Q4 Physical activity is known to confer a number of health benefits and remaining active is a key component of healthy aging. However, many individuals do not meet the recommended guidelines for engaging in higher intensity activities (moderate-to-vigorous physical activity, MVPA). A randomized trial was conducted in an older, sedentary population where participants' were randomized to one of two trial arms. In the intervention arm of the trial, participants were enrolled in an educational program about the benefits of regularly engaging in MVPA and provided with an exercise program tailored to their physical status (i.e. exercises which avoid exacerbating existing injuries and are compatible with any pre-existing physical limitations). After the intervention, participants were sent home and instructed to wear a wrist-worn accelerometer which tracked their MVPA for one year. The wearable accelerometer summarized participants' physical activity in minutes of MVPA every three days, resulting in a maximum of 122 repeated observations per person. All participants began wearing the accelerometer on January 1 of the same year. The goal of the study is to estimate the effect of the intervention on participants' physical activity.

The amount and type of physical activity individuals engage in is associated with the weather, and thus season. There is therefore a presumed seasonal effect in the physical activity trends. **The investigators believe that the intervention may have an effect on average MVPA which varies both in overal average and by season**. MVPA every three days is believed to follow a Poisson distribution.

For the following questions, use Poisson regression with a log link function. Suppose population average the seasonal effect takes the form of $\sin(2\pi s)$ for $s \in [0, 1]$ where $s = \text{day}/365$ (day of the year re-scaled to be between 0 and 1) on the scale of the linear predictor (link scale), and that the presumed seasonal intervention effect takes the same functional form. Note that observations within an individual (MVPA over the year) are correlated. In particular, individuals' physical activity trends are also believed to have seasonal variation in their physical activity which can be characterized on the linear predictor scale as: $b_{i1}\sin(2\pi s)+b_{i2}\cos(4\pi s)$, $[b_{i1}, b_{i2}]^t \sim N(\mathbf{0}_{2\times 1}, \mathbf{\Sigma}_b)$ with $\mathbf{\Sigma}_b = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$. Answer the following questions using the "MVPA_seasonal.csv" data file provided with this exam. The data contains four columns: ID (participant identifier), Y (minutes of MVPA), day (day of year), trt (0 control, 1 intervention).

(a) Write out the mean model which allows investigators to test their hypotheses of interest (the model which does not include subject-specific effects).

(b) Write out the conditional (conditional on subject-specific effects) implied by the prompt.

(c) As mentioned in the prompt, these data are correlated within individuals. Visualize the within person correlation structure of the observed data (correlation **and** covariance on the response scale). Comment on your findings.

(d) Not all participants have data complete data due to a combination of device malfunction and non-compliance with wearing of the accelerometer. Following the trial, the investigators surveyed participants and discovered that participants reported fatigue from wearing the device every day. In addition, some participants noted that the strap used to secure the device to the wrist broke following high intensity movement. Investigate patterns of missing data and comment on your findings. Addresses whether missing data patterns vary by trial arm, follow-up, and historical data. Justify your response using 1-2 tables and/or figures prepared as if for a scientific journal using captions and labels as appropriate (i.e. do not just copy and paste software/regression output).

(e) Two common methods for analyzing correlated count data are generalized estimating equations (GEE) and generalized linear mixed models (GLMM). Given the question prompt and your findings from the previous questions, write out both a GEE and GLMM model which would be appropriate for the data. Which model would you recommend based on the data and your findings in the previous questions? Justify your decision.

(f) Estimate both GEE and GLMM models provided in your previous answer using the your software of choice (e.g. `geepack` and `lme4` packages in `R`, respectively). For GEE, choose from one of: independence, exchangeable, or AR(1) (**do not chooce unstructured covariance for this problem**) and justify your choice, given the available options. Write out both models mathematically. Present your results as if for a scientific journal. Summarize and interpret the output from both models.

(g) Choices for the correlation/covariance structure for GEE models is limited in current software. Assume the GLMM model correctly characterizes the association structure on the linear predictor scale. This implies a covariance structure on the response scale which may be used in GEE estimation.

   i. Treating the output from the GLMM as the fixed "truth", derive the implied covariance (and correlation) on the response scale separately for each treatment group. Note: ****Do NOT attempt this here**, but this result could be used to create a custom set of GEE for estimating model parameters**.

   ii. Does your finding from (i.) match with the observed covariance/correlation in the data? Comment on any differences you see and propose an explanation for said differences. Justify your response using 1-2 tables and/or figures prepared as if for a scientific journal using captions and labels as appropriate.

(h) Write a brief paragraph summarizing your methods and results for the study investigators. Present specific results from the method you believe is most appropriate for the goals of the investigators.