UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL 2011 EXAMINATIONS

STA 303 H1S / STA 1002 HS

Duration - 3 hours

Examination Aids: Calculator

LAST NAME: SOLUTIONS **FIRST NAME:** ____________

STUDENT NUMBER: ____________

- There are 23 pages including this page.
- Pages 14 to 21 contain SAS output.
- The last page (page 23) is a table of formulae that may be useful. For all questions you can assume that the results on the formula page are known.
- Some quantiles from the standard normal distribution and a table of the chi-square distribution can be found on page 22.
- Total marks: 90

1abc	1def	2abcd	2efgh	2ij	3abc	3de

4abc	4d	5	6a(i,ii)	6a(iii)bc

1. A study was carried out on mice to see how their diet affects their lifetime, with particular focus on the effect of restricting caloric intake. Three hundred and forty-nine female mice were randomly assigned to one of the following six diet groups:

   (1) `N/N85` – Mice in this group were fed normally before weaning and then afterwards they were restricted to 85 kilocalories per week.
   (2) `N/R40` – Mice in this group were fed normally before weaning and then afterwards they were restricted to 40 kilocalories per week.
   (3) `N/R50` – Mice in this group were fed normally before weaning and then afterwards they were restricted to 50 kilocalories per week.
   (4) `NP` – Mice in this group ate as much as they pleased of a standard diet.
   (5) `R/R50` – Mice in this group were fed a diet restricted to 50 kilocalories per week both before and after weaning.
   (6) `lopro` – This group had a similar diet to `N/R50` but the protein content was restricted.

   Lifetimes, in months, for the mice were recorded. Some output from SAS is given on pages 14 to 15. The questions below relate to this output.

   (a) (1 mark) Why is the `Model DF` equal to 5?

   There are 6 diets so the model needs 5 indicator explanatory variables.

   (b) (1 mark) The least squares mean for diet `N/N85` has been replaced by `X`'s. What is it?

   $$39.69 - 6.99 = 32.7$$

   (c) (2 marks) Explain, in practical terms, what you can conclude from the 6 t-tests on the top of page 15. (The 6 tests you should be considering are the tests with test statistics: 44.47, -5.57, 4.38, 2.19, -9.40, and 2.54.)

Test statistic	*Practical conclusion*
44.47	*The mean lifetime of the lopro group is not 0.*
$-5.57, \ldots, 2.54$	*There is evidence (ranging from strong to moderate) that the mean lifetime on lopro differs from the mean lifetimes on each of the other diets.*

(Question 1 continued)

(d) (2 marks) Suppose that we are particularly interested in the comparison in mean lifetime between diets `N/R40` and `N/R50`. Using the first formula on the formula sheet, we could construct a pooled two-sample t-test for this comparison with $\overline{y}_1 = 45.12$, $\overline{y}_2 = 42.30$, $n_1 = 60$, $n_2 = 71$, and $s_p = \sqrt{((60-1)6.70^2 + (71-1)7.77^2)/((60-1)+(71-1))}$. Will the resulting p-value for this pooled two-sample t-test be 0.0166? (0.0166 is taken from the matrix of p-values on page 15.) Explain why or why not.

No. The t-test with p-value $= 0.0166$ *compares the same means but it uses as s.d. the pooled s.d. for all 6 diets (which is* $\sqrt{MSE} = 6.68$*) and has a corresponding different d.f. (343). The d.f. for the proposed test is* $60 - 1 + 71 - 1 = 129$*.)*

(e) (2 marks) On page 15, there is the following note in the SAS output:

```
NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

What is the purpose of this note from SAS? What should you do and why?

Because the pairwise comparisons are many (15) tests, the chance of at least one Type I error is high. You should control for it by using Tukey's or Bonferroni's method.

(f) (4 marks) On page 15 you are given a plot of the standardized residuals versus the predicted values and a normal quantile plot of the standardized residuals. What are you looking for in each plot? What do you conclude? How do your conclusions affect your answers to the previous questions?

Plot of standardized residuals versus predicted values:
Look for outliers and approximately equal variance in each group. There are some large negative residuals, but otherwise no problems.

Normal quantile plot:
Look for a straight line. The plot is curved indicating that the distribution of errors is skewed rather than normal.

The p-values in parts (c) and (d) are not correct (or only approximately correct) so the conclusions to the previous questions may be wrong.

2. For this question, we will consider the data from assignment 2. For children born in 1990 in South Africa, their race (black or white) and whether or not their mother had medical aid was recorded. Attempts were made for follow-up medical evaluations in 1995 and the data includes whether or not the children participated in the follow-up, recorded as yes or no in the variable `Traced`. We are interested in the relationship among race, medical aid status, and whether or not a child had a follow-up. Output from SAS is given on pages 16, 17, and 18 for 3 models fit to these data. The variable `Count` is the number of children in each category.

   (a) (3 marks) In the output for model 1, a few numbers have been replaced by `X`'s. Find the values of the following:

   BIC = $-2(-219.7333) + 4\log(8) = 447.78$

   Lower limit for the missing Wald 95% Confidence Interval = $1.6998 - 1.96(0.1120) = 1.48$

   The missing Wald Chi-Square = $\left(\frac{1.6998}{0.1120}\right)^2 = 230.33$

   (b) (2 marks) Write the **model** that was fit for model 1, defining all terms.

   $$\log(\mu_{ijk}) = \beta_0 + \beta_1 I_{[\texttt{Traced}=\texttt{No}],ijk} + \beta_2 I_{[\texttt{MedicalAid}=\texttt{No}],ijk} + \beta_3 I_{[\texttt{Race}=\texttt{Black}],ijk}$$

   where $I_{[condition]} = 1$ *if the condition holds and is* 0 *otherwise and* μ_{ijk} *is the mean of the corresponding Poisson distribution.*

   (c) (2 marks) For model 1, give a practical interpretation of the coefficient whose estimate is 1.7223 (assuming the model is appropriate).

   For people with the same race and follow-up status, the odds that they do not have medical aid are $\exp\{1.7223\} = 5.6$.

   (d) (3 marks) From model 2, what are the estimated odds of being traced for a child with medical aid (assuming the model is appropriate)?

   If race is white, odds of being traced for a child with medical aid are
   $\exp\{-2.3452\} = 0.096$.
   If race is black, odds of being traced for a child with medical aid are
   $\exp\{-2.3452 + 1.4204\} = 0.396$.

(e) (3 marks) From model 3, what is the odds ratio of being traced, comparing a black child to a white child (assuming the model is appropriate)?

Since there is no interaction between traced and medical aid status in the model, the answer is the same regardless of medical aid status.
For a black child, the odds of being traced are $\exp\{-2.3514 + 1.3982\}$.
For a white child, the odds of being traced are $\exp\{-2.3514\}$.
So the required odds ratio is $\exp\{1.3982\} = 4.05$.

(f) (3 marks) For model 1, the deviance is large. Ignoring what you learn from the other models, give at least three reasons why the deviance might be large when fitting a model of this type to data.

(1) The model is not appropriate; interaction terms are needed.
(2) There are outliers.
(3) There is extra-Poisson variation.

(g) (1 mark) For model 3, under the `Criteria for Assessing Goodness of Fit`, why is `DF` equal to 2?

There are 8 *observations.*
There are 6 *parameters (β's) in the model.*
$8 - 6 = 2$

(h) (4 marks) Is it possible to carry out a Likelihood Ratio Test comparing the fits of models 1 and 3? If not, explain why not. If yes, carry it out, giving each of the following: (I) the test statistic, (II) the distribution of the test statistic under the null hypothesis, (III) the p-value, (IV) the conclusion.

(I) Test statistic: $G^2 = -2\,(-219.73 - (-23.22)) = 393$ *(using full log-likelihood; can get the same answer from log-likelihood or deviance)*
(II) chi-square with $4 - 2 = 2$ *d.f.*
(III) p-value < 0.005
(IV) Strong evidence in favour of model 3, i.e. the coefficients of the interaction terms are not 0

(i) (2 marks) Wald tests for the model parameters for each of these models use chi-square distributions to calculate the p-values. Explain why chi-square is the appropriate distribution.

The estimates are found by maximum likelihood estimation. MLEs are asymptotically normally distributed. When standardized and squared (giving the Wald test statistic), their distribution is chi-square.

(j) Choosing from the 3 models for which you are given SAS output, pick the model that you think is most appropriate for these data.

i. (2 marks) Which of the 3 models did you choose? Why?

Both models 2 and 3 have small deviance. Choose the simpler of these models. So choose model 3.

ii. (2 marks) For the model that you chose in part i., characterize in practical terms what you conclude about the relationship among race, medical aid status, and whether or not a child had a follow-up.

Traced and medical aid status are conditionally independent, conditional on race.

iii. (2 marks) When you analysed these data in assignment 2, one of the analyses treated `Traced` as a response variable and fit a logistic regression model with `Race` and `MedicalAid` as explanatory variables. Explain how the model you chose in part i. can tell you which variables were statistically significant predictors in the logistic regression.

From part ii., we know that medical aid status will not be related to traced status for a model that also includes race. So race will be the only statistically significant predictor of traced.

3. In this question, we will consider the data from assignment 1. The data were weights collected on 72 girls suffering from anorexia. The girls were randomly assigned to receive one of three therapies: cognitive behavioural (coded `b`), family (coded `f`), or the control therapy (coded `c`). The girls' weights were measured at the beginning of the study and after following the therapy for a period of time. Therapies are considered successful if girls gain weight on the therapy.

   For this question, our interest is whether a girl gained or lost weight (and not how much). A new variable `gained` is defined as 1 if a girl gained weight and 0 otherwise.

   Some edited output from SAS for an analysis of these data is on page 19. Some numbers have been replaced by `X`'s.

   (a) (2 marks) From what you are given, do you have any concerns about the appropriateness of the inferences from the logistic regression model that was fit? What else would you like to see?

   There is nothing to be concerned about in what we are given. This could be conceived as a binomial response model and then we could look at the residuals for potential outliers. (Deviance GOF test wouldn't make sense since using indicator variables for therapy <u>is</u> the saturated model.)

   (b) (1 mark) What is the estimated probability that a girl on therapy `c` gains weight?

$$\hat{\pi}_c = \frac{\exp\{1.1787 - 1.4888\}}{1 + \exp\{1.1787 - 1.4888\}} = 0.42$$

   (c) (4 marks) Carry out an hypothesis test with null hypothesis that the log-odds of gaining weight are the same for all three therapies; include: (I) the test statistic, (II) the distribution of the test statistic under the null hypothesis, (III) the p-value, (IV) the conclusion.

   Likelihood ratio test comparing the fitted to the null model:
   (I) Test statistic: $97.804 - 92.472 = 5.3$
   (II) chi-square with 2 d.f.
   (III) $0.05 < p < 0.1$
   (IV) weak evidence that the coefficient of the indicator variables are not both 0 *so there is weak evidence of differences in the log-odds of gaining weight among diets.*

(d) In the SAS output, you are given odds ratio estimates for therapy `b` versus therapy `f` and for therapy `c` versus therapy `f`.

i. (1 mark) What is the odds ratio estimate for therapy `b` versus therapy `c`?

$$\exp\{-0.6862 + 1.4888\} = 2.23$$

ii. (2 marks) Calculate the missing 95% Wald Confidence Interval for the odds ratio of therapy `b` versus therapy `f`.

95% CI for coefficient of $I_{[therapy=b]}$*:* $-0.6862 + 1.96(0.6880) = (-2.0347, 0.6623)$
CI for odds ratio: $(e^{-2.0347}, e^{0.6623}) = (0.131, 1.94)$

iii. (2 marks) Explain how the confidence interval in part ii. is consistent with one of the p-values in the output.

The CI for the odds ratio includes 1 so there is no evidence of a difference in the odds between therapies b and f.
The Wald test for the coefficient of the indicator variable for therapy b relative to f has $p = 0.3186$ *so this test also gives no evidence of a difference between the therapies.*

(e) (2 marks) The table below gives the counts of the numbers of girls who did or did not gain weight for each therapy.

	Therapy		
	b	c	f
Gained weight	18	11	13
Did not gain weight	11	15	4

An alternative analysis for these data would test whether the row and column variables in this table are independent. Do you prefer this proposed analysis or the analysis that you are given in the SAS output for this question? Why?

Prefer the logistic regression since there is a clear response.

4. In this question, we will again consider the data from assignment 1. The data were weights collected on 72 girls suffering from anorexia. The girls were randomly assigned to receive one of three therapies: cognitive behavioural (coded `b`), family (coded `f`), or the control therapy (coded `c`). The girls' weights were measured at the beginning of the study and after following the therapy for a period of time. Therapies are considered successful if girls gain weight on the therapy.

   For this question, we will use the weight of the girls as the response variable, with two measurements on each girl. The variable `when` is equal to `baseline` if the weight was measured at the beginning of the study and is equal to `end` if the weight was measured after the therapy period.

   Some edited output from SAS for an analysis of these data is given on pages 20 and 21. The fitted model assumes variances and covariances are the same for all subjects and includes a random effect for subject.

   (a) (3 marks) From the output that you are given, what can you conclude about the relative effectiveness of the therapies? Support your answer with appropriate numbers from the SAS output.

   There is strong evidence ($p = 0.0065$) that differences among therapies are different at baseline than at the end (`when*therapy` interaction).
   Although we aren't given all pairwise comparisons, from the tests of the coefficients we can see that therapy c at baseline differs from c at the end ($p = 0.0016$) but therapy b does not ($p = 0.0684$).

   (b) (3 marks) Write the model being fit; define all terms. State clearly which parts of the model are random and which are not random.

$$\begin{aligned} Y_{ijjk} &= \beta_0 + \beta_1 I_{[therapy=b],ijk} + \beta_2 I_{[therapy=c],ijk} + \beta_3 I_{[when=baseline],ijk} \\ &\quad + \beta_4 I_{[therapy=b],ijk} * I_{[when=baseline],ijk} + \beta_5 I_{[therapy=c],ijk} * I_{[when=baseline],ijk} \\ &\quad + u_{ij} + e_{ijk} \end{aligned}$$

   where Y_{ijk} is weight, $I_{[condition],ijk}$ is 1 if condition is true for observation ijk and 0 otherwise, u_{ij} is $N(0, \sigma_u^2)$ random effect for subject and e_{ijk} is $N(0, \sigma_e^2)$ noise.

   (c) (3 marks) What is the estimated variance-covariance matrix of the 144 observed weights?

   It is the 144×144 block diagonal matrix $\begin{pmatrix} \hat{\mathbf{D}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{D}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \hat{\mathbf{D}} \end{pmatrix}$ *where $\hat{\mathbf{D}}$ is the 2×2 matrix* $\begin{pmatrix} 23.3387 + 11.8019 & 11.8109 \\ 11.8109 & 23.3387 + 11.8019 \end{pmatrix}$

(d) The table of the standard deviations suggests that it may be worth considering a model that has different variances for baseline and end measurements, and that estimates a different variance-covariance matrix for each therapy group.

i. (1 mark) How many variance-covariance parameters would need to be estimated to accommodate this structure?

$$3 * 3 = 9$$

ii. (2 marks) How could you compare whether this proposed model fits the data better than the model fit in the SAS output?

Can compare AIC or conduct a likelihood ratio test

5. Suppose people are categorized by three variables. Variable 1 has I categories, variable 2 has J categories, and variable 3 has K categories. Thus there are $I \times J \times K$ categories in total. We observe y_{ijk}, the count of the number of people for whom variable 1 is i, variable 2 is j, and variable 3 is k. We will assume that the y_{ijk} can be considered observations from Poisson distributions with means μ_{ijk} and use Poisson regression. We will fit a model that assumes that variables 1, 2, and 3 are independent.

   (a) (4 marks) Show that the deviance is

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

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

   The likelihood function is

$$\prod_i\prod_j\prod_k \frac{e^{-\mu_{ijk}}\mu_{ijk}^{y_{ijk}}}{y_{ijk}!}$$

   The log-likelihood function is

$$\sum_i\sum_j\sum_k (-\mu_{ijk} + y_{ijk}\log(\mu_{ijk}) - \log(y_{ijk}!))$$

   For the saturated modek, $\hat{\mu}_{ijk} = y_{ijk}$
   The deviance is

$$2\log(L_{sat}) - 2\log(L_{fitted}) =$$
$$2\sum_i\sum_j\sum_k (-y_{ijk} + y_{ijk}\log(y_{ijk}) - \log(y_{ijk}!) + \hat{\mu}_{ijk} - y_{ijk}\log(\hat{\mu}_{ijk}) + \log(y_{ijk}!))$$

   and $\sum\sum\sum \hat{\mu}_{ijk} = \sum\sum\sum y_{ijk}$
   So the deviance is $2\sum\sum\sum y_{ijk}\log\left(\frac{y_{ijk}}{\hat{\mu}_{ijk}}\right)$

   (b) (2 marks) What are the estimated values of μ_{ijk} from the fitted model? Give how they can be calculated from the observed counts; you do not need to derive them.

   Since the fitted model is the independence model

$$\hat{\mu}_{ijk} = n\frac{y_{i\cdot\cdot}}{n}\frac{y_{\cdot j\cdot}}{n}\frac{y_{\cdot\cdot k}}{n}$$

   where $y_{i\cdot\cdot} = \sum_j\sum_k y_{ijk}$, *etc. and* $n = \sum_i\sum_j\sum_k y_{ijk}$

6. In this course, we have studied the following (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) (12 marks (4 each)) Three scenarios (below and on the next page) relate 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 cereals in the study, it was recorded whether the cereal was on the lower, middle, or upper shelf. For each scenario indicate:
   (I) which of the 6 types of generalized linear model is appropriate
   (II) the model you would use for the analysis, defining all terms
   (III) the null and alternative hypotheses for the test that addresses the question of interest.

      i. The cereals were examined for their content of various vitamins and minerals. The researcher believes that stores may tend to put healthier cereals on the upper shelf since they are more likely to appeal to adults. We are interested in whether the content (in grams per serving) of three specific vitamins in the cereals are useful in predicting whether a cereal is displayed on the upper shelf.

      (I) binary logistic regression
      (II) Model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1\texttt{vit1} + \beta_2\texttt{vit2} + \beta_3\texttt{vit3}$$

      where π *is the probability of being on the upper shelf and* `viti` *is the content of vitamin* `i`
      (III) Test $H_0:\ \beta_1 = 0$ *versus* $H_a:\ \beta_1 \neq 0$ *and* $H_0:\ \beta_2 = 0$ *versus* $H_a:\ \beta_2 \neq 0$ *and* $H_0:\ \beta_3 = 0$ *versus* $H_a:\ \beta_3 \neq 0$

      ii. We are interested in learning whether there are differences in the average sugar content (in grams per serving) of the cereals depending on their placement on the lower, middle, or upper shelves.

      (I) one-way analysis of variance
      (II) Model:

$$Y = \beta_0 + \beta_1 I_{[lower\ shelf]} + \beta_2 I_{[middle\ shelf]} + e$$

      where Y *is the sugar content and* $I_{[condition]}$ *is* 1 *if the cereal is on the shelf indicated and* 0 *otherwise*
      (III) Test $H_0:\ \beta_1 = \beta_2 = 0$ *versus* $H_a:$ *at least one of* $\beta_1,\ \beta_2 = 0$ *is not* 0

iii. Many of the cereals come with an incentive to buy them such as a free toy in the box or a chance to win a prize. We count the number of cereals with and without an incentive on each of the lower, middle, and upper shelves. We are interested in learning if shelf placement and whether or not a cereal has an incentive are related.

(I) Poisson regression
(II) Model:

$$\begin{aligned}\log(\mu) &= \beta_0 + \beta_1 I_{[lower\ shelf]} + \beta_2 I_{[middle\ shelf]} + \beta_3 I_{[incentive]} \\ &\quad + \beta_4 I_{[lower\ shelf]} * I_{[incentive]} + \beta_5 I_{[middle\ shelf]} * I_{[incentive]}\end{aligned}$$

where μ is the mean of the Poisson distribution and $I_{[condition]}$ is 1 if the condition holds and 0 otherwise
(III) Test H_0 : $\beta_4 = \beta_5 = 0$ versus H_a : at least one of β_4, $\beta_5 = 0$ is not 0

(It is possible to answer this question by giving a model without the interaction terms and then saying that a deviance goodness-of-fit test should be carried out.)

(b) (3 marks) Of the 6 models we have studied (as identified at the beginning of this question), which have random error terms in their models? Why do some models need the random error term and some models do not?

There are random error terms in the two analysis of variance models and in the mixed model. In these models we are modelling the data so we need to include a term for the noise; in the others we are modelling a parameter of the underlying probability distribution.

(c) (2 marks) In order to carry out inference about the coefficients of the explanatory variables, which of the 6 models we have studied require a large sample size? Why is a large sample size necessary for these models?

The tests for the two logistic regression models and the Poisson regression model rely on the large sample properties of MLEs.