York University

MATH 4939 – Midterm

Instructor: Georges Monette

February 15, 2019 – 10:30 am to 11:20 am (50 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 80.

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

WARNING

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

1. Describe how variable selection strategies should be affected by the purpose of an analysis and the way variables were obtained. *(10 points)*

Purpose: causal or predictive
Data: observational or experimental
i.e was there random assignment

Elaborate on variable selection
in relevant combinations of above.

Be ready for: Q: What are the first 3 questions you should ask before starting an analysis?

2. 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)*

The test for a possible relationship with colour is a joint test that all the parameters involving colour are jointly 0. You can't carry out a joint test using p-values for individual components of that test — as we have repeatedly seen.

3. (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 third line of the anova table and in the sixth line of the anova table. What is the 'real-world' interpretation of these tests? *(10 points)*

$H_0: \beta_0 + \beta_1\,\text{colour} + \beta_2\,\text{sex} + \beta_4\,\text{colour} \times \text{sex}$

$H_A: \beta_0 + \beta_1\,\text{colour} + \beta_2\,\text{sex} + \beta_3\,\text{employed} + \beta_4\,\text{colour} \times \text{sex}.$

OR $H_0: \beta_3 = 0$ in the full model:

$\beta_0 + \beta_1\,\text{colour} + \beta_2\,\text{sex} + \beta_3\,\text{employed} + \beta_4\,\text{colour} \times \text{sex}$

4. Consider the general linear model with the usual notation. Let $\eta_1 = L_1\beta$ and $\eta_2 = L_2\beta$. Suppose that within both L_1 and L_2 the rows are linearly independent and that the rows of L_1 can be expressed as linear combinations of the rows of L_2. Show that the Wald tests for $\eta_1 = 0$ and for $\eta_2 = 0$ come to identical conclusions. *(10 points)*

Error!!. This needs to go both ways. Otherwise the row space of L_1 could be a subspace of the row space of L_2

Since the rows of L_1 can be expressed as LCs of the rows of L_2, there exists a matrix A such that

$$L_1 = AL_2$$

Here I made a mistake because we need to know that A^{-1} exists.

The test for $\eta_1 = 0$ has numerator

$$(L_1\hat\beta)'(L_1(X'X)^{-1}L_1')^{-1}L_1\hat\beta$$

$$= \hat\beta' L_1'(L_1(X'X)^{-1}L_1')^{-1}L_1\hat\beta$$

5. Write a generic function `tran` and a set of methods so that `tran(x, a, b)` replaces every instance of the value `a` in `x` with `b`. e.g. `tran(c(1,2,2), 2, 3)` should return the vector `1, 3, 3`. The function should work with numeric, character and factor objects and should return an object of the same type. *(10 points)*

$$= \hat\beta'(AL_2)'(AL_2(X'X)^{-1}(AL_2)')^{-1}AL_2\hat\beta$$

$$= \hat\beta' L_2'A'(AL_2(X'X)^{-1}L'A')^{-1}AL_2\hat\beta$$

$$= \hat\beta' L_2'A'A'^{-1}(L_2(X'X)^{-1}L_2')^{-1}A^{-1}AL_2\hat\beta$$

$$= \hat\beta L_2'(L_2(X'X)^{-1}L_2')^{-1}L_2\hat\beta$$

which is the numerator in the test for $\eta_2 = 0$

6. Consider of vector of strings containing names of people. Each string contains one name which can be in various formats: 'Mary Ellen Brown' (i.e. first name followed by middle name if any) and by last name), 'Brown, Mary Ellen' (last name, followed by first and middle names), 'Paul Smith' (if there is no middle name) or 'Paul, Smith'. Write a function in R that takes two arguments: a vector of such strings and a single character string. The function counts how often the second argument occurs as a last name in the vector that is the first argument. *(10 points)*

7. Consider the following function in R:

```
f <- function(x = {y <- 5; 2}, y = 10) { x + y }
```

State what the following expression will return and explain why:

a) `f()`
b) `f(20)`
c) `f(-10,2)`
d) `f(y = 20)`
e) `f(x = 21)`
(10 points)

this does not happen until R needs to get the default value for X. If the function has a value for x, either by position or by name, then this never happens (i.e. is never evaluated)

8. The following is some output from a linear regression of life expectancy in a number of countries regressed on HE (health expenditures per capita per year in dollars US), smoke (cigarettes per capita per year), hiv and special, that are two indicator variable to identify anomalous countries.

```
fit.hiv2 <- lm( LifeExp ~ log(HE) * (smoke + I(smoke^2)) + hiv+special, dd ,
                na.action = na.exclude)
summary(fit.hiv2)
```

```
|
|   Call:
|   lm(formula = LifeExp ~ log(HE) * (smoke + I(smoke^2)) + hiv +
|       special, data = dd, na.action = na.exclude)
|
|   Residuals:
|       Min      1Q  Median      3Q     Max
|   -9.0373 -2.3005  0.2043  2.0760  9.7344
|
|   Coefficients:
|                        Estimate Std. Error t value Pr(>|t|)
|   (Intercept)         3.283e+01  2.674e+00  12.280  < 2e-16 ***
|   log(HE)             6.091e+00  5.024e-01  12.124  < 2e-16 ***
|   smoke               3.642e-02  7.520e-03   4.844 3.31e-06 ***
|   I(smoke^2)         -1.518e-05  3.946e-06  -3.846 0.000181 ***
|   hiv                -7.351e-01  7.593e-02  -9.681  < 2e-16 ***
|   special            -1.822e+01  2.137e+00  -8.526 2.11e-14 ***
|   log(HE):smoke      -4.878e-03  1.155e-03  -4.223 4.30e-05 ***
|   log(HE):I(smoke^2)  2.007e-06  5.726e-07   3.504 0.000614 ***
```

a) Construct a hypothesis matrix to estimate the predicted difference in life expectancy associated with an increase of 1,000 cigarettes per capita per year for a country with a level of health expenditures equal to 2,000 and cigarette consumption equal to 1,000.
b) Construct a hypothesis matrix to estimate the predicted difference in life expectancy associated with an increase of 1,000 dollars in health expenditures per capita per year for a country with a level of health expenditures equal to 2,000 and cigarette consumption equal to 1,000.
(10 points)

a) Since non-linear in smoke, can use difference

$L_2 = [\,1 \quad \log(2000) \quad 2000 \quad 2000^2 \quad hiv \quad special \quad \log(2000)\times 2000 \quad \log(2000)\times 2000^2\,]$

$L_1 = [\,1 \quad \log(2000) \quad 1000 \quad 1000^2 \quad hiv \quad special \quad \log(2000)\times 1000 \quad \log(2000)\times 1000^2\,]$

$L_{diff} = L_2 - L_1$

$= [\,0 \quad 0 \quad 1000 \quad 2000^2 - 1000^2 \quad 0 \quad 0 \quad \log(2000) \times 1000 \quad \log(2000)[2000^2 - 1000^2]\,]$

b) Similar