Introduction to Applied Multivariate Analysis with R

Presented by Geoffrey Hubona

The Georgia R School

Session 1 Agenda: MV Data and Analysis

- **Introduction to MV Data and Analysis**
  - What is Multivariate Data?
  - Types of Variables and Missing Values
  - Example Multivariate Datasets
  - Covariances, Correlations, and Distances
  - The Multivariate Normal Density Function

Multivariate Data

We typically write multivariate datasets in a rectangular format:

Unit	Variable 1	...	Variable q
1	x_{11}	$\ldots$	x_{1q}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	$\ldots$	x_{nq}

where n is the number of units, q is the number of variables recorded on each unit, and x_{ij} denotes the value of the jth variable for the ith unit. The observation part of the table above is generally represented by an $n \times q$ *data matrix*, $\mathbf{X}$. In contrast to the *observed* data, the theoretical entities describing the univariate distributions of each of the q variables and their joint distribution are denoted by so-called *random variables* $X_1, \ldots, X_q$.

hypo Data

hypo data. Hypothetical set of Multivariate data:

individual	sex	age	IQ	depression	health	weight
1	Male	21	120	Yes	Very good	150
2	Male	43	NA	No	Very good	160
3	Male	22	135	No	Average	135
4	Male	86	150	No	Very poor	140
5	Male	60	92	Yes	Good	110
6	Female	16	130	Yes	Good	110
7	Female	NA	150	Yes	Very good	120
8	Female	43	NA	Yes	Average	120
9	Female	22	84	No	Average	105
10	Female	80	70	No	Good	100

`measure` Data

`measure` data. Chest, waist, and hip measurements on 20 individuals (inches):

chest	waist	hips	gender	chest	waist	hips	gender
34	30	32	male	36	24	35	female
37	32	37	male	36	25	37	female
38	30	36	male	34	24	37	female
36	33	39	male	33	22	34	female
38	29	33	male	36	26	38	female
43	32	38	male	37	26	37	female
40	33	42	male	34	25	38	female
38	30	40	male	36	26	37	female
40	30	37	male	38	28	40	female
41	32	39	male	35	23	35	female

pottery Data

pottery data (partial). Romano-British pottery data:

Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO	kiln
18.8	9.52	2.00	0.79	0.40	3.20	1.01	0.077	0.015	1
16.9	7.33	1.65	0.84	0.40	3.05	0.99	0.067	0.018	1
18.2	7.64	1.82	0.77	0.40	3.07	0.98	0.087	0.014	1
16.9	7.29	1.56	0.76	0.40	3.05	1.00	0.063	0.019	1
17.8	7.24	1.83	0.92	0.43	3.12	0.93	0.061	0.019	1
18.8	7.45	2.06	0.87	0.25	3.26	0.98	0.072	0.017	1
16.5	7.05	1.81	1.73	0.33	3.20	0.95	0.066	0.019	1
18.0	7.42	2.06	1.00	0.28	3.37	0.96	0.072	0.017	1
15.8	7.15	1.62	0.71	0.38	3.25	0.93	0.062	0.017	1
14.6	6.87	1.67	0.76	0.33	3.06	0.91	0.055	0.012	1
13.7	5.83	1.50	0.66	0.13	2.25	0.75	0.034	0.012	1
14.6	6.76	1.63	1.48	0.20	3.02	0.87	0.055	0.016	1

`exam` Data

`exam` data. Exam scores for five psychology students:

subject	maths	english	history	geography	chemistry	physics
1	60	70	75	58	53	42
2	80	65	66	75	70	76
3	53	60	50	48	45	43
4	85	79	71	77	68	79
5	45	80	80	84	44	46

USairpollution Data

USairpollution data (partial). Air pollution in 41 US cities:

	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
Chicago	110	50.6	3344	3369	10.4	34.44	122
Cincinnati	23	54.0	462	453	7.1	39.04	132
Cleveland	65	49.7	1007	751	10.9	34.99	155
Columbus	26	51.5	266	540	8.6	37.01	134

Covariances

The *covariance* of two random variables is a measure of their *linear* dependence. The population (theoretical) covariance of two random variables, X_i and X_j, is defined by

$$\mathsf{Cov}(X_i, X_j) = \mathsf{E}(X_i - \mu_i)(X_j - \mu_j),$$

where $\mu_i = \mathsf{E}(X_i)$ and $\mu_j = \mathsf{E}(X_j)$; E denotes expectation.

The covariance of X_i and X_j is usually denoted by σ_{ij}.

Covariances

In a multivariate data set with q observed variables, there are q variances and $q(q-1)/2$ covariances. These quantities can be conveniently arranged in a $q \times q$ symmetric matrix, $\mathbf{\Sigma}$, where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1q} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \dots & \sigma_q^2 \end{pmatrix}.$$

Note that $\sigma_{ij} = \sigma_{ji}$. This matrix is generally known as the *variance-covariance matrix* or simply the *covariance matrix* of the data.

For a set of multivariate observations, perhaps sampled from some population, the matrix $\mathbf{\Sigma}$ is estimated by

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top,$$

Correlations

The covariance is often difficult to interpret because it depends on the scales on which the two variables are measured; consequently, it is often standardised by dividing by the product of the standard deviations of the two variables to give a quantity called the *correlation coefficient*, ρ_{ij}, where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j},$$

where $\sigma_i = \sqrt{\sigma_i^2}$.

Euclidean Distance

$$d_{ij} = \sqrt{\sum_{k=1}^{q} (x_{ik} - x_{jk})^2},$$

where x_{ik} and $x_{jk}, k = 1, \ldots, q$ are the variable values for units i and j, respectively. Euclidean distance can be calculated using the `dist()` function in R.