# Homework 2: Due on March. 2

1. The instructor of math 5072 would like to evaluate the performance students in various departments and from various years. A balanced random sample were taken. Let $y_{ijk}$ be the math score of the $k$-th student from the $j$-th department in the $i$-th year, where $i = 1, \cdots, m, j = 1, \cdots, n, k = 1, \cdots, r$. Write the following linear mixed effects models in matrix notation. In each case derive the marginal mean and variance of the response variable.

   (a) Two-way random model: The $n$ departments and $m$ years in the sample were randomly selected from the population. Hence one may want to treat both of them as random effects. That is,

   $$y_{ijk} = \mu + a_i + b_j + e_{ijk},$$

   where $a_i \overset{\text{iid}}{\sim} N(0, \sigma_a^2)$, $b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2)$, $e_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$, and $a_i$, $b_j$ and $e_{ijk}$ are independent.

   (b) Two-way mixed model: While individual year effect, $a_i$, is not of interest, the individual department effect may be of interest and could be treated as fixed effect, $\beta_j$. That is,

   $$y_{ijk} = \mu + a_i + \beta_j + e_{ijk},$$

   where $a_i \overset{\text{iid}}{\sim} N(0, \sigma_a^2)$, $e_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$, and $a_i$ and $e_{ijk}$ are independent.

   (c) Two-way mixed model with interaction: Consider that the difference among departments may vary over years. One may extend the previous model by including the interaction between department and year, $g_{ij}$. This should be treated as random since year effect is random. That is,

   $$y_{ijk} = \mu + a_i + \beta_j + g_{ij} + e_{ijk},$$

   where $a_i \overset{\text{iid}}{\sim} N(0, \sigma_a^2)$, $g_{ij} \sim N(0, \sigma_g^2)$, $e_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$, and $a_i$, $g_{ij}$ and $e_{ijk}$ are independent.

2. Consider the growth curve model (example: sleep deprivation data),

   $$y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + b_{1i} x_{ij} + e_{ij}, \quad i = 1, \cdots, n, j = 1, \cdots, m_i.$$

   Here $\beta_0$ and $\beta_1$ are fixed intercept and slope, respectively, and $b_{0i}$ and $b_{1i}$ are random intercept and slope. Assume that $b_{0i} \overset{\text{iid}}{\sim} N(0, \sigma_0^2), b_{1i} \overset{\text{iid}}{\sim} \sigma_1^2, e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$, and $b_{0i}, b_{1i}$ and $e_{ij}$ are independent. Derive the distribution of the response $y$ and write the marginal mean and variance in matrix form.

3. Derive the ML and REML estimate of $\sigma_e^2$ in the following models. In each case, $i = 1, \cdots, n$ and $j = 1, \cdots, m$.

(a) $y_{ij} = \mu + e_{ij}, e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

(b) $y_{ij} = \mu + \alpha_i + e_{ij}, e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. (One-way ANOVA with fixed effect.)

(c) $y_{ij} = \mu + a_i + e_{ij}, a_i \overset{\text{iid}}{\sim} N(0, \sigma_a^2), e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and $a_i$ and $e_{ij}$ are independent. (One-way ANOVA with random effect.)

(d) $y_{ij} = \mu + \beta x_i + e_{ij}, e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. (Simple linear regression.)

(e) $y_{ij} = \mu + a_i + \beta x_i + b_i x_i + e_{ij}, a_i \overset{\text{iid}}{\sim} N(0, \sigma_a^2), b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2), e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and $a_i, b_i$ and $e_{ij}$ are independent. (A special case of growth curve.)

4. A car company was interested to learn the prognostic factors of the discount rate offered by their dealers. Over ninty thousands cars were sold by over two hundreds and fifty dealers in 2019. Sale data and dealer information of 2019 were collected and can be found on the course website. Please fit a linear mixed effect model to investigate what factors affect the discount rate and how.

   Please write your analysis in the form of a report and follow the guideline on the next page.

   Hint: You may first focus on only one selected type of car, for example "L550", since different type of cars may have different discount reasons. Furthermore, not all factors are needed and many redundant information can be excluded from your model.

# Statistical Report Guideline

- **Confidentiality Agreement**

  Note the posed data is only for educational propose and should not be used outside of this class. Please include the following confidentiality agreement at the BEGINNING of your report.

  "I understood and agreed to that the data used in this report must be kept confidential and must not be shared or distributed outside this class. The statistical analysis, the result and the report were agreed to be only used in this class and will not be posted or published without the authorization of Prof. Jimin Ding. "

- **Format**

  Your statistical report should contain introduction (background of the study and overall scientific aims), methodology and model setup (review of general statistical methods and rationales of your model choice), analysis result, conclusion and discussion. You should clearly label all your figures and tables. All the included figures and tables should be referred in the main text and you can highlight some numbers in your tables to make your findings more noticeable. Remember that the readers of your report may not be all statisticians but someone who has only limited statistical knowledge. So please write your report toward general audience, explain your main statistical findings in layman's terms, and interpret your results in the scientific context.

- **Length**

  The report, including the tables and figures, should be NO LONGER THAN 10 PAGES. The margins should be at least 1.5 inches and the front size should be at least 12 points.

- **Coding**

  Please use R to analyze the data. Please save all your program codes and we may ask you to submit your codes later. It is also wise code with comments to make your codes easy to understand and more readable.