Workshop 1

Callum Abbott

02/10/2020

Problem 3

In this exercise we will simulate data and investigate the impact of different mechanisms of missingness.

Consider $Y = (Y1, Y2, Y3)$, to be simulated from a standard trivariate normal distribution with correlations $\rho_{1,2}$, $\rho_{1,3}$ and $\rho_{2,3}$ all equal to 0.5.

Missingness will then be imposed on $Y2$, while $Y1$ and $Y3$ remain fully observed. Additionally, let R be the missingness indicator, taking the value 1 for observed values and 0 for missing values. In the following consider $n = 500$.

```
set.seed(42)
require(MASS)

# Set parameters given by question.
n <- 500;
mu1 <- mu2 <- mu3 <- 0;
sigma1 <- sigma2 <- sigma3 <- 1;
rho <- 0.5

# Form the covariance matrix
Sigma = matrix(
  c(sigma1^2, rho*sigma1*sigma2, rho*sigma1*sigma3,
    rho*sigma2*sigma1, sigma2^2, rho*sigma2*sigma3,
    rho*sigma3*sigma1, rho*sigma3*sigma2, sigma3^2),
    nrow=3, ncol=3, byrow=TRUE)

# Sample from multivariate Gaussian.
Y <- mvrnorm(n, mu=c(mu1, mu2, mu3), Sigma=Sigma)

# Look at first 10 rows of dataset.
Y[1:10,]
```

```
##              [,1]       [,2]       [,3]
##  [1,]  0.2229903 -2.3051398 -1.2759991
##  [2,]  0.7636762 -0.1476143  0.7671605
##  [3,]  0.2639601 -0.5754916 -0.5779478
##  [4,] -0.2990845 -0.6935578 -0.5575482
##  [5,] -0.9050861  0.3174943 -0.4026593
##  [6,] -0.2583066  0.3581909  0.1600666
##  [7,] -1.1387446 -0.7672521 -1.7964609
```

```
##  [8,] -1.6134683  1.4061453  0.4391894
##  [9,] -2.1375796 -0.7928578 -2.0136709
## [10,]  0.5122688 -0.5974295  0.2387782
```

```
# Storing and rounding the simulated values in three variables
Y1 <- round(Y[,1]); Y2 <- round(Y[,2]); Y3 <- round(Y[,3])
mean(Y1); mean(Y2); mean(Y3); sd(Y1); sd(Y2); sd(Y3)
```

```
## [1] 0.02
```

```
## [1] 0.024
```

```
## [1] 0.026
```

```
## [1] 1.010764
```

```
## [1] 1.032257
```

```
## [1] 1.052397
```

Exercise 3a

We now wish to impose the MCAR missing data mechanism on the data by considering $Pr(R = 0|Y1, Y2, Y3, \beta) = 0.65$ and then observe the densitites of the complete $Y2$ values, the observed $Y2$ values (after imposing MCAR) and the missing $Y2$ values. We note that R is a binary value and hence can be simulated using a binomal distribution with the probability parameter set to 0.65.

```
theta <- 0.65
r_mcar <- rbinom(n, size=1, prob=theta)
```