BASKIN SCHOOL OF ENGINEERING
DEPARTMENT OF STATISTICS

2020 First Year Exam, Take Home Question

Due by 5:30 PM, Wednesday June 10th, 2020

Instructions:

Please work individually on this problem. You are allowed to consult any material you wish, but do not share with any other individual any information or comments about your findings or the models and methods you use. You are required to email your report as **one pdf file** to the graduate director at `juheelee@soe.ucsc.edu`

by 5:30 PM, Wednesday June 10th, 2020

Please organize and present the material in the best possible way. Be informative but concise. You should include a summary of your work at the beginning of the report, include and annotate all relevant figures and tables in the body of the report, write your conclusions in a separate section, and list your references (if any). You are required to write your report in LaTeX, using the template from

`https://users.soe.ucsc.edu/~juheelee/FYE-take-home/`

Your report should consist of no more than 10 letter-size pages (typeset with 11pt or larger font and margins on all four sides of at least 1 inch), including all figures, tables, and appendices (but excluding the numerical codes); answers longer than 10 pages will lose credit for excess length. You must include your R code for both problems at the end of your report; the codes do not count toward the 10-page limit.

Exam Problem:

The dataset `schoolgirls`, which is available from

`https://users.soe.ucsc.edu/~juheelee/FYE-take-home/`

shows the heights of 20 pre-adolescent girls measured on a yearly basis from age 6 to 10. The dataset contains variables: **height:** the height in centimeters; **child:** a numerical label for each of the 20 girls; **age:** the age in years; **group:** a variable that indicates whether the girl's mother was categorized as short (1), medium (2) or tall (3).

Part A. You can use `R` for all the questions of part A; we do not expect you to fit Bayesian models for this part of the problem.

A.1 Using quantitative and graphical tools summarize the main features of this dataset.

A.2 Consider the total 100 height observations for the 3 groups without taking into account the subject index or the age. In other words, let y_{ij} denote the j-th height observation in group i, with $i = 1, 2, 3$, and $j = 1, ..., n_i$, where $n_1 = 30$, $n_2 = 35$ and $n_3 = 35$.

(a) Using the `lm` function in `R` fit a model of the form:

$$y_{ij} \quad = \quad \mu + \delta_i + \epsilon_{ij}$$

with $\epsilon_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$ and $\delta_{i_0} = 0$ for some i_0. Based on your results provide the estimates for all the model parameters.

(b) Can you conclude that there is a statistically significant difference in the mean heights across groups? If this is the case which groups are statistically different? Justify your answer.

(c) Is this a reasonable model for explaining the heights of the girls? Justify your answer.

A.3 Now fit a regression model of the form

$$y_{ijt} = \alpha_i + \beta_i (5 + t) + \epsilon_{ijt}$$

with y_{ijt} the height of girl j in group i at age $5 + t$, for $i = 1, 2, 3$, $j = 1, ..., m_i$, and $t = 1, ..., 5$, where $m_1 = 6$, $m_2 = 7$, $m_3 = 7$. Assume $\epsilon_{ijt} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$.

(a) Summarize the fit of your model and provide the estimates of the model parameters. Explicitly include parameter restrictions if there are any. Explain the effects of the age variable, the group variable, and their interaction, and determine if such effects are statistically significant.

(b) Based on your results, is there a different linear model that you would suggest that uses both variables, age and group? Write such model explicitly, fit the model using the `lm` function in `R`, and provide the estimates of its parameters along with a discussion of your results.

(c) Finally, perform a residual analysis of the model you proposed.

Part B. For this part of the problem, you will explore Bayesian models for the `schoolgirls` data. Note that you are required to write your own code for implementing the models in questions B.1 and B.2 below.

B.1 Consider the data excluding the information on the group variable. Fit the following Bayesian model:

$$y_{kt} = \gamma_0 + \gamma_1 (5+t) + \epsilon_{kt}; \quad \epsilon_{kt} \mid \sigma^2 \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2), \text{ for } k = 1, ..., 20 \text{ and } t = 1, ..., 5.$$

Here, y_{kt} is the height observation for the k-th child at age $(5+t)$, for $k = 1, ..., 20$ and $t = 1, ..., 5$. Discuss your priors for model parameters (γ_0, γ_1) and σ^2, and study the fit of the model.

B.2 Implement a Bayesian hierarchical extension of the model in B.1 that introduces random effect parameters for the schoolgirls. In particular, consider the model:

$$\begin{aligned} y_{kt} \mid (\gamma_{0k}, \gamma_{1k}), \sigma^2 &\overset{ind.}{\sim} \mathrm{N}\left(\gamma_{0k} + \gamma_{1k}(5+t), \sigma^2\right), \quad k = 1, ..., 20; \ t = 1, ..., 5 \\ (\gamma_{0k}, \gamma_{1k}) \mid \boldsymbol{\mu}, \Sigma &\overset{i.i.d.}{\sim} \mathrm{N}_2(\boldsymbol{\mu}, \Sigma), \quad k = 1, \ldots, 20 \end{aligned}$$

where $\mathrm{N}_2(\boldsymbol{\mu}, \Sigma)$ is the bivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Discuss your prior specification for σ^2, and for the parameters of the random effects distribution, $\boldsymbol{\mu}$ and Σ. Explore how this model captures differences in the growth curves of the different girls. Perform Bayesian model checking for your model.

B.3 Perform model comparison for the models in B.1 and B.2. Discuss your conclusions from such model comparison.

B.4 Discuss Bayesian model formulations for the full dataset that incorporate the group variable. You do **not** need to implement any of the more general models you discuss. However, explain why and how you expect them to improve model fit and predictions.