STAT 437 - Midterm Take-home Assessment
Due: Sunday, March 13 on Crowdmark
No Re-submission

This assessment covers the materials contained in **Lecture 001** through to **Lecture 028**. You are **not** permitted to discuss to these problems with classmates. 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 statements, simply indicate whether the answer is True or False. If you feel as though there is insufficient information provided, you may answer NEI (not enough information). Note, you **do not** need to give a justification for your answer. Each question is worth 1 mark, and there are no penalties for incorrect answers.

TF 1. A longitudinal study is said to be balanced if, for all individuals, we observe exactly k observations.

TF 2. Suppose that a model is fit using an AR(1) correlation pattern. If we observe that $\text{cor}(Y_{i1}, Y_{i2}) = 0.6$, then we know that $\text{cor}(Y_{i5}, Y_{i8}) = 0.126$.

TF 3. One individual in our dataset is observed for 5 different time points. At each time point we take measurements of their blood pressure (the outcome of interest). We also measure their age, height at baseline, in addition to an indicator as to whether or not they have a family history of heart disease. For this individual, Y_i will be a 5×1 vector, and X_i will be a 5×4 matrix, if we include an intercept.

TF 4. Generalized marginal models fit using the option `corstr = "ind"` can be fit using a call to `glm`.

TF 5. If we have a linear marginal models for a continuous variate, we are unable to perform likelihood ratio tests with it.

TF 6. We can use the fact that an exchangeable correlation structure is nested within an unstructured correlation structure in order to test whether or not the exchangeable assumption is appropriate using likelihood ratio procedures.

TF 7. The REML estimators are obtained by optimizing a modified version of the log-likelihood, which involves subtracting a penalty term from the true log-likelihood.

TF 8. If we have a marginal model, testing the hypotheses that $\beta_1 = \beta_2$, $\beta_4 = 5$, and $\beta_3 = 2\beta_5$ simultaneously involves a comparison to a χ^2_3 distribution.

TF 9. Generalized estimating equations use M-estimation to provide estimators which are consistent so long as the structure for the mean is correctly specified. As a result, there is no purpose in correctly specifying the covariance structure.

TF 10. We are unable to use generalized linear marginal models to determine the impact of a child's height on their weight as they age, even if we had a well-designed longitudinal study that measured both height and weight for children overtime.

TF 11. If we wish to make assessments at the population-level, we cannot use a mixed effects model.

TF 12. When doing model selection for mixed effects models, we tend to prefer those models which have a lower `AIC` and `BIC`. However, we cannot use these same criteria when doing model selection for generalized linear marginal models.

TF 13. When fitting a mixed effects model we estimate parameters for the population-level effects, typically denoted β, and we estimate parameters for the individual-level effects, typically denoted b_i.

TF 14. When predicting individual outcomes from a mixed effects model we use a weighted average between the estimated population mean and the individual observation. The population average receives less weight when the estimated error variance is large.

TF 15. Suppose that $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ give are the parameters corresponding to $\text{var}(b_{0,i})$, $\text{var}(b_{1,i})$, $\text{cov}(b_{0,i}, b_{1,i})$, and σ^2, respectively, in a mixed effects model. The null distribution for a test of $H_0 : \theta_3 = 0$ and the null distribution for a test of $H_0 : \theta_2 = 0$ are different.

TF 16. In a transition model with the first-order Markov assumption, we can conclude that

$$P(Y_t = \ell | Y_{t-1} = m, Y_{t-2} = m', \cdots, Y_1 = k) = P(Y_t = \ell | Y_{t-1} = m) = P(Y_{t+1} = \ell | Y_t = m).$$

TF 17. A third-order, time-homogeneous model for transition probabilities requires 8 regression coefficients to be estimated, assuming no covariates are included in the model.

TF 18. Suppose that we are using automated data collection instruments to record wind speed information, daily, at several different locations. Because of a calibration issue, if wind speeds exceed 60 kilometers per hour, the data are not correctly sent to the server, and are recorded as missing instead. In this situation, multiple imputation would be a valid means of correcting for the missingness.

TF 19. When observations are MCAR, running a complete case analysis is valid in terms of consistency and bias, but it makes us less likely to detect significant effects of interest.

TF 20. The choice of whether we will use a marginal model or a mixed effects model for a particular problem should be made primarily based on statistical concerns regarding how well the model fits the data.

Part 2: Matching Response (10 Marks)

For each of the following scenarios, provide an answer from the set of options provided for that group of statements. The same answer can be given for multiple statements. If there are multiple appropriate answers, select the one which is most appropriate with the given information. If you feel as though there is insufficient information provided, you may answer NEI (not enough information). Note, you **do not** need to give a justification for your answer. Each question is worth 1 mark, and there are no penalties for incorrect answers.

Longitudinal Model Selection

For each of the following scenarios, answer with either Generalized Linear Marginal Models (**GLMM**), Mixed Effects Models (**MEM**), or Transition Models (**TM**), based on which is the most appropriate.

MR 1. A research team wishes to investigate the longitudinal effect of the pharmaceutical treatment of depression. The primary outcome of interest is a patient's QIDS score, which is a numeric score based on survey results assessing symptoms of depression, which is commonly treated as a continuous variate. The researchers are interested in comparing the relative impact of two different kinds of treatment, while taking into account baseline age, baseline socioeconomic status, gender identity, as well as the amount of sleep that the individual has averaged per night between clinical visits.

MR 2. Suppose that you wish to determine the factors that impact a child's academic performance throughout elementary school. In particular, you are interested in their development of mathematical skills, and so you decide to summarize each child's performance through the letter grade that they received in Math. Previous research suggests that children tend to have high correlation between their grades over years; your interest is in determining the factors that may predict larger positive or negative shifts in a child's performance.

MR 3. Researchers are attempting to study the longitudinal development of blood pressure (a continuous variate). They are considering behavioural effects, demographic factors, as well as the impact of specific non-pharmaceutical interventions as explanatory factors in their models. Their hope is that the results of this study will be used to inform public policy in Ontario.

MR 4. You have a friend who has collected longitudinal data that they are trying to analyze, but which they will not tell you the particular contents of. You note that the outcome they are trying to analyze is continuous, but it appears heavily skewed, and transformations do not appear to return symmetry. The data set is very large, with hundreds of thousands of observations. Your friend is not clear on what exactly they hope to do with the results of the analysis, but they claim to have high confidence in what variates impact the mean response.

MR 5. The Korean government runs a longitudinal study which explores the development of family income in an attempt to inform their governmental welfare policies. They control for the region that the individual's live in, educational attainment, and other demographic factors. You are given this data, and are asked to analyze it to determine which regions in the country may require further investment from the government.

Missingness Model Response

For each of the following scenarios, answer with either Missing Completely at Random (**MCAR**), Mixed at Random (**MAR**), or Not Missing at Random (**NMAR**), based on which is the most appropriate for the missingness that is discussed.

MR 6. In political polling research it is well-known that individuals with stronger political beliefs (further from the center of the political spectrum) are more likely to not respond to political questions. Researchers are conducting a longitudinal study to determine how an individual's political opinions on controversial subjects change over time, as they age, and as they are exposed to different ideas in public. In collecting responses, however, they find a substantial amount of missingness in individuals' responses to these political questions.

MR 7. An experimental drug is being tested for its ability to improve lung function in patients with primary ciliary dyskinesia (PCD). PCD is a genetic condition in which cilia in the respiratory system have defective function, which prevents the clearance of mucous from the lungs, and can lead to frequent respiratory infections. Owing to ethical considerations, whenever the lung function for a patient on the experimental treatment worsens past a pre-determined threshold, they are removed from the study, put back onto the standard treatment regime, and their subsequent values are considered missing.

MR 8. The health of patients in end-of-life care is being studied to determine the best interventions to improve patient well-being in this setting. Questionnaires are given to patients over the course of the study, which ask them for their current health status (the main outcome), as well as several questions relating to the treatment that they have received. The researchers running this study find that with some frequency, entire questionnaires are missed by some patients, which is the primary cause of missingness in their data.

MR 9. You are considering the results of an analysis which has been performed on data which you know has missing values. The analysis seems to make no correction for the missingness at all, and instead just simply uses all of the observed data, as given. You have confidence in the individual who performed the analysis that what they did was statistically valid.

MR 10. A longitudinal study investigating individuals food consumption habits is being conducted. Once a month, for a year, participants are asked to fill out a food journal which reports what they ate over the past 24 hours. Each month, a random sample of individuals in the study are contacted a second time, where the researcher works through the food journal with them for a second date. This is only conducted on a portion of the individuals owing to cost constraints, but it is generally regarded as higher quality information. In analyzing the data, researchers treat those who were not contacted for the secondary questionnaire in a given month as having missing responses.

Part 3: Short Answer Questions (20 Marks)

For each of the following questions, write a short answer justifying your response. The number of marks for each problem is specified. Partial marks will be awarded for correct work with an incorrect final answer.

Problem 1. (2 Marks) Suppose that you have fit the following linear marginal model

$$E[Y_{ij}|X_{ij}] = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{height}_{ij} + \beta_3 \text{income}_{ij} + \beta_5 \text{treatment}_{ij} + \beta_4 t_{ij} \text{treatment}_{ij}.$$

According to the underlying theory you expect that the impact of height on the outcome should be 100 times that of income, and that the the impact of treatment is not mediated by time. Write down a corresponding L matrix to test this hypothesis as $H_0 : L\beta = 0$, and specify the corresponding distribution.

Problem 2. (3 Marks) A study is run investigating the incidence of respiratory infection in children. The outcome is taken to be $Y_{ij} = 1$ if the child has an infection, and $Y_{ij} = 0$ otherwise. The observed two-step transition counts for the dataset are given in the following table.

		Y_{ij}	
$Y_{i,j-2}$	$Y_{i,j-1}$	0	1
0	0	855	64
0	1	81	12
1	0	587	56
1	1	47	8

Based on this, compute the maximum likelihood estimates for the transition probabilities under a time-homogeneous, first-order Markov model.

Problem 3. (2 Marks) Nine different models are fit using generalized estimating equations to Poisson data. The specific model forms, as well as output regarding the QIC for each model is provided in the following table. In the displayed models, `Age` represents the age at baseline in the study, `Trt` represents the treatment that the individual received, and `t` represents the time. Based on the presented information, is there a model which can be selected as preferable?

Model	Correlation Structure	Mean Structure	QIC	QICu
1	Unstructured	`Y ~ Age*Trt*t`	-400	-415
2	Unstructured	`Y ~ Age+Trt+t`	-800	-810
3	Unstructured	`Y ~ Trt*t`	-860	-860
4	AR(1)	`Y ~ Age*Trt*t`	-880	-885
5	AR(1)	`Y ~ Age+Trt+t`	-960	-985
6	AR(1)	`Y ~ Trt*t`	-980	-990
7	Exchangeable	`Y ~ Age*Trt*t`	-100	-100
8	Exchangeable	`Y ~ Age+Trt+t`	-280	-270
9	Exchangeable	`Y ~ Trt*t`	-760	-735

PROBLEM 4. (3 Marks) A study is investigating the how the skulls of rats grow overtime, and how this growth is impacted by a hormone therapy drug *Decapeptyl*. Rats are randomized to one of three groups: low-dose (group 1), high-dose (group 2), or control (group 3) and their skull heights are measured every 10 days. We assume that time is a continuous factor, measured by the visit number (so that $t_{ij} = 2$ on day 20, for instance). Define $x_{ij1} = I(\text{group}_{ij} = 1)$, $x_{ij2} = I(\text{group}_{ij} = 2)$, and $x_{ij3} = I(\text{group}_{ij} = 3)$, and let t_{ij} be the time of each measurement. Suppose that a random intercept and slope model is fit to the data such that

$$Y_{ij} = \beta_0 + \beta_1 x_{ij1} t_{ij} + \beta_2 x_{ij2} t_{ij} + \beta_3 x_{ij3} t_{ij} + b_{0,i} + b_{1,i} t_{ij} + \epsilon_{ij}.$$

From this model we get the following output:

Random Effects

	Std. Dev.	Corr
(Intercept)	1.88810	(Intr)
time	0.00015	0.00001
Residual	1.20199	

Fixed Effects

	Value	Std. Error	DF
(Intercept)	68.607	0.33123	199
x1:t	7.507	0.22516	199
x2:t	6.871	0.22761	199
x3:t	7.314	0.28077	199

Based on these results:

(a) What is the expected skull height for a rat on day 20 in the high-dose group?

(b) What is the covariance between measurements on days 20 and 30 for a rat in the low-dose group?

(c) If we wanted to test whether the growth rates among rats is heterogeneous (i.e., differs by individuals overtime) using an applicable likelihood ratio test, what would the critical value for such a test be at a 5% significance level?

PROBLEM 5. (5 Marks) In trying to assess the impact of income on perceived quality of life, you decide to fit a linear mixed effects model. You use an unstructured correlation matrix, and specify the form to be

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{Income}_{i1} + \beta_3 \text{Income}_{ij} + b_{0,i} + b_{1,i} t_{ij} + \epsilon_{ij}.$$

Here, income is measured in 1000s of dollars. The fixed effects model summary is given by

```
             Value       Std.Error
(Intercept)  5.7467003  0.76762391
t            2.5748823  0.15246301
income1      3.5489572  0.06902625
income       0.9154423  0.60609418
```

Give point estimates for both the cross-sectional and longitudinal effects of income on life satisfaction. For the longitudinal effect, include a 95% confidence interval. Is the longitudinal effect statistically significant?

PROBLEM 6. (5 Marks) Suppose that we wish to fit a GEE to continuous data which takes only positive values. Suppose that we use a log link function, $g(\mu) = \log(\mu)$ and an identity variance function $V(\mu) = \mu$. Further, suppose we constrain the dispersion parameter, $\phi = 1$. If we have $k = 3$ time points measured, X_{ij} has $p = 2$ components, and make an exchangeable correlation assumption, what are the forms of D_i and V_i?

Solution Template: True or False

For your convenience, feel free to indicate your answers to Part 1 on this sheet and upload the results to Crowdmark. Circle exactly one option.

Question		
Q1	True	False
Q2	True	False
Q3	True	False
Q4	True	False
Q5	True	False
Q6	True	False
Q7	True	False
Q8	True	False
Q9	True	False
Q10	True	False
Q11	True	False
Q12	True	False
Q13	True	False
Q14	True	False
Q15	True	False
Q16	True	False
Q17	True	False
Q18	True	False
Q19	True	False
Q20	True	False

Solution Template: Matching

For your convenience, feel free to indicate your answers to Part 2 on this sheet and upload the results to Crowdmark. Circle exactly one option.

Question			
Q1	GLMM	MEM	TM
Q2	GLMM	MEM	TM
Q3	GLMM	MEM	TM
Q4	GLMM	MEM	TM
Q5	GLMM	MEM	TM
Q6	MCAR	MAR	NMAR
Q7	MCAR	MAR	NMAR
Q8	MCAR	MAR	NMAR
Q9	MCAR	MAR	NMAR
Q10	MCAR	MAR	NMAR