STAT 437 - Assignment 2
Due: Friday, February 11 on Crowdmark
Resubmit: Friday, March 4 on Crowdmark

This assignment covers the materials contained in **Lecture 010** through to **Lecture 020**. Reminder that you are permitted to discuss to these problems with classmates, but every student must submit their own solutions which are their own work (including any code, figures, etc.). **Please indicate any students that you discussed solutions with on your submission**. Please ensure that your submissions on Crowdmark are legible, and separated based on the problems included at the submission link. Submissions can be handwritten or typeset.

Part 1: True or False (20 Marks)

For each of the following problems indicate whether the statement is true or false, and give a short justification for your answer. Correct answers without justification will receive only partial credit.

PROBLEM 1. Suppose that a linear mixed effects model is specified for a continuous outcome, Y_{ij}, such that

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 A_i + b_{0,i} + b_{1,i} A_i + \epsilon_{ij}.$$

Here, t_{ij} represents the time, treated as a continuous variable and $A_i \in \{0, 1\}$ is a binary treatment indicator. Assume that $G_i = \sigma^2 I$. **True or false: The within-person correlation structure can be written as $A_i \mathbf{R}(\rho_1) + (1 - A_i)\mathbf{R}(\rho_2)$, where $\mathbf{R}(\rho)$ is a correlation matrix which assumes compound symmetry.**

PROBLEM 2. Suppose that we have observed data $\{W_i, Z_i\}$ for an independent sample, $i = 1, \ldots, n$ where W_i is a binary indicator and Z_i is a discrete random variable, with values $\{-1, 0, 1\}$. Consider the M-estimator given by

$$U(\theta) = \sum_{i=1}^{n} \begin{pmatrix} I(Z_i = -1)(W_i - \theta_1) \\ I(Z_i = 0)(W_i - \theta_2) \\ 1 - \theta_1 - \theta_2 - \theta_3 \end{pmatrix}.$$

Denote $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3)$ as the solution to $U(\widehat{\theta}) = 0$. **True or false: $\widehat{\theta}_3$ is consistent for $P(W_i = 1 | Z_i = 1)$.**

PROBLEM 3. You have a friend who is not a particularly strong calculus student. In attempting to implement GEE for a marginal model, they **incorrectly** find that

$$D_i = \frac{\partial}{\partial \beta} g^{-1}(X_i \beta) = \mathbf{1},$$

(the correct sized matrix of all 1's). They also make the **incorrect** assumption of working independence ($V_i = \sigma^2 I$). Despite this, they have correctly specified the mean model and link function. **True or false: the estimators produced using the GEE under their specification will be consistent for the true β.**

PROBLEM 4. Consider a hypothetical dataset that contains information on the following variables:

- **Smoking Status**: A binary indicator, Y_{ij}, measured for each individual i at each time point j.
- **Age**: A continuous variate denoting time, t_{ij}, recorded for each individual i at each time point j.
- **Baseline Income**: A continuous variate, Inc_i, recorded for each individual i at the baseline.
- **Employment Status**: A binary indicator, E_{ij}, measured for each individual i at each time point j.

Suppose a marginal model is fit to this data, which specifies a logistic link function, binomial variance pattern, and a linear predictor given by

$$\text{logit}(E[Y_{ij}|X_{ij}]) = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{Inc}_i + \beta_3 E_{ij}.$$

True or false: This is a valid marginal model which can be fit using GEE.

PROBLEM 5. A professional basketball team has hired a new data analyst to try to help guide decision making on the team. As a first project, the analyst fits a generalized linear marginal model to data on the league's players. In particular, the analyst fits a model which takes as the outcome the median number of points scored per game for each player, and controls for various factors (the player's height and weight, the team they play for, whether they were injured, and so forth). The model is fit with a longitudinal trend representing the player's age. The idea is to use this model to consider what the impact of aging is on the quality of play that players manage. **True or false: This model can be used to predict the impact of aging for a specific individual. (For instance, the model can be used to estimate the aging curve for team's star player.)**

PROBLEM 6. Suppose that a linear mixed effects model is fit, given by

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0,i} + b_{1,i} t_{ij} + \epsilon_{ij}.$$

True or false: We can interpret β_1 as the expected change (across the whole population) in the outcome for a unit increase in t_{ij}.

PROBLEM 7. Suppose that the following marginal model is fit to data

$$E[Y_{ij}|X_{ij}] = \beta_0 + \beta_1 \text{Age}_{ij} + \beta_2 \text{Income}_{ij} + \beta_3 Z_i + \beta_4 \text{Age}_{i1} + \beta_5 \text{Income}_{i1}.$$

True or False: Testing the null hypothesis $H_0 : \beta_4 = \beta_5 = 0$ is equivalent to a test of the hypothesis that the longitudinal effects of age and income are equal to the cross-sectional effects of age and income, respectively.

PROBLEM 8. A linear mixed effects model is fit which include 5 $(b_0, b_1, b_2, b_3, b_4)$ random effects terms. An analyst wants to test whether 2 of those random effects can be dropped, and so they fit the nested model with the terms removed, including only (b_0, b_1, b_2). They compute a likelihood ratio statistic of $\Lambda = 8.5$. **True or False: At a 5% significance level, the analyst rejects the null hypothesis that** $H_0 : \sigma^2_{b_3} = \sigma^2_{b_4} = 0$**.**

PROBLEM 9. Suppose that a marginal model is fit to the data from Problem 4 (see above) where we take

$$\text{logit}(E[Y_{ij}|X_{ij}]) = \beta_0 + \beta_1 t_{ij} + \beta_2 t^2_{ij} + \beta_3 \text{Inc}_i.$$

True or False: $\exp(1000\beta_3)$ **represents the odds ratio associated with propensity to smoke for a** 1000 **increase in baseline income.**

PROBLEM 10. Your friend, who is not taking STAT 437, is attempting to make conclusions regarding the impact of aging in their favourite esports league. They know that you're taking STAT 437, and so they tell you the following: "I was interested in determining how aging impacts competitors. As such, I fit a model using GEEs which included the age and character for each of the competitors, and the interaction between these terms. I then fit the model dropping the interaction term. Using a likelihood ratio test, I rejected the null hypothesis that the interaction term was zero. As a result, I concluded that what character a player uses changes their success rate as they age." **True or False: The procedure outlined by your friend, as well as their conclusions, are valid.**

Part 2: Conceptual Question (10 Marks)

For the following questions, provide your answers with justification and clear communication. The answers do not need to be long, but correct responses without complete justification will receive only partial credit.

You are asked to analyze a longitudinal dataset where the goal of the study is to compare two treatments (a new experimental drug, with $A_i = 1$ as compared to an existing medication with $A_i = 0$) based on their ability to lower the diastolic blood pressure over time in a group of sick individuals. The individuals in the survey are similar (except for their assigned treatment). Suppose that, in total, there are $n = 500$ patients, with each treatment receiving 250 individuals. Further suppose that blood pressure measurements are taken at times $\{0, 2, 4, 6, 8\}$, measured in days.

Problem 11. (2 Marks) Suppose that you initially decide to fit a linear mixed effects model, with random effects for the intercept, time (which is treated as a continuous variable during this analysis), and treatment terms, and assume $G_i = \sigma^2 I$. We saw in lecture that $\text{var}(Y_i) = Z_i D Z_i' + G_i$. Write down Z_i for patient $i = 1$, who is receiving the experimental drug, and for patient $i = 2$ who is receiving the existing treatment.

Problem 12. (3 Marks) How many parameters are needed to estimate the variance structure in this model? Under what condition will $\text{var}(Y_i)$ be the same for all individuals in the study?

Problem 13. (2 Marks) Suppose that we find that the random effects terms included in the model do not adequately explain the data. Without seeing any data, does it seem reasonable to drop the random effects terms from the model without otherwise modifying the assumed covariance structure?

Problem 14. (3 Marks) Instead of using a linear mixed effects model, you consider fitting a linear marginal model with the same mean structure used for the previous fixed effects, an identity link function, $V(\mu) = \sigma^2$, and an unstructured correlation. Your coworker says that this was a waste of time since you already had fit this model, arguing the linear mixed effects models are also marginal models. Is your coworker correct in this situation? Why?

Part 3: Theoretical or Applied Question (20 Marks)

Please **pick one** of the following two problems and solve it. **If you solve both** only the first problem will be looked at. The first problem corresponds to an **application** problem, while the second is a **theoretical question**. For the application question, please provide the code and relevant output (consider using a software like RMarkdown, or being highly selective with what output you copy to ensure your solution is legible). For the theoretical question, please include enough of your work to justify the steps you have taken.

PROBLEM 15. (Application) On the course website you will find a data file `schoolgirls.csv`. This reports a study of height growth for 20 girls who were followed from age 6 to age 10. The study also records the height of each girl's mother at birth, grouping them into 1 = short, 2 = medium, and 3 = tall.

1. First, consider a model (`model1`) which is a linear mixed effects model with random slope and intercept, and where the marginal mean can vary based on the mother's height, in addition to the girl's age. Fit `model1` using R. Then, write down this model mathematically, identify the key assumptions for it to be valid, and report the estimated parameters from your model. (Note: you should report the parameter values alongside your mathematical notation, perhaps using a table, rather than simply printing the summary output).

2. Provide an estimated 95% confidence interval for: (a) the impact on average height comparing mothers who were tall to those who were short, (b) the variation in the random slope, and (c) the expected height for a 12 year old girl with a medium-height mother.

3. In addition to `model1`, consider fitting `model2` which drops the random slope term, and `model3` which drops the random intercept term. Decide which of these three models is most appropriate for the data, explicitly stating any hypothesis tests that you run.

4. Fit a corresponding linear marginal model, using GEE, deciding whether an unstructured, exchangeable, or autoregressive correlation structure is most appropriate for these data.

5. Using your selected optimal models from (3) and (4): predict the subject level and population level response for `id=7` at `age=10`. Interpret these values.

PROBLEM 16. In class we discussed the best linear unbiased predictor (BLUP) for random effects. We stated that the best predictor (in terms of MSE) for b_i is going to be given by

$$E[b_i|Y_i] = DZ_i'V_i^{-1}(Y_i - X_i\beta).$$

In this question you will prove that this is true!

1. Suppose that $g(W_i)$ is a predictor for Ω_i, where Ω_i and W_i are arbitrary random variables. Prove that the MSE of such a predictor is minimized when $g(W_i) = E[\Omega_i|W_i]$. Recall that the MSE is defined as

$$E\left\{[\Omega_i - g(W_i)]^2\right\}.$$

   Hint: Recall the law of iterated expectation.

2. Demonstrate that, assuming correct specification of a linear mixed effects model as

$$Y_i = X_i\beta + Z_ib_i + \epsilon_i,$$

   we will have that the joint distribution of (Y_i, b_i) is a multivariate normal. Recall that

$$\begin{aligned} Y_i|b_i &\sim N(X_i\beta + Z_ib_i, \sigma^2 I) \\ b_i &\sim N(0, D). \end{aligned}$$

   Here we take Y_i to be $K \times 1$ and b_i to be $q \times 1$.

   Hint: it may help to note that the moment generating function of a multivariate normal, $W \sim N(\mu, \Sigma)$, is

$$M_W(t) = E[e^{t'W}] = \exp\left(\mu't + \frac{1}{2}t'\Sigma t\right).$$

   Further, if (W, Ω) are jointly multivariate normal then the MGF, $M_{(W,\Omega)}(t) = E[\exp(t_1'W + t_2'\Omega)]$, where $t = (t_1', t_2')'$. Finally, if two random quantities have the same MGF, then they have the same distribution.

3. Appealing to parts (1) and (2) of the question, demonstrate that the best predictor of b_i (as a function of Y_i) is given by

$$DZ_i'V_i^{-1}(Y_i - X_i\beta),$$

   where $V_i = Z_iDZ_i' + \sigma^2 I$. Moreover, indicate the variance of the BLUP (e.g. $\text{var}(b_i|Y_i)$).

   Hint: It may be helpful to know that if

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right),$$

   then $Y_1|Y_2$ will follow a normal distribution with mean $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2)$ and variance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Here the mean and variance are partitioned based on the sizes of Y_1 and Y_2.