STAT 437 - Assignment 1
Due: Friday, January 21 on Crowdmark
Resubmit: Friday, February 4 on Crowdmark

This assignment covers the materials contained in **Lecture 002** through to **Lecture 009**. 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; 2 Marks Each)

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. In an attempt to study the effects of long-term exposure to birds on individual health, researchers compared the rates of lung disease in long-term bird owners (15+ years) to the rates of lung disease in an (otherwise similar) group of new bird owners (<1 year), and a demographically-matched control group. They estimated that long-term bird owners were nearly twice as likely to have developed serious lung diseases. **True or false: assuming that the impact on health was correctly computed, this is an example of a longitudinal effect.**

Problem 2. Suppose that it is known that, as individuals age, their personality (as measured by the Big 5 Personality traits) changes independently of their past-self. That is, the correlation of these values (when measured distantly from one another) is expected to be zero, due to these measures being independent. **True or false: a study which follows individuals from childhood through adulthood, considering the development of their personality traits, can be analyzed using standard (e.g., linear or GLM) regression methods.**

Problem 3. You are given a dataset in long format. The dataset contains information on 100 individuals. It contains a column for their ID, for the time index that the measurement was taken, their outcome, and a treatment indicator. The first several rows of this dataset are shown below, **sorted by ID**.

ID	Time	Treatment	Outcome
1	1	1	30.5
1	2	1	35
1	3	1	35
2	2	0	26
2	3	0	24.5
2	5	0	20
⋮	⋮	⋮	⋮

True or false: when translated into wide format, the data frame will have 100 rows and 5 columns.

PROBLEM 4. **True or false: in a linear regression model, in order to produce *valid* Wald-type confidence intervals for the coefficients (e.g., $\widehat{\beta} \pm 1.96 \times \text{s.e.}(\widehat{\beta})$ or similar), we must assume normality of the outcome (conditional on the covariates).**

PROBLEM 5. Consider the following expressions for the conditional expectation of Y given X:

$$E[Y|X] = \beta_0 + \beta_1 X \tag{1}$$
$$E[Y|X] = \beta_0 + \beta_1 \log(X) \tag{2}$$
$$E[Y|X] = \beta_0 + \exp(\alpha_1 + X) \tag{3}$$
$$E\left[\log(Y)|\, X\right] = \exp(\alpha_0) + \beta_1 X^2 \tag{4}$$
$$E[Y|X] = \exp(\alpha_0 + \alpha_1 X) \tag{5}$$

True or false: all these models, except for (5), express a linear relationship in the conditional mean which is estimable with standard (OLS) linear regression.

PROBLEM 6. Suppose we have fit a linear marginal model, with a structure given by

$$\begin{aligned} E[Y_{ij}|X_{ij}] = {} & \beta_0 + \beta_1 I(t_j = 2) + \beta_2 I(t_j = 3) + \beta_3 I(t_j = 4) + \beta_4 \text{Trt}_i + \beta_5 I(\text{Age}_i > 50) \\ & + \beta_6 I(t_j = 2)\text{Trt}_i + \beta_7 I(t_j = 3)\text{Trt}_i + \beta_8 I(t_j = 4)\text{Trt}_i \\ & + \beta_9 I(t_j = 2) I(\text{Age}_i > 50) + \beta_{10} I(t_j = 3) I(\text{Age}_i > 50) + \beta_{11} I(t_j = 4) I(\text{Age}_i > 50). \end{aligned}$$

Here, we have t_j representing the time index (that is, $I(t_j = 2)$ is 1 if it is time 2, and is 0 otherwise), the Trt_i is a binary indicator for active treatment, and $I(\text{Age}_i > 50)$ is 1 for any individual who was over 50 years old when the study began. Take

$$L = \begin{array}{c} \begin{array}{cccccccccccc} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 & \beta_9 & \beta_{10} & \beta_{11} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

True or false: testing the hypothesis $L\widehat{\beta} = \mathbf{0}$ performs a test of whether or not the time trend associated with treatment is equivalent to the time trend associated with being over 50 years old, and this hypothesis should be tested against a χ^2_3 distribution.

PROBLEM 7. Two linear marginal models are fit on the same data, using the same conditional mean structure. The first makes an auto-regressive assumption for the correlation pattern matrix (i.e., $\text{cor}(Y_{ij}, Y_{i\ell}) = \rho^{|j-\ell|}$) and the second makes the exchangeable assumption (i.e., $\text{cor}(Y_{ij}, Y_{i\ell}) = \rho$). Both models are fit using maximum likelihood. The log-likelihood for the first model is found to be -1225 and the log-likelihood for the second model -1260.

True or false: we can test the hypothesis that the second model is adequate using a LRT, where the test statistic is $\Lambda = 70$, with a distribution of χ^2_1, resulting in a p-value < 0.001, and a rejection of the null hypothesis.

PROBLEM 8. **True or false: The Poisson distribution (with parameter λ) is an exponential family distribution with: (i) canonical parameter $\theta = \log(\lambda)$, (ii) $b(\theta) = e^{\theta}$, (iii) $a(\phi) = 1$, (iv) $c(y, \phi) = -\log(y!)$, and (v) canonical link $g(x) = \log(x)$.**

PROBLEM 9. Suppose that we have count data which are known to be highly over-dispersed, relative to the Poisson assumption (that is, $\text{var}(Y)$ is much larger than $E[Y]$). Suppose further that $E[Y|X] = \exp(\beta_0 + \beta_1 X)$, for unknown β_0 and β_1. A very large sample of independent realizations is taken. Consider

$$U(\beta) = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \{y_i - \exp(\beta_0 + \beta_1 x_i)\}.$$

True or false: Solving $U(\widehat{\beta}) = \mathbf{0}$, will produce a consistent estimator for β.

PROBLEM 10. A new, experimental treatment is being tested as a means of delaying the onset of dementia. Suppose a marginal linear model has been fit, using A_i to represent the treatment assignment (binary indicator, with $A_i = 0$ for placebo and $A_i = 1$ for experimental treatment) for individual i, and where t_j is time, treated as a continuous variate. Consider the model given by

$$E[Y_{ij}|A_i, t_j] = \beta_0 + \beta_1 t_j + \beta_2 A_i t_j.$$

True or false: β_1 is interpreted as the expected change in outcome as time passes (i.e., for a 1 unit increase in time) for those in the placebo group and β_2 represents the expected change in outcome as time passes (i.e., for a 1 unit increase in time) for those in the experimental treatment group.

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.

Suppose that the viral load (a continuous outcome) is repeatedly measured on HIV positive subjects from three treatment groups: a placebo (control) group, a low-dose group, and a high-dose group. We treat time, t_{ij} as measuring the number of days since baseline enrollment in the study. At baseline, individuals had no discernible differences in viral load and were assigned to their treatment groups completely at random.

Suppose that a marginal linear model is fit to the data such that,

$$E[Y_{ij}|X_{ij}, t_{ij}] = \beta_0 + \beta_1 x_{i1} t_{ij} + \beta_2 x_{i2} t_{ij} + \beta_3 x_{i3} t_{ij},$$

where $x_{i1} = 1$ if the patient is in the control group, $x_{i2} = 1$ if the patient is in the low-dose group, and $x_{i3} = 1$ if the patient is in the high-dose group (with these indicators taking 0 otherwise). The following questions all relate to this model.

Problem 11. (2 Marks) There is no main effect of treatment included in this model. Explain why this is a natural choice and indicate what the inclusion of (non-zero) treatment effects would indicate about the underlying scenario.

Problem 12. (3 Marks) Suppose that we wish to test whether or not the response of the low-dose group changes over time. Write down the hypothesis in terms of model parameters, as a general linear hypothesis of the model (e.g., $L\beta = \mathbf{c}$ for suitable L and $\mathbf{c}$) and specify distribution under the null.

Problem 13. (3 Marks) Suppose that we wish to test whether the responses of the low-dose and high-dose groups change at the same rate over time, while the control group exhibit no time-trend. Write down the hypothesis in terms of model parameters, as a general linear hypothesis of the model (e.g., $L\beta = \mathbf{c}$ for suitable L and $\mathbf{c}$) and specify distribution under the null.

Problem 14. (2 Marks) Suppose that we consider two individuals, one in the low-dose group, and one in high-dose group. If it is found that $\widehat{\beta}_2 = 0.5\widehat{\beta}_3 < 0$, what is the predicted difference in viral load between these two patients after 10 days? Suppose that the first individual remains on treatment for 60 days. How long would the second individual need to remain on treatment to have an equivalent viral load prediction?

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 `dental.csv`. This reports a study of dental growth where measurements of the distance (in mm) from the center of the pituitary gland to the pteryomaxillary fissure were obtained on 11 girls and 16 boys at ages 8, 10, 12, and 14. The data file contains these observations in wide format, recording the ID, Sex, and outcomes (Y_1, Y_2, Y_3, Y_4).

1. Generate relevant plots for the data that illustrate how the data evolve overtime, to help inform modelling. Explain what information your plots show.

2. Fit a fully saturated, marginal linear model to the data, with time and sex both treated as discrete. Use an unstructured correlation matrix and allow for variances to change over each week. Based on your fitted saturated model, is the time trend the same between the two groups?

3. Based on your model, provide an estimated 95% confidence interval for the mean distance for 8 year old boys, and 14 year old girls.

4. Determine whether it is possible to simplify the assumed correlation and variance structure. Consider: (1) a model with assumed constant variance but unconstrained correlation, (2) a model with compound symmetry but differing variance, and (3) a model which assumes constant variance and uses an auto regressive correlation structure. Are any of these models appropriate?

5. A doctor is interested in the distance measurements for 13 year old girls, and 15 year old boys. Provide 95% confidence interval for the mean predictions of these categories, and indicate whether or not these predictions are appropriate.

Problem 16. (Theoretical) Consider a marginal linear model where we take the mean to be given by $\mu_i = X_i\beta$ and the variance $\Sigma_i = \sigma^2\mathbf{R}(\rho)$. Recall that we stated that the MLE for β are given by

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i'\Sigma_i^{-1}X_i\right)^{-1}\sum_{i=1}^{n} X_i'\Sigma_i^{-1}Y_i.$$

1. If $\Sigma_i = \sigma^2\mathbf{R}(\rho)$, show that the MLE of β does not depend on σ^2.

2. Derive the maximum likelihood estimator for σ^2.

3. What is the expression for the profile log-likelihood of ρ (can be in terms of $\widehat{\beta}$ and $\widehat{\sigma}^2$)?

4. Prove that the MLE for the marginal linear model simplify to the OLS estimators if we assume constant variance and independent correlation.

5. Show that, if the data are balanced with equal spacing between follow-ups, assuming a first-order auto regressive correlation structure is equivalent to assuming an exponential correlation structure.