Assignment 2 - PH242C/STAT247C

John Semerdjian

September 28, 2015

1. Notation

1.1 What is the value of m (number of independent units)?

10

1.2 What is a good notation for the outcome, Y, including the appropriate indices, and the range of each index, e.g., $i = 1, ..., m, j = 1, ...,$? (This may take more subscripts that we've used in examples in class.)

Let Y_{ijkl} denote the outcome of visit l for child k in household j in village i.

$$\begin{aligned} \text{village} = i &\in \{1...10\} \\ \text{household} = j &\in \{1...35\} \\ \text{children} = k &\in \{1...m_{ik}\} \\ \text{visit} = l &\in \{1, 2, 3\} \end{aligned}$$

1.3 Provide a logistic model that relates the probability of diarrhea for a particular child at one of the visits given the two explanatory variables listed above. Assume that all observations have the same relationship (coeffcients) with the explanatory variables.

$$\text{logit}\{\text{E}(Y_{ijkl} = 1|X_{ijkl})\} = \beta_0 + \beta_1^{water} x_{1,ij} + \beta_2^{travel} x_{2,i}$$

1.4 Expand this model to allow the associations of water treatment to differ randomly by household, and the association of movement to vary randomly by village. (However you do this, be explicit about the definitions of your notation)

$$\begin{aligned} \text{logit}\{\text{E}(Y_{ijkl} = 1|X_{ijkl})\} = &\beta_0 + \beta_1^{water} x_{1,ij} + \beta_2^{travel} x_{2,i} \\ &+ \gamma_1^{water} d_{1,ij}^{house} + ... + \gamma_{350}^{water} d_{350,ij}^{house} \\ &+ \nu_1^{travel} d_{1,i}^{village} + ... + \nu_{10}^{travel} d_{10,i}^{village} \end{aligned}$$

I used dummy variable notation for adding water treatment and movement variables to variable for household and village, respectively.

- $\gamma_{ij}^{water} d_{ij}^{house}$ represents the effect of water treatment on household j in the i^{th} village. There are a total of 350 individual households to estimate.
- $\eta_i^{travel} d_i^{village}$ represents the effect of travel/movement in village i. There are 10 independent villages to estimate.

2. Data-generating distributions for repeated measures data

2.1 Simulate data in STATA from the model implied by the data-generating description in Question 1.4. Assume the random variables are normally distributed (every thing else is provided). Also, assume the following parameters:

- $\sigma_\alpha^2 = 0.5$
- $cor(Y_{ij}, Y_{ij'}) = \rho = 0.3$
- $\mu = EY_{ij} = 10$

2.2 Simulate data at sample sizes of m = 20, 100 and 1000 always with $n_1 = n_2 = ...n_m$ For each of these simulations, estimate ρ and μ. Turn in the following:

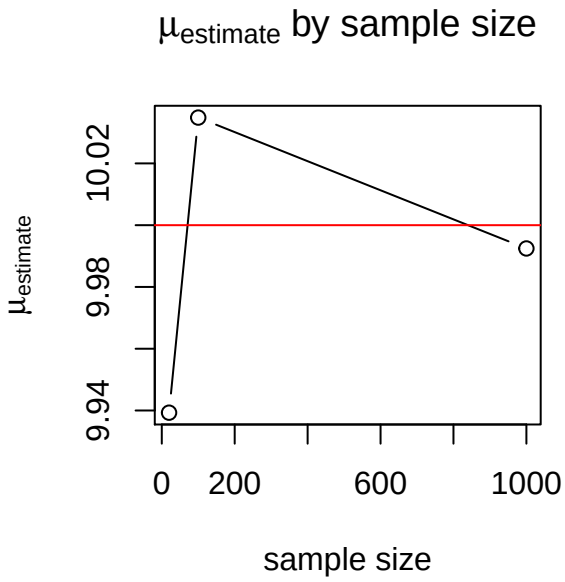
1. Stata code used to generate simulation and estimates of ρ and μ.
2. Plot of these estimates versus the sample size, m, separately for ρ and μ including putting a horizontal line for the true value of these.
3. Short explanation of what these plots show.

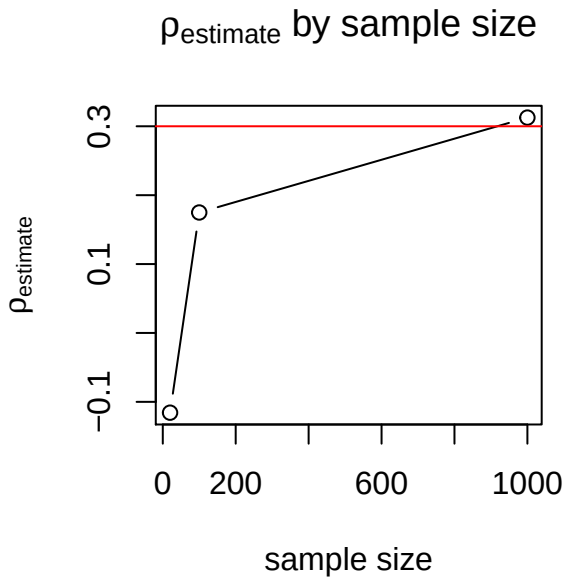
The plots show the relationship between increasing sample size and the effects on μ and ρ. The red line indicates their true values. As we increase sample size, the closer our results get to the true value. See attached code. Output is below:

```
set.seed(247)

assignment2 = function(n=5000, numlist=c(20, 100, 1000)) {
  id = 1:n
  mu = 10
  rho = 0.3
  sigma_alpha = sqrt(0.5)
  sigma_e = sqrt((sigma_alpha^2/rho) - sigma_alpha^2)

  alpha = rnorm(n, 0, sigma_alpha)
  e1 = rnorm(n, 0, sigma_e)
  e2 = rnorm(n, 0, sigma_e)

  Y_time1 = mu + alpha + e1
  Y_time2 = mu + alpha + e2

  cor = sigma_alpha^2/(sigma_alpha^2 + sigma_e^2)
  cat("cor = ", cor, "\n")

  # combine data into wide format
  df = data.frame(id, alpha, e1, e2, Y_time1, Y_time2)

  mu_est_vec = NULL
  rho_est_vec = NULL
  for(i in 1:length(numlist)) {
    k = numlist[i]
    cat("m =", k, "\n")

    mu_est = mean(c(df[1:k, "Y_time1"], df[1:k, "Y_time2"]))
    mu_est_vec[i] = mu_est
```

```
    cat("  mu_est  =", mu_est, "\n")

    rho_est = cor(df[1:k, "Y_time1"], df[1:k, "Y_time2"])
    rho_est_vec[i] = rho_est
    cat("  rho_est =", rho_est, "\n")
  }

  # plots
  par(mfrow=c(1,2))
  plot(x=numlist,
       y=mu_est_vec,
       type="b",
       main=expression(paste(mu[estimate], " by sample size")),
       xlab="sample size",
       ylab=expression(mu[estimate]))
  # add true value
  abline(h=mu, col="red")

  plot(x=numlist,
       y=rho_est_vec,
       type="b",
       main=expression(paste(rho[estimate], " by sample size")),
       xlab="sample size",
       ylab=expression(rho[estimate]))
  # add true value
  abline(h=rho, col="red")
}

assignment2()
```

```
## cor =  0.3
## m = 20
##   mu_est  = 9.939293
##   rho_est = -0.1157341
## m = 100
##   mu_est  = 10.03484
##   rho_est = 0.1748438
## m = 1000
##   mu_est  = 9.992465
##   rho_est = 0.3126366
```


μestimate by sample size
μestimate
10.02
9.98
9.94
0
200
600
1000
sample size


ρestimate by sample size
ρestimate
0.3
0.1
−0.1
0
200
600
1000
sample size