RESEARCH
Monograph

Publishing Nutrition Research: A Review of Multivariate Techniques—Part 3: Data Reduction Methods

Philip M. Gleason, PhD; Carol J. Boushey, PhD, MPH, RD; Jeffrey E. Harris, DrPH, MPH, RD, LDN; Jamie Zoellner, PhD, RD

ARTICLE INFORMATION

Article history:
Submitted 1 September 2014
Accepted 6 March 2015
Available online 30 April 2015

Keywords:
Principal component analysis
Factor analysis
Reduced rank regression
Cluster analysis
Eigenvalue

http://dx.doi.org/10.1016/j.jand.2015.03.011

ABSTRACT
This is the ninth in a series of monographs on research design and analysis, and the third in a set of these monographs devoted to multivariate methods. The purpose of this article is to provide an overview of data reduction methods, including principal components analysis, factor analysis, reduced rank regression, and cluster analysis. In the field of nutrition, data reduction methods can be used for three general purposes: for descriptive analysis in which large sets of variables are efficiently summarized, to create variables to be used in subsequent analysis and hypothesis testing, and in questionnaire development. The article describes the situations in which these data reduction methods can be most useful, briefly describes how the underlying statistical analyses are performed, and summarizes how the results of these data reduction methods should be interpreted.
J Acad Nutr Diet. 2015;115:1072-1082.

THIS ARTICLE REPRESENTS THE NINTH IN A SERIES exploring the importance of research design, statistical analysis, and epidemiologic methods in nutrition and dietetics research. The purpose of this series is to help nutrition and dietetics practitioners apply and interpret analytic and scientific principles consistent with high-quality nutrition research in their own work. Research is the foundation of the dietetics profession, providing the basis for decisions in practice and policy. This series uses examples relevant to the field of nutrition and dietetics. An effort is made to appeal to the seasoned researcher as well as the nutrition research novice.

The purpose of this monograph is to provide an overview of data reduction methods, and it represents the third installment in a set of three articles on multivariate statistical techniques.[1,2] Nutrition researchers are often faced with the challenge of summarizing with a few simple variables complex concepts, such as diet quality, for which there are a large number of individual measures. Data reduction methods help researchers face this challenge by creating new variables that more efficiently summarize the large quantity of information originally available or use that information efficiently in subsequent analysis. One example outside of nutrition where these techniques are used is in the analysis of clusters of genes. There are thousands of genes that could be studied; data reduction methods have enabled researchers to reduce this number to smaller groups for further focus or analysis. Four major data reduction techniques will be covered in this article: principal components analysis (PCA), factor analysis (FA), reduced rank regression, and cluster analysis. Figure 1 provides a glossary of the relevant terminology.

Data reduction methods are set apart from other quantitative statistical methods covered in this series in that these methods themselves do not fall within the category of inferential statistics that would include statistical significance testing. Rather, these methods represent techniques researchers use to manipulate data they have collected for further analysis. The techniques include different algorithms that group variables of interest or sample members into underlying correlated or clustered groups according to well-defined rules set a priori by the investigator. In each of these analyses, the researcher plays an important role in guiding the selection and manipulation of variables and/or grouping of sample members.

In the field of nutrition, data reduction methods can be used for three general purposes. First, researchers often use these techniques for descriptive purposes alone; to summarize dietary patterns; or analyze the relationships among multiple foods, nutrients, or combinations of foods and nutrients. Research using the concept of dietary patterns rather than the analysis of individual nutrients is useful as people consume combinations of foods containing multiple nutrients vs individual nutrients alone. The dietary patterns paradigm represents a more comprehensive characterization

To take the Continuing Professional Education quiz for this article, log in to www.eatrightPRO.org, go to the My Account section of the My Academy Toolbar, click the "Access Quiz" link, click "Journal Article Quiz" on the next page, and then click the "Additional Journal CPE quizzes" button to view a list of available quizzes.

of the diets of individuals or groups. Examples of research using data reduction methods to identify and summarize dietary patterns are provided by Nettleton and colleagues[4] and Reedy and colleagues.[5]

Second, data reduction methods can be used to create variables to be used in subsequent analysis and hypothesis testing. Although it might be impractical to test hypotheses involving dozens of measures of dietary intake, for example, it is more feasible to test hypotheses involving a single variable representing an overall dietary pattern created using data reduction methods.

A third common use of these methods is in questionnaire development. PCA and cluster analysis can be applied when developing a questionnaire to reduce the number of questionnaire items (variables) to a subset of items that best captures variation within the target population in the underlying concepts of interest. When considering questionnaire development, there may be many salient items of interest that could be included in the questionnaire; for example, those ascertained from focus groups. PCA and cluster analysis can help identify those items that correlate well with each other and with the underlying concepts of interest and guide elimination of those that do not. The factors that remain make good candidates for a questionnaire. For example, Glanz and Steffen[6] used cluster analysis as part of the development process of a questionnaire to assess psychosocial constructs related to calcium consumption among adolescents.

UNDERSTANDING AND PERFORMING THE TECHNIQUES

Principal Components Analysis and Factor Analysis

When researchers have a large number of potential variables to analyze and would like to summarize the information contained in those variables as efficiently as possible, PCA and FA are two closely related options for doing so. In each case, the method begins with a large number of "input variables" and ends with a much smaller number of variables—referred to as "principal components" or "factors"—that summarize the information in the input variables. This section describes PCA and briefly summarizes FA and its similarities to and differences with PCA.

PCA and FA might be used for a number of different reasons. One key distinction to make between different uses of these techniques involves whether the method is applied before or after the study's main data collection effort. A researcher or research team might use a data reduction technique like PCA or FA before collecting information from the full study sample in order to determine which questions to include in a survey instrument to best capture a particular construct of interest. In a situation like this, there may be a large number of candidate questions that capture some key aspect of the underlying concept, and the researcher would collect data on all candidate variables from a small subsample and conduct PCA or FA to identify the candidate questions that best "hang together" and reflect the underlying construct. Then, these questions would be included in a final survey instrument that could be administered to the full sample of interest.

In other situations, the researcher may be working with data that have already been collected on the full sample of interest, but wishes to whittle down a large group of input variables so that any subsequent analysis can be conducted more efficiently. Often, a small number of summary measures can provide more useful and easy-to-digest descriptive information about some underlying construct than a large number of input variables. If the constructs are intended to be used as covariates in a statistical technique such as a regression model, reducing the number of covariates can minimize the likelihood of a statistical problem known as multicollinearity, whereby high correlations among covariates make it difficult to identify their true relationship with the model's dependent variable.

Nettleton and colleagues[4] provide an example of a study that used PCA on already collected data in a sample of adults. The input variables in their analysis were dietary intake measures that were systematically condensed to 47 different food groups. Using PCA, the researchers created four principal components that they labeled as representing different dietary patterns. For example, the first principal component represented diets high in "fats and processed meats." They used these principal components both for descriptive purposes and in an analysis of the relationship between dietary patterns and markers of subclinical atherosclerosis.

Performing Principal Components Analysis. PCA can be performed with many statistical software packages. This section first describes the basic process for performing PCA and then provides several examples that illustrate these procedures. The starting point for researchers is a dataset that includes a large number of input variables that are distinct yet related in some way, such as numerous variables reflecting individuals' nutritional attitudes. The software will produce output that can be used to create a set of principal components, or summary variables that reflect some combination of the input variables included in the PCA. The researcher must interpret this output in order to determine the number of principal components to be created and determine how these principal components will be created and interpret the meaning of each, if possible.

Suppose the researcher begins with eight input variables and the goal is to summarize these variables with a smaller number of summary variables, or principal components. A useful preliminary step is to generate a correlation matrix that shows correlations between all of the input variables. This will provide a general sense of whether particular subsets of input variables tend to "hang together"; that is, to be highly correlated with one another. These are variables that will likely contribute most importantly to the same principal components.

A key step in the analysis involves the researcher determining the number of principal components to "retain." In other words, he or she must determine how many principal components will be used to summarize the full set of input variables. Technically, PCA will generate the same number of principal components as there are input variables in the analysis (eg, eight in the example given). However, each additional component identified will explain a smaller amount of the variation, and thus will become increasingly less useful. As shown here, researchers will typically "retain" the principal components that explain the most variation and drop from the analysis those that explain the least. Each principal component is a linear combination of the input