York University

MATH 4330 – Final Exam

Instructor: Georges Monette

December 11, 2018 – 2:00 pm to 4:00 pm (120 minutes)

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:

Be sure to read questions 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 marks for each questions are shown at the end of the question. The sum of the marks is 110. The exam will be graded out of 100 so that you can potentially earn 10 bonus points.

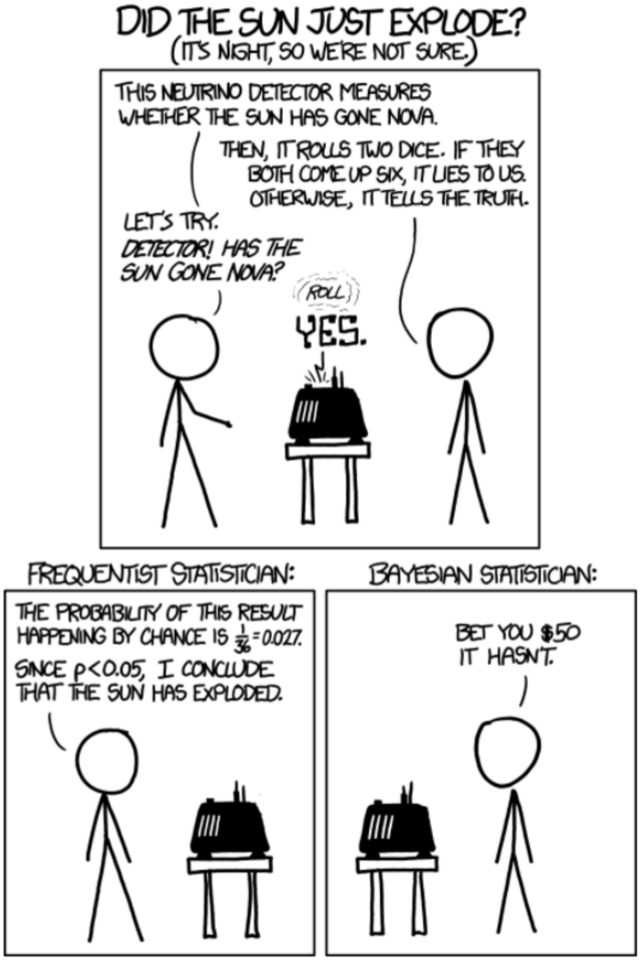
Aids allowed: Non-programmable calculator, ruler, pencils, pens, erasers.

WARNING

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

1. Let Y be the number of fatal traffic accidents in one day in Toronto. Suppose the expected number of fatal traffic accidents depends on a number of variables and that, if all of these variables were taken into account, the conditional distribution of Y would be Poisson. Prove that, assuming that the variables are not all constant and that the conditional mean of Y depends non-trivially on these variables, the unconditional distribution of Y itself is overdispersed relative to that of a Poisson distribution. *(10 points)*

2. The following XKCD cartoon shows two statisticians interpreting the same data: one who uses a frequentist approach unquestioningly and one who uses a Bayesian approach.


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

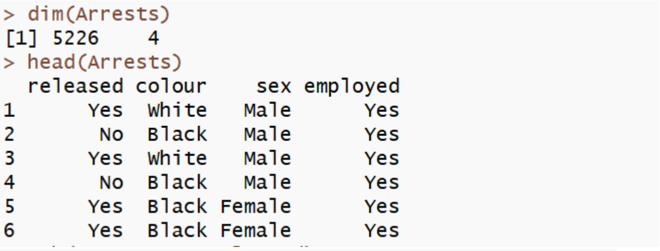
Discuss whether the frequentist statistician is justified in saying that $p < 0.05$ from a frequentist point of view. Make some reasonable assumptions, stating them explicitly, and calculate a reasonable value for the Bayesian statistician's posterior probability that corresponds to the frequentist's p-value? If there is a difference between them, briefly discuss why. Under what circumstances would you expect a p-value to be close to a corresponding posterior probability? *(20 points)*

3. Suppose you have observational data and you are interested in the causal effect of a variable X on a variable Y. You also have data on a variable Z which is highly correlated with X. You notice that including Z in your model greatly increases the standard error of the coefficient for X. You do some research on the internet and find discussions on multicollinearity and variance inflation factors that suggest that, in the presence of high multicollinearity, you should consider dropping Z to improve your estimate of the effect of X or you should consider using principal components analysis. Discuss how you would go about deciding on the right course of action. *(10 points)*

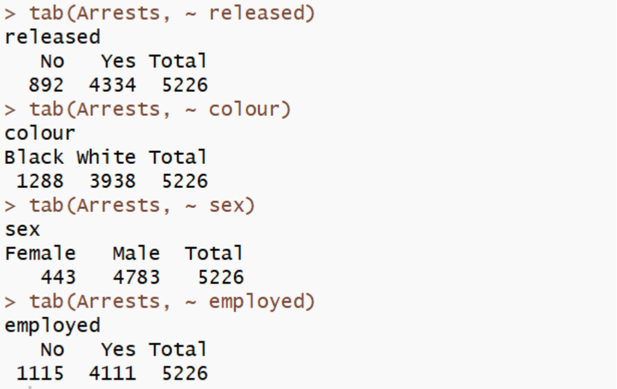
4. Mary is choosing between restaurants A and B to take her friend, John, out for dinner. Restaurant A has an average rating of 4.1 and restaurant B of 4.3. But looking at ratings by gender, among men, A has a rating of 4.0 and restaurant B of 3.8. Among women, restaurant A has a rating of 4.6 and restaurant B of 4.4. It seems that men and women separately prefer restaurant A but together prefer restaurant B. Draw a Paik-Agresti diagram and explain how this apparent contradiction could arise. Which restaurant should Mary choose and why? Briefly describe the principle that guides your analysis. *(10 points)*

5. (continued from the previous question) Suppose Mary is choosing between restaurants A and B as above and the overall ratings are the same as in the previous question but they are classified, not by the gender of the customer, but according to the waiter at each restaurant. Each restaurant has a good waiter and a bad waiter. Customers can't choose the waiter they get. Ratings among clients getting the good waiter are 4.6 at restaurant A and 4.4 at restaurant B. Ratings among clients getting the bad waiter are 4.0 at restaurant A and 3.8 at restaurant B. Suppose that your chances of getting each type of waiter at either restaurant are similar to the proportions in the ratings. Explain how this apparent contradiction could arise. Which restaurant should Mary choose and why? Briefly describe the principle that guides your analysis. *(10 points)*

6. A study of arrests for posession of marijuana in Toronto in the early 2000s recorded data for 5,226 arrests by Toronto police over a period of approximately 2 years. For each arrest we consider the variables: colour of the person arrested (Black or White), sex (Male or Female), employed (Yes or No) and ‘released’ (Yes or No) according to whether the person arrested was released directly on the spot by the police or whether they were taken to jail before being released on bail.

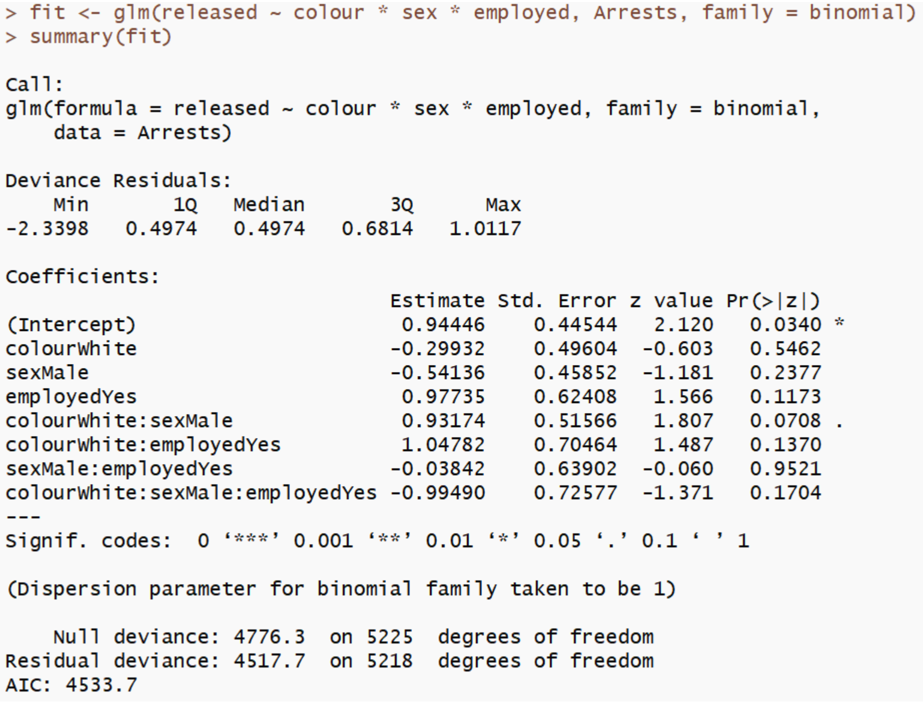


```
> dim(Arrests)
[1] 5226    4
> head(Arrests)
  released colour    sex employed
1      Yes  White   Male      Yes
2       No  Black   Male      Yes
3      Yes  White   Male      Yes
4       No  Black   Male      Yes
5      Yes  Black Female      Yes
6      Yes  Black Female      Yes
```



```
> tab(Arrests, ~ released)
released
   No   Yes Total
  892  4334  5226
> tab(Arrests, ~ colour)
colour
Black White Total
 1288  3938  5226
> tab(Arrests, ~ sex)
sex
Female   Male  Total
   443   4783   5226
> tab(Arrests, ~ employed)
employed
   No   Yes Total
 1115  4111  5226
```

The following is some output from a logistic regression of ‘released’ on the other variables:



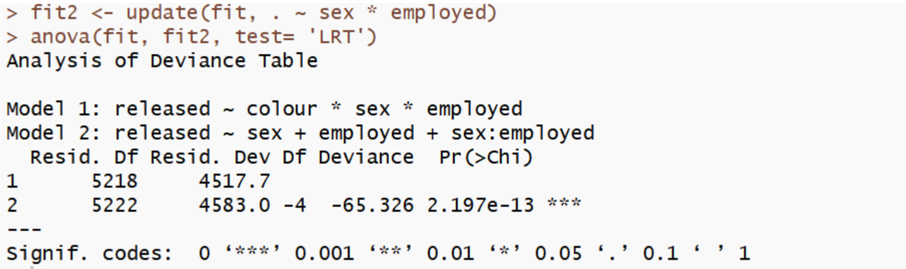
```
> fit <- glm(released ~ colour * sex * employed, Arrests, family = binomial)
> summary(fit)

Call:
glm(formula = released ~ colour * sex * employed, family = binomial,
    data = Arrests)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3398   0.4974   0.4974   0.6814   1.0117

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      0.94446    0.44544   2.120   0.0340 *
colourWhite                     -0.29932    0.49604  -0.603   0.5462
sexMale                         -0.54136    0.45852  -1.181   0.2377
employedYes                      0.97735    0.62408   1.566   0.1173
colourWhite:sexMale              0.93174    0.51566   1.807   0.0708 .
colourWhite:employedYes          1.04782    0.70464   1.487   0.1370
sexMale:employedYes             -0.03842    0.63902  -0.060   0.9521
colourWhite:sexMale:employedYes -0.99490    0.72577  -1.371   0.1704
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4776.3  on 5225  degrees of freedom
Residual deviance: 4517.7  on 5218  degrees of freedom
AIC: 4533.7
```

A colleague of yours notes that none of the coefficients are significant and concludes that there is no evidence that ‘colour,’ in particular, is related to the probability of release. What would you say to your colleague? *(10 points)*

7. (continued from the previous question) What is the predicted probability of being released for an unemployed black male? What is the predicted probability of being released for an unemployed white male? *(10 points)*

8. (continued from the previous question) Write the linear hypothesis matrix you would use to carry out a Wald test of the hypothesis that there is no effect of ‘employed’ status. *(10 points)*

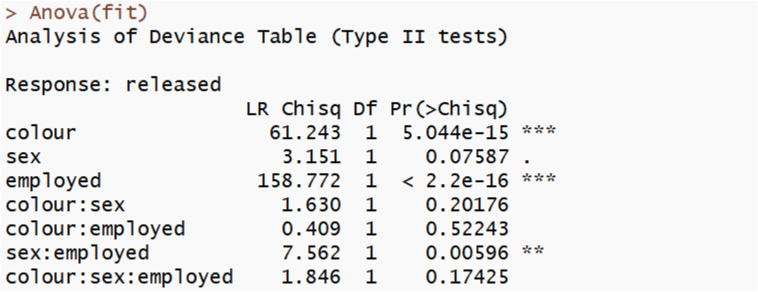
9. (continued from the previous question) Consider the following output:



```
> fit2 <- update(fit, . ~ sex * employed)
> anova(fit, fit2, test= 'LRT')
Analysis of Deviance Table

Model 1: released ~ colour * sex * employed
Model 2: released ~ sex + employed + sex:employed
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      5218     4517.7
2      5222     4583.0 -4  -65.326 2.197e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Describe the implications of this ouput. *(10 points)*

10. (continued from the previous question) Consider the following output:



```
> Anova(fit)
Analysis of Deviance Table (Type II tests)

Response: released
                   LR Chisq Df Pr(>Chisq)    
colour               61.243  1  5.044e-15 ***
sex                   3.151  1    0.07587 .  
employed            158.772  1  < 2.2e-16 ***
colour:sex            1.630  1    0.20176    
colour:employed       0.409  1    0.52243    
sex:employed          7.562  1    0.00596 ** 
colour:sex:employed   1.846  1    0.17425    
```

Describe unambiguously the null and alternative hypotheses tested in the first line of the anova table and in the sixth line of the anova table. *(10 points)*