# *Introduction and Motivation*

*Arnab Maity*

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

*Contents*

## *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 methods will be explored but not fully developed.
2. To be able to select one or more appropriate methods 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*

This iris dataset[1], available in R as `iris`, consists of the measurements of the variables sepal length and width, and petal length and width (in centimeters), respectively, for fifty flowers from each of three species (setosa, versicolor and virginica) of iris.

| Species | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---:|---:|---:|---:|
| setosa | 5.1 | 3.5 | 1.4 | 0.2 |
| setosa | 4.9 | 3.0 | 1.4 | 0.2 |
| versicolor | 7.0 | 3.2 | 4.7 | 1.4 |
| versicolor | 6.4 | 3.2 | 4.5 | 1.5 |
| virginica | 6.3 | 3.3 | 6.0 | 2.5 |
| virginica | 5.8 | 2.7 | 5.1 | 1.9 |

*Table:* *Snapshot of the 'iris' dataset.*

The distributions of the four variables across the three species are shown in Figure 1. While we can compare individual variables across species, such an analysis does not reveal how the four variables behave together, and whether such behavior varies across the three species. Are the means of the four variables different from species to species? If a difference is found among the means, how do we estimate such differences?

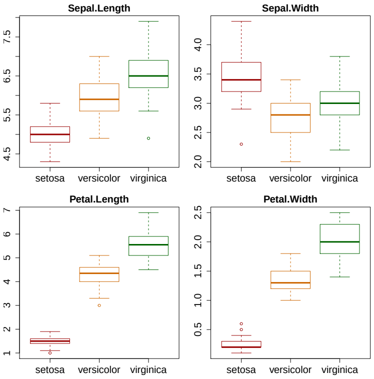

Figure 1: Boxplots of each variable in the iris dataset for each species.

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

### *Example 2: The Romano-British pottery data*

The pottery dataset[2], available as `pottery` in the `HSAUR3` library, consists of 45 observations on the nine chemicals on specimens of Romano-British pottery. The variable `kiln` shows the region where the specimen was made.

| Al2O3 | Fe2O3 | MgO | CaO | Na2O | K2O | TiO2 | MnO | BaO | kiln |
|---|---|---|---|---|---|---|---|---|---|
| 18.8 | 9.52 | 2.00 | 0.79 | 0.4 | 3.20 | 1.01 | 0.077 | 0.015 | 1 |
| 16.9 | 7.33 | 1.65 | 0.84 | 0.4 | 3.05 | 0.99 | 0.067 | 0.018 | 1 |
| 18.2 | 7.64 | 1.82 | 0.77 | 0.4 | 3.07 | 0.98 | 0.087 | 0.014 | 1 |
| 16.9 | 7.29 | 1.56 | 0.76 | 0.4 | 3.05 | 1.00 | 0.063 | 0.019 | 1 |

*Table:* *Snapshot of the ‘pottery‘ dataset.*

A correlation plot of the pottery dataset is shown in Figure 2. The numeric values displayed in the boxes correspond to the correlation coefficient between the corresponding variables. We can see that some of the variables have a moderate to high correlation. So do we need to examine all the chemicals or just a few summaries that are sufficient to capture variation in the data? Can we separate the specimens into different regions just by using the variables or their summaries?

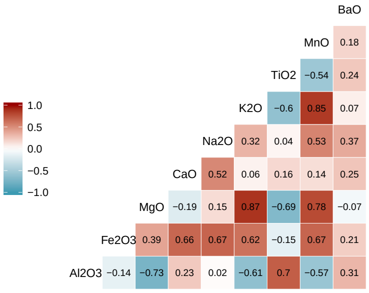

Figure 2: Correlation plot of the pottery data

### *Example 3: Students' ability data*

Let us consider the ability data[3], where the following six variables were recorded for 556 eighth-grade students:

SCA: self-concept of ability;
PPE: perceived parental evaluation;
PTE: perceived teacher evaluation;
PFE: perceived friend's evaluation;
EA: educational aspiration;
CP: college plans.

The correlation matrix of these variables is avalilable in the `MVA` package.

Calsyn and Kenny (1977) postulated that there are two latent variables (known as factors) that they designated as `ability` and `aspiration`, are responsible for the pattern observed in the correlation matrix. Specifically, the four variables, SCA, PPE, PTE, and PFE, are indicative of ability; the last two variables, EA and CP, are indicative of aspiration. The postulated model is shown in Figure 3.

How to formally test whether a pre-specified model fits the covariance among the variables well enough? If we do not have a postulated model yet, how do we explore the data to uncover such a model?

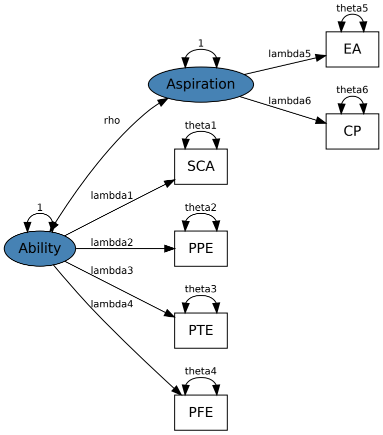

Figure 3: Postulated ability factor model.

[2] A. Tubb and N. J. Parker and G. Nickless (1980), The analysis of Romano-British pottery by atomic absorption spectrophotometry. Archaeometry, 22, 153-171.

[3] Calsyn, R. J., & Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement? Journal of Educational Psychology, 69(2), 136–145.

*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 differences 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 oversimplifying 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: Treatment of Lead Exposed Children (TLC) Trial*

The TLC trial[4] was a placebo-controlled, randomized study of a chelating agent (succimer) in children with blood lead levels of 20-44 micrograms/dL. Let us only consider a subsample of size 50 (N=50) from the children who received succimer. The dataset consists of four repeated measurements of blood lead levels obtained at baseline (week 0), week 1, week 4, and week 6 on each of the 50 children.

| ID | Week 0 | Week 1 | Week 4 | Week 6 |
|---|---|---|---|---|
| 1 | 26.5 | 14.8 | 19.5 | 21.0 |
| 2 | 25.8 | 23.0 | 19.1 | 23.2 |
| 3 | 20.4 | 2.8 | 3.2 | 9.4 |
| 4 | 20.4 | 5.4 | 4.5 | 11.9 |
| 5 | 24.8 | 23.1 | 24.6 | 30.9 |
| 6 | 27.9 | 6.3 | 18.5 | 16.3 |
| 7 | 35.3 | 25.5 | 26.3 | 30.3 |
| 8 | 28.6 | 15.8 | 22.9 | 25.9 |

*Table:* *A snapshot of TLC trial data.*

How does the average response (blood lead level) change over time? How can we predict the future of a new subject given previous measurements?

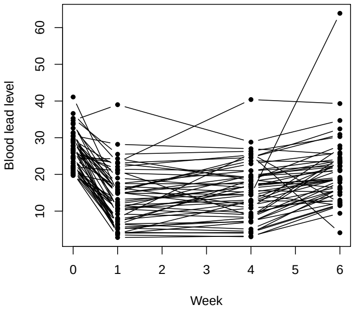

Figure 4: Blood lead levels of 50 Children over weeks

### *Example 2: Weight versus age of chicks on different diets*[5]

The dataset is available in R as `ChickWeight`. There were four groups of chicks on different protein diets. 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. The growth curves are shown in Figure 5.

| 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.*

Are the mean growth of the four groups the same at all the time points? If there is a difference, how does the difference change over time? In general, is there any relationship between growth and diet?

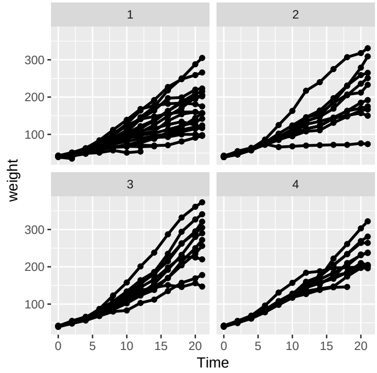

Figure 5: Figure: Growth curves of chickens – each panel corresponds to a particular type of diet

[4] The dataset and its description are available at [`https://content.sph.harvard.edu/fitzmaur/ala2e/`].

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

### *Characteristics of longitudinal data*

The same outcome/response is measured repeatedly on each unit (e.g., an individual, a plant, etc.). The condition of measurement is called generically *time*.[6]

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 the same results. Nevertheless, one could employ methods from multivariate statistics to analyze longitudinal data. Common questions of interest are

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.

[6] Do note that the repeated measures may correspond to other conditions than time, such as drug dosage (e.g., diastolic blood pressure measurements for several dose levels of an antihypertensive drug on the same subject), or height (diameter measurements at several height levels on the same tree).