UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL 2016 EXAMINATIONS

STA303H1S / STA1002H1S

Duration - 3 hours

Examination Aids: Scientific Calculator

STA 303/1002
Winter 2016
Final Exam
April 2016
Time Limit: 3h

Last Name (Print): _____

First Name: _____

Student Number: _____

Check one: STA303 ☐ STA1002 ☐

This exam contains 19 pages (including this cover page) and 6 problems. Check to see if any pages are missing. Enter all requested information on the top of this page.

- You may *not* use your books or notes on this exam. You may use a scientific calculator and the formulae and tables at the end of the exam.
- MLE stands for Maximum Likelihood Estimate.
- You are required to show your work on each problem on this exam, except for the problems containing missing output in R. Please carry all possible precision through a numerical question, and give your final answer to four (4) decimals, unless they are trailing zeroes.
- You may use a benchmark of 5% for all inference, unless otherwise indicated. Round DF down to the nearest integer if not available on the table.
- When quoting effects, please give the magnitude, direction and evidence of the effect.
- Do not write in the table to the right.

Problem	Points	Score
1	10	
2	25	
3	35	
4	20	
5	10	
6	0	
Total:	100	

1. (10 points) For each of the following situations, write down the statistical model you will fit to the data, in scalar form. Be sure to define all terms in the model (except the coefficients).

   (a) (3 points) A two-way additive (no interaction) Analysis of Variance model with 2 levels for factor A (A1 and A2) and three levels for factor B (B1, B2, B3).

   (b) (3 points) A logistic regression model with one covariate (X) and one factor with 2 levels (A and B); no interactions.

   (c) (4 points) A longitudinal model with one fixed effect (factor A, levels A1 and A2) and measures repeated over 3 time periods (T1, T2, T3) for all subjects. Please include interaction terms.

2. (25 points) Some statistics were compiled on collision, derailment and overrun accidents on British railway lines over several years. The response variable is the number of accidents (CDO_acc), while potential predictors are the number of years before 2016 (2016 - year) and millions of kilometers of track used (Mkm). Two GLMs were fit to the data, and some relevant R code is shown below.

```
## Model 1 ##
> summary(fit1)

Call:
glm(formula = CDO_acc ~ I(2016 - year), family = poisson, data = rail)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.563510   0.256575     [A]   0.0281
I(2016 - year)  0.038084   0.004794   7.944 1.95e-15

(Dispersion parameter for poisson family taken to be 1)

Null deviance:    [B]    on [C]  degrees of freedom
Residual deviance:  53.051  on 56  degrees of freedom
AIC: 214.1


## Model 2 ##
> summary(fit2)

Call:
glm(formula = CDO_acc ~ I(2016 - year) + Mkm, family = poisson,
data = rail)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.354837   0.532614   0.666   0.5053
I(2016 - year)  0.052485   0.009274   5.659 1.52e-08
Mkm            -0.003139   0.001646  -1.907   0.0565

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 123.750  on 57  degrees of freedom
Residual deviance:  49.295  on [D]  degrees of freedom
AIC: [E]
```

(a) (5 points) Some values have been replaced with letters. Fill in those values.

(A) (B) (C)

(D) (E)

(b) (4 points) Perform a drop-in-deviance test to determine whether the second model (the one with Mkm and year) is necessary, or if the first model would suffice. Give the test statistic, df and the most accurate p-value you can, and state your conclusion.

(c) (3 points) Perform a Goodness-of-Fit test for the first model. Give the test statistic, df and the most accurate p-value you can, and state your conclusion.

(d) (3 points) Using the first model, predict the number of accidents that occurred in the year 1995.

(e) (5 points) Using the first model, say something about the effect of year on the number of accidents. Make sure to quote the relevant statistic, df and p-value. Give an approximate 95% CI for the effect of year.

(f) (5 points) For a loglinear model for a 3-way contingency table, show that the deviance is

$$2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}\log\left(\frac{y_{ijk}}{\hat{\mu}_{ijk}}\right)$$

3. (30 points) In the National Football League (NFL), a team may score three points by kicking a field goal successfully. It's generally accepted that the farther away the kick is taken, the less likely it is to be made. Some data from the 2003 season are analyzed below. In addition to the yardage (yards) and the success of the kick (1 if kick was successful, 0 if not), we also have the week in which the game occurred (1 - 17).

   We have dichotomized week into seasonHalf by splitting at the 9 week mark ("First" or "Second"), and yards into range which is "Short" if the kick is less than 46 yards, and "Long" otherwise.

```
                    ## Model 1 ##
Call:
glm(formula = success ~ yards + week, family = binomial, data = nfl)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.29969    0.51105  12.327   <2e-16
yards       -0.11274    0.01076 -10.477   <2e-16
week        -0.05243    0.01832  -2.862   0.0042

Null deviance:     955.38  on 947  degrees of freedom
Residual deviance: 808.86  on 945  degrees of freedom
AIC: 814.86

                    ## Model 2 ##
Call:
glm(formula = success ~ yards + seasonHalf, family = binomial)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       5.99944    0.47681  12.583   <2e-16
yards            -0.11220    0.01072 -10.467   <2e-16
seasonHalfSecond -0.41742    0.17683  -2.361   0.0182

Residual deviance: 811.59  on 945  degrees of freedom
AIC: 817.59

                    ## Model 3 ##
Call:
glm(formula = success ~ range + seasonHalf, family = binomial)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.4917     0.1570   3.131  0.00174
rangeShort         1.5285     0.1759   8.689  < 2e-16
seasonHalfSecond  -0.3652     0.1706  -2.140  0.03232

Residual deviance: 878.50  on 945  degrees of freedom
AIC: 884.5
```

(a) (2 points) Which of models 1-3 do you prefer, and why?

(b) (3 points) Using the first model, say something about the effect of yards on the chance of making a successful kick, controlling for week. Make sure to quote the relevant statistic, df and p-value.

(c) (3 points) Using the second model, give the ratio of the odds of making a successful kick in the first half of the season, compared to the second (controlling for yards), and provide an approximate 95% CI for this ratio.

(d) (3 points) Using the third model, predict the probability of a successful kick from short range in the first half of the season.

(e) (3 points) Perform a deviance goodness-of-fit test for using *just the sample proportion of successful kicks* to predict the success of a kick. Give the test statistic, df, p-value and a conclusion in words.

(f) (3 points) Perform a drop-in-deviance test for using range and seasonHalf (the two dichotomized variables) to predict the success of a kick. Give the test statistic, df, p-value and a conclusion in words. Is it significantly better than just using the sample proportion?

Here is some more output to consider:

```
> with(nfl, ftable(success, seasonHalf, range))
                   range Long Short
success seasonHalf
0       First              48    43
        Second             43    58
1       First              77   329
        Second             50   300

> fit4 <- glm(V1 ~ success * range * seasonHalf, family= poisson,
              data= ddply(nfl, .(range, seasonHalf, success), nrow))
> fit5 <- glm(V1 ~ success * range, family= poisson,
              data= ddply(nfl, .(range, seasonHalf, success), nrow))
```

```
> anova(fit4, test= "LRT")
Analysis of Deviance Table
Model: poisson, link: log
Response: V1
Terms added sequentially (first to last)

                         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                         7     732.56
success                   1   358.83         6     373.73  < 2e-16
range                     1   291.84         5      81.90  < 2e-16
seasonHalf                1     2.23         4      79.67  [-----]
success:range             1    72.27         3       7.40  < 2e-16
success:seasonHalf        1     2.44         2       4.96  0.11827
range:seasonHalf          1     4.92         1       0.04  0.02654
success:range:seasonHalf  1     0.04         0       0.00  [-----]


> summary(fit5)
Call:
glm(formula = V1 ~ success * range, family = poisson, data = ddply(nfl,
.(range, seasonHalf, success), nrow))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.8177     0.1048  36.419   <2e-16
success             0.3333     0.1373   2.427   0.0152
rangeShort          0.1043     0.1445   0.721   0.4707
success:rangeShort  1.4957     0.1742   8.585   <2e-16

Null deviance:     732.5620  on 7  degrees of freedom
Residual deviance:   9.6326  on 4  degrees of freedom
AIC: 67.566


> anova(fit5, fit4, test= "LRT")
Analysis of Deviance Table

Model 1: V1 ~ success * range
Model 2: V1 ~ success * range * seasonHalf

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4     9.6326
2         0     0.0000  4   9.6326  [-----]
```

(g) (2 points) What is the probability that a random kick is successful?

(h) (2 points) What is the probability that a random kick taken in the second half of the season is from long range?

(i) (3 points) Among kicks taken in the **first half of the season**, what is the odds ratio of success, for short kicks relative to long kicks? (ie. how many times higher are the odds of making a short kick compared to a long kick?)

(j) (3 points) How many times **more likely** is a kicker to make a short kick compared to a long kick, at any point in the season?

(k) (3 points) What is the **absolute risk reduction** of missing a kick if the team can get from long range to short range? Consider the entire season.

(l) (3 points) Does knowing the half of the season in which the kick was taken (first or second) significantly improve prediction of the counts of success/failure when you already know the range of the kick? Give the test statistic, df, p-value and a conclusion in words.

(m) (2 points) Explain the relationship between the three variables `success`, `range` and `seasonHalf` using terminology from lecture.

4. (15 points) A 2007 study by Beata et al. examined the effect of a Zylkene treatment on the anxiety score (higher is better) for cats. The Score was measured over 5 time points, and you can assume the relationship of score to time is linear. Demographic information on the cats was also recorded (age, weight, and gender). Some potentially useful R output is shown below. Every cat received a unique treatment (Tx).

```
> str(cats.long)
'data.frame': 170 obs. of  8 variables:
$ ID        : Factor w/ 34 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
$ Tx        : Factor w/ 2 levels "Placebo","Zylkene": 2 1 2 1 2 1 2 2 2 2 ...
$ weight    : num  4 4 6 3.5 6.2 2 3.5 3.7 6 4 ...
$ age       : int  15 67 55 78 50 10 9 54 64 30 ...
$ gender    : Factor w/ 3 levels "Female","NeutFemale",..: 3 3 3 2 3 2 3 3 3 2 ...
$ Time      : num  1 1 1 1 1 1 1 1 1 1 ...
$ Score     : int  8 9 9 9 10 6 14 13 12 5 ...
$ TimeFactor: Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...

> with(cats.long, tapply(Score, list(Tx, TimeFactor), mean))
            1     2     3     4     5
Placebo  9.06 10.18 10.88 11.41 11.41
Zylkene 10.94 12.12 13.59 15.18 16.12

> anova(lm(Score ~ Tx*Time + weight + age + gender, data= cats.long))
Analysis of Variance Table
Response: Score
           Df  Sum Sq Mean Sq F value    Pr(>F)
Tx          1  382.50  382.50 26.6191 7.164e-07
Time        1  318.36  318.36 22.1551 5.365e-06
weight      1  115.26  115.26  8.0214   0.00521
age         1    8.12    8.12  0.5650   0.45336
gender      2   78.16   39.08  2.7196   0.06891
Tx:Time     1   47.44   47.44  3.3013   0.07107
Residuals 162 2327.84   14.37

> summary(aov(Score ~ Tx*Time + weight + age + gender + Error(ID), data= cats.long))
Error: ID
          Df Sum Sq Mean Sq F value Pr(>F)
Tx         1  382.5   382.5   5.548 0.0257
weight     1  115.3   115.3   1.672 0.2066
age        1    8.1     8.1   0.118 0.7341
gender     2   78.2    39.1   0.567 0.5737
Residuals 28 1930.4    68.9

Error: Within
           Df Sum Sq Mean Sq F value   Pr(>F)
Time        1  318.4   318.4   107.3  < 2e-16
Tx:Time     1   47.4    47.4    16.0 0.000104
Residuals 134  397.4     3.0
```

(a) (1 point) How many cats participated in this study?

(b) (2 points) Is weight a useful predictor of anxiety score, controlling for treatment? Give test statistic, df, p-value and a conclusion.

(c) (3 points) Do cats treated with Zylkene differ on average anxiety score from cats treated with placebo? Give test statistic, df, p-value and a conclusion.

(d) (3 points) Did the cats anxiety scores improve significantly, in general, over time? Give test statistic, df, p-value and a conclusion.

(e) (4 points) The main research question is to assess whether or not cats treated with Zylkene improved significantly faster over time, compared to the placebo group. What conclusion do you draw? Give the test statistic, df and p-value that supports your findings.

(f) (2 points) What correlation structure was imposed on the model for this analysis?

(g) (3 points) Name three other correlation structures other than the one above. If you don't know the name, you can sketch out the covariance matrix and describe it, but only the name is required.

(h) (2 points) Explain the difference between a **within-group** factor and a **between-group** factor.

5. (10 points) In the Winter Olympics speed skating event, two skaters start side-by-side on a track (one in the inner lane and one in the outer lane). Halfway through the race, they cross over so that they skate the same distance. Despite this, some viewers felt that the skaters in the outer lane had an advantage, and had faster times. Some potentially useful R analysis is shown below, with some output deleted.

```
> t.test(time ~ lane, data= skate)

Welch Two Sample t-test

data:  time by lane
t = 0.64716, df = 31.835, p-value = 0.5222
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.071528  2.069175
sample estimates:
mean in group Inner mean in group Outer
          121.0412            120.5424


> t.test(time ~ lane, var.equal= T, data= skate)

Two Sample t-test

data:  time by lane
t = 0.64716, df = 32, p-value = 0.5221
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.071210  2.068857
sample estimates:
mean in group Inner mean in group Outer
          121.0412            120.5424


> t.test(time ~ lane, paired= T, data= skate)

Paired t-test

data:  time by lane
t = 0.88139, df = 16, p-value = 0.3912
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.7009392  1.6985863
sample estimates:
mean of the differences
              0.4988235
```

```
> summary(aov(time ~ lane + Error(race/lane), data= skate))
Error: race
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals --    ---   -----

Error: race:lane
          Df Sum Sq Mean Sq F value Pr(>F)
lane       1   2.12   2.115     [B]    [C]
Residuals [A] -----   -----
```

(a) (3 points) Some values have been replaced with letters. Fill in those values. You do not need to show any work for this part.

(A) (B) (C)

(b) (5 points) Is there a (statistically significant) difference in winning times between the two lanes? Give the effect and quote the relevant test statistic, df and p-value to support your answer.

(c) (2 points) What assumption(s) did you make in your conclusion from the previous part?

6. BONUS (5 points): In class we showed how to find the unrestricted MLEs of the expected values in a contingency table, using the method of Lagrange multipliers. Recall that the method involves finding a stationary point of the Lagrangian $\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \sum_1^M \lambda_k g_k(\mathbf{x})$ where the $g_k(\mathbf{x})$ are the M constraint functions. Derive the restricted MLEs for an arbitrary IxJ contingency table under the assumption of row and column independence.

Some formulae:

Pooled t-test

$$t_{obs} = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Linear Regression

$$b_1 = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2} = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sum x_i^2 - n\overline{x}^2} \qquad b_0 = \overline{y} - b_1\overline{x}$$

One-way analysis of variance

$$\text{SSTO} = \sum_{i=1}^{N}(y_i - \overline{y})^2 \qquad \text{RSS} = \sum_{g=1}^{G}\sum_{(g)}(y_i - \overline{y}_g)^2$$

$$\text{SSReg} = \sum_{g=1}^{G} n_g(\overline{y}_g - \overline{y})^2$$

Bernoulli and Binomial distributions

If $Y \sim \text{Bernoulli}(\pi)$
$\text{E}(Y) = \pi$, $\text{Var}(Y) = \pi(1-\pi)$

If $Y \sim \text{Binomial}(m, \pi)$
$\text{E}(Y) = m\pi$, $\text{Var}(Y) = m\pi(1-\pi)$

Logistic Regression with Binomial Response formulae

$$\text{Deviance} = 2\sum_{i=1}^{n}\{y_i\log(y_i) + (m_i - y_i)\log(m_i - y_i) - y_i\log(\hat{y}_i) + (m_i - y_i)\log(m_i - \hat{y}_i)\}$$

$$D_{res,i} = \text{sign}(y_i - m_i\hat{\pi}_i)\sqrt{2\left\{y_i\log\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - m_i\hat{\pi}_i}\right)\right\}}$$

$$P_{res,i} = \frac{y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}} \qquad \text{where } \hat{\pi}_i = \hat{\pi}_{M,i}$$

Log-linear models for count data

$$logL(\pi_{ij}) = \sum_{ij} y_{ij} log(\pi_{ij}) + f(y) \qquad G^2 = -2\ log(\frac{L_R}{L_F})$$

Critical values of the χ^2 distribution. Upper tail area is across the top.

DF	0.995	0.99	0.975	0.95	0.9	0.5	0.1	0.05	0.025	0.01	0.005
1	0.0	0.0	0.0	0.0	0.0	0.5	2.7	3.8	5.0	6.6	7.9
2	0.0	0.0	0.1	0.1	0.2	1.4	4.6	6.0	7.4	9.2	10.6
3	0.1	0.1	0.2	0.4	0.6	2.4	6.3	7.8	9.3	11.3	12.8
4	0.2	0.3	0.5	0.7	1.1	3.4	7.8	9.5	11.1	13.3	14.9
5	0.4	0.6	0.8	1.1	1.6	4.4	9.2	11.1	12.8	15.1	16.7
6	0.7	0.9	1.2	1.6	2.2	5.3	10.6	12.6	14.4	16.8	18.5
7	1.0	1.2	1.7	2.2	2.8	6.3	12.0	14.1	16.0	18.5	20.3
8	1.3	1.6	2.2	2.7	3.5	7.3	13.4	15.5	17.5	20.1	22.0
9	1.7	2.1	2.7	3.3	4.2	8.3	14.7	16.9	19.0	21.7	23.6
10	2.2	2.6	3.2	3.9	4.9	9.3	16.0	18.3	20.5	23.2	25.2
11	2.6	3.1	3.8	4.6	5.6	10.3	17.3	19.7	21.9	24.7	26.8
12	3.1	3.6	4.4	5.2	6.3	11.3	18.5	21.0	23.3	26.2	28.3
13	3.6	4.1	5.0	5.9	7.0	12.3	19.8	22.4	24.7	27.7	29.8
14	4.1	4.7	5.6	6.6	7.8	13.3	21.1	23.7	26.1	29.1	31.3
16	5.1	5.8	6.9	8.0	9.3	15.3	23.5	26.3	28.8	32.0	34.3
18	6.3	7.0	8.2	9.4	10.9	17.3	26.0	28.9	31.5	34.8	37.2
20	7.4	8.3	9.6	10.9	12.4	19.3	28.4	31.4	34.2	37.6	40.0
24	9.9	10.9	12.4	13.8	15.7	23.3	33.2	36.4	39.4	43.0	45.6
28	12.5	13.6	15.3	16.9	18.9	27.3	37.9	41.3	44.5	48.3	51.0
32	15.1	16.4	18.3	20.1	22.3	31.3	42.6	46.2	49.5	53.5	56.3
36	17.9	19.2	21.3	23.3	25.6	35.3	47.2	51.0	54.4	58.6	61.6
40	20.7	22.2	24.4	26.5	29.1	39.3	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	49.3	63.2	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	46.5	59.3	74.4	79.1	83.3	88.4	92.0
70	43.3	45.4	48.8	51.7	55.3	69.3	85.5	90.5	95.0	100.4	104.2
80	51.2	53.5	57.2	60.4	64.3	79.3	96.6	101.9	106.6	112.3	116.3
100	67.3	70.1	74.2	77.9	82.4	99.3	118.5	124.3	129.6	135.8	140.2
150	109.1	112.7	118.0	122.7	128.3	149.3	172.6	179.6	185.8	193.2	198.4
200	152.2	156.4	162.7	168.3	174.8	199.3	226.0	234.0	241.1	249.4	255.3
300	240.7	246.0	253.9	260.9	269.1	299.3	331.8	341.4	349.9	359.9	366.8
400	330.9	337.2	346.5	354.6	364.2	399.3	436.6	447.6	457.3	468.7	476.6
500	422.3	429.4	439.9	449.1	459.9	499.3	540.9	553.1	563.9	576.5	585.2
600	514.5	522.4	534.0	544.2	556.1	599.3	644.8	658.1	669.8	683.5	693.0
700	607.4	615.9	628.6	639.6	652.5	699.3	748.4	762.7	775.2	790.0	800.1
800	700.7	709.9	723.5	735.4	749.2	799.3	851.7	866.9	880.3	896.0	906.8
900	794.5	804.3	818.8	831.4	846.1	899.3	954.8	970.9	985.0	1001.6	1013.0
1000	888.6	898.9	914.3	927.6	943.1	999.3	1057.7	1074.7	1089.5	1107.0	1118.9