UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL 2014 EXAMINATIONS

STA 303 H1S / STA 1002 HS

Duration - 3 hours

Examination Aids: Calculator

LAST NAME:_____ FIRST NAME:_____

STUDENT NUMBER: _____

INSTRUCTIONS:

- There are 6 questions - answer all questions.
- There are 19 pages in total (including this page and output/table/formula sheet). Make sure you have all pages before starting the test.
- Pages 13 to 17 contain R output. Pages 17-18 are statistical tables. You may remove these pages from your exam but hand them in with your exam. Do not write answers on the pages of R output/tables/formula sheet. Answer all questions on the question papers.
- **Show your work and justify answers to earn full marks**. Correct answers with no justifications will not receive any marks.
- Round your answers to 4 decimal places where appropriate.
- You may copy numbers from output to answer questions unless the question specifically asks you to calculate from scratch.
- For all hypothesis tests, include the following steps: state H_0, H_a, the value of the test statistic and its distribution under the null hypothesis, p-value, and a conclusion.
- For all questions, consider p-values < 0.1 as statistically significant.
- Make sure to interpret and give practical conclusions to answer the questions of interest.
- The last page (page 19) is a table of formulae that may be useful. For all questions you can assume that the results on the formula page are known.
- Total marks: 100

Question	1	2	3	4	5	6	Total:
Value	24	13	28	7	8	20	100
Mark Earned							

1. The Physician's Health Study is a famous experiment in which male physicians between 40 and 84 years old were randomly assigned to take an aspirin or placebo every day. They were then followed and the number of myocardial infarctions (heart attacks) in each group was recorded. The main question of interest is if the type of drug taken is associated with getting a myocardial infarction. The data are summarized in the table below:

	Myocardial Infarction	
Group	Yes	No
Placebo	189	10,845
Aspirin	104	10,933

(a) *[4m]* Conduct from scratch a hypothesis test to see if the probability of a myocardial infarction in each treatment group is the same.

(b) *[1m]* What assumption(s) are required to do the test in (a)?

(c) *[5m]* Using output, conduct a hypothesis test to answer the question of interest using the Pearson Chi-square Test of Independence. What assumptions are required?

Continued

(d) *[1m]* What is the relationship between the test statistic in (c) and the test statistic in (a)?

(e) *[4m]* R was used to conduct a Multinomial Likelihood Ratio Test for the question of interest. Referring to the output, write down all the steps for the hypothesis test.

(f) *[2m]* Explain how you would use Poisson regression to test for the question of interest. What assumptions are required to use this method?

(g) *[5m]* Use the relevant Poisson regression output to test the question of interest. Write out any models that are being fitted using proper notation.

(h) *[2m]*
Explain why the deviance of 'model2' from the output is 0. What is the statistical terminology for this type of model?

Continued

2. Suppose subjects are categorized by three variables. Variable 1 has I categories, variable 2 has J categories, and variable 3 has K categories. We observe y_{ijk}, the count of the number of subjects for whom variable 1 has level i, variable 2 has level j, and variable 3 has level k. We will assume that the y_{ijk} can be considered observations from Poisson distributions with means λ_{ijk} and use Poisson regression.

   (a) *[4m]* We fit a model that assumes that variables 1, 2, and 3 are independent. Show that the deviance is

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

   where $\hat{\lambda}_{ijk}$ are the estimated values of λ_{ijk} from the fitted model.

   (b) *[2m]* For the model assuming complete independence, what is the estimated value of λ_{124} (the mean of the count in cell $(1,2,4)$) from the fitted model? Briefly explain.

   (c) *[3m]* Explain what the Uniform Association model would be for these data. Specify the type of generalized linear model, the response, and predictors.

Continued

(d) *[2m]* Suppose you use use the Goodness of Fit Test to compare the Uniform Association model to the saturated model. How many degrees of freedom does the test statistic have? Explain briefly.

(e) *[2m]* Suppose the GOF test from above has a very large p-value (say bigger than 0.5). What conclusion would you make about the model choice and association between the three variables?

3. The data we will consider were collected from an experiment to investigate the effects of tree resin on termites. A resin was derived from tree bark and dissolved in a solvent in two different concentration doses: 5 mg and 10 mg (variable name: 'Dose'). For each dose, eight dishes were used with 25 termites in each dish. After 15 days, the number of termites still alive in each dish were recorded (variable name: 'Number'). The explanatory variable is Dose and we are interested in determining if the dose affects the odds of survival for termites. The data are in the table below:

		Number	
Dish	Dose	Day 1	Day 15
1	5	25	11
2	5	25	11
3	5	25	12
4	5	25	12
5	5	25	5
6	5	25	9
7	5	25	6
8	5	25	10
9	10	25	16
10	10	25	13
11	10	25	1
12	10	25	0
13	10	25	0
14	10	25	0
15	10	25	0
16	10	25	3

Some edited output from R is given in the output section. In model1, dose is modelled as a categorical variable and in model2, dose is modelled as a quantitative variable.

Continued

(a) *[3m]* In model1, some of the numbers in the output have been replaced with letters. Give the values of the letters below:

(A) = _____

(B) = _____

(C) = _____

(b) *[4m]* Write out the model that is being fit by R in model1. Write out the model that is being by R in model2. Define any variables/notation you introduce.

(c) *[4m]* Give a practical interpretation of the estimate of β_1 in model1 and then in model2.

(d) *[2m]* For model1, what is the estimated probability that a termite in a dish with 5 mg dose of resin is dead on day 15?

Continued

(e) *[2m]* Suppose you wanted to predict the probability of survival for a termite in a dish with a 30 mg dose of resin. Is it appropriate to use model1 or model2 or neither? State which model and then give the prediction. If neither, explain why.

(f) *[3m]* Find a 95% confidence interval for the odds ratio of termite survival for a dose of 5 mg vs 10 mg. (Use the appropriate model).

(g) *[2m]* There are only 16 observations. Is this a concern? Why or why not?

Continued

(h) *[4m]* For model2, conduct a likelihood ratio test to determine if dose significantly affects the odds that a termite is alive on day 15.

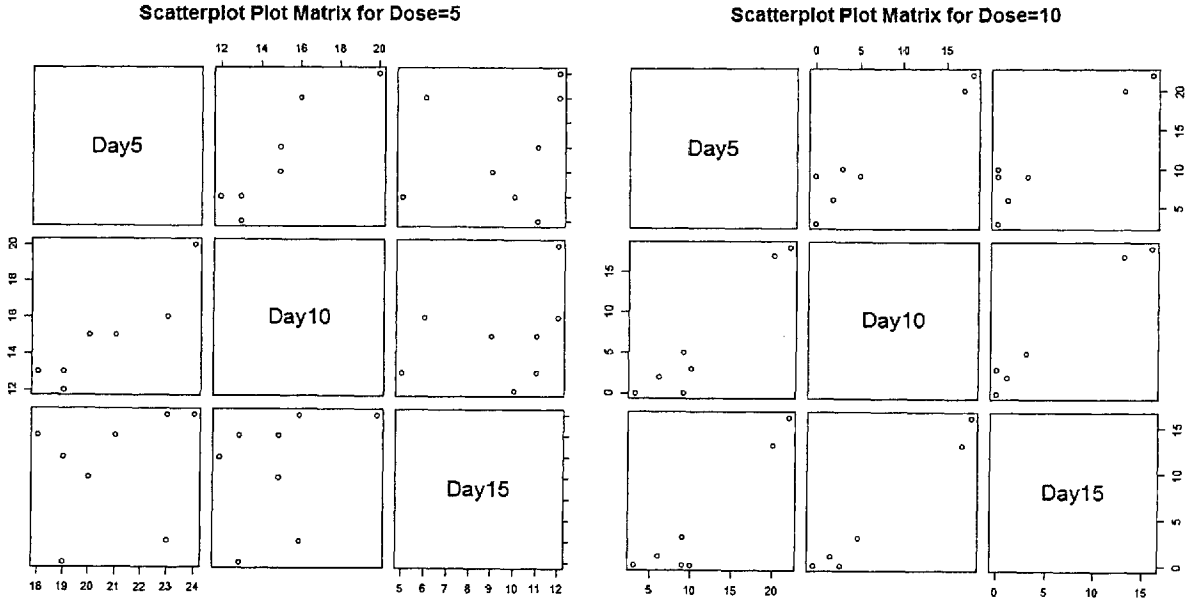
(i) *[4m]* For model2, conduct another test (other than the LRT) to determine if dose significantly affects the odds that a termite is alive on day 15. Name the method/test you are using. Compare your conclusion to the previous part.

Continued

4. Once again consider the data about termites and resin. On day 1, 25 termites were put in each of eight dishes of each of two resin doses (5 mg and 10 mg). In this question we will consider the number of termites alive in each dish on each of days 5, 10, 15. Both dose and days are considered categorical variables.

   (a) *[3m]* What type of model would you use to for this scenario? Justify your choice.

   (b) *[4m]* Given below are scatterplots of the numbers of termites alive on one day in each dish versus the number alive on another day in the dish for each pair of days. What do you learn from these plots?



Continued

5. For a one-way ANOVA with G factor levels, consider the following model:

$$Y_{gi} = \beta_g + e_{gi} \quad ; \quad \text{for } g = 1, 2, \ldots, G \text{ and } i = 1, 2, \ldots, n_g.$$

(a) *[4m]* Show that the Least Squares estimates of the regression parameters are given by: $\hat{\beta}_g = \bar{y}_g$; where as usual $\bar{y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{gi}$.

(b) *[2m]* Interpret each of the βs. Are the interpretations the same as it would be for a model with $G - 1$ indicator variables? If yes, state yes. If not, explain the differences.

(c) *[2m]* If you wanted to test for differences in the group means, what are H_0 and H_a in terms of the regression parameters?

Continued

6. In this course, we have studied the following six (generalized) linear models:
(1) one-way analysis of variance, (2) two-way analysis of variance, (3) binary logistic regression,
(4) binomial logistic regression, (5) Poisson regression, and (6) mixed models.

(a) ***[12m - 4 each part]*** For each of the scenarios given below, state: (I) which among the above 6 types of generalized linear models is most appropriate, (II) the model that you would use for analysis, carefully defining any terms you use, (III) the null and alternative hypotheses to test the question of interest.

(i) This scenario relates to a study of 73 breakfast cereals sold at a large grocery store. In marketing a cereal, a consideration is whether or not it is displayed at eye level on the grocery store shelf. For each of the 73 cereals in the study, it was recorded whether the cereal was on the lower, middle, or upper shelf. In order to appeal to children, the researcher thinks that stores tend to put sugary cereals on the low shelf. We are interested in whether the sugar level (measured as low, medium, or high) is useful in predicting whether the cereal is placed on the low shelf.

(ii) We count the number of cereals with high sugar content on each of the lower, middle, and upper shelves. We are interested in learning if shelf placement and whether or not a cereal has high sugar are associated.

(iii) We now wish to determine if there are differences in sugar content (measured in grams per serving) between the low, middle, and upper shelves.

Continued

(b) *[8m - 2 each part]* This question relates to assumptions necessary for each of the models. Answer the questions and briefly justify.

(i) For which of the 6 models is a large sample size required for valid inference?

(ii) For which of the 6 models is constant variance of the response assumed?

(iii) For which of the 6 models is a Normal probability plot / QQ-plot useful?

(iv) Which of these models are most appropriate for count data? For each of the models you name, what must be true about the way the data were collected in order for the model to be appropriate.

Continued

Physician's Health Study (Heart Attack): R OUTPUT

```
> tbl = rbind(c(189, 10845),c(104, 10933))
> tbl
     [,1]  [,2]
[1,]  189 10845
[2,]  104 10933

> chisq.test(tbl,correct=FALSE)

        Pearson's Chi-squared test

data:  tbl
X-squared = 25.0139, df = 1, p-value = 5.692e-07


> likelihood.test(tbl)

        Log likelihood ratio (G-test) test of independence without correction

data:  tbl
Log likelihood ratio statistic (G) = 25.372, X-squared df = 1, p-value = 4.727e-07


> heartattack = read.csv("heartattack.csv")

> heartattack
     Drug  MI  Count
1 Placebo Yes    189
2 Placebo  No 10,845
3 Aspirin Yes    104
4 Aspirin  No 10,933

> attach(heartattack)

> model1 = glm(Count ~ Drug + MI, family="poisson")
> summary(model1)

Deviance Residuals:
      1        2        3        4
 3.3610  -0.4078  -3.7072   0.4072

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.2956443  0.0095506  973.30   <2e-16 ***
DrugPlacebo -0.0002718  0.0134623   -0.02    0.984
MIYes       -4.3084830  0.0588123  -73.26   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```


Continued

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 27507.580  on 3  degrees of freedom
Residual deviance:    25.372  on 1  degrees of freedom
AIC: 67.203

Number of Fisher Scoring iterations: 4

> model2 = glm(Count ~ Drug*MI, family="poisson")
> summary(model2)

Call:
glm(formula = Count ~ Drug * MI, family = "poisson")

Deviance Residuals:
[1]  0  0  0  0

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        9.299541   0.009564 972.369  < 2e-16 ***
DrugPlacebo       -0.008082   0.013553  -0.596    0.551
MIYes             -4.655150   0.098523 -47.249  < 2e-16 ***
DrugPlacebo:MIYes  0.605438   0.122842   4.929 8.28e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2.7508e+04  on 3  degrees of freedom
Residual deviance: 2.3315e-12  on 0  degrees of freedom
AIC: 43.831
```


Continued

Termites: R OUTPUT

```
> termites = read.csv("termitesdata.csv")
> termites
   Dish Dose Day1 Day15
1     1    5   25    11
2     2    5   25    11
3     3    5   25    12
4     4    5   25    12
5     5    5   25     5
6     6    5   25     9
7     7    5   25     6
8     8    5   25    10
9     9   10   25    16
10   10   10   25    13
11   11   10   25     1
12   12   10   25     0
13   13   10   25     0
14   14   10   25     0
15   15   10   25     0
16   16   10   25     3

> attach(termites)

> Dose
 [1]  5  5  5  5  5  5  5  5 10 10 10 10 10 10 10 10

> DoseLevel = factor(Dose)
> DoseLevel
 [1] 5  5  5  5  5  5  5  5  10 10 10 10 10 10 10 10
Levels: 5 10

> Number = Day15
> Number
 [1] 11 11 12 12  5  9  6 10 16 13  1  0  0  0  0  3

################################
# model1: Dose as categorical #
################################

> model1 <- glm(Number/Day1 ~ DoseLevel, family=binomial, weights=Day1)
> summary(model1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0027  -2.2246  -0.4191   0.7137   5.3135

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4895     0.1457   -3.36 0.000778 ***
DoseLevel10  -1.1319     0.2398    (A)  2.36e-06 ***
---
```


Continued

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 117.829  on 15  degrees of freedom
Residual deviance:  94.018  on (B)  degrees of freedom
AIC: 138.78

Number of Fisher Scoring iterations: 5

> anova(model1,test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: Number/Day1

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         15    117.829
DoseLevel  1    (C)           14     94.018  1.063e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

################################
# model2: Dose as quantitative #
################################

> model2 <- glm(Number/Day1 ~ Dose, family=binomial, weights=Day1)

> summary(model2)

Call:
glm(formula = Number/Day1 ~ Dose, family = binomial, weights = Day1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0027  -2.2246  -0.4191   0.7137   5.3135

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.64239    0.34811   1.845    0.065 .
Dose        -0.22639    0.04796  -4.720 2.36e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 117.829  on 15  degrees of freedom
Residual deviance:  94.018  on 14  degrees of freedom
AIC: 138.78
```


Continued

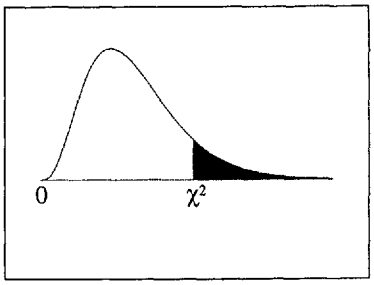
```
> anova(model2,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Number/Day1

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    15    117.829
Dose  1   23.811        14     94.018 1.063e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> termites2 = read.csv("termitesdata2.csv")
> termites2
   Dish Dose Day1 Day5 Day10 Day15
1     1    5   25   18    13    11
2     2    5   25   21    15    11
3     3    5   25   23    16    12
4     4    5   25   24    20    12
5     5    5   25   19    13     5
6     6    5   25   20    15     9
7     7    5   25   23    16     6
8     8    5   25   19    12    10
9     9   10   25   22    18    16
10   10   10   25   20    17    13
11   11   10   25    6     2     1
12   12   10   25   10     3     0
13   13   10   25    9     0     0
14   14   10   25    3     0     0
15   15   10   25    3     0     0
16   16   10   25    9     5     3

> attach(termites2)

> pairs(termites2[Dose==5,4:6],main="Scatterplot Plot Matrix for Dose=5")
> pairs(termites2[Dose==10,4:6],main="Scatterplot Plot Matrix for Dose=10")
```

Percentiles of the standard normal distribution

Probability to left of quantile	0.95	0.975	0.99	0.995
Quantile	1.645	1.960	2.326	2.576

Continued

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Continued

SOME FORMULAE:

Pooled t-test:

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

Test for two proportions:

$$z = (\hat{\pi}_1 - \hat{\pi}_2) \Big/ \sqrt{\hat{\pi}_p(1-\hat{\pi}_p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

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:

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

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

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:

$$\text{Deviance} = 2\sum_{i=1}^{n}\{y_i\log(y_i) + (m_i-y_i)\log(m_i-y_1) - y_i\log(\hat{y}_i) + (m_i-y_i)\log(m_i-\hat{y}_i)\}$$

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

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

Multinomial distribution for 2×2 table:

$$\Pr(\mathbf{Y}=\mathbf{y}) = \frac{n!}{y_{11}!y_{12}!y_{21}!y_{22}!}\pi_{11}^{y_{11}}\pi_{12}^{y_{12}}\pi_{21}^{y_{21}}\pi_{22}^{y_{22}}$$

Poisson distribution:

$$\Pr(Y=y) = \frac{\lambda^y e^{-\lambda}}{y!},\ y = 0,1,2,\ldots$$

$$\text{E}(Y) = \lambda,\ \text{Var}(Y) = \lambda$$

Two-way contingency tables (easily generalizable to three-way tables):

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(y_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \qquad G^2 = 2\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ij}\log\left(\frac{y_{ij}}{\hat{\lambda}_{ij}}\right)$$

$$D_{res,ij} = \text{sign}(y_{ij}-\hat{\lambda}_{ij})\sqrt{2\left\{y_{ij}\log\left(\frac{y_{ij}}{\hat{\lambda}_{ij}}\right) - y_{ij} + \hat{\lambda}_{ij}\right\}}$$

$$P_{res,ij} = \frac{y_{ij}-\hat{\lambda}_{ij}}{\sqrt{\hat{\lambda}_{ij}}}$$

Model Fitting Criteria:

$$\text{AIC} = -2\log(L) + 2(p+1) \qquad \text{BIC} = -2\log(L) + (p+1)\log(N)$$

Total pages: 19
Total marks: 100