BASKIN SCHOOL OF ENGINEERING
DEPARTMENT OF STATISTICS

2022 First Year Exam, Take Home Question

Due by 5PM, Thursday June 9th, 2022

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 5PM, Thursday June 9th, 2022

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 codes for all problems at the end of your report; the codes do not count toward the 10-page limit.

Exam Problems

Please download the dataset `data_pines.csv` from:

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

The dataset `data_pines.csv` contains data from leaves of two pine species (20 trees in total) that were sampled throughout their canopy. Eight samples were taken at various heights in each tree, with the objective of investigating whether or not there is a pattern of higher leaf thickness (higher LMA) toward the top of the trees. In total, there are 160 rows and the following variables: **ID:** an ID of the individual tree, **species:** a categorical variable with two levels, Pinus ponderosa and Pinus monticola, **dfromtop:** a numerical variable corresponding to the distance from top of tree to where leaf sample was taken in meters, **height:** a numerical variable corresponding to the height from the ground where sample was taken in meters, and **LMA:** a numerical value corresponding to the leaf mass per area in g/m^2.

1. (10 points) Use quantitative and graphical tools to summarize the main features of this dataset. Discuss your main findings.

2. (5 points) Consider the problem of constructing a new categorical variable using **dfromtop**. Your collaborator in this project says that the usual way this variable is analyzed is in categories, but adds that there is no standard way to construct them. How would you construct three categories out of **dfromtop**? Describe the process and create a new variable named **catdfromtop.**

3. (15 points) Consider the total 160 observations. Let $y_{i,j,k}$ denote the LMA observation from leaf k in tree j from species i. That is, $i = 1, 2$ corresponds to Pinus ponderosa or Pinus monticola, respectively, and $j = 1, \ldots, n_i$, with $n_1 = 8$, and $n_2 = 12$. Let $x_{i,j,k}$ denote the **dfromtop** for leaf k in tree j from species i.

   (a) (5 points) Use the `lm` function in `R` to fit a model of the form:

   $$y_{i,j,k} = \mu + \beta_1 \mathbb{1}_{c_1 < x_{i,j,k} \leq c_2} + \beta_2 \mathbb{1}_{c_2 < x_{i,j,k}} + \epsilon_{i,j,k} \tag{1}$$

   with $\epsilon_{i,j,k} \sim N(0, \sigma^2)$. $\mathbb{1}$ denotes the indicator function and c_1, c_2, c_3 are the cutoffs identified in question 2. Based on your results, provide the estimates for all the model parameters.

(b) (5 points) Is there enough statistical evidence to conclude that there is a significant difference in the mean LMA across **catdfromtop** groups? If this is the case which groups are significantly different? Justify your answer.

(c) (5 points) Perform a residual analysis for this model. Are there any issues with the residuals? Would you suggest any changes to your model (e.g., transformations) based exclusively on your residual analysis? Is this a reasonable model for explaining the response variable LMA? Justify your answer.

4. (20 points) Use the `lm` function in `R` to fit a regression model of the form:

$$y_{i,j,k} = \mu_i + \beta_i x_{i,j,k} + \epsilon_{i,j,k} \tag{2}$$

with the same notation as before but species-specific parameters.

(a) (10 points) Summarize the fit of your model and provide estimates of the model parameters. Explicitly include assumptions if there are any. Explain the interpretation for each of the coefficients and determine if there is statistical evidence to say that they are different from 0.

(b) (5 points) Perform a residual analysis of this model.

(c) (5 points) How does this model differ from the model in (1)? What are the advantages and disadvantages of each model, which one would you prefer? Justify your answer.

5. (50 points) Using the same dataset and notation, consider a Bayesian approach for the following model that further accounts for tree-level variation:

$$\begin{aligned}
y_{i,j,k} &= \mu_{ij} + \beta_i x_{i,j,k} + \epsilon_{i,j,k} && (3)\\
\epsilon_{i,j,k} &\sim N(0, \sigma^2) && (4)\\
\mu_{ij} &\sim N(\mu_{i0}, \tau^2), \quad i = 1, 2;\ j = 1, ..., n_i && (5)\\
\beta_i &\sim N(0, \phi^2), \qquad i = 1, 2 && (6)
\end{aligned}$$

Choose appropriate priors for the model parameters σ^2, $\{\mu_{i0}\}_{i=1,2}$, and τ^2, and an appropriate value for ϕ^2.

(a) (20 points) Discuss your prior specification and show the posterior conditionals for your posterior inference. For implementation of the model, you may implement your own MCMC or use existing software. Summarize and study the fit of this model.

(b) (5 points) Summarize the posterior predictive distribution obtained for the Bayesian model above for the LMA that a leaf sampled at 2.330 meters away from the top of one of the Pinus ponderosa would obtain, and that for the LMA of another leaf of the same tree sampled at 10.742 meters away from the top.

(c) (20 points) Consider the following model now with additional tree-level slopes:

$$y_{i,j,k} = \mu_{ij} + \beta_{ij} x_{i,j,k} + \epsilon_{i,j,k} \tag{7}$$

Discuss how you would specify the priors for the parameters and conduct posterior inference. Summarize and study the fit of the model.

(d) (5 points) Which model would you choose, the model you fitted in part (a) or the model you fitted in part (c) above? Justify your answer.