**STOR 590: Spring 2020**
**Take-home Midterm Exam**

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 **6pm Friday February 28**. 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.

Please acknowledge you accept these 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. A number of patients were given a new drug which unfortunately has some undesirable after-effects. Table 1 shows how many patients were given the drug at each of seven doses, and how many of them developed after-effects.

| Dose | 0.9 | 1.1 | 1.8 | 2.3 | 3.0 | 3.3 | 4.0 |
|---|---|---|---|---|---|---|---|
| Number of patients | 46 | 72 | 118 | 96 | 84 | 53 | 38 |
| Number with after-effects | 17 | 22 | 52 | 58 | 56 | 43 | 30 |

Table 1: Data for Question 1

[The data are available in a file Drugs.csv, in the Data folder under resources on sakai. To load it from the file, first copy to a directory on our own machine, then use some command like `Drugs=read.csv('Drugs.csv')`. You may need to insert a directory path in front of the file name.]

   (a) Draw a plot that shows how the probability of after-effects is related to dose. Describe the shape of the plot. **[4 points.]**

   (b) Using the glm command, construct a model for the probability of getting after-effects as a function of dose. Show the parameter estimates, standard errors and deviance. **[6 points.]**

   (c) Does the model fit the data? Use the standard diagnostics such as deviance and Pearson statistics, residual plots, leverage and influence. Summarize your conclusions. **[6 points.]**

(d) Do alternative forms of model fit the data better? Consider, in particular, standard alternatives such as including a quadratic term in the Drug variable, and overdispersion. Summarize your conclusions. **[6 points.]**

(e) Using the model from part (b), calculate the probability of after-effects for a dose of 2.6, with a 95% confidence limit. **[6 points.]**

(f) At what level of dose would you estimate the probability of after-effects to be 0.5? **[5 points.]**

**Bonus question:** Calculate a 95% confidence interval for your answer to (f).

2. Backache is a common complaint in pregnant women. To investigate this, researchers conducted a survey among mothers who had just given birth. The data file Backache.csv records the results of the survey. (You can read this in R by first downloading the data to your computer, and then a command of the structure `Back=read.csv('.../Backache.csv')`, as for the previous question.) Just for clarification, "Severity" denotes the severity of the perceived backache on a scale of 0,1,2,3, the variables from "Age" through "PrevBackache" are explanatory variables that related to the mother's background, the variables from "Tablets" through "Walking" are factors that are perceived to relieve backache, while variables from "Fatigue" through "Walking2" are factors perceived to aggravate backache. All the variables that essentially have yes-no responses are recorded 0 for a negative response and 1 for a positive response. Thus, for example, the variable "Walking" is 1 if walking is perceived to relieve the pain and "Walking2" is 1 if walking is perceived to aggravate the pain.

(a) Create two new variables as follows: (a) instead of the "Severity" variable as given, create a binary response y which is 0 if Severity=0 or 1, or $y = 1$ if Severity=2 or 3. (This is more convenient for logistic regression.) (b) Instead of treating "Weightstart" and "Weightend" (meaning mother's weight at the beginning and end of the pregnancy) as two separate variables, create a variable "Weightgain" that represents weight gain during the pregnancy, and then use "Weightstart" and "Weightgain" (but not "Weightend") as covariates in the subsequent analysis. **[2 points.]**

(b) Do you think there are potential outliers or even erroneous observations in the data? Construct plots or tables (as appropriate) to illustrate your answer, and make adjustments to the data if required. **[5 points.]**

(c) Use appropriate plots to examine the relationship between backache and (i) age, (ii) weight gain, (iii) number of previous children. Based on visual analyses alone, say which one(s) you think are important. **[5 points.]**

(d) Now conduct a formal logistic regression analysis, using the variables from "Age" through "PrevBackache" as predictors. You should decide which variables are significant using any standard variable selection technique. Summarize your conclusions. **[6 points.]**

(e) For the analysis you did in (d), construct suitable diagnostic plots to judge how well the model fits the data. Also, construct the Hosmer-Lemeshow test to judge overall fit, and summarize your conclusions. **[6 points.]**

(f) Of the eight variables from "Tablets" to "Walking" that are identified as relieving backache, which ones seem to have an effect? Use any appropriate statistical procedure. **[3 points.]**

(g) Of the eight variables from "Fatigue" to "Walking2" that are identified as relieving backache, which ones seem to have an effect? Use any appropriate statistical procedure. **[3 points.]**

(h) Write a short plain-English summary of your conclusions, that would be understandable to a doctor who has never taken a statistics course. **[4 points.]**

3. (See text, question 8.2.) An experiment was conducted as part of an investigation of the effect of certain toxic agents. The survival time of rats depended on the type of poison and the treatment. The data are in `rats` in the Faraway package.

(a) Make plots of the data and comment on the different between treatments and poisons. **[4 points.]**

(b) Fit a standard (normal-theory)linear model with an interaction between the two predictors. Use a Box-Cox transformation to determine an optimal transformation of the response. Can this optimal transformation be rounded to a more interpretable response? **[6 points.]**

(c) Refit the model using your chosen transformation. Draw suitable plots to determine whether the model is a good fit. **[6 points.]**

(d) Is the interaction statistically significant? Simplify the model if justified and determine which combination of poison and treatment will result in the shortest survival time. **[5 points.]**

(e) Build an inverse Guassian GLM for this data. Based on your previous modeling, select an appropriate link function. Based on this model, which combination of poison and treatment will result in the shortest survival time? **[7 points.]**

(f) Compare the predicted values on the original scale of the response for the two models. Do the two models produce a similar fit? **[5 points.]**