ST 437/537: Applied Multivariate and Longitudinal Data Analysis

Introduction and Motivation

Arnab Maity
NCSU Department of Statistics
SAS Hall 5240 919-515-1937 amaity[at]ncsu.edu

Objective

The goal of this course is to provide an introduction to the statistical methods used to analyze *multivariate data* and *longitudinal data*; these are data where multiple observations are collected for each sampling unit (subject or object) of many.

There are three main objectives for this course:

1. Gain a thorough understanding of the details of various multivariate and longitudinal techniques. The theoretical basis for the techniques will be explored but not fully developed.
2. To be able to select one or more appropriate techniques for a given multivariate or longitudinal data set.
3. To be able to interpret the results of a computer analysis of a multivariate/longitudinal data set.

Multivariate statistical analysis

Multivariate statistical analysis refers to advanced techniques for examining relationships among multiple variables at the same time. The need often arises in science, medicine, engineering, law, religion, and social science (business, management).

Example 1: Fisher's or Anderson's Iris data

Source: Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, 179-188.

Data set: Available in R as `iris`

This data set consists of the measurements of the variables sepal length and width, and petal length and width (in centimeters), respectively, for 50 flowers from each of three species (setosa, versicolor and virginica) of iris.

ID	Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	setosa	5.1	3.5	1.4	0.2
2	setosa	4.9	3.0	1.4	0.2
51	versicolor	7.0	3.2	4.7	1.4
52	versicolor	6.4	3.2	4.5	1.5
101	virginica	6.3	3.3	6.0	2.5
102	virginica	5.8	2.7	5.1	1.9

Table: Snapshot of the `iris` dataset.

Example 2: Violent Crime Rates by US State

Source: McNeil, D. R. (1977) Interactive Data Analysis. New York: Wiley.

Data set: Available in R as `USArrests`

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas (`UrbanPop`).

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5

Table: Snapshot of the `USArrest` dataset.

[an error occurred while processing this directive] [an error occurred while processing this directive]

Example 3: Air Pollution in US Cities

Source: R. R. Sokal and F. J. Rohlf (1981), Biometry, W. H. Freeman, San Francisco (2nd edition).

Data set: Available in R as `USairpollution` in the `HSAUR2` package

The annual mean concentration of sulphur dioxide, in micrograms per cubic metre, and 6 aspects of climate and human ecology of 41 US cities are measured.

	SO2	temp	manu	popul	wind	precip	predays
Albany	46	47.6	44	116	8.8	33.36	135
Albuquerque	11	56.8	46	244	8.9	7.77	58
Atlanta	24	61.5	368	497	9.1	48.34	115
Baltimore	47	55.0	625	905	9.6	41.31	111
Buffalo	11	47.1	391	463	12.4	36.11	166
Charleston	31	55.2	35	71	6.5	40.75	148

Table: Snapshot of the `USairpollution` dataset.

Characteristics of multivariate data

In a multivariate dataset, several variables are measured for each subject or object. These variables are not necessarily ordered.

There are four main types of research questions:

1. Degree of relationships between the variables
2. Measure significant difference between group means
3. Predicting membership of subjects/objects into two or more groups based on two or more variables
4. Explaining underlying structure

Analysis of multivariate data requires more advanced statistical techniques that account for joint modeling and dependence among the multiple measurements recorded on the same subject/object. Ignoring them has the risk of providing an over simplifying picture of the problem and may lead to inaccurate results.

Longitudinal data analysis

The multiple observations correspond to a single variable observed at *multiple follow-up times*. Longitudinal data analysis refers to statistical techniques for studying the behavior of the variable over time. The need often arises in agriculture and the life sciences, medical and public health research, and physical science and engineering, among other fields.

Example 1: Weight versus age of chicks on different diets

Source: Crowder, M. and Hand, D. (1990), Analysis of Repeated Measures, Chapman and Hall (example 5.3)

Data set: Available in R as `ChickWeight`.

The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets.

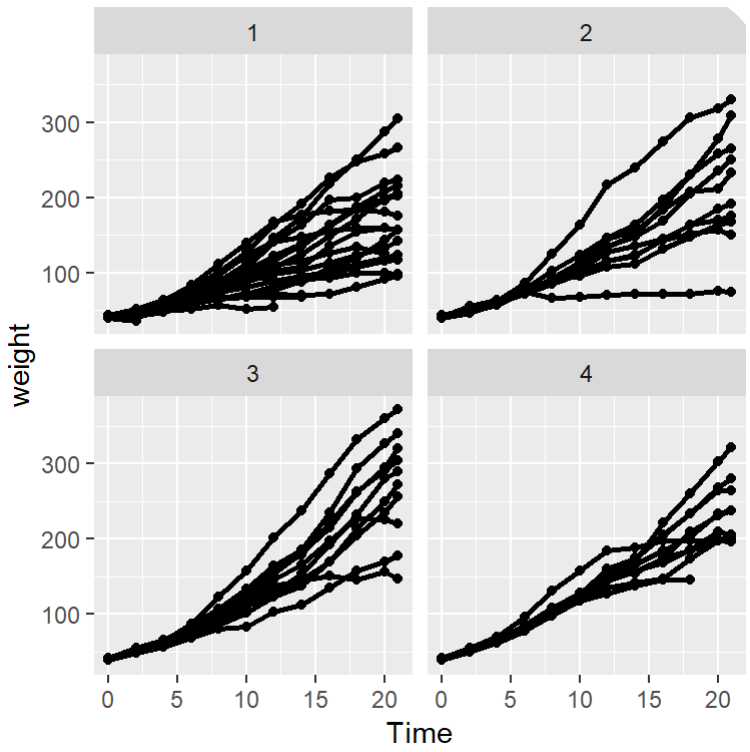

Figure: Growth curves of chickens – each panel corresponds to a particular type of diet

weight	Time	Chick	Diet
42	0	1	1
51	2	1	1
59	4	1	1
64	6	1	1
76	8	1	1
93	10	1	1
106	12	1	1
125	14	1	1

Table: A snapshot of chicken growth data.

Example 2: Growth of Loblolly pine trees

Source: Kung, F. H. (1986), Fitting logistic growth curve with predetermined carrying capacity, in Proceedings of the Statistical Computing Section, American Statistical Association, 340–343.

Data set: Available in R as `Loblolly`.

The Loblolly data frame has 84 rows and 3 columns of records of the growth of Loblolly pine trees.

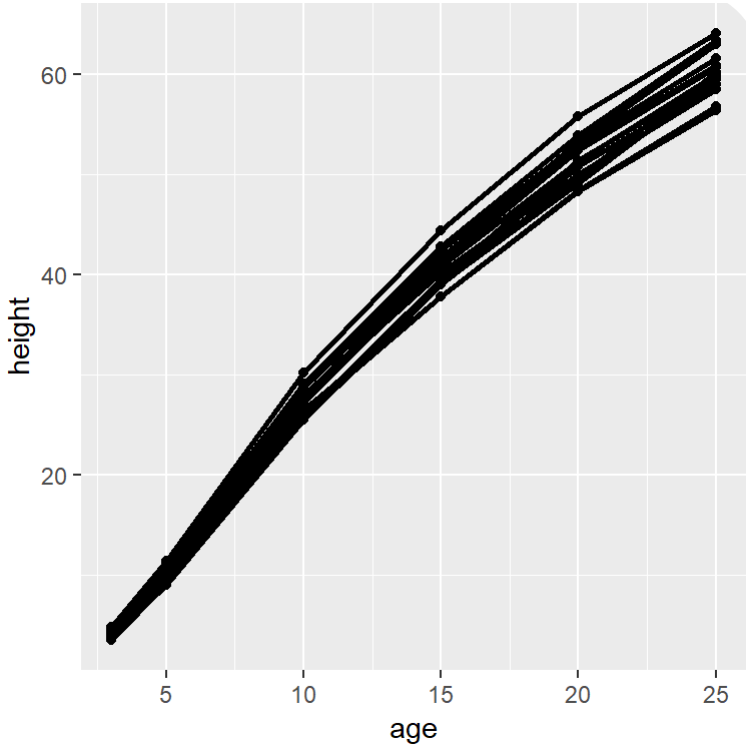

Figure: Growth curves of pine trees.

height	age	Seed
4.51	3	301
10.89	5	301
28.72	10	301
41.74	15	301
52.70	20	301
60.92	25	301
4.55	3	303
10.92	5	303

Table: A snapshot of pine tree growth data.

Characteristics of longitudinal data

Same outcome/response is measured repeatedly on each unit (eg. individual, plant, etc.). The condition of measurement is called generically *time*. Do note that the repeated measurements may correspond to other conditions than time, such as drug

dosage (eg. diastolic blood pressure measurements for several dose levels of a anti-hypertensive drug on the same subject), or height (diameter measurements at several height levels on the same tree).

Longitudinal data are different from multivariate data, as the **order of the repeated measurements is essential in the analysis of longitudinal data**, whereas permuting the order of the variables in multivariate analysis yields same results. Nevertheless one could employ methods from multivariate statistics to analyze longitudinal data.

Common questions of interest:

1. How does the typical response (mean response) vary over time? How does the rate of change of the typical response (mean response) vary over time?
2. If groups of subjects are followed over time, then how does the rate of change in the mean response vary across groups?
3. If additional covariates are available, then what is their effect on the response?

Analysis of longitudinal data requires sophisticated statistical techniques because the repeated measurements on the same subject are typically correlated. This must be recognized in the inferential process to obtain valid inferences.

Main page: **ST 437/537: Applied Multivariate and Longitudinal Data Analysis (https://maityst537.wordpress.ncsu.edu/)**