UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL 2012 EXAMINATIONS

STA 303 H1S / STA 1002 HS

Duration - 3 hours

Examination Aids: Calculator

LAST NAME: SOLUTIONS **FIRST NAME:**

STUDENT NUMBER:

• There are 30 pages including this page.
• Pages 15 to 29 contain SAS output. You may remove these pages from your exam but hand them in with your exam. Do not write answers on the pages of SAS output.
• The last page (page 30) 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: 90

1a	1b	2abc	2def	2gh	3a	3b

4ab	4cd	5ab(i)	5b(ii,iii, iv)	6	7

1. 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 doses: 5 mg and 10 mg (variable name: `dose`). For each dose, eight dishes were used with 25 termites in each dish. The variable of interest is the number of termites still alive; counts were taken daily for 15 days. In this question, our response variable is the number of termites alive in a dish on day 15 (variable name: `number`).

   The data are in the table below. The Day 1 counts are not used in the analysis.

		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 SAS is given on page 15. `dose` is treated as a categorical variable. The questions below relate to this output.

   (a) You are given SAS output from `proc glm` but `proc ttest` could also have been used. If `proc ttest` had been used and you considered the part of the output from `proc ttest` that corresponds to the output from `proc glm`:

   i. (2 marks) Should you look at the t-test that is labelled "Satterthwaite" or "pooled"? Why?

   You should look at the test labelled "pooled" as that test is carried out under the assumption that the variances in the two groups are the same, corresponding to the constant variance assumption of the linear model which is made in the output from `proc glm`.

   ii. (1 mark) What would be the value of the corresponding p-value on the `proc ttest` output?

   $p = 0.0492$

(Question 1 continued)

(b) For the given output from `proc glm`:

i. (2 marks) What is the practical meaning of the coefficient whose estimate is 5.3750?

On average, 5.375 more termites are alive at day 15 in dishes with dose 5 mg than in dishes with does 10 mg.

ii. (2 marks) What can you conclude from the Type III SS F-test? State your answer in practical terms.

The p-value is 0.0492. So we have evidence that the mean number of termites still alive is different between the two doses.

iii. (3 marks) Show how the standard error of the estimate of β_1 can be calculated from other numbers on the output. (*Hint:* How can you calculate the estimate from other numbers on the output?)

$\hat{\beta}_1 = \overline{y}_{10} - \overline{y}_5$
So $Var(\hat{\beta}_1) = \frac{\sigma^2}{8} + \frac{\sigma^2}{8}$ *since observations in different dishes are independent.*
The estimate of the error variances is 24.9196.
So the estimate of the standard error of $\hat{\beta}_1$ *is* $\sqrt{24.9196/4} = 2.496$.

2. In this question, we will again use the data from question 1. Twenty-five termites were put in each of eight dishes containing resin with concentration 5 mg and in each of eight dishes containing resin with concentration 10 mg. In this question, the counts of the number of termites still alive in each dish at day 15 are modelled as Binomial random variables (variable name: `number`). The explanatory variable is the concentration of the resin (variable name: `dose`).

   Some edited output from SAS is given on pages 16 to 18. In MODEL 1 `dose` is modelled as a categorical variable and in MODEL 2 `dose` is modelled as a quantitative variable.

   (a) (3 marks) In MODEL 1, some of the numbers in the output have been replaced with letters. Give the values of the letters below.

   (A) = $468.584 + (1)\log(400) = 474.58$

   (B) = $\left(-1.16215/0.1905\right)^2 = 72.45$

   (C) = $\exp\left(1.1319 + 1.96(0.2398)\right) = 4.96$

   (b) (4 marks) Give a practical interpretation of the estimate of β_1:

   i. For MODEL 1:

   $\exp(\hat{\beta}_1) = \exp(1.1319) = 3.10$
   The estimated odds of a termite being alive on day 15 are 3.1 times greater if the termite was exposed to the 5 mg dose than if it was exposed to the 10 mg dose.

   ii. For MODEL 2:

   $\exp(\hat{\beta}_1) = \exp(-0.2264) = 0.797$
   For an increase in dose of 1 mg, the estimated odds of a termite being alive at day 15 are 79.7% (about 20% less) of what they were before.

   (c) (2 marks) For MODEL 1, what is the estimated probability that a termite in a dish with 5 mg dose of resin is alive on day 15?

$$\hat{\pi} = \frac{\exp(-1.6215 + 1.1319)}{1 + \exp(-1.6215 + 1.1319)} = 0.38$$

(d) (3 marks) For MODEL 1, calculate the Pearson residual for the first dish. The first dish had a dose of 5 mg of resin and 11 termites were alive at day 15.

$$\frac{11 - 25(0.38)}{\sqrt{25(0.38)(1-0.38)}} = 0.62$$

where the 0.38 is from part (c)

(e) (3 marks) The researchers would like an estimate of the dose of resin that results in a 50% chance of a termite being alive on day 15. Should you use MODEL 1 or MODEL 2? Why? Estimate the required dose.

Use MODEL 2 since it treats dose as a quantitative variable so it can be applied to more doses than just 5 and 10 mg.
If $\pi = 0.5$,

$$\begin{aligned} 0 &= 0.6424 - 0.2264\,\widehat{\texttt{dose}} \\ \widehat{\texttt{dose}} &= 2.84\ mg \end{aligned}$$

(f) (2 marks) There are only 16 observations. Is this a concern? Why or why not?

It is not a concern. We need large sample sizes for the Wald and Likelihood Ratio tests to be valid, but the assumed sample size here is $25 \times 16 = 400$ *(which is large) since there are 25 termites per dish.*

(g) (4 marks) For MODEL 2, conduct a likelihood ratio test to determine if dose statistically significantly affects the odds that a termite is alive on day 15. Specify (I) the null and alternative hypotheses, (II) the value of the test statistic, (III) the distribution of the test statistic under the null hypothesis, (IV) an appropriate conclusion.

(I) $H_0: \beta_1 = 0$ *versus* $H_a: \beta_1 \neq 0$ *where* β_1 *is the coefficient of* `dose`
(II) Test statistic: $468.584 - 444.773 = 23.811$
(III) Chi-square(1)
(IV) 23.811 will be far in the right tail of a Chisquare(1) distribution so the p-value will be very small. Thus we have strong evidence that β_1 *is not 0, do dose statistically significantly affects the odds that a termite is alive on day 15.*

(h) (2 marks) For MODEL 2, explain how the 95% Wald confidence interval for the odds ratio agrees with your conclusion in part (g).

The confidence interval is $(0.726, 0.876)$ *and does not include 1. An odds ratio value of 1 implies that the odds of being alive are the same regardless of dose. So the fact that the confidence interval for the odds ratio does not include 1 is consistent with the conclusion in part (g) that dose is a statistically significant predictor of the odds of being alive.*

3. In this question we will again consider the data about termites and resin. As before, 25 termites were put in each of eight dishes of each of two resin doses, 5 mg and 10 mg (variable name: `dose`, treated as a categorical variable for this question). On day 15, the termites were classified as dead or alive (variable name: `status`). The form of the data used in this question is summarized in the table below.

	status	
dose	alive	dead
5	76	124
10	33	167

SAS output for the analysis of the data in this form is given on pages 19 and 20.

(a) For ANALYSIS I:

i. (1 mark) What are the null and alternative hypotheses for the test with test statistic 23.3173?

H_0 : `dose` *and* `status` *are independent*
versus
H_a : `dose` *and* `status` *are not independent*

ii. (4 marks) Give a complete practical conclusion.

$p < 0.0001$ *so we have strong evidence that* `dose` *and* `status` *are not independent. 38% of termites exposed to a dose of 5 mg are alive at day 15 but only 16.5% of termites exposed to a dose of 10 mg are alive at day 15.*

(b) For ANALYSIS II:

i. (1 mark) Why is the deviance 0?

The model fit is the saturated model and perfectly estimates the counts in the table.

ii. (3 marks) What do you conclude from the Wald test for the coefficient of the interaction term? Give your answer in practical terms.

$p < 0.0001$ *so we have strong evidence that the coefficient of the interaction term is not 0. So we have strong evidence that the difference in the mean counts of dead versus alive termites differs with the dose exposure.*

iii. (3 marks) Give the estimated mean and variance for the number of termites alive on day 15 who were exposed to a resin dose of 5 mg.

$\hat{\mu} = \exp(4.8203 - 0.4895) = 76.0$
$\widehat{variance} = 76.0$

iv. (3 marks) What are the estimated odds of a termite being alive on day 15 for a termite exposed to a resin dose of 5 mg?

$\exp(-0.4895) = 0.613$

4. In this question we will again consider the data about termites and resin. As before, 25 termites were put in each of eight dishes of each of two resin doses, 5 mg and 10 mg (variable name: `dose`). The number of termites still alive in each dish was counted each day, and in this question we will consider the number of termites alive in each dish (the response variable, variable name: `number`) on each of days 5, 10, 15 (variable name: `day`).

The data are in the table below. The counts on Day 1 are not included in the analysis.

		Number of termites			
Dish	Dose	Day 1	Day 5	Day 10	Day 15
1	5	25	18	13	11
2	5	25	21	15	11
3	5	25	23	16	12
4	5	25	24	20	12
5	5	25	19	13	5
6	5	25	20	15	9
7	5	25	23	16	6
8	5	25	19	12	10
9	10	25	22	18	16
10	10	25	20	17	13
11	10	25	6	2	1
12	10	25	10	3	0
13	10	25	9	0	0
14	10	25	3	0	0
15	10	25	3	0	0
16	10	25	9	5	3

Some edited SAS output for this question is on pages 21 to 24. Two models were fit to the data: MODEL 1 and MODEL 2. In all of the analysis for this question, `dose` and `day` are treated as categorical variables.

(a) (2 marks) On page 21 there are two plots of the least squares means of the number of termites alive for each dose and day. The solid line connects the means for dishes containing 5 mg of resin and the dashed line connects the means for dishes containing 10 mg of resin. Which plot is the plot for MODEL 1 and which plot is the plot for MODEL 2? How can you tell?

The second plot is for MODEL 2. The lines are parallel indicating no interaction and there is no interaction in MODEL 2.

(b) (2 marks) The output for MODEL 2 includes both the means and least squares means and they are equal. Is this always the case for the type of model being fit here? Explain why it is always the case or why it is the case for these data.

They are only equal if the numbers of observations are the same in each level of the categorical variables.

(c) (6 marks) Give a complete, practical conclusion based on all of the output given for question 4. Support your answer with numbers from the SAS output. If there are any numbers that you would like to have that are not available in the SAS output, indicate what they are and why you want them.

There is no evidence of interaction between `day` *and* `dose` *(*$p = 0.3535$*) so how dose affects the number of termites alive does not depend on the day and thus we should use MODEL 2 to investigate the main effects.*
From MODEL 2:
- There is strong evidence that the mean number of termites alive differs with dose ($p < 0.0001$*). On average, fewer termites are alive on a dose of 10 mg (mean of 6.67 termites) than on a dose of 5 mg (mean of 15.125 termites).*
- There is strong evidence that the mean number of termites alive differs with day ($p = 0.0001$*). On average, 15.56 termites are alive on day 5 but only 10.3 termites are alive on day 10 and 6.8 on day 15.*

We would like to have pairwise comparisons between days to see if all of the differences among days are statistically significant.

(d) (5 marks) In order for the inferences for MODEL 2 to be valid, certain conditions (assumptions) must hold. Evaluate whether these conditions hold for this model based on what you are given.

Conditions we need and evaluation of whether or not they hold:

- *Independent observations*
  This is not the case since observations over time on the same dish are correlated.
- *Constant variance*
  From the plot of the studentized residuals versus the predicted values, there appear to be differences.
  From the table on page 21, the ratio of the largest to the smallest s.d. is $7.54/2.23 > 2$*. So there is non-constant variance.*
- *Normally distributed residuals (with no outliers)*
  From the normal quantile plot, the distribution of the residuals appears to be skewed (although it's OK to conclude that it is close enough to straight considering the robustness of the tests based on means).
  Some studentized residuals are > 2 *but there are no extreme outliers.*

5. In this question, we will 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 (the response variable, variable name: `number`) on each of days 5, 10, 15 (variable name: `day`). The data being used in the analysis for this question are the same as in question 4. (Again, the counts on Day 1 are not included in the analysis.)

   Some edited output from SAS for this question is given on pages 25 to 29. `dose` and `day` are treated as categorical variables.

   (a) (3 marks) On page 25, for each dose you are given 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?

   For dose of 10 mg, there seems to be approximately equal correlation between observations on different days.
   For dose of 5 mg, there are weaker correlations among observations on days 5 & 15 and 10 & 15 than on days 5 & 10.
   So compound symmetry covariance structure may not be appropriate (should allow covariance to differ between pairs of days) and we should use a different covariance matrix for each dose.

   (b) You are given output from two models which differ by the covariance structure used for counts from the same dish. The covariance structures are:
   MODEL 1: Compound symmetry covariance structure, the same for each dose
   MODEL 2: Unstructured covariance structure, different for each dose

   i. (4 marks) Is it possible to carry out a statistical test to compare the fit of MODEL 1 to MODEL 2? If so, carry out the test. If not, indicate why not.

   Use a Likelihood Ratio Test.
   H_0 *: the restrictions on the covariance parameters to turn MODEL 2 into MODEL1 hold*
   H_a *: the restrictions don't hold*
   Test statistic: $221.2 - 196.3 = 24.9$
   Under H_0*, this is an observation from a chisquare distribution with* $12 - 2 = 10$ *degrees of freedom.*
   The resulting p-value will be small.
   So we have evidence that the restrictions don't hold and MODEL 2 is better for these data.

ii. (3 marks) What are the values of AIC for the two models? Do the values of AIC give you any additional information for determining which model is better? Explain.

MODEL 1: AIC $= 221.2 + 2(2) = 225.2$
MODEL 2: AIC $= 196.3 + 2(12) = 220.3$
Since AIC is smaller for MODEL 2, this criterion picks MODEL 2 as more appropriate for these data.

iii. (3 marks) For each model, what is the estimate of the covariance between observations on days 5 and 10 from the same dish?

MODEL 1: 25.46 *for both doses*
MODEL 2: 5.0 *for dose 5 mg and* 51.68 *for dose 10 mg*

iv. (5 marks) For the model you prefer (either MODEL 1 or MODEL 2), give a complete practical conclusion. Support your answer with appropriate numbers from the SAS output.

Using MODEL 2 (as chosen in parts i. and ii.), there is evidence (p $= 0.0029$*) that the differences in the mean number of termites alive across days differs with dose.*
For dose 5 mg, the mean number of termites alive differs significantly across all days. On days 5, 10, and 15, the mean number of termites alive in dishes with dose 5 mg are 20.9, 15.0, and 9.5, respectively. The Tukey adjusted p-values all show strong evidence of differences between each pair of means (between days 5 & 10 p < 0.0001*, between days 5 & 15 p* < 0.0001*, between days 10 & 15 p* $= 0.0002$*).*
For dose 10 mg, the mean number of termites alive on days 5, 10, and 15 are 10.3, 5.6, and 4.1, respectively. There is evidence of that the means differ between days 5 & 10 (p < 0.0001*) and days 5 & 15 (p* < 0.0001*) but only weak evidence of a difference in the means between days 10 & 15 (p* $= 0.0859$*).*
Alternatively, we could look at how the mean number of termites alive differs between doses on each day. On day 15, there is no evidence of a difference between the two doses (means are 9.5 for dose of 5 mg and 4.1 for dose of 10 mg, p $= 0.2905$*). On day 10, there is evidence that the mean number of termites differs between doses (means are 15.0 for dose of 5 mg and 5.6 for dose of 10 mg, p* $= 0.0262$*). On day 5, there is strong evidence that the mean number of termites differs between doses (means are 20.9 for dose of 5 mg and 10.3 for dose of 10 mg, p* $= 0.0050$*). (Again, the p-values for the differences between pairs of means are Tukey-adjusted.)*

6. (3 marks) The analyses in questions 1 through 5 used data from the same experiment. Which analysis do you prefer? Why?

   I prefer the analysis in question 5 (the repeated measures analysis using a mixed model). This analysis allows us to consider the effect of dose over time, while appropriately dealing with correlations between pairs of observations on the same dish.

   (It is possible to also make a case for the binomial logistic regression model if you argue that you don't care about how fast the termites die. There is no compelling case for the other models. There is a clear response, so the Poisson model is unnecessarily difficult to interpret. Note that you can't compare AIC among different types of models. The two-way analysis of variance has serious problems as noted in 4(d).)

7. 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) (2 marks) For which of these models would a normal quantile plot be a useful tool?

   (1) one-way analysis of variance
   (2) two-way analysis of variance
   (6) mixed models

   (b) (2 marks) For which of these models is a large sample size essential for inference to be valid?

   (3) binary logistic regression
   (4) binomial logistic regression
   (5) Poisson regression
   (6) mixed models for inference on the covariance parameters

   (c) (2 marks) Both binomial logistic and Poisson regression models can be used when the outcome is a count. What must be true of the manner in which the data were collected if it is appropriate to use a Poisson regression model but it is not appropriate to use a binomial logistic regression model?

   For binomial logistic regression, we need to have a count of events in a fixed number of trials. In Poisson regression, there is no fixed number of trials.