2001) recommend using the ratio of 40 by 1 as a rule of thumb, that is, 40 observations per each explanatory variable and even more if the data will be used to run a cross-validation check on the applicability of the model.

- Observations should be independent

In the traditional RA, the observations should be independent. Practically speaking, RA should not be used in longitudinal designs with dependent variables which are usually analyzed by using repeated measures.[6] The pretest-posttest design with dependent samples is traditionally analyzed by using the analysis of covariance (ANCOVA),[7] a specific type of RA. One needs to be careful also with the nested or clustered samples, such as students in schools types of samples, without using multilevel modeling (MML),[8] a specific type of RA, the estimations tend to be biased.

- No inflated correlations between the predictors

When running a traditional RA, it is assumed that explanatory variables correlate reasonably strongly with the dependent variable, but not too strongly with each other. Without sufficiently high correlations, it is impossible to form proper models. On the other hand, inflated correlations among the predictor variables result in

[6] See Chapter 10 (10.5), Section II, of Volume 3.
[7] See Section 10.3 of Volume 2.
[8] See Section III of Volume 3.

CHAPTER 10

Family of Analysis of Variance

Goals of the Chapter

1. To know the research questions when ANOVA, MANOVA, and ANCOVA are relevant.
2. To understand the main differences between the methods.
3. To be able to use ANOVA, ANCOVA, and MANOVA when relevant.
4. To deepen the skills and knowledge of SPSS in the area of analysis of variance.

10.1 Introduction

The family of analysis of variance (ANOVA) is generally used to investigate whether the group means differ in a statistically significant manner. The central concept is to compare differences in several groups' means while considering the size of error related to each mean. We can quantify the magnitude of the error (standard error of mean, S.E.M.) directly by the variance of the variable. The standard error is expressed as $s/\sqrt{n}$, where s is the standard deviation and n is the sample size.[1] The variation in the data (described by the variance) can be decomposed into separate components. Hence, the analysis methods are called "analysis of variance."

10.1.1 A Historical Note

According to Stephen Stigler (2000, 154), a historian, the original idea of ANOVA is credited to Pierre-Simon Laplace, who studied the universe as far back as in the 1780s. The idea was formalized as a concept of *least-squares* by Adrien-Marie Legendre (1805) and Carolo Friderico Gauss (1809; 1821) in their astronomical studies. Later, the method was applied in social sciences by Adolphe

[1] For details, consult Section VI in Volume 1 and Chapter 35 therein.

Short history:
1. 1800: Least squares
2. 1885: Analysis of Variance Components

3. 1908: *t*-distribution and *t*-test
4. 1918: ANOVA
5. 1925: ANCOVA
6. 1925: *t*-test for the same variances
7. 1935: *t*-test for different variances

8. 1950s–1960s: GLM

Quetelet (1827) and Francis Edgesworth (1883) and developed as *ANOVA components* by Edgesworth (1885).[2] However, ANOVA, as it is known today, was formalized and popularized by Sir R. A. Fisher (1918; 1921; 1925a; 1935a). In addition, he proposed analysis of covariance (ANCOVA; Fisher 1925a, 1935a). William S. Gosset[3]—better known by his pen name "Student"—developed, with assistance from Karl Pearson, the t-distribution for the small samples (Student, 1908a; 1908b; 1909). This distribution was found preferable and hence Fisher (1925b) demonstrated that it can be used for comparing two means (*Fisher's t-test*[4], or just the *t-test*) as well as for testing whether the slope of the regression line is zero. Later, Fisher (1935b) proposed another *t*-test on the basis of W. V. Behrens' (1929) work for unequal variances;[5] nowadays it is called *Behrens–Fisher t-test* or just the *t*-test for unequal variances. It was weak in hypothesis testing and so B. L. Welch (1938; 1947) proposed a correction; this is now known as *Welch's t-test* or Welch's degrees of freedom (*df*) for *t*-test.

General linear modeling (GLM), the common framework for regression analysis, ANOVA, and ANCOVA, has evolved softly but

[2] Stigler (2002, pp. 109–110) reminds that at the same conference of British Association of Advancement of Science where Sir Galton presented the famous idea of regression analysis (*Regression toward mediocrity*), two days later Francis Edgeworth presented the basic idea of ANOVA.

[3] Gosset was not a mathematician in an ordinary sense; he had a degree in chemistry and statistics and worked at the Guinness brewery. Gosset noticed that the Normal theory promoted by Fisher did not work when the sample sizes were small and hence he developed the *t*-distribution. For mathematical finesses, Gosset got help from Karl Pearson. Though Gosset himself thought he was just a "student" among two giants, Pearson and Fisher, he was highly appreciated by the latter, his friend and "the star." Fisher doubted that Gosset had fully realized his own contribution and wrote: "In spite of his many activities it is the 'Student' … who has won, and deserved to win, a unique place in the history of scientific method" (Fisher, 1939, 8). According to an anecdote, Gosset, a modest man, responded to the gratitudes: "oh, that's nothing—Fisher would have discovered it all anyway" (Cunliffe, 1976, p. 4). Note that the seminal article of the "Student" (1908) was almost totally ignored by the scientific community but Fisher addressed its topics two decades later (Aldrich, 2008).

[4] From early on, the *t*-test was attributed to Fisher (see Rulon, 1943) because the idea of *t*-test was developed by Fisher; Student showed that the *z*-test follows the new *t*-distribution rather than the Normal distribution when the sample size is small. However, nowadays the test is usually referred to as Student's *t*-test to honor Gosset's role in the development. Gosset also created the tables for assessing the statistical significance of the *t*-test (in Pearson, 1914; see Aldrich, 2008; Student, 1925).

[5] This modification of the *t*-test is one of the solutions for the Behrens-Fisher problem: the distribution of two independent means with different variances is actually unknown because it depends on the unknown variances. Several other solutions have also been proposed (see the review of Sawilowsky, 2002).

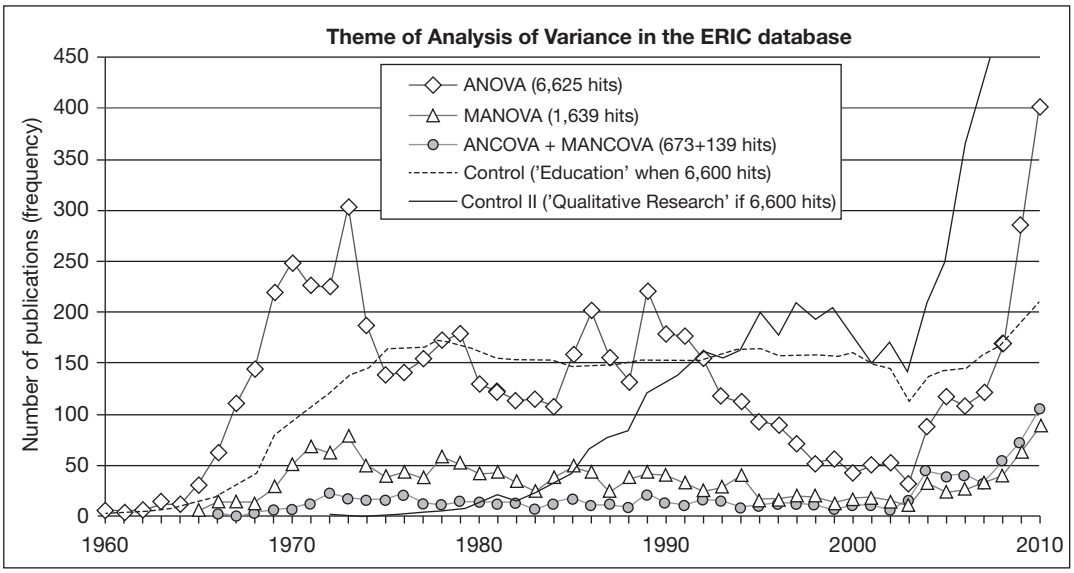


Figure 10.1 Historical Trends in the Theme of Analysis if Variance in the ERIC Database

unmistakably through the years. No one knows when the paradigm shift happened. However, in practice, GLM has completely superseded, for example, the multiway or multifactor ANOVA, which has become just a special case of regression analysis. Though the timelines are still unclear, we known that regression analysis and ANOVA/ANCOVA have been developed and used independent from each other for a long time (see "pocket history" of it in Rutherford, 2001, pp. 2–3); The concept "GENLIN" was first used in the ERIC database at 1968 (Bashaw & Findley, 1968) in reference to a conference held on 1965; the acronym GLM was used in 1974 (Pohlmann & McShane, 1974), although this cannot be the very first time. In any case, GLM is such a remarkable innovation that Trochim (2006) notes: "The GLM is one of the most important tools in the statistical analysis of data. It represents a major achievement in the advancement of social research in the twentieth century."

Figure 10.1 presents the historical trends of the methods discussed in this chapter. Based on the number of hits, ANOVA is the most popular method and the other ones are relatively rare. Another curious trend is that since 1973 the volume of the publications has been declining. Especially after 1989, the drop is remarkable in all publications as well as in indexed journals. It may be a coincidence, but it seems that the more publications with the keyword "Qualitative Research" appear, the less there are publications with the keywords "ANOVA" or "analysis of variance." It is noteworthy that, since 2003, the number of publications has soared at the same rate as the "'Qualitative research" ones. For some

reason it follows the same trend as the keywords "'Correlation" (see Figure I.5 in *Introduction* in Volume 1), "Regression Analysis" and "Factor Analysis" (see Figure I.3 in *Introduction, in Volume 1*).

10.1.2 Variations of ANOVA

One grouping variable and one dependent variable: → ANOVA
Several grouping variable and one dependent variable: →Multiway ANOVA

Methods for comparing two means, such as the *t*-test, are covered in other chapters of the book.[6] In this chapter, the discussion is expanded to comparing *more than two means*.[7] If the analysis involves only a single grouping variable, the method to be employed is the *one-way* ANOVA, which is not an actual multivariate method in itself. If the analysis involves two or more grouping variables, the analysis method is called a Two-, Three-, or Four- (etc.) way ANOVA—or, in general, a *multiway ANOVA*.[8] The question addressed in a one-way ANOVA might involve comparing job satisfaction in different occupational groups—assuming that there are more than two. A two-way ANOVA involves the question of whether there are statistically significant differences in job satisfaction in (a) different occupational groups between (b) the genders. These issues are addressed in section 10.2.

One (ore several) grouping variable, one (or several) controlling variables, and one dependent variable: →ANCOVA

Sometimes we need to control for a certain factor. Respondents with different compensation levels might differ in terms of job satisfaction. In this case, the compensation level might be controlled by including it in the analysis as a covariate. The ANCOVA could be used to investigate whether there are differences between the two genders in job satisfaction when the compensation levels are controlled. These analysis methods are usually discussed from the GLM point. In this case, the idea is that the grouping variable explains the variation in the dependent variable. These issues are discussed in section 10.3.

Several grouping variables and several explained variables → MANOVA, MANCOVA

Multivariate ANOVA (MANOVA) is used when there are several dependent and independent variables. If there are, in addition, one or several controlling variables, the method is *Multivariate* ANCOVA (MANCOVA). In Chapter 10, this method is used to investigate how a person's gender and a choice of hobby explain variation on the four dimensions revealed in *principal component analysis*. If there are, in addition, one or several controlling variables, the method is *Multivariate* ANCOVA (MANCOVA).

---

[6] See Section V in Volume 1 and Chapter 34 therein.

[7] ANOVA is discussed in detail also in Section II, in Volume 3 which serves as an introduction to ANOVA, as well as here. You find some similarities in the text of both sections.

[8] Therefore, the SPSS software includes two different ANOVA methods; in the this chapter ANOVA by GLM is used. Section II in Volume 3 contains an introduction to GLM.

One should not mind that ANOVA and ANCOVA are discussed here in the section related to multivariate analysis even though they are not actually *multivariate* methods.[9] For the sake of giving an overall picture of this family, they are justified in this section, too.

10.2 One- and Multiway Analysis of Variance (ANOVA)

One of the most common research questions in human sciences is whether the means differ between groups. When there are only two groups to compare, such as males and females, the *t*-test is traditionally run. When there are several groups to compare, the *F*-test is used. Comparing only one grouping variable at a time is the simplest case of ANOVA; usually the design of the experimental study is more complicated.[10] If the research design is thorough, even very small samples can be reliably analyzed.

Fisher originally developed ANOVA for agricultural studies. Hence, many design names (such as "plots," "split-plot," or "Latin Square") and other concepts (such as fixed-effect) come from there. Different experimental designs are discussed in detail in Section II in Volume 3 and are not repeated here. However, keep in mind that, traditionally, ANOVA had been used when, say, different amounts of fertilizer were added to different arable areas. Then, the question is, what amount of fertilizer produces the best results. This is a typical *fixed-effect model* when all the treatments (fertilizers) are included in the analysis. In human sciences, the simplest research question might involve finding out whether different districts/organizations/occupational groups/and so on differ in terms of the mean of a central variable. If *all* the possible groups of interest are involved in the analysis, the fixed-effect model is used. If, however, there are many groups available but only some of them are included in the analysis—for example, if 20 out of 100 schools or six out of ten districts and the researcher is willing to generalize the results into the larger population—the design is called *random-effect model*: only a randomly chosen part of all the possible groups is included in the analysis. When there are both the fixed effect(s) and random effect(s), it is called the *mixed-effect*

Different models:
- fixed-effect model

- random-effect model
- mixed-effect model

[9] I wish to thank professor Esko Leskinen for these remarks—as well as for his well-informed comments overall—for he, perhaps unintentionally, steered the whole Chapter on ANOVA in the right direction.

[10] Although the one-way ANOVA is not a multivariate method, we discuss it in this chapter. One finds a more explicit description of the different variations of ANOVA in Section II in Volume 3 and Chapter 10 therein.

model. The computations are mainly similar in each model but the interpretations of the results differ.

10.2.1 Suitability

ANOVA examines whether the group means are equal…
… or what kind of an effect the grouping variables have on the explained variable

The one-way ANOVA is used to investigate whether the means of more than two groups differ from each other. This can be also expressed from the viewpoint of regression analysis: "What is the *effect* of the grouping variable on the dependent variable?" ANOVA tests the hypothesis that *the means are equal in the different groups of a certain variable*.[11] The multiway ANOVA—that is, two-way, three-way, and so on—is used to investigate the simultaneous effects of two or more factors on a single dependent variable. A simple example of a two-way analysis would be comparing job contentment in different occupational groups between men and women. From regression analysis and GLM point, the question involves investigating the effect of occupational group and gender on job satisfaction.

10.2.2 Limitations and Assumptions

The general form of ANOVA has three basic assumptions: (a) cases are independent of each other, (b) the populations related to each of the groups are distributed (sufficiently) Normally, and (c) variances in all groups are equal.

Assumptions of ANOVA:

- independence of observations

The *independence of the observations* is a basic prerequisite and it is achieved when the design and sampling are performed adequately. When the cases are intentionally non-independent, it is called *repeated measures ANOVA*—the same cases are measured several times.[12] Another type of dependent situation, typical in school research, is the so-called clustered or nested sample: the students in the same school tend to be more similar than totally randomly drawn sample. These kinds of samples are analysed usually by using *multilevel modeling* (see Section III in Volume 3).

- Normality in the population and in the groups

The *Normality* of the population and in the groups may prove more challenging than the independency of the cases. Especially, if the sample size is small, the claim of Normality may be at risk. In this case, it is best to use nonparametric methods such as the Kruskal–Wallis test.[13] It is advisable to perform a *visual inspection*

[11] For further discussion on hypothesis testing, see Section VI in Volume 1 and Chapter 39 therein.

[12] Repeated measures design is discussed in Section II in Volume 3 and Chapter 10 therein.

[13] In many cases, the nonparametric methods are preferable to the parametric methods because the latter involve distributional assumptions, though they are rarely fully realized. The differences between parametric and non-parametric

CHAPTER 11

Family of Cluster and Classification Analysis

Goals of the Chapter

1. To know the research questions when Discriminant analysis, Cluster analysis, and Decision tree analysis are relevant.
2. To understand the main differences between the methods.
3. To be able to use DA, CA, and DTA when relevant.
4. To deepen the skills and knowledge of SPSS in the area of cluster- and classification analyses.

11.1 Introduction

In this chapter, three different families of methods for clustering or classifying the observations are compiled into one. In research, sometimes the basis for the case classification is *known in advance*: for example, some of the observations are from staff and some are from managers, or some belong to one hobby group and the rest to other groups. Based on other variables, it might be interesting to create a combination of variables by which the observations can be classified based on the very groups to which they are known to belong. This is a basic aim of discriminant and classification analyses. In section 11.2, the variables that characterize sport enthusiasts in comparison with people who are engaged in visual arts as their leisure-time activities are investigated. *discriminant analysis* (DA) has the longest history—from Hotelling (1931; 1936), Wilks (1932), and Fisher (1940)—and hence it is introduced first in section 11.2.

Discriminant and Classification Analyses when the classification basis is known

Cluster analysis when the classification basis is *not* known.

If the basis for classification is *not known* but we want to find out which observations and variables go together based on the measured variables, *cluster analysis* (CA) is selected; these methods are presented in section 11.3. CA has become one of the main methods in data mining, that is, simultaneous analysis of large amount of data. In section 11.3, three different methods—the hierarchical clustering, *k*-means clustering, and two-step clustering—are introduced and used to cluster motivational items to find out which of them carry common elements. We are expecting to find the result close to that produced with principal component analysis (PCA; see section 8.2). CA might deserve more in-depth discussion than the one presented here, for it has not only become one of the main methods in modern data mining, but it is also used in image analysis, pattern recognition, information retrieval, bioinformatics, and machine learning.

Decision tree analysis is efficient at finding cut-off points for hierarchical variables.

The third set of methods, *decision tree analysis*, is presented in section 11.4. It is also used in data mining; it finds accurately and effectively cut-off points amongst the *hierarchical* explanatory variables by which the classifying variable can be reliably clustered. In the presented example, the analysis finds out of 54 variables of motivation three that identify 12 drop-outs among 700 hobbyists.

Based on the historical trends and the number of hits in the ERIC database with more than 1 million publications, it seems that out of these three methods, CA is the most popular (Figure 11.1). Its golden era was in mid-1970s; Blashfield & Aldenderfer (1978) were correct in their computation that the research literature

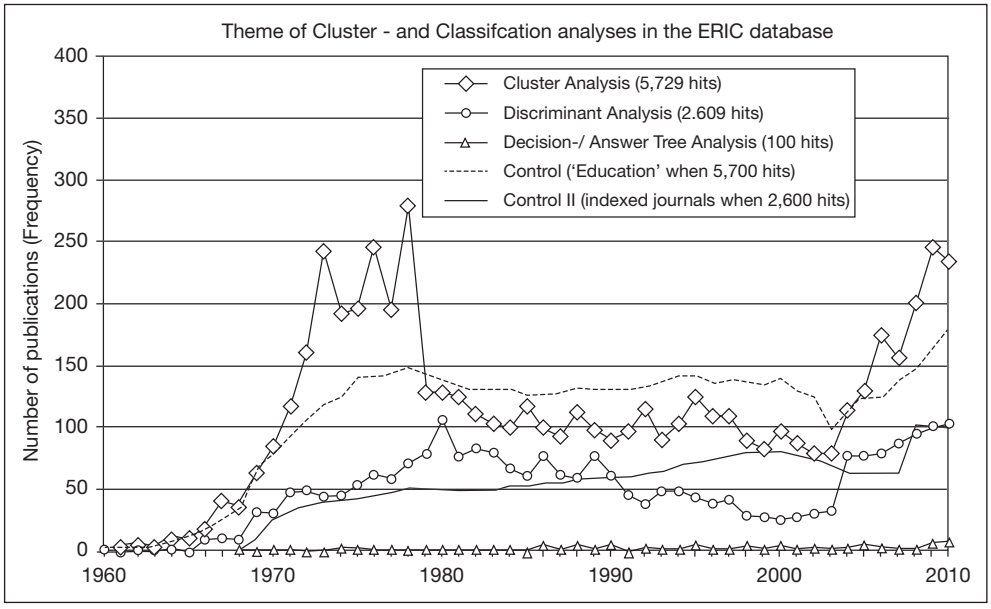


Figure 11.1 Historical Trends of Three Methodological Tools Connected with the Family of Cluster and Classification Analyses

connected with CA doubled every three years. During the 1980s, the use of CA was in a moderate decline but since 2003 the number of publications has sky-rocketed again. The historical trend of DA has been more moderate than that of CA—there are neither great peaks nor valleys in the trend. Every year around 50 to 100 publications appear with this keyword. Decision or answer tree analysis has not produced much research literature as a method: 2 to 5 publications appear steadily per year that deal with this methodological tool.

11.2 Discriminant and Classification Analyses

Discriminant Analysis (DA) or discriminant function analysis (DFA), or linear discriminant analysis (LDA), and the classification analysis (CA), usually performed along with DA, are the methods of finding the combination—or technically speaking, a mathematical function—among the explanatory variables that discriminate among groups the best. When the mathematical rule is found, new observations, if the values of their explanatory variables are known, can be classified into the correct groups. We can, for example, classify the hobbyists into different hobby groups based on their motives—as it was done in Metsämuuronen (1995 and 1997).

A mathematical rule is created in DA; the rule is used in CA to classify the observations into various groups

There are two phases in DA. In the *first* phase, the discriminant (or classification) functions are formed and tested. In the *second* phase, we examine the extent of the formed functions abilities to predict to which group the observations from the (same or different) dataset belong. Practically, in the simplest case, the discriminant function means that an explanatory variable X is used to classify the explained, categorically, variable Y into two groups ($Y=1$, e.g., male and $Y=2$ for female). Formally, a linear (or non-linear) function connects variables Y and X as follows:

Phases of DA:
1. Create a discriminant function
2. Predict and classify the observations

$$Y=aX$$

When aX gets a value higher than the criterion value T, the case is categorized into the group 1 (i.e., the case is a male) and when the value is lower than T, the case is categorized into the group 2 (i.e., the case is a female). The function does not have to be linear, and usually there are more than one explanatory variables in a model.

DA and MANOVA are technically identical methods of analysis.[1] While in MANOVA the main question is the differences in group means, in DA it is the prediction of group membership and the dimensions that differentiate groups. In DA, the interest lies in the simultaneous operation of the whole group of explanatory variables,

[1] Multivariate analysis of variance (MANOVA) is discussed in section 10.4.

while in MANOVA the focus is on the individual variables. On the other hand, the logic of DA is very close to regression analysis and its logic; in section 9.4, it is stated that DA is a special case of canonical correlation analysis.

In DA, as in many other multivariate methods, the observations in the dataset are predicted based on the data. However, when the model has been "trained" to observe things based on the specific data, we do not know whether the model works for the cases that are not part of the data. Then it is recommendable to analyze only a *part* of the cases in the modeling phase and use the remaining ones to test whether the model works on all observations. This method is called *cross-validation*.[2]

In this section, DA is addressed in a rather nonmathematical manner. In addition to the general books on multivariate analysis (see the list in Chapter 7), some recent publications on DA are McLachlan (2004), Huberty and Olejnik (2006) and online, Garson (2011a)[3] and Statsoft, Inc. (2011)[4] in a more condensed manner.

11.2.1 Suitability

DA functions well in non-experimental designs

DA works best for the studies of one categorical variable and several explanatory factors. The categorical variable can be on a nominal scale or have a finite number of values. The aim is to find a combination of explanatory variables that can best group the observations into various classes. DA is developed for situations where there are no actual experimental settings (cf. ANOVA and MANOVA); the groups are formed naturally and the group sizes are not necessarily the same.

11.2.2 Limitations and Assumptions

Assumptions of DA:
- multinormality

DA assumes (as do many of multivariate methods) that the observations come from a multinormal population. Although—as Tabachnick and Fidell (2012) point out—if the objective is only to classify observations, the distribution assumption is not strict: that assumption is more relevant from the point of statistical inference. If the classification fails or it is unsatisfactory, it could be because the assumption of Normality does not hold. A more essential assumption than multinormality is the *assumption* that *there are no outliers in the data* (see also section 7.2.2).

- no outliers

The minimum technical requirement for a sample size is that there must be at least as many observations in the smallest group as

[2] In the context of regression analysis in section 9.2.4, there is also a discussion on cross-validation.

[3] http://www.statisticalassociates.com/discriminantfunctionanalysis.pdf

[4] http://www.statsoft.com/textbook/discriminant-function-analysis/?button=1

there are explanatory variables. Therefore, if there are 10 X variables, then there should be at least 10 cases in the smallest group. However, in that case the risk of over-fitting is great: each observation gets its "own variable" and the group is perfectly classified. To gain stable results that can be generalized, there must be considerably more observations in the smallest group than there are explanatory variables. A good rule of thumb is that there should be at least 5 times as many cases as there are explanatory variables, or at least 20 observations with a small number of variables.

- a sufficient number of observations

11.2.3 Briefly on Theory and Concepts

In the simplest case, there are two variables: the classified variable Y and the explanatory variable X. The variables are connected with a discriminant function that classifies the cases into groups. The basic idea of DA is illustrated in a graphical manner in section 11.2.3.1, conceptual matters are handled in section 11.2.3.2, procedures for selecting the variables in section 11.2.3.3, diagnostics in section 11.2.3.4, and concepts related to the classification analysis in section 11.2.3.5.

11.2.3.1 The Basic Idea behind DA Illustrated Graphically

Figures 11.2a and 11.2b show in a simplified form the basic concept behind DA when there are only two explanatory variables X. In the illustration, the symbols + and *o* are used to denote the observations in two groups, the values of which differ from each other, but not necessarily at a statistically significant level for either variable X_1 or X_2.

A *contour* curve is drawn around the groups; depending on the desired precision, 60 – 80% of all the observations remain within the ellipse. The black dot • marks the center of gravity or centroid, which illustrates the mean of the observations. The role of DA is to find such a mathematical function that rotates both dimensions X_1 and X_2 into different positions for the maximum separation between the groups in regard to both variables. These new dimensions are represented in the illustration by the oblique lines "X_1" and "X_2":

11.2.3.2 Concepts Related to DA

The computation of DA is similar to MANOVA; the actual testing is based on comparing the cross-product matrix between the groups ($\mathbf{S}_{between}$) and within the groups ($\mathbf{S}_{within}$, see also section 10.4.3.) To simplify, the total variation is divided—like in MANOVA—into systematic variation and error variation:

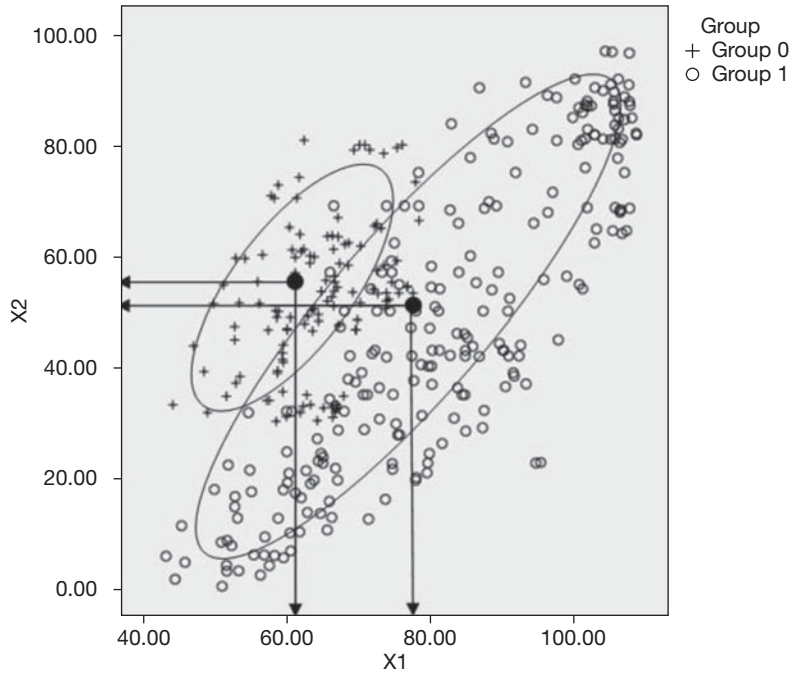


Figure 11.2a The Start of Discriminant Analysis

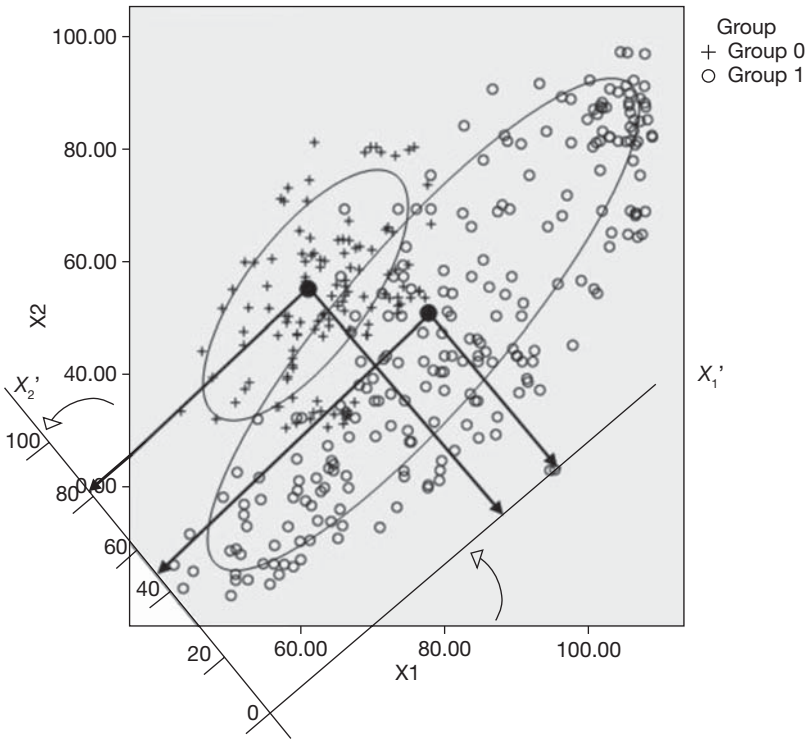


Figure 11.2b The Basic Concept of Discriminant Analysis