UNIVERSITY OF TORONTO
FACULTY OF ARTS AND SCIENCE

FINAL EXAMINATION, AUGUST 2016

DURATION: 3 hours

STA 303 H1S/STA 1002 H1S — Methods of Data Analysis II

Aids allowed: Non-programmable calculators
Examiner(s): M. Guerzhoy

Please detach aid sheet if necessary

Student Number: ⌊_|_|_|_|_|_|_|_|_|_|_|_|_⌋

Family Name(s): _____

Given Name(s): _____

---

*Do **not** turn this page until you have received the signal to start.*
In the meantime, please read the instructions below *carefully*.

This final examination paper consists of 8 questions on 36 pages (including this one), printed on both sides of the paper. *When you receive the signal to start, please make sure that your copy is complete, fill in the identification section above, and write your student number where indicated at the bottom of every odd-numbered page (except page 1).*

Answer each question directly on this paper, in the space provided, and use the reverse side of the previous page for rough work. If you need more space for one of your solutions, use the reverse side of a page or the pages at the end of the exam and *indicate clearly the part of your work that should be marked.*

Write up your solutions carefully! In particular, use notation and terminology correctly and explain what you are trying to do—part marks *will* be given for showing that you know some aspects of the answer, even if your solution is incomplete.

MARKING GUIDE

# 1:	_____/ 25
# 2:	_____/ 25
# 3:	_____/ 10
# 4:	_____/ 10
# 5:	_____/ 10
# 6:	_____/ 10
# 7:	_____/ 15
# 8:	_____/ 10
TOTAL:	_____/115

Good Luck!

OVER...

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Question 1. [25 MARKS]

The Donner Party was a large group of people who were travelling to California in a wagon train. The Donner Party was trapped in the Sierra Nevada mountain range for more than three months because of heavy snowfall. Their food supplies were exhausted, and of the 87 members of the party, only 48 survived. For 45 of the members of the Donner Party (30 men and 15 women), we have data on their age, sex, and status (i.e., whether they survived or didn't survive.)

Here are several lines from the data table.

```
> donner
   AGE    SEX   STATUS
1   23   MALE     DIED
2   40 FEMALE SURVIVED
...
44  24   MALE     DIED
45  25 FEMALE SURVIVED
```

Part (a) [5 MARKS]

We are interested in using logistic regression (with the logit link function) in order to predict whether a member of the party survived or died. What is the logistic regression model for the data? Define all variables.

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (b) [5 MARKS]

What are the model assumptions of logistic regression? For each assumption, state whether the assumption is likely satisfied for the Donner Party data, based on the description of the dataset, and explain your reasoning. If it is impossible to say anything about whether the assumption is likely satisfied or not based on the description of the dataset, say so, and briefly explain your reasoning.

Use this page for rough work—clearly indicate any section(s) to be marked.

The following model was fit to the Donner Party data, predicting the survival of a member of the Party (i.e., $y_i = 1$ means person i survived.)

```
Call:
glm(formula = STATUS ~ SEX + AGE, family = "binomial", data = donner)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7445  -1.0441  -0.3029   0.8877   2.0472

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.63312    1.11018   1.471   0.1413
SEXFEMALE    1.59729    0.75547   2.114   0.0345 *
AGE         -0.07820    0.03728  -2.097   0.0359 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 51.256  on 42  degrees of freedom
AIC: 57.256

Number of Fisher Scoring iterations: 4
```

Part (c) [6 MARKS]

Based on the output above, and on the information about the dataset provided on the previous page, what is the **probability** of survival for a 30-year-old man? Provide a point estimate and a 95% prediction interval.

Part (d) [4 MARKS]

Suppose that a logistic regression model were fit to the same data, but predicting the death of a member of the Party (i.e., $y_i = 1$ would mean that person i died.) Furthermore, suppose that the reference sex were female, so that a `SEXMALE` coefficient would have to be estimated. What would be the `(Intercept)`, `SEXMALE`, and `AGE` coefficients in that case? You only need to provide the estimates, not the standard errors. Show your work.

`(Intercept)`:

`SEXMALE`:

`AGE`:

Use this page for rough work—clearly indicate any section(s) to be marked.

The following model was fit, using the same data as before.

```
Call:
glm(formula = STATUS ~ SEX + AGE + SEX * AGE, family = "binomial",
    data = donner)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2279  -0.9388  -0.5550   0.7794   1.6998

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.31834    1.13103   0.281   0.7784
SEXFEMALE      6.92805    3.39887   2.038   0.0415 *
AGE           -0.03248    0.03527  -0.921   0.3571
SEXFEMALE:AGE -0.16160    0.09426  -1.714   0.0865 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 47.346  on 41  degrees of freedom
AIC: 55.346

Number of Fisher Scoring iterations: 5
```

Part (e) [5 MARKS]

Suppose you want to conduct a Likelihood Ratio Test to determine whether there is evidence for an interaction between SEX and AGE in the model. You can only use the output provided to you and the `pchisq` function (you can also use the table if you precisely state how you used it) Describe precisely how to determine whether there is evidence for an interaction between SEX and AGE in the model.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 2. [25 MARKS]

A study was conducted in order to determine whether age influences the number of successful matings of male elephants. Some of the data obtained looks as follows.

```
   Age Matings
1   27       0
2   28       1
...
38  45       5
39  47       7
40  48       2
41  52       9
```

The following model was fit to the data:

```
glm(formula = Matings ~ Age, family = poisson(link = log), data = elephants)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.58201    0.58590  -2.700   0.0102 *
Age          0.06869    0.01479   4.645 3.81e-05 ***
```

Part (a) [5 MARKS]

What is the model for the data that is assumed when using the R code above? Use as many numerical values as possible.

Part (b) [2 MARKS]

What is the interpretation of the coefficient that corresponds to `Age`?

Part (c) [3 MARKS]

What is the interpretation of the intercept in the model?

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (d) [5 MARKS]

Sketch the residual plot that you would expect to see for the data if the Poisson regression regression model fits the data well. Reminder: in a residual plot, the x-axis is the *predictions*, and the y-axis is the *difference between the observed values and the predictions* (Note: those are not the deviance or Pearson residuals). Briefly **state what properties of the plot you tried to visualize**. Label the axes.

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (e) [5 MARKS]

Recall that in simple linear regression, the coefficients are obtained by finding β_0 and β_1 such that $\sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$ is minimized. How are the coefficients obtained for Poisson Regression?

Part (f) [5 MARKS]

Assuming the Poisson Regression model is correct, is there evidence that age influences the number of matings? Briefly state how you got your answer from the R output.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 3. [10 MARKS]

Part (a) [5 MARKS]

The dispersion parameter ψ for the elephant data from the previous question was estimated to be 1.15. Using only functions such as `tf`, `pf`, and `qf`, how would you estimate the p-value for the null hypothesis that age does not influence the number of matings? (Alternatively, state precisely how you would use probability tables.)

Part (b) [5 MARKS]

Describe a scenario (i.e., a story about the data in the question) that would explain the presence of overdispersion in the elephant data.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 4. [10 MARKS]

Part (a) [5 MARKS]

Recall that you can generate k Bernoulli random variables with success probability p using `rbinom(k, size=1, prob=p)`. Write R code to generate a Binomial random variable with success probability p and 3 trials. You can only use `rbinom` to generate random variables, and you are only allowed to use `size=1`.

Part (b) [5 MARKS]

Now, write R code to generate a random variable whose distribution is very close to $N(0, 1)$. You can only use `rbinom` to generate random variables (in particular, you may *not* use `rnorm`, and you are only allowed to use `size=1`).

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 5. [10 MARKS]

Suppose that we are running an observational study. The goal of the study is to understand the relationship between students' grades in STA302 and their year of study and Program of Study (POSt). In our study, the students are all in one of the Statistics major, the Mathematics major, or the Economics major. For each student, the year of study is either First Year, Second Year, Third Year, or Fourth Year (in the data, the year of study is an integer). The model is

$$y_i = \beta_0 + \beta_1 I_{stat,i} + \beta_2 I_{math,i} + \beta_3 I_{econ,i} + \beta_4 Year_i + \beta_5 I_{stat,i} Year_i + \beta_6 I_{math,i} Year_i + \beta_6 I_{econ,i} Year_i + \epsilon_i.$$

Here, y_i is the STA302 grade of student i.

Part (a) [5 MARKS]

Describe a scenario (i.e., a plausible story about the data in the question) where there is an interaction between the major and the year of study.

Part (b) [5 MARKS]

Draw two interaction plots: one where there is an interaction between Major and Year in the data, and one where there is not an interaction Major and Year in the data. Label the axes in the plots. State which plot indicates the presence of an interaction, and which plot indicates the absence of an interaction.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 6. [10 MARKS]

Part (a) [4 MARKS]

What are two purposes of cross-validation?

Part (b) [6 MARKS]

Explain how k-fold cross-validation works. In your explanation, explain how to compute at least two different cost functions (note: explain using math or English, not R code) used when using cross-validation.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 7. [15 MARKS]

Data was collected on the price of various items in several grocery stores. Some of the rows in the data table look as follows:

```
> prices
                item  store price
1            lettuce storeA  1.17
...
12          potatoes storeB  1.98
13              milk storeB  1.69
...
```

Here is some R output:

```
Linear mixed model fit by REML ['lmerMod']
Formula: price ~ item + (1 | store)
   Data: prices

Random effects:
 Groups   Name        Variance Std.Dev.
 store    (Intercept) 0.01503  0.1226
 Residual             0.04495  0.2120
Number of obs: 40, groups:  store, 4

Fixed effects:
                      Estimate Std. Error t value
(Intercept)             4.8600     0.1225   39.69
itembread              -3.0950     0.1499  -20.64
itemcereal             -1.7700     0.1499  -11.81
itemeggs               -4.0050     0.1499  -26.71
itemground.beef        -2.7225     0.1499  -18.16
itemlaundry.detergent   1.3550     0.1499    9.04
itemlettuce            -3.4775     0.1499  -23.20
itemmilk               -3.2200     0.1499  -21.48
itempotatoes           -2.9275     0.1499  -19.53
itemtomato.soup        -4.2075     0.1499  -28.06
```

Part (a) [5 MARKS]

What is the model that was fit to the data in the R code above? Define all the variables, and use numerical values as possible from the R code above. For ease of writing, you can use "..." and only consider the price of bread and cereal.

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (b) [2 MARKS]

What does the model imply about the differences in prices between different stores? (See Part (c) for a hint.)

Part (c) [3 MARKS]

If we happenned to know that some stores are more expensive than other stores by a constant factor (e.g., the prices at Loblaws are 10% larger than at Metro), how could we modify the R code to better model those kinds of differences? Answer this question with the modified R code.

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (d) [5 MARKS]

We are interested in whether the prices for lettuce, milk, and potatoes are different from each other (that is, if the prices for one of the itmes is different from the price for another of the two items.) Describe how to conduct a statistical test to answer this question at 95% confidence. You may use R functions such as `qnorm` and `pnorm`, but not other functions.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 8. [10 MARKS]

Recall that in the Pygmalion Effect dataset, we had platoons, each of which belonged to a company, and each of which either had or had not the Pygmalion Effect treatment. In a context of a dataset similar to the Pygmalion Effect dataset, describe a scenario (i.e., a story about the data) where the MSR (Mean Square Regression) is expected to be **smaller** than the MSE (Mean Square Error). You may use any reasonable simplification (for example, you can assume that there are only two companies.) Write code to randomly generate a dataset where you expect the MSR to be smaller than the MSE.

Use this page for rough work—clearly indicate any section(s) to be marked.

This page was intentionally left blank

Use this page for rough work—clearly indicate any section(s) to be marked.

This page was intentionally left blank

PLEASE WRITE NOTHING ON THIS PAGE