Biostatistics 140.655

Lab 1

**Topics:**

* Multivariate Gaussian distribution
* Simulating multivariate Gaussian data
* Simulating multilevel Gaussian data from a multi-level longitudinal model
* Making autocorrelation plots for longitudinal data (MLLD)
* Estimating within-cluster measures of association (correlations)

**Learning Objectives:**

Students who successfully complete this lab will be able to:

* Work with the multivariate Gaussian distribution to represent means, variances, covariances and correlations among continuous random variables.
* Simulate multi-level (e.g. longitudinal) data with particular correlation structures
* Check simulated data to assure they have the designed means and covariances
* Make a pairs plot that displays the degree of association among repeated measurements on individuals over time
* Estimate and display an autocorrelation matrix and function ("correlogram")

**Multivariate Gaussian (aka "normal" - it is not) distribution:** the univariate Gaussian distribution can be extended to handle a vector of two ("bivariate") or more ("multivariate") random variables of length n. In this case, the mean vector also has length n and the variance/covariance matrix is n x n (n rows and n columns). The n diagonal elements of the square covariance matrix contain the variances for the n individual random variable, and the off-diagonal elements are the covariances for each of the "n choose 2" (n(n-1)/2) possible pairs of random variables.

For instance, the multivariate Gaussian distribution for is given by:

*1. What do the values and represent? Hand draw the distribution of labeling the mean and standard deviation in your figure. (10 pts)*

*2. What does the value represent? (10 pts)*

*3. Interpret the values , and ? (10 pts)*

Consider the multivariate Gaussian model for our application where, for each individual, we have 5 repeated measurements of the SF-36 mental health score at hospital discharge and then monthly for 4 months. The parameters associated with this multivariate (*n* = 5) Gaussian distribution comprise: 5 means, 5 variances and pairwise correlations.

Let be the SF-36 mental health score for subject i = 1, …, 100, at visit j = 1, 2, 3, 4, 5 corresponding to hospital discharge (*j = 1),* and subsequent monthly assessments (*j = 2, 3, 4, 5)*.

Assume the data for subject *i* are generated at random from a multivariate Gaussian distribution with: means 35, 38, 43, 49, 48 for j = 1, 2, 3, 4, 5; equal variances over time of 100; and correlation ρjk for times j and k of ρ12 = 0.85, ρ13 = 0.80, ρ14 = 0.72, ρ15 = 0.69, ρ23 = 0.85, ρ24 = 0.80, ρ25 = 0.72, ρ34 = 0.85, ρ35 = 0.80, ρ45 = 0.85.

4. *In the space below, practice your understanding of the notation by writing out the multivariate (n=5) Gaussian distribution. (10 pts)*

5. *Describe in a sentence or two the pattern you observe in the correlations for pairs of observations. Use the "autocorrelation matrix" (ACM) above to obtain the "autocorrelation function "(ACF): . Briefly describe the approach/method you used to obtain the ACF from the ACM. Given your lay understanding of the SF-36 measure, how is nature likely to work to produce the ACF pattern you observe. (10 pts)*

**Simulating multivariate Gaussian data.** Below find STATA and R code for generating n independent random draws from a multivariate Gaussian distribution with the mean and covariance matrix you specify.

**STATA:**

\* Set the random seed so we can replicate our generated data

**set** seed 12275

\* Define the mean, standard deviation and correlation matrix

**matrix** **m** = (35, 38, 43, 49, 48)

**matrix** sd = (10,10,10,10,10)

**matrix** C = (1, 0.85, 0.80, 0.72, 0.69, 1, 0.85, 0.80, 0.72,

1, 0.85, 0.80, 1, 0.85, 1)

\* Generate the k-variate normal data

**drawnorm** y0 y1 y2 y3 y4, **n**(100) corr(C) cstorage(upper) **means**(**m**) sds(sd)

**gen** id = \_n

**R:**

**library**(mvtnorm)

set.seed(02022022) # Important for reproducibility

mm <- c(35, 38, 43, 49, 48)

C <- matrix(c(1.00,0.85,0.80,0.72,0.69,

0.85,1.00,0.85,0.80,0.72,

0.80,0.85,1.00,0.85,0.80,

0.72,0.80,0.85,1.00,0.85,

0.69,0.72,0.80,0.85,1.00),

nrow = 5)

sigma <- C \* 100

y <- rmvnorm(n=100, mean=mm, sigma=sigma)

id <- seq(1,100)

dat <- as.data.frame(cbind(y,id))

names(dat) <- c("y0","y1","y2","y3","y4","id")

6. *Simulate 100 random draws from the multivariate Gaussian distribution for the SF-36 data whose true mean and covariance matrix is specified above (and in the code). Report the first 5 draws in the space below. (10 pts)*

7. *Display your simulated data using a spaghetti plot and a pairs plot.* *Calculate the mean and covariance matrix for your sample of n=100 simulated vectors. Use these plots and estimates to describe the patterns of potential scientific interest in the SF-36 data. (10 pts)*

8. *Calculate the autocorrelation matrix and autcorrelation function for your simulated data. Describe what else these estimates contribute to our understanding of the SF-36 data. (10 pts)*

*9. Now generate a new data set in a multilevel manner, producing a 100 x 5 matrix, call it Y. (10 pts)*

To generate each row, 5x1 vectors for each person , we use the equation

where is the mean value for theth random variable common to all 100 people (no subscript ) (we will use from above), is a scalar "random intercept", common to all 5 observations for person , but different among people, and is a residual or "error" specific to observation for person . We assume that: (1) the 100 ’s are independent and identically distributed ("IID") draws from a Gaussian distribution with mean 0 and variance 120; (2) the 500=100x5 ’s are iid Gaussian with mean 0 and variance 100, and that (3) the ’s are independent of the ’s.

(1) generate a 100 x 5 matrix of ’s by drawing 100 realizations from a 5-dimensional Gaussian distribution with mean 0 and covariance matrix that is diagonal with 100 on the main diagonal and 0 off the diagonal. Because the covariances are all 0 and the data are multivariate Gaussian, the repeated observations for an individual are independent of one another.

(2) generate a vector of 100 independent univariate Gaussian variates with mean zero and variance 120. These values are the 100 "random intercepts", one for each person (row).

(3) Create the 100 x 5 matrix of Y values by adding the mean vector (repeated over rows), random intercepts (repeated over columns), and the random errors matrix.

9. *Again,* *display your simulated data using a spaghetti plot and pairs plot. (10 pts)*

*10. Calculate the autocorrelation matrix and autocorrelation function for these simulated data. (10 pts)*

*11. Using the SF-36 example, explain the scientific meaning of the means, random intercepts, variance of the random intercepts, residuals, and variance of the residuals. (10 pts)*