York University

MATH 4330 – Sample Final Questions

Professor Georges Monette

December 2017

WARNING

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE INSTRUCTED TO DO SO

Student number: _____

Family name: (in block letters)_____

Given name: (in block letters)_____

Signature_____

Information: This exam has 32 questions. Make sure you complete every question. For multiple-choice questions, select the best answer. More than one may be somewhat correct, but one answer will be the "best" or most precise response to the question. For other questions, make sure to read the question closely. Some may ask for multiple pieces of information. Make sure to respond completely. If you need more space to answer, write "**OVER**" and continue the answer on the back of the page. The point value is shown at the end of each question. The sum of the points is NA.

Aids allowed: Non-programmable calculator, ruler, pencils, pens, erasers. One letter-size (8.5" x 11") sheet of paper with formulas and notes (both sides).

WARNING

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE INSTRUCTED TO DO SO

1. In order to assess factors related to reckless driving behaviors, investigators ran a study in which observers at different intersections with a stop sign recorded the number of cars that did not stop properly along with various information on each violator, including gender of the driver, type of car (sedan, sports utility, mini-van, wagon, truck, other), and approximate age of the driver (under 30, 30-40, 40-50, 50+). Pooling information from several intersections and for different observers, investigators recorded the number of violators in each category.
   Describe a generalized linear model for analyzing these data (specify a reasonable, possibly non-linear, model and its transformation to linear form as well as a conditional distribution and link function) and outline a specific approach for assessing the following questions of interest using frequentist methods: (1) is gender an important predictor? (2) is type of car an important predictor, and if so which types are predictive of a greater frequency of violations? (3) is there a trend with age in the frequency of violations? *(30 points)*

2. In marketing research, it is widely believed that subliminal messages in advertisements can be effective in improving a consumer's impression of a product. To test whether the type and frequency of the subliminal message has an impact, investigators ran a study in which they enrolled male and female graduate students. Study subjects were all shown an outwardly-identical taped advertisement for a soft drink and then asked to rank their impression of the soft drink after watching the tape on a scale from 1-5 (1=strongly negative, 2=mildly negative, 3=no opinion, 4=mildly positive, 5=strongly positive). Tapes varied in the type of subliminal message (1=none, 2=attractive female face, 3=attractive male face) and (for tapes with a subliminal message) the frequency (1=low, 2=medium, 3=high).
   Describe a regression model for relating gender, type of subliminal message, and frequency to an individual's impression. Describe the specifics of a Bayesian approach for addressing the following questions of interest to the investigator: (1) do subliminal messages have an effect overall? (2) does this effect vary depending on the gender of the observer? (3) does the effect depend on the frequency? and (4) do males and females respond differently depending on the type of subliminal message? Detail the prior used, the form of the regression model, and the likelihood. Provide an outline of the method used for posterior computation and inferences on the above questions of interest. *(30 points)*

3. Suppose that a random variable Y has a Poisson distribution with mean λ were $ln(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Assuming that X_2 is not constant and that $\beta_2 \neq 0$, show that Y does not have a Poisson distribution if one fits a model using only X_1. *(20 points)*

4. Uterine fibroids are a common reproductive tract tumor. To study factors related to fibroid incidence (that is, the rate of onset for women who do not have fibroids), women aged 20-30 who did not have fibroids at a baseline examination were enrolled in a prospective study. These women were then given a screening examination approximately every 5 years (though the specific ages at examination varied) to assess whether fibroids had yet developed. Each woman was followed for 15 years (3 examinations) or until she either developed fibroids or dropped out of the study. Various information was collected for each women, including race (white, black, other), age at menarche, age at entry into the study, age at each examination, and whether the woman was a smoker.
   Assuming that fibroids do not go away once they have developed, describe a regression model for these data that allows fibroid incidence to vary with age and other factors. Show the observed data likelihood under this regression model, and develop a Bayesian approach for estimating age-specific fibroid incidence for women in different groups. Detail the prior used and the form of the posterior. Outline an approach for posterior computation and describe specifically how you obtain point and interval estimates for the probability of developing fibroids by a given age for women in different groups. *(30 points)*

5. For adults who sailed on the Titanic on its fateful voyage, the odds ratio between gender (female, male) and survival (yes, no) was 11.4. What is wrong with the interpretation, "The probability of survival for females was 11.4 times that for males." . *(5 points)*

6. *(continued from previous question)* When would the quoted interpretation be approximately correct? Why?. *(5 points)*

7. *(continued from previous question)* The odds of survival for females equaled 2.9. For each gender, find

the proportion who survived. *(5 points)*

8. Explain what is meant by overdispersion, and explain how it can occur for Poisson generalized linear models for count data. *(10 points)*

9. Explain two ways in which the generalized linear model extends the ordinary regression model that is commonly used for quantitative response variables.. *(5 points)*

10. Each of 100 multiple-choice questions on an exam has five possible answers but one correct response. For each question, a student randomly selects one response as the answer. Specify the probability distribution of the student's number of correct answers on the exam, identifying the parameter(s) for that distribution. Would it be surprising if the student made at least 50 correct responses? Explain your reasoning.. *(10 points)*

11. Suppose `y`, `x`, and `z` are numerical variables in the R data frame `dd`.
Explain the difference between a linear model fitted with the formula `y ~ x*z` in comparison with the formula `y ~ I(x*z)`.. *(5 points)*

12. There is a lab test for a rare disease D that has a specificity of .98 and a sensitivity of .95. Suppose the prevalence of the disease is .01%. Explain how, if one thinks of the lab test as a hypothesis test for the null hypothesis of no disease, a positive result produces a p-value of 0.01.. *(5 points)*

13. *(continued from previous question)* If someone selected at random (for example if the test is used to screen for the disease) is given the test and gets a positive result, what is the probability that they have the disease? *(5 points)*

14. An article states that the PSA blood test for detecting prostate cancer stated that, of men who had this disease, the test fails to detect prostate cancer in 1 in 4, and, of men who do not have it, approximately 2/3 received false positive results. Let D (D^c) denote the event of having (not having) prostate cancer and let Pos (Neg) denote a positive (negative) test result. What is the sensitivity and specificity of this test? (from Agresti, 2007. *(3 points)*

15. *(continued from previous question)* Of men who take the PSA test, 1% have the disease. Find the cell probabilities in the 2×2 table for the joint distribution for having (not having) the disease versus a positive or negative test result. *(4 points)*

16. *(continued from previous question)* Find the probability of having prostate cancer given a positive test result and find the probability of not having the disease given a negative test result. *(4 points)*

17. A lab test for a rare disease D has a specificity of .95 and a sensitivity of .95. Suppose the prevalence of the disease is .01%. What proportion of the time is the test in error? *(4 points)*

18. *(continued from previous question)* Given a positive test result, what is probability of error? *(4 points)*

19. *(continued from previous question)* Given a negative test result, what is probability of error? *(4 points)*

20. *(continued from previous question)* How do you reconcile the 3 preceding results? *(4 points)*

21. A British study in 1998 report that, of smokers who get lung cancer, "women were 1.7 times more vulnerable than men to get small-cell lung cancer". What kind of statistic is the figure '1.7' reported here? (Agresti, 2007. *(4 points)*

22. A National Cancer Institute study about tamoxifen and breast cancer reported in 1998 that the women taking the drug were 45% less likely to experience invasive breast cancer compared with the women taking placebo. Find the relative risk for (i) those taking the drug compared to those taking placebo, (ii) those taking placebo compared to those taking the drug. (Agresti, 2007. *(4 points)*

23. In the United States, data reported in 1993 indicates that the annual probability that a woman over the age of 35 dies of lung cancer equals 0.001304 for current smokers and 0.000121 for nonsmokers. Calculate and interpret the difference of proportions, the relative risk and the odds ratio. Which is more informative? Why? (Agresti, 2007. *(4 points)*

24. *(continued from previous question)* Are the odds ratio and relative risk similar or dissimilar? Why? *(2 points)*

25. *(continued from previous question)* Give a real world example of three variables, X, Y, and Z, for which expect X and Y to be marginally associated but conditionally independent controlling for Z. (Agresti, 2007. *(4 points)*

26. With a 2×2 table what is the effect of interchanging two rows on the odds ratio? On the log-odds ratio? *(2 points)*

END OF EXAM