STOR 590: Spring 2020
Take-home Final Exam With Grade Scheme

Answer all questions.

This is a take-home exam that you are expected to do in your own time and hand in no later than **6:00 pm Thursday April 30**. The exam should be submitted via the "Assignments" tab of the course sakai page.

Rules of the Exam. All course resources including text, personal notes and resources available through R or R-Studio are permitted. Your submitted answers should include full verbal answers to the questions, illustrated where appropriate by R code, tables or figures. Very long-winded answers are discouraged; greatest credit will be given for full but concise answers to the questions. Solutions may be submitted in R-Markdown but this is not required. (A fully acceptable alternative is if you submit a Word document into which you cut and paste R output as appropriate; however, I recommend you "save as" a pdf file for the final submission.) Other web resources may be used if fully acknowledged and referenced. Discussion among yourselves or with an outside party is not permitted; you are allowed to email the instructor if you find the question ambiguous or if you think there is an error, but the instructor will not give advice how to solve the problems.

The datasets are posted under the "Resources" tab in sakai. Please download the data **first** and contact the instructor immediately if you have any problem with this step. The datafiles are all in "csv" format and can be loaded into R by typing a command of the form `soappads=read.csv('.../soappads.csv')` where you should insert your own path to identify the file.

If you don't understand or can't do one part of a question, feel free to attempt the later sections first and then go back to the one you skipped. There is no requirement to do the part-questions in the order they appear on the exam. If you feel that some part of a question does not have a clear-cut answer, give your best interpretation.

Please acknowledge you accept the conditions by copying out and signing:

PLEDGE: I will neither give nor receive unauthorized aid in this exam.

SIGNED: (A typed signature will be accepted)

1. The "soappads" dataset contains the result of an experiment to evaluate the quality of soap pads which differ in three respects: amount of detergent (d), coarseness (c) and solubility (s). Each has two possible levels resulting in eight possible treatment combinations labelled 0, d, c, s, dc, ds, cs, dcs according to which variables are applied in which specimen (0 means all three are at their base level). The experiment is conducted over four replicates (1 through 4), two days (1 or 2) and 16 judges (1–16). Each judge evaluates two types of soap pad on each of two days, and gives them a score of 1 through 5, where 1 is best. The ultimate objective of the study is to decide which of the eight possible treatments is best. The data is in the file "soappads.csv".

   (a) For each of the eight possible treatments, calculate the mean score given by the judges. Which treatment(s) come out best by this measure? **[4 points.]**

(b) A possible (fixed effects) model would be to take each of the variables `Judge`, `Replicate`, `Day` and `Treat` as factor variables, and fit a linear regression with `Score` as a continuous (normally distributed) response. Explain why that method would not work. **[4 points.]**

(c) Now consider the variant on part (b) where we drop `Replicate` and `Day` and consider `Judge` and `Treat` as fixed-effects factor variables. Fit the resulting analysis of variance model and say which treatment now shows up best in the sense of minimizing the mean score. Explain in words why this gives a different result from (a). **[6 points.]**

(d) Suppose in part (c) we treat `Judge` as a random effect instead of a fixed effect. Fit the model under this assumption and again say how to interpret the result in terms of which treatment is best. **[6 points.]**

(e) Now consider the alternative viewpoint in which a score of 1 or 2 is considered "Success" and any other score a failure. Refit the model of (c) as a logistic regression model and state which of the eight possible treatments gives the largest probability of a successful result. What difficulties arise in applying this model? **[6 points.]**

(f) Now return to your answer from (d) and state (i) whether you think the treatment you selected as best is indeed better than the others, and (ii) whether there is any statistically significant difference among the eight treatments. State clearly what assumptions or statistical tests you are using to support your answer. **[7 points.]**
[33 points for the whole question.]

2. The "spruce" dataset documents the growth of 79 spruce trees divided among four chambers, labelled 1, 2, 3, 4. Two of the chambers (1 and 2) have a high ozone environment, marked by `tx=0`, and the other two a low-ozone environment (`tx=1`). The `y` variable is the size of the tree (the logarithm of an estimate of total tree volume) and the `day` variable marks the day within the year (all measurements are taken during the summer of a single year). The variable `id` (values 1–79) is a indicator of which tree is which.

(a) Draw a line plot that shows the growth of each tree against time, in two panels where one panel represents the high-ozone environment and the other panel represents the low-ozone environment. Based on the plot, would you say that either environment is beneficial to (i) the overall size, (ii) the rate of growth, of a tree? **[4 points.]**

(b) Analyze the data using an ordinary linear regression, using `day`, `tx` and `chamber` as covariates, but ignoring the fact that there are repeated measures on each tree. Decide which (if any) of these variables should be treated as factor variables, and also which (if any) interactions are appropriate. Use the results of your analysis to answer the questions posed in part(a), and also draw suitable plots to illustrate (i) how well the model fits the data, and (ii) whether the model fit could be improved by a Box-Cox transformation. Summarize your conclusions. **[10 points.]**

(c) Now do a random effects analysis in which `id` is nested within `chamber`, both as random effects. Repeat the analysis of (b), calculating the estimates and standard errors of the fixed-effect terms that depend on `tx`. How do your conclusions compare with those of part (b)? **[7 points.]**

(d) Use the `PBmodcomp` and `exactRLRT` functions to test for the statistical significance of the two random effects, and report your conclusions. **[5 points.]**

(e) Using whichever random effects model you consider appropriate based on (d), conduct a more formal test of the statistical significance of the `tx` effects using a Kenward-Roger test. What is your conclusion? **[3 points.]**

(f) Combining all your answers to the preceding parts, what would you say are the advantages or disadvantages of doing a random effects analysis for this dataset? **[4 points.]**
[33 points for the whole question.]

3. The "schiz" dataset concerns the progress of schizophrenia patients. There are five variables: `ID` is the patient id, `Y` is the symptom indicator (1 means symptoms observed, 0 means no symptoms), `MONTH` is the number of months since the patient was admitted to hospital (0 through 11), `GENDER` is the patient's gender (1=female, 0=male) and `AGE` is an indicator of the patient's age (1 if age < 20, 0 if age ≥ 20). The patients were all admitted to hospital in month 0 and there is a general decline over time of the proportion of patients showing symptoms.

(a) For each month (0 through 11), calculate the proportion of patients showing symptoms in that month and draw a scatterplot. Briefly describe the shape of the scatterplot. **[3 points.]**

(b) On top of the scatterplot you drew in (a), show a fitted smooth curve using (i) the Bowman-Azzalini method (kernel regression with cross-validation choice of bandwidth), (ii) a regression splines model with 3DF. How many DF are needed in the regression splines model for the appearance of the curve to be approximately the same as that of the Bowman-Azzalini estimator? **[7 points.]**

(c) The main question of interest here is the influence of age and gender on the proportion of patients showing symptoms. First, do an analysis using the `glm` command with `MONTH`, `AGE` and `GENDER` all treated as a factor variables. Qualitatively describe the resulting conclusions, e.g. do women show symptoms more frequently than men, and does the effect vary according to age? Do the results change substantially if you use a quasibinomial model instead of binomial? **[5 points.]**

(d) Now repeat the same analysis, treating `ID` as a random effect, using the Gauss-Hermite quadrature method. In what respects do your conclusions differ from those of part (c)? **[5 points.]**

(e) Repeat the analysis, treating `ID` as a random effect, using the GEE approach with AR1 correlation structure. In what respects do your conclusions differ from those of parts (c) and (d)? **[5 points.]**

(f) Repeat the analysis, treating `ID` as a random effect, using the INLA approach. Use 95% posterior intervals as a measure of statistical significance for each of the fixed effects. In what respects do your conclusions differ from those of parts (c) through (e)? **[5 points.]**

(g) Summarize your conclusions. How sensitive are the results of the analysis to the choice of analysis method? **[4 points.]**
[34 points for the whole question.]