BIOS6643 Fall 2022 HW1 Due Friday, Sep. 16, 5pm (upload to Canvas)

General expectations for homework

Since I have taught BIOS6643 for over a decade, I have developed many homework exercises and some I reuse from year to year, possibly after some modification. Although I usually mix in new homework questions as well, some of the long-standing exercises have been 'battle tested', refined and are still interesting and relevant. I like to add newer questions as well, although the difficulty with new questions is that they have not been refined and tested, so I often have kinks to work out compared with the longer-standing ones. <u>I expect all students in the current class to not review any homework solutions from previous years, from this point forward (honor code)</u>. However, I don't mind students in the current course discussing homework questions and possibly 'bouncing ideas' of each other. This does not mean that you 'watch' someone's solutions and repeat it, I expect that you will all be active participants in completing your own homework. The more you put into it, the more you get out of it.

General instructions for format and submission

You can use an editing program of your choice (e.g., Word of Google Docs) to compile typed solutions and graphs and output using SAS or R. For your main turn-in, please streamline computer output, including only the relevant pieces. You can include SAS or R code and more detailed output in an Appendix (which may be for your benefit as much as mine, in terms of review later). You can embed output or tables into your text answers. If you need to attach any handwritten material, that is fine as long as it is clearly legible; just scan and attach it, keeping the questions in order as much as you can. <u>Please submit your entire homework in one document, and upload a PDF to Canvas</u>.

A. <u>For practice (do not turn in): A random walk model</u>. Consider the random walk defined by $Y_t = Y_{t-1} + B_t$, where $B_t = 1$ with probability ½ and –1 with probability ½ (B_t, t=1,2,… are *iid*) and $Y_0 = 0$. Let t and h be nonnegative integers.
   a. Determine $E(Y_t)$
   b. Determine $Cov(Y_t, Y_{t+h})$
   c. Determine $Corr(Y_t, Y_{t+h})$
   d. Is $\{Y_t\}$ a stationary process?
   a. Determine $E(Y_t)$
   b. How do answers in a-d change when considering $0 \le p \le 1$ rather than just p=½?

<u>To turn in</u>:

(1) Consider a first-order autoregressive process, $\varepsilon_t = \phi\ \varepsilon_{t-1} + Z_t$, where $Z_t \sim N(0, \sigma^2)$, where t is an integer for discrete units of time (e.g., days), and $|\phi|<1$. In order to derive the quantities below, say that this is an 'infinite process' (i.e., t extends backwards in time to infinity). First, by iteration we can show that $\varepsilon_t = Z_t + \phi Z_{t-1} + ... + \phi^k Z_{t-k} + \phi^{k+1}\varepsilon_{t-k-1}$. If we keep going, we get the expression $\varepsilon_t = \sum_{j=0}^{\infty}\phi^j Z_{t-j}$. [We can show that this equality holds since $\sum_{j=0}^{k}\phi^j Z_{t-j}$ is mean-square convergent as $k\to\infty$:
$E(X_t - \sum_{j=0}^{k}\varphi^j Z_{t-j})^2 = \varphi^{2k+2}\ E(X_{t-k-1}^2) \xrightarrow{k\to\infty} 0$ since $E(X_t^2)$ is constant over t.]
   a. Determine $E(\varepsilon_t)$
   b. Determine $Cov(\varepsilon_t, \varepsilon_{t+h})$
   c. Determine $Corr(\varepsilon_t, \varepsilon_{t+h})$
   d. Is $\{\varepsilon_t\}$ a stationary process?

(2) The subject-level linear mixed model is $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$. Consider a special case of this, the 'random intercept model' that can be expressed at the observation level as $Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_{p-1}x_{p-1,ij} + b_i + \varepsilon_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + b_i + \varepsilon_{ij}$, where i denotes subject and j denotes time, and $\mathbf{X}_{ij}$ is the row in $\mathbf{X}$ corresponding to the ij^{th} observation. Errors $\varepsilon_{ij} \sim iid\ N(0, \sigma_\varepsilon^2)$ are independent of $b_i \sim iid\ N(0, \sigma_b^2)$. Note that $Var(\boldsymbol{\varepsilon}_i) = \mathbf{R}_i$, based on the covariance structure defined. Review the marginal and conditional distributions shown in the slides in order to determine the following. It is sufficient to write out the 1st, 2nd, and last (or r^{th}) elements of the matrices. For example, at the subject level, we can write

$$\mathbf{R}_i = \begin{pmatrix} \sigma_\epsilon^2 & 0 & \cdots & 0 \\ 0 & \sigma_\epsilon^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_\epsilon^2 \end{pmatrix}$$

Note that the general formulas for the quantities below are easily derived by considering the mean and variance of the distributions of interest (either $\mathbf{Y}_i$ or $\mathbf{Y}_i \mid \mathbf{b}_i$). You then just customize the quantities for the random intercept model.

a. $E(\mathbf{Y}_i)$
b. $E(\mathbf{Y}_i \mid \mathbf{b}_i)$
c. $Var(\mathbf{Y}_i)$
d. $Corr(\mathbf{Y}_i)$
e. $Var(\mathbf{Y}_i \mid \mathbf{b}_i)$

A few notes: (i) For a vector of responses (e.g., $r\times 1$ vector $\mathbf{Y}_i$ for subject i and their repeated measures), the covariance matrix has dimension $r\times r$; (ii) The covariance matrix is often either called '*Var*' or '*Cov*'; they are used interchangeably; (iii) The correlation matrix, '*Corr*' is simply a standardized version of the covariance matrix; (iv) There are useful identities/formulas for mean, variance, covariance and correlation that you learn in 6631/32 (theory) that will be useful here. These should essentially be review questions for those courses. If you did not take that sequence or need a refresher, we can discuss more in class and/or office hours.