Solutions to Assignment 2
Longitudinal Data (PH242C: STAT247C)

1 Notation

Consider a clustered, longitudinal study of diarrheal disease in children in Ecuador [Eisenberg et al., 2006]. Though not precisely how this study was done, assume that it was a study a a) random sample of 10 *independent* villages, b) within those villages, a random sample of 35 households that contained children under the age of 5, and c) within the households, all children under the age of 5. At three annual visits, households data were collected on:

1. the (binary) outcome of diarrhea over the last week for each child
2. whether or not any water treatment was currently being used at the household level, and
3. the proportion of villagers that traveled to the closest main town over the past month (village-level "movement" variable).

Being as close as you can to the notation in chapter 1 (though, you can define things differently as long as your are explicit about the definitions)

- What is the value of m (number of independent units)?
  - 10 villages.
- What is a good notation for the outcome, Y, including the appropriate indices, and the range of each index, e.g., $i = 1, ..., m, j = 1, ...,$?
  - The outcome, water treatment and travel could be notated as $(Y_{ijkl}, X_{ij.l}, Z_{i..l})$, where the "." notation indicates that it does not change over that index, with:
    - $i \in \{1, \ldots, 10\}$ is the village identifier

    - $j \in \{1, \dots, 35\}$ is the household identifier
    - $k \in \{1, \dots, n_{ij}\}$ is the child identifier (in that house in that village)
    - $l \in \{1, 2, 3\}$ is the annual visit number

- Provide a (symbolic) logistic (logit-linear) model that relates the probability of diarrhea for a particular child at one of the three visits given the two explanatory variables listed above - assume that all observations have the same relationship (coefficients) with the explanatory variables.
  - $logit[\mathbb{E}(Y_{ijk}|X_{ij.l}, Z_{i..l}] = \beta_0 + \beta_1 X_{ij.l} + \beta_2 Z_{i..l}$

- Expand this same model to allow the associations of water treatment to differ randomly by household, and the association of movement to vary randomly by village.
  - $logit[\mathbb{E}(Y_{ijk}|X_{ij.l}, Z_{i..l}] = \beta_0 + \beta_{1ij} X_{ij.l} + \beta_{2i} Z_{i..l}$

2 Data-generating distributions for Repeated Measures Data

- Question 1.3 in chapter 1 (Jewell and Hubbard).

We want to show that $\text{cor}(Y_{i1}, Y_{i2}) = \rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}$. We are given that $\text{cor}(Y_{i1}, Y_{i2}) = \frac{\text{cov}(Y_{i1}, Y_{i2})}{\text{SD}(Y_{i1})\text{SD}(Y_{i2})}$.

Let's first show that $\sigma_\alpha^2 = \text{cov}(Y_{i1}, Y_{i2})$:

$$\text{cov}(Y_{i1}, Y_{i2}) = \mathbb{E}\left[(Y_{i1} - \mathbb{E}Y_{i1})(Y_{i2} - \mathbb{E}Y_{i2})\right].$$

Since $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$,

$$\mathbb{E}(Y_{ij}) = \mathbb{E}(\mu + \alpha_i + \epsilon_{ij}) = \mathbb{E}(\mu) + \mathbb{E}(\alpha_i) + \mathbb{E}(\epsilon_{ij}),$$

because the expectation of the sum is the sum of the expectations. Thus $\mathbb{E}(Y_{ij}) = \mu$, since $\mathbb{E}(\mu) = \mu$, $\mathbb{E}(\alpha_i) = 0$, and $\mathbb{E}(\epsilon_{ij}) = 0$. Note that the expectation of Y_{ij} is the same for all i and j so $\mathbb{E}(Y_{i1}) = \mathbb{E}(Y_{i2}) = \mu$. Note also that $Y_{i1} = \mu + \alpha_i + \epsilon_{i1}$ and $Y_{i2} = \mu + \alpha_i + \epsilon_{i2}$. Thus

$$\text{cov}(Y_{i1}, Y_{i2}) = \mathbb{E}\left[(\mu + \alpha_i + \epsilon_{i1} - \mu)(\mu + \alpha_i + \epsilon_{i2} - \mu)\right]$$

$$= \mathbb{E}\left[\alpha_i^2 + \alpha_i\epsilon_{i1} + \alpha_i\epsilon_{i2} + \epsilon_{i1}\epsilon_{i2}\right]$$

$$= \mathbb{E}(\alpha_i^2) + \mathbb{E}(\alpha_i\epsilon_{i1}) + \mathbb{E}(\alpha_i\epsilon_{i2}) + \mathbb{E}(\epsilon_{i1}\epsilon_{i2}).$$

Note that since $0 = \text{cov}(\alpha_i, \epsilon ij) = \mathbb{E}(\alpha_i\epsilon_{ij}) - \mathbb{E}(\alpha_i)\mathbb{E}(\epsilon_{ij}) \Rightarrow \mathbb{E}(\alpha_i\epsilon_{ij}) = 0$ and thus the middle two terms in the expression are zero. Similarly, since $0 = \text{cov}(\epsilon_{i1}, \epsilon_{i2})$ the last term is also zero. Thus

$$\text{cov}(Y_{i1}, Y_{i2}) = \mathbb{E}(\alpha_i^2).$$

Recall that $\text{var}(X) = \text{cov}(X, X)$ and so

$$\text{var}(\alpha_i) = \text{cov}(\alpha_i, \alpha_i)$$

$$= \mathbb{E}(\alpha_i\alpha_i) - \mathbb{E}(\alpha_i)\mathbb{E}(\alpha_i)$$

$$= \mathbb{E}(\alpha_i^2) - 0 * 0$$
$$= \mathbb{E}(\alpha_i^2).$$

Thus

$$\text{cov}(Y_{i1}, Y_{i2}) = \text{var}(\alpha_i) = \sigma_\alpha^2.$$

Now let's show that $\text{SD}(Y_{i1})\text{SD}(Y_{i2}) = \sigma_\alpha^2 + \sigma_\epsilon^2$.

$$\text{SD}(Y_{i1})\text{SD}(Y_{i2}) = \sqrt{\text{var}(Y_{i1})}\sqrt{\text{var}(Y_{i2})}.$$

Note that $\text{var}(Y_{ij}) = \text{var}(\mu + \alpha_i + \epsilon_{ij})$. Since constants do not change the value of the variance this is equal to $\text{var}(\alpha_i + \epsilon_{ij})$. Since $\text{cov}(\alpha_i, \epsilon_{ij}) = 0$, $\text{var}(\alpha_{ij} + \epsilon_{ij}) = \text{var}(\alpha_i) + \text{var}(\epsilon_{ij})$. Therefore we have

$$\text{SD}(Y_{i1})\text{SD}(Y_{i2}) = \sqrt{\text{var}(\alpha_i) + \text{var}(\epsilon_{i1})}\sqrt{\text{var}(\alpha_i) + \text{var}(\epsilon_{i2})}$$
$$= \sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2}\sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2}$$
$$= \sigma_\alpha^2 + \sigma_\epsilon^2.$$

Putting together the numerator and denominator we have proven our desired result.

- Simulate data in STATA from the model implied by the data-generating description in Question 1.3. Assume the random variables are normally distributed (every thing else is provided). Also, assume the following parameters:

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

- Simulate data at sample sizes of $m = 20, 100$ and 1000 always with $n_1 = n_2 = \ldots n_m = 2$. 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.

```
**** Simulation of Simple Random Effects Model
 clear
 set obs 5000 /*m=1000 */
**** Generate Random N(0,sigma_alpha) variable
 scalar sigmaalpha=sqrt(0.5)
 scalar rho = 0.3
 scalar sigmae = sqrt((sigmaalpha^2/rho)-sigmaalpha^2)
 gen alpha = rnormal(0,sigmaalpha)
**** Generate id
 gen id = _n
**** Give two observations for every person
 expand 2
**** Make "time" variable
 sort id
 by id: gen time = _n
**** Grand mean = 10
 gen mu = 10
**** independent random errors
  ** Generate Random independent eij~N(0,sigmae) variables (see solutions for sigma_e =
 gen epsilon = rnormal(0,sigmae)
**** Make outcome variables, Yij
 gen Y = mu+alpha+epsilon
 ** Distri of outcome
 twoway hist Y
*** True Correlation (should equal rho)
scalar cor = sigmaalpha^2/(sigmae^2+sigmaalpha^2)
display "cor = " cor
**** Break up samples chunks so that we can get estimate for m=10,50,100,1000 by using I
```

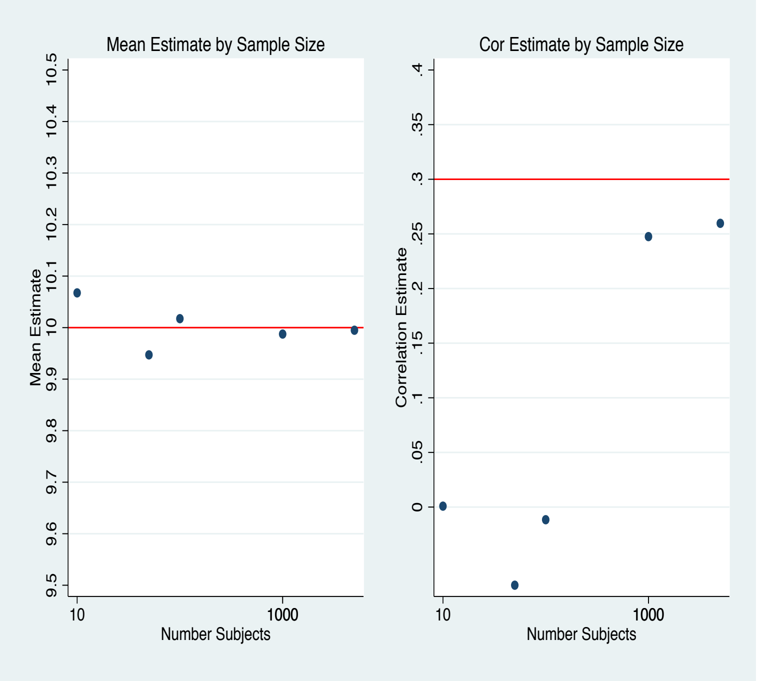
```
*************** (A little trick)
gen muest = -1000
gen corest = -1000
foreach k of numlist 10 50 100 1000 5000 {
display " m = " ‘k’
capture drop Ytemp
gen Ytemp = Y
**** Make all values > m of interest = . to restrict sample size
replace Ytemp = . if id > ‘k’
capture drop mutemp
**** Estimate mean, and store value at id = m of interest
egen mutemp = mean(Ytemp)
replace muest = mutemp if id == ‘k’
**** Estimate cor, and store value at id = m of interest
reshape wide Y Ytemp epsilon, i(id) j(time)
cor Ytemp1 Ytemp2
scalar cortemp = r(rho)
replace corest=cortemp if id == ‘k’
reshape long Y Ytemp epsilon, i(id) j(time)
}
reshape wide Y Ytemp epsilon, i(id) j(time)
*** Get rid of all observations except those with estimates to plot
keep if muest > -999
label variable id "Number Subjects"
label variable corest  "Correlation Estimate"
label variable muest "Mean Estimate"
scatter muest id, t1("Mean Estimate by Sample Size") saving(graph1, replace) xlabel(10 1
scatter corest id, t1("Cor Estimate by Sample Size") saving(graph2, replace) xlabel(10 1
gr combine graph1.gph graph2.gph
```

The figure shows that on average, the estimates of μ and ρ get closer to the true values, 10 and .3, respectively, as the number of subjects increases.

References

Joseph NS Eisenberg, William Cevallos, Karina Ponce, Karen Levy, Sarah J Bates, James C Scott, Alan Hubbard, Nadia Vieira, Pablo Endara, Mauricio Espinel, et al. Environmental change and infectious disease: how new roads affect the transmission of diarrheal pathogens in rural ecuador. *Proceedings of the National Academy of Sciences*, 103(51):19460–19465, 2006.

Figure 1: Estimate of μ and ρ by m


Mean Estimate by Sample Size
Mean Estimate
Number Subjects
Cor Estimate by Sample Size
Correlation Estimate
Number Subjects