ScienceDirect

Multivariate Analysis

Through multivariate analysis, regional science helps regions confront deleterious aspects such as climate change, sea level rise, and other unprecedented phenomena, increasingly present in the status quo of the Anthropocene.

From: International Encyclopedia of Human Geography (Second Edition), 2020

Related terms:

High School Student, Regression Analysis, Mental Health, Mental Diseases, Job Satisfaction, Research Workers, Questionnaires

View all Topics

Multivariate Analysis: Overview

I. Olkin, A.R. Sampson, in International Encyclopedia of the Social & Behavioral Sciences, 2001

1 Introduction

Multivariate analysis is conceptualized by tradition as the statistical study of experiments in which multiple measurements are made on each experimental unit and for which the relationship among multivariate measurements and their structure are important to the experiment's understanding. For instance, in analyzing financial instruments, the relationships among the various characteristics of the instrument are critical. In biopharmaceutical medicine, the patient's multiple responses to a drug need be related to the various measures of toxicity. Some of what falls into the rubric of multivariate analysis parallels traditional univariate analysis; for example, hypothesis tests that compare multiple populations. However, a much larger part of multivariate analysis is unique to it; for example, measuring the strength of relationships among various measurements.

Although there are many practical applications for each of the methods discussed in this overview, we cite some applications for the classification and discrimination methods in Sect. 6.5. The goal is to distinguish between two populations, or to classify a new observation in one of the populations. Examples are (a) solvent and insolvent companies based on several financial measures; (b) nonulcer dyspeptics versus normal individuals based on measures of anxiety, dependence, guilt, and perfectionism; (c) Alaskan vs. Canadian salmon based on measures of the diameters of rings. For other such applications, see Johnson and Wichern (1999).

Multivariate analysis, due to the size and complexity of the underlying data sets, requires much computational effort. With the continued and dramatic growth of computational power, multivariate methodology plays an increasingly important role in data analysis, and multivariate techniques, once solely in the realm of theory, are now finding value in application.

> Read full chapter

Analysis and Interpretation of Multivariate Data

D.J. Bartholomew, in International Encyclopedia of Education (Third Edition), 2010

Multivariate analysis is concerned with the interrelationships among several variables. The data may be metrical, categorical, or a mixture of the two. Multivariate data may be, first, summarized by looking at the pair-wise associations. Beyond that, the different methods available are designed to explore and elucidate different features of the data. The article briefly summarizes the scope and purpose of the following methods: cluster analysis, multidimensional scaling, principal components analysis, latent class analysis, latent profile analysis, latent trait analysis, factor analysis, regression analysis, discriminant analysis, path analysis, correspondence analysis, multilevel analysis, and structural equation analysis.

> Read full chapter

Multivariate Analysis: Discrete Variables (Overview)

Alan Agresti, in International Encyclopedia of the Social & Behavioral Sciences (Second Edition), 2015

Abstract

This article deals with discrete multivariate analysis of categorical response variables. A categorical variable is one for which the measurement scale is a set of categories.

For studies having two or more categorical variables, a contingency table displays the counts for their cross-classification of categories. Modeling approaches are vital for investigating association and interaction structure. Logistic regression is an analog of ordinary regression for binary response variables. Loglinear models, by contrast, are relevant for analyses analogous to correlation analyses, studying the association structure among a set of categorical response variables. Correspondence analysis is a related descriptive and graphical method. Multivariate analyses are relevant when there are several categorical response variables, such as in repeated measurement (e.g., longitudinal) studies. Recent advances here include models that address marginal components of a multivariate response and models that use cluster-specific random effects to describe the joint distribution. Other recent advances in discrete multivariate analysis include specialized models for ordinal responses and exact small-sample methods. Finally, the unifying concept of generalized linear models connects the primary categorical modeling procedures – logistic regression models and loglinear models – with long-established regression and analysis of variance methods for continuous response variables.

> Read full chapter

Conclusions and Recommendations

Brent E. Turvey, in Forensic Fraud, 2013

Hierarchical multiple regression analyses

In Chapter 9, "Multivariate Analysis of Forensic Fraud, 2000–2010," hierarchical multiple regression analyses revealed significant correlations between employer, job description, and employee variables related to examiner approach, the impact of fraud, and evidence affected.[8]

Employer independence

Significant correlations regarding Employer Independence, relevant to the theses of this study, are as follows (see also discussion provided in Chapter 9):

1. As mentioned, Simulators represent the most frequent approach to committing forensic fraud (90%; *n* = 90).
2. Increased employer independence from law enforcement is associated with a significantly reduced frequency of Simulators.
3. Conversely, increased law enforcement dependence (i.e., affiliation) is associated with a significantly increased frequency of Simulators.

4. Independence from law enforcement is not associated with a significant increase or reduction in the frequency of other approaches to fraud.

These findings provide explicit empirical support for the thesis that those working on behalf of the state, specifically the police and the prosecution, are responsible for a substantial amount, if not the majority, of known cases of forensic fraud. These findings also support the assertion that the culture of law enforcement has a significant and potentially corrupting effect on the forensic examiners that it employs.

Laboratory accreditation

Significant correlations regarding Laboratory Accreditation, relevant to the theses of this study, are as follows (see also discussion provided in Chapter 9):

1. If a forensic laboratory is accredited, fraudulent examiners are significantly more likely to exaggerate, embellish, lie about, or otherwise misrepresent results.
2. Laboratory accreditation is significantly correlated to increased falsification of only one kind of physical evidence: DNA.
3. Accredited laboratories are significantly less likely to impose severe consequences on fraudulent examiners.

These findings provide explicit empirical support for the thesis that forensic fraud tends to be the result of cultural, pathological, and systemic causes rather than the narrow motives of single individuals, as the circumstances surrounding it must be allowed to develop and persist by those in the immediate forensic environment.

Internal audits

Significant correlations regarding Internal Audits, relevant to the theses of this study, are as follows (see also discussion provided in Chapter 9):

1. Internal Audits are significantly correlated with an increase in the number of cases under review; as this is generally the purpose of an audit, this finding was expected.
2. Significantly more Simulators are revealed in association with Internal Audits, demonstrating their effectiveness with identifying this type of forensic fraud.
3. Significantly fewer Pseudoexperts, and related Education and Experience fraud, are revealed in association with Internal Audits. This may reflect that audits are more often focused on reviewing cases and protocols, and not hiring practices or examiner resumes. Alternatively, this may suggest that the kinds of forensic laboratories imposing Internal Audits are less likely to hire examiners with phony qualifications in the first place.

4. While Internal Audits are significantly correlated with identifying more Drug1 evidence-related fraud (e.g., weights and amounts), they are significantly correlated with identifying fewer cases of fraud related to Money and Biological (non-DNA) evidence.

These findings provide explicit empirical support for the thesis that forensic fraud tends to be the result of cultural, pathological, and systemic causes rather than the narrow motives of single individuals, as the circumstances surrounding it must be allowed to develop and persist by those in the immediate forensic environment.

History of addiction

Significant correlations regarding Examiner History of Addiction, relevant to the theses of this study, are as follows (see also discussion provided in Chapter 9):

1. Examiners with a history of addiction were significantly correlated with Drug1 evidence related to fraud (e.g., weights and amounts).
2. Examiners with a history of addiction were significantly correlated with an increased number of overturned cases.

A history of addiction is something that can be screened for by an employer both prior to and during examiner employment. Therefore, these findings provide explicit empirical support for the thesis that forensic fraud tends to be the result of cultural, pathological, and systemic causes rather than the narrow motives of single individuals, as the circumstances surrounding it must be allowed to develop and persist by those in the immediate forensic environment.

Drug2 and biological evidence

There was a noteworthy absence of any significant correlations between the Independent variables in this study, and Drug2 and Biological (non-DNA, aka BioEV) evidence (detailed discussion provided in Chapter 9; see Table 9-6). This may suggest that Drug2- and BioEV-related fraud exist as cultural or systemic problems within the forensic science community, distributed evenly regardless of employer, job description, or employee-related variables. The failure to significantly correlate either of these types of evidence suggests empirical support for the thesis that forensic fraud tends to be the result of cultural, pathological, and systemic causes rather than the narrow motives of single individuals.

> Read full chapter

Multivariate Normal Distribution

J. Tacq, in International Encyclopedia of Education (Third Edition), 2010

An Application: Hotelling's T_2-Test and Mahalanobis' Distance *D2*

In statistics, especially in multivariate analysis, there are many applications in which multivariate normal distribution plays an important role. Of course, linear regression analysis and its extension, structural equation models, are the example, because normality lies at the heart of these techniques. Other examples are discriminant analysis, multivariate analysis of variance, and canonical correlation analysis. Take for example discriminant analysis. For a number of groups, two or more than two, we will examine whether there is a significant difference between the centroids of the groups. If there are two groups, this is an extension of Student's *T*-test; if there are more than two groups, it is an extension of Fisher's *F*-test. We concentrate on the case of two groups (in which Hotelling's T_2 will be developed as a special case of Wilks' □) and we emphasize testing (rather than predicting group membership).

In Student's *T*-test there are two groups and only one variable. In such a univariate case, we know that the sampling distribution of the sample mean is normal if the distribution of the population from which the sample is taken is normal, and even if the distribution of the population is not normal, on the condition that the sample is sufficiently large.

In Student's *T*-test, the *t* statistic is calculated as , in which n_0 and n_1 are the group sizes and □w_2 (which has to be estimated!) is the pooled average of the two variances (the variance □$_{(0)2}$ of *X* in group 0 and the variance □$_{(1)2}$ of *X* in group 1) and is the standard error of the sampling distribution of differences of means.

The formula of its square t_2 can be written as follows:

Hotelling constructed a statistic, called Hotelling's T_2, in which multiple discriminating variables are included, in which a difference of group means now becomes a difference of group centroids, because there are several variables, and in which division by the estimated variance is replaced by multiplication by the inverse of a covariance matrix containing not only the dispersions but also the mutual associations between the discriminating variables:

Here n_0 and n_1 are again the group sizes and Σw is the pooled average of the two covariance matrices Σ_0 and Σ_1. The vector **d** is the difference vector between group centroids. The part **d**□ Σw_{-1} **d** in the formula of Hotelling's T_2 is Mahalanobis' distance D_2.

In this multivariate case it holds, analogously, that the sampling distribution of the sample centroid is multivariate normal if the population is multivariate normal and/or if the sample is sufficiently large (which is just an extension of the central limit theorem). Under this assumption of multivariate normality, Hotelling has proven that the value $[(n - p - 1)/p(n - 2)]$ T_2 is distributed as F with p and $n - p - 1$ degrees of freedom. So it becomes straightforward to test whether there is a significant difference between the centroids of two groups, that is, between the means of the many variables, taken together and taking their variances and correlation coefficients into consideration. The assumption of multivariate normality can be guaranteed by a large sample. For small samples it has to be tested.

> Read full chapter

Statistical Analysis, Special Problems of: Transformations of Data

D. Ruppert, in International Encyclopedia of the Social & Behavioral Sciences, 2001

5 Multivariate Transformations

Most of the classical techniques of multivariate analysis assume that the population has a multivariate normal distribution, an assumption that is stronger than that the individual components are univariate normal. Andrews et al. (1971) generalize the Box–Cox model to multivariate samples. They transform each coordinate of the observations using the same transformation family, e.g., power transformations for all coordinates, but with coordinate-specific transformation parameters. It is assumed that within this family of multivariate transformations, there exists a transformation to multivariate normality. All parameters in this model are estimated by maximum likelihood. See Gnanadesikan (1997) and Ruppert (2001) for further discussion.

> Read full chapter

Crude Oil and Refined Product Fingerprinting: Applications

Zhendi Wang, Jan H. Christensen, in Environmental Forensics, 1964

17.7.3.3 Variable Selection or Weighted Least Squares PCA

Similar to case study 3 the multivariate analysis can be refined by deselecting the most uncertain variables or scale variables with regard to their analytical uncertainties. These two approaches have been successfully used to identify spilled oils (Christensen et al., 2004) and it was found that both approaches improved the resolution power of the PCA by increasing the variance described by the model compared with the variability of replicate samples. It was concluded that although variable selection improved the model, the selection process is not trivial and it contradicts the original aim of minimizing the subjectivity in the data analysis. Furthermore, it was concluded that fitting the PCA model according to a WLS criterion represents a more objective alternative and such an approach also improves the resolution power of the PCA. Both approaches have advantages and selection of the best approach depends generally on the specific case.

However, the results from the WLS-PCA with mean-centering shows that, in spite of weathering processes (mainly evaporation and water washing) for up to 14 days, the Baltic Carrier oil spill samples and the corresponding source oil are clustered in PC1 through PC4. Likewise, the round-robin spill samples, spill I and spill II, are grouped in the plot with the corresponding sources, Oseberg East (E) and Oseberg Field Centre (FC) (Faksness et al., 2002). Oseberg southeast (SE) lies close to Oseberg E along PC1, PC2 and PC3, but are well separated along PC4 which is a minor component describing only 6.3% of the total variation in the calibration set (Figure 17.7.6).

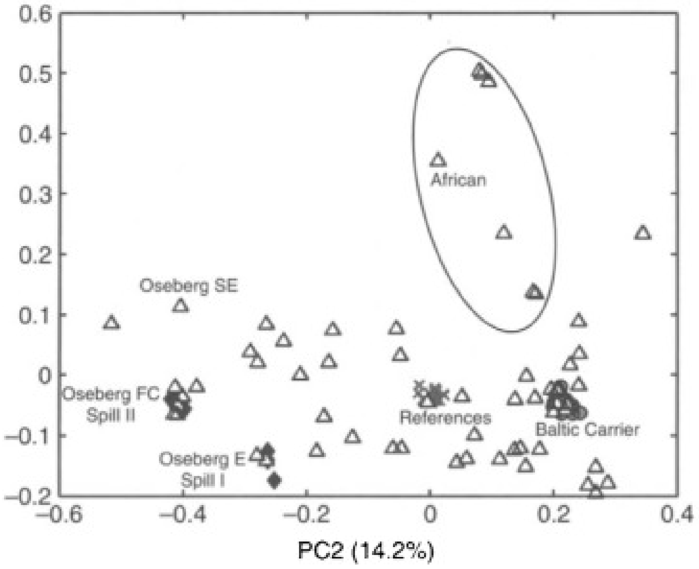

Figure 17.7.6. PCA score plots of PC2 vs PC4 using WLS-PCA.

> Read full chapter

Using CO2 Emission Allowances in Equity Portfolios

Michael Graham, ... Jarkko Peltomäki, in Handbook of Environmental and Sustainable Finance, 2016

17.4 Results

Table 17.3 presents the results from multivariate analysis of carbon emission future returns to stock market returns for the whole sample. The results show a positive relation between the returns on Vanguard Total World Stock Exchange Traded Fund and carbon emission futures returns. The coefficient for the MSCI EMETF is not statistically significant. The positive sign and statistical significance for Vanguard Total World Stock Exchange Traded Fund does not change when we reestimate the model without the statistically insignificant Exchange Traded Fund coefficients. As the coefficients for MSCI EMETF are not statistically significant, these results suggest in line with the correlation statistics in Table 17.2 that carbon emission returns are not particularly important with respect to emerging market equity index returns.

Table 17.3. Multivariable Analysis of Carbon Emission Allowance Returns

	CEFR: NAV		**CEFR: Mkt Returns**	
Variable	**Coefficient**	***t*-Statistic**	**Coefficient**	***t*-Statistic**
C	−0.066	−0.72	−0.026	−0.31
WETF	0.474▯▯▯	5.54	0.349▯▯▯	2.59
EMETF	−0.057	−0.96	0.060	0.79
HML	−0.245▯	−1.83	−0.251	−1.35
SMB	0.067	0.49	0.045	0.27
Adj. R2	0.03		0.02	
F-statistic	10.70		7.59	
Probability	0.000		0.000	
N	1511		1511	

This table reports ordinary least squares model results for the relation between carbon allowance futures returns and equity returns. The returns are in excess of the 1-month T-bill rate. The standard errors are both heteroskedasticity and autocorrelation robust. CEFR is the carbon emission returns; NAV and Mkt Returns are the net-asset-value (NAV) and market returns of iPath Global Carbon Exchange Traded Note, respectively; EMETF is return on the iShares MSCI Emerging Markets Exchange Traded Funds; and WETF is return on the Vanguard Total World Stock Exchange Traded Funds. ▯ refers to statistical significance at the 0.1 level; ▯▯ refers to statistical significance at the 0.05 level; ▯▯▯ refers to statistical significance at the 0.01 level. HML and SMB are the value and small firm premiums, respectively.

The value premium is shown to relate negatively to carbon emission future returns, indicating that short positions in carbon emission allowance futures would associate with the value investment strategy. From another point of view, firms that take long positions in carbon emission allowance futures have a negative exposure to the value premium. Additionally the finding means that corporate policies that reduce any need to hedge against rising carbon emission prices can reduce costs which arise from hedging against higher carbon emission prices. Thus environmentally responsible management policies that aim at reducing carbon emissions could return reduced hedging costs related to the value premium. This association between the value premium and carbon emission returns is a new finding and has not been discussed in previous studies such as Byun and Cho (2013) and Kanamura (2013), and lends itself to new lines of empirical investigations.

In further analyses, we investigate how carbon emission future returns relate to stock market returns at different levels of stock market stress (see Table 17.4). We show the positive relation between higher thresholds of stock market stress. The coefficient for the MSCI EMETF is not statistically significant except for the highest threshold (CFSI > 20).

Table 17.4. Carbon emission allowances returns and stock Market stress

Variable	**CFSI&lt;10**		**CFSI&gt;10**		**CFSI&gt;15**		**CSFI&gt;20**	
	Coefficient	***t*-Statistic**	**Coefficient**	***t*-Statistic**	**Coefficient**	***t*-Statistic**	**Coefficient**	***t*-Statistic**
CEFR: NAV								
C	0.030	0.15	–0.126	–1.42	–0.215□□	–2.16	–0.204	–0.87
WORLDM	–0.306	–0.58	0.560□□□	6.50	0.587□□□	6.90	0.684□□	2.42
EMR	0.053	0.18	–0.068	–1.38	–0.083□	–1.83	0.134	1.11
HML	0.532	1.04	–0.346-□□□	–2.66	–0.359□□	–2.55	–1.048-□□□	–3.11
SMB	0.990□	1.79	–0.057	–0.45	–0.013	–0.09	0.68□□7	2.36
Adj. R2	0.00		0.06		0.08		0.14	
F-statistic	0.79		18.07		17.48		8.97	
Probability	0.534		0.000		0.000		0.000	
N	478		1000		732		190	
CEFR: Mkt Returns								
C	0.181	0.95	–0.117	–1.31	–0.206□□	–2.15	–0.257	–1.08
WORLDM	0.395	0.61	0.342□□	2.39	0.367□□	2.48	–0.443	–0.98
EMR	0.100	0.26	0.065	0.86	0.072	0.88	0.632□□	2.43
HML	–0.862	–1.22	–0.153	–0.74	–0.216	–0.99	–0.280	–0.59
SMB	0.627	0.85	–0.080	–0.47	–0.070	–0.37	0.411	1.10
Adj. R2	0.00		0.03		0.04		0.02	
F-statistic	1.32		8.65		9.21		2.19	
Probability	0.260		0.000		0.000		0.072	
N	478		1000		732		190	

This table reports ordinary least squares model results for the relation between carbon allowance futures returns and equity returns in different states of stock market stress. The returns are in excess of the 1-month T-bill rate. The standard errors are both heteroskedasticity and autocorrelation robust. CEFR is the carbon emission returns; NAV and Mkt Returns are the net-asset-value (NAV) and market returns of iPath Global Carbon Exchange Traded Note, respectively; EMETF is return on the iShares MSCI Emerging Markets Exchange Traded Funds; and WETF is return on the Vanguard Total World Stock Exchange Traded Funds. □ refers to statistical significance at the 0.1 level; □□ refers to statistical significance at the 0.05 level; □□□ refers to statistical significance at the 0.01 level. HML and SMB are the value and small firm premiums, respectively.

Thus the exposures of carbon futures returns to equity return increases with stock market stress. This finding is similar to the results of Kanamura (2013) which suggest that the correlation between financial and carbon asset increases in financial turmoil.

> Read full chapter

Case Studies

Tom Tullis, ... Fiona Tranquada, in Measuring the User Experience (Second Edition), 2013

10.1.4 Discussion

Thus, we found that running the multivariate analysis showed that the user experience contributed 36% to increasing product recommendations. At Year 2, we hadn't met our target of increasing Likelihood to Recommend our product by 5%, but by investing in ease of use and in a few key features we were able to improve the Likelihood to Recommend by 3%. The Net Promoter model had provided us with a way to define and prioritize investment in user experience design and had given us a way to track the return of that investment year after year.

We wanted to test the Net Promoter model further. Could the model be used as a predictor of sales growth, as it was originally intended (Reichheld, 2003)? We know the average sales price of our products. We know, from the multivariate analysis, that interface design contributes 36% to motivating users to recommend our product. If we knew how many promoters refer the product actively, we could estimate the revenue gains associated with improved user experience of our software.

What we did next is determine if there is a link between “promoters” and an increase in customer referrals. In our survey, we asked if the respondent—all were existing customers—had referred the product to a friend in the last year (Owen & Brooks, 2008). From these data we derived the proportion of customers obtained through referrals and who likely refer others. This allowed us to approximate the number of referrals necessary to acquire one new customer (see Figure 10.4). Data used to derive this number are proprietary. For the purpose of this chapter, we use the number eight: we need eight referrals to acquire one new customer. In the NPS model, it is *promoters* who refer a product actively. But we didn’t want to assume that every respondent who answered 9 or 10 to the *likelihood to recommend* question, that is, every promoter, had referred our product actively. The actual percentage of promoters who referred our product actively within the last year was 63%. From this, we derived that the total number of promoters needed to acquire one new customer was 13.

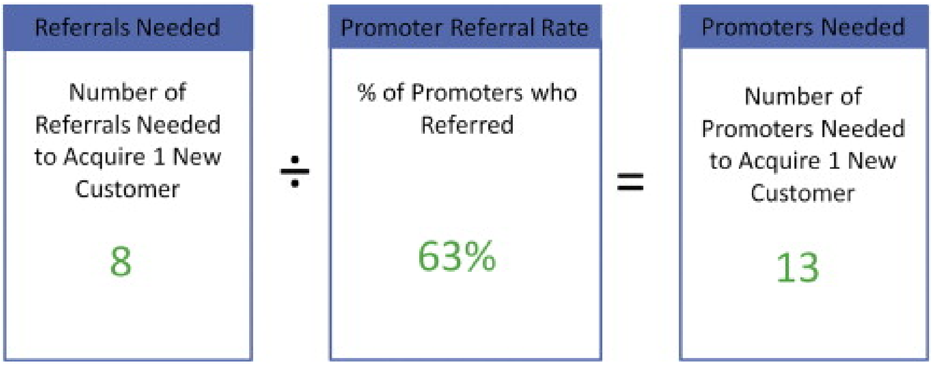
Referrals Needed
Number of Referrals Needed to Acquire 1 New Customer
8
÷
Promoter Referral Rate
% of Promoters who Referred
63%
=
Promoters Needed
Number of Promoters Needed to Acquire 1 New Customer
13

Figure 10.4. How many promoters are necessary to acquire one new customer?

> Read full chapter

Factor Analysis and Latent Structure: Overview

D.J. Bartholomew, in International Encyclopedia of the Social & Behavioral Sciences, 2001

7 Future Developments

The advent of massive computer power has changed the practice of multivariate analysis radically, and of latent variable analysis in particular. The limiting factor is no longer computing power but of getting data of sufficient quality and quantity to fit the very complicated models which the theory provides and computers can handle. The precision with which models with many parameters can be estimated is often very low unless the sample size runs into thousands. This makes it all the more important to estimate the sampling variability of the estimates of the parameters on which the interpretation depends. Traditionally this has been done by finding asymptotic standard errors but these can be very imprecise. However, it is now possible to supplement these results by resampling methods such as the bootstrap.

There are other fields in which latent variable models are used which currently exist in isolation. An obvious generalization is to latent time series. Some work has been done for the case where the latent process is a Markov chain. In this area the term 'hidden' is used instead of 'latent' which helps to conceal the family connections (see *Neural Networks and Related Statistical Latent Variable Models*) (for an introduction see MacDonald and Zucchini 1997). An application in a more traditional time series context will be found in Harvey and Chung (2000). Also, there is work by economists on unobserved heterogeneity as it is called which, essentially, involves the introduction of latent variables into econometric models.

> Read full chapter