Chapter 1

Overview and Examples of Multivariate Methods

Section 1.1

Introduction and Examples of Multivariate Statistics

Objective

- Recognize appropriate analyses for a variety of multivariate research questions.

3

Univariate versus Multivariate Statistics

Univariate statistics

- consider only one dependent variable (DV) at a time.
  - Examples: sample mean, *t*-test, ANOVA

Multivariate statistics

- consider more than one dependent variable at a time.
  - Examples: vector of sample means, Hotelling's T^2, MANOVA

4

Advantages of Multivariate Methods

Univariate statistics

- increase the risk of type-I error with many DVs
- only demonstrate the relationships between independent variables (IVs) and a DV, but miss relationships among the DVs.

Multivariate statistics

- control type-I error by considering a set of dependent variables in multidimensional space
- account for relationships among the DVs as well as the relationships between IVs and DVs.

5

Applications of Multivariate Statistics

Multivariate statistics can be used to address a wide variety of research questions.

Consider several examples of multivariate statistical applications in scientific research.

6

A Comparison of Advertising Strategies

Evaluate the effectiveness of three different commercials.

What makes an effective commercial?

- Product recognition
- Product recall
- Product liking
- Price willing to pay for product

This example has

- three levels of the independent variable
- four response variables.

The Effectiveness of a Drug

A drug company wants to compare the effectiveness of two different drug formulations (Old, New) across different dosages (50, 100, 200 mg).

How is effectiveness evaluated?

- Score on a depression scale
- Scores on two different obsessive-compulsive behavior scales.

2 × 3 factorial design, three responses.

Multivariate Analysis of Variance: MANOVA

MV analogy to ANOVA.

Tests for significant differences between groups on two or more related dependent variables simultaneously while accounting for the correlation among the dependent variables.

Research question: “Are there significant differences between two or more groups on a set of responses?”

Corporate Training Example

A company wants to compare the effectiveness of three employee training methods in a repeated measures study.

Effectiveness is defined as:

Score on a test of corporate policies

Score on a test of job-specific skills.

Employees are tested at three time intervals (2 weeks, 4 weeks, and 6 weeks).

Doubly Multivariate Repeated Measures

Tests for significant group differences over time across a **set of response variables** measured at each time while accounting for correlation among the responses.

Research question:

“Do the factors have an effect on a set of responses over time?”

“Do changes in an independent variable over time predict changes in the set of responses?”

The Diagnostic Usefulness of an Instrument

How well does a new psychological instrument perform compared to an established instrument?

- The established instrument, based on diagnostic criteria, contains twelve items and must be administered by a trained interviewer.
- The test instrument contains twenty items and can be completed by the respondent with pen and paper.

This example has

- twelve continuous predictors
- twenty continuous responses.

Multivariate Multiple Regression

Test for significant linear relationship between a set of predictors and a set of responses while accounting for the correlations among the responses.

Research Question:

"Does variation in a set of continuous independent variables adequately predict a set of continuous responses?"

Canonical Correlation Analysis

Canonical correlation analysis tests the same hypotheses as multivariate regression, but also allows you to

- interpret how the predictors are related to the responses
- interpret how the responses are related to the predictors
- examine how many dimensions the variable sets share in common.

Pathological Gambling Example

Researchers want to use responses to questionnaire items to classify people identified as steady gamblers, binge gamblers, and control/non-gamblers.

A twelve-item questionnaire is administered to three groups of participants.

Question: What linear combination of responses accounts for most of the variation in classification of gamblers?

Customer Profiling and Prediction

A credit card company is interested in using financial information to decide whether potential customers represent good or bad risk before offering a credit card.

An analyst is interested in understanding what combination of demographic variables best predict whether a customer prefers one of several different marketing strategies.

Discriminant Function Analysis

Discriminant function analysis (DFA) is a dimension reduction method that can be used to identify a linear combination of variables that produces the greatest distance between categories. DFA is conceptually similar to logistic regression for multivariate data, and it is computationally similar to MANOVA.

Bird Habitat Example

A researcher is interested in understanding the habitat of a species of bird. Twenty characteristics are measured for each habitat. Many of these measures are associated.

These variables will be used in regression, discriminant, and cluster analyses.

The researcher wants to reduce the total number of variables from 20 to something smaller and eliminate potential collinearity problems.

Principal Components Analysis

A dimension reduction technique

- creates new variables that are linear combinations of a set of correlated variables
- does not assume an underlying latent factor structure.

Practical question:

"How can I reduce the set of 20 correlated variables to a more manageable number of uncorrelated variables?"

Perceptions of Mathematics in School

An researcher wants to know whether students' self-perceptions in math reflect several underlying latent factors or one single factor.

Questionnaire items from several instruments intended to measure mathematics-related perceptions are administered to 4000 students.

- Exploratory analysis identifies possible underlying factors.
- Confirmatory analysis tests hypotheses about factors.

Factor Analysis

Exploratory factor analysis is a variable identification technique with superficial resemblance to principal components analysis, but with some important distinctions.

Factor analytic methods are used when an underlying factor structure is presumed to exist but cannot be represented easily with a single value.

Factor Analysis Research Questions

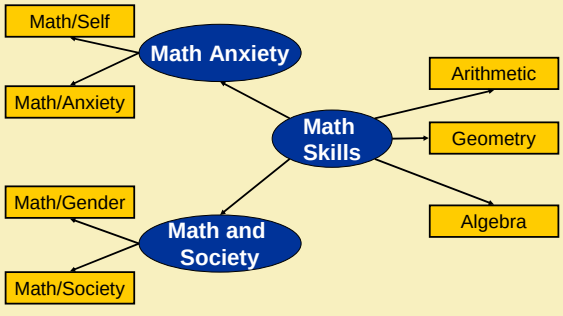
- "Are self-perceptions in mathematics the result of a single self-view, or are there multiple underlying constructs that each contribute to self-perceptions in mathematics?"
- "Is growth of an organism the result of a single growth process, or are there several latent constructs that separately contribute to the growth of an organism?"
- "Is economic growth a single construct or the result of several latent variables that can interrelate to produce economic climate?"

Success in Math

A research team wants to know whether math anxiety and students' perceptions of math in society are associated with mathematical ability among high school students.

Math anxiety, perceptions of math in society, and mathematical ability are all **latent variables**: they cannot be measured directly but must be estimated with a combination of measured variables.

Success in School



Structural Equation Modeling (SEM)

SEM is used to investigate regression-type relationships among a set of observed or latent responses and a set of observed or latent predictors.

Multivariate multiple regression, path analysis, and confirmatory factor analysis are all special cases of structural equation models.

Richness versus Simplicity

Just as multivariate analysis takes into account more complex, multidimensional relationships among variables, MV statistics can also be difficult to interpret.

- There's a reason UV methods are commonly used – they are easier to understand!
- Spending more time and effort in understanding your multivariate relationships can be profitable, well worth the effort. BUT…
- sometimes reality is so complex that you must return to univariate analyses in order to make sense of them at all!

A Silver Bullet? Definitely Not

Finally, keep in mind that your analysis can never be better than the data that went into it.

- Specify your research questions.
- Design the right study to examine the research questions.
- Define and document your sampling plan.
- Operationalize your variables appropriately.
- Check the data carefully for anomalies and errors.

Do not expect "Garbage In, Roses Out."
(Tabachnik and Fidell 2001)

sas | Education

Section 1.2

Review of Univariate Statistics

Objectives

- Review univariate model fitting concepts.
- Discuss concept of analysis of variance.
- Discuss methods for linear regression analysis.

Univariate Linear Models

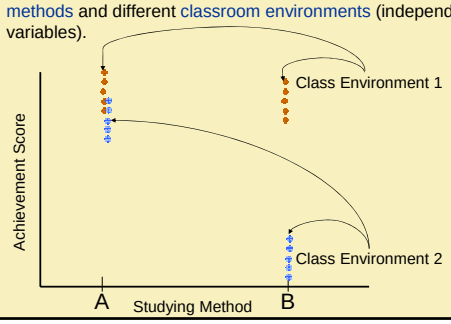
If a dependent variable can be written as a linear function of one or more continuous or categorical independent variables, then that expression is known as a linear model.

Examples of linear models:

- analysis of variance (ANOVA)
- linear regression
- analysis of covariance (ANCOVA).

Analysis of Variance (ANOVA)

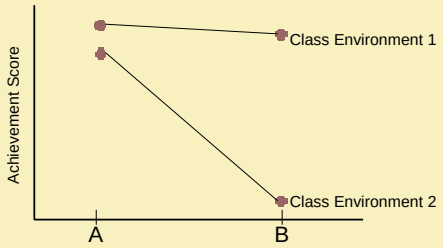
Example: Determine whether there are significant differences in achievement (dependent variable) as a result of different studying methods and different classroom environments (independent variables).



31

ANOVA Model

Linear model: $Y_{ijk} = \mu + \alpha_i + \tau_j + \alpha\tau_{ij} + \varepsilon_{ijk}$



32

ANOVA Table

Source of Variation	df	SS	MS	$F_{(p-1,\ N-p)}$
Model (*b*)	p-1	$\sum n_{ij}(\bar{Y}_j - \bar{\bar{Y}})^2$	$\frac{SS_b}{df_b}$	$\frac{MS_b}{MS_w}$
Error (*w*)	N-p	$\sum (Y_i - \bar{Y}_j)^2$	$\frac{SS_w}{df_w}$	
C.Total	N-1	$\sum (Y_i - \bar{\bar{Y}})^2$		

33

The GLM Procedure

General form of the GLM procedure:

PROC GLM <*options*>;
  CLASS *class-variables*;
  MODEL *dependents* = *independents* <*options*>;
RUN;

34

Linear Regression

Example: Find a straight line that best predicts profit (dependent variable) from expenditures (independent variable).

35

Linear Regression Model

Linear model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

36

ANOVA Table for Regression

Source of Variation	df	SS	MS	$F_{(p-1,\ N-p)}$
Model (*m*)	p-1	$\sum(\hat{Y}_{ij}-\bar{\bar{Y}})^2$	$\frac{SS_m}{df_m}$	$\frac{MS_m}{MS_e}$
Error (*e*)	N-p	$\sum(Y_i-\hat{Y}_{ij})^2$	$\frac{SS_e}{df_e}$	
C.Total	N-1	$\sum(Y_i-\bar{\bar{Y}})^2$		

Notice that *m* and *e* notation in regression are equivalent to the *b* and *w* notation from ANOVA.

37

The REG Procedure

General form of the REG procedure:

```
PROC REG <options>;
    MODEL dependents = independents <options>;
RUN;
```

38