STA303/1002- MIDTERM TEST SOLUTIONS

March 1, 2018

NAME : _____

STUDENT NUMBER: _____

Instructions:

- This test consists of 4 questions on 14 pages. The total marks are distributed as follows:

Question	1	2	3	4	Total
Marks	11	13	20	6	50

- Time allowed: **90 minutes**
- Aids allowed: a non-programmable calculator
- **Answer all questions, in the space provided. Work written on the back of pages will NOT be graded.**
- Where appropriate, round your final answers to **2 decimal places**.
- Relevant Tables and Formulas are provided.
- Use a benchmark significance level of 5%, unless otherwise stated.
- Keep calm and do your best!

1. (11 marks) Consider the question in assignment 2, "Does baby birth weight change with gestational maturity?" The data included 409 measurements of babies birth weight (`bwt`) in ounces and we created a variable named `maturity`. Maturity level 1 corresponded to premature babies, level 2 corresponded to babies of normal maturity level and level 3 corresponded to postmature babies. Using the edited R codes and output below, answer the following questions.

```
> summary(fitm)

Call:
lm(formula = bwt ~ maturity)

Residuals:
    Min      1Q  Median      3Q     Max
-47.832 -13.832  -0.832  11.168  55.168

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   99.907      1.835  54.431  < 2e-16 ***
maturity2     18.925      2.324   8.145 4.73e-15 ***
maturity3     27.980      2.352  11.895  < 2e-16 ***

Residual standard error: 18.08 on 406 degrees of freedom
Multiple R-squared:  0.2599,Adjusted R-squared:  0.2562
F-statistic: 71.28 on 2 and 406 DF,  p-value: < 2.2e-16

> anova(fitm)
Analysis of Variance Table

Response: bwt
           Df Sum Sq Mean Sq F value    Pr(>F)
maturity    2  46586 23293.1  XXXXX  < 2.2e-16 ***
Residuals 406 132680   326.8
```

   (a) (2 marks) What are the averages of baby weight for premature and postmature babies? Show your work. (*1 mark each*)

      1. Premature: **99.91 ounces**

      3. Postmature: **99.907+27.980= 127.89 ounces**

(b) (3 marks) Is there evidence of a difference in the mean weight among babies of differing maturity levels? Answer this question by completing the following parts.

i. Null and Alternative hypotheses:

$H_0 : \beta_1 = \beta_2 = 0$, H_a: at least one of β_1, β_2 is not zero

or

$H_0 : \mu_1 = \mu_2 = \mu_3$, $H_a : \mu_i \neq \mu_j$ for some $i \neq j$, $i, j = 1, 2, 3$
(at least one pair is different)

ii. Test Statistic: $\dfrac{23293.1}{326.8} = 71.28$

iii. Distribution of test statistic under H_0: $\underline{F(2, 406)}$

iv. P-value (or range of p-value): $\underline{p < 0.0001}$ **or** $\underline{p \approx 0}$

v. Conclusion: **There is strong evidence that babies mean birth weight differ by maturity level.**

1 mark each for parts i, (ii, iii) and (iv, v).

(c) (3 marks) While controlling the Type I error rate at the individual level, Bonferroni's method can be used to find all three pairwise confidence intervals to compare mean baby birth weight by maturity level. Give the following three values of the formula below to verify that the difference between mean birth weight at maturity level 2 and mean birth weight at maturity level 1 is between 13.3 and 24.5 ounces, by Bonferroni's method. (*1 mark each*)

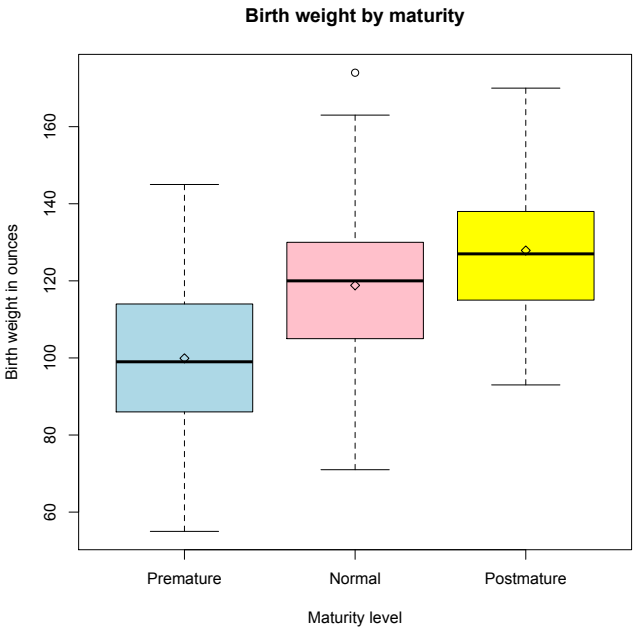
$$(\bar{x}_2 - \bar{x}_1) \pm t_{\frac{\alpha^*}{2}, df} \sqrt{s^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}$$

1. Observed difference, $\bar{x}_2 - \bar{x}_1$: $\underline{\textbf{18.93 ounces}}$

2. Individual confidence level: $\underline{(1 - \frac{0.05}{3})100\% = 98.33\%}$

3. Pooled variance, s^2: $\underline{\textbf{326.8}}$

(d) (3 marks) Given the side-by-side box plots and the table of group sizes, n and group variances, s^2 below, determine whether or not the assumptions of the method are satisfied. The diamond in the box plots correspond to the group means.



Maturity level	s^2	n
Premature	452.02	97
Normal	306.40	161
Postmature	268.41	151

Assumption	Discussion	Determination
Normal populations	The group means and group medians are very similar. No obvious outliers. The two whiskers of each box plot are similar in length.	Assumption satisfied
Equal Population variances	From the box plots, the ranges of the samples are similar. Using the Rule of Thumb, $\frac{452.02}{268.41} = 1.68 < 4$.	Assumption satisfied
Independence	Populations are independent since each birth weight fall into one and only one category. Assume that there are no twins or related babies in the data.	Assumed satisfied

1 mark for each assumption.

2. (13 marks) In assignment 2, we considered how baby birth weight varied with gestational maturity (level 1, 2 or 3) and maternal smoking status (smoker or non-smoker). Using the following R codes and output, answer the questions that follow.

```
> fitsm=lm(bwt~maturity*smoke)
> summary(fitsm)

Call:
lm(formula = bwt ~ maturity * smoke)

Residuals:
    Min      1Q  Median      3Q     Max
-50.893 -12.839  -0.732  11.161  51.161

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       105.893      2.342  45.209  < 2e-16 ***
maturity2          16.946      2.965   5.716 2.13e-08 ***
maturity3          23.458      2.942   7.974 1.60e-14 ***
smoke1            -14.161      3.603  -3.931 9.97e-05 ***
maturity2:smoke1    4.675      4.561   1.025   0.3059
maturity3:smoke1   10.070      4.673   2.155   0.0318 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 17.53 on 403 degrees of freedom
Multiple R-squared:  0.3093,Adjusted R-squared:  0.3007
F-statistic: 36.09 on 5 and 403 DF,  p-value: < 2.2e-16

> anova(fitsm)
Analysis of Variance Table

Response: bwt
                Df Sum Sq Mean Sq F value    Pr(>F)
maturity         2  46586 23293.1 75.8139 < 2.2e-16 ***
smoke            1   7395  7394.8 24.0683 1.352e-06 ***
maturity:smoke   2   1467   733.5  2.3874   0.09317 .
Residuals      403 123818   307.2
```

(a) (2 marks) Circle A, B, C, D or E to specify which statistical method was used in the R output above? **Explain why this method was chosen based on the information given in the introductory paragraph.**

(C) two-way analysis of variance

This model was appropriate because there is a continuous response variable (baby birth weight) and two categorical explanatory variables (maturity level and maternal smoking status).

(b) (3 marks) Write the model that was fit, carefully defining all terms.

$$Y_i = \beta_0 + \beta_1 I_{mat2,i} + \beta_2 I_{mat3,i} + \beta_3 I_{smoke,i} + \beta_4 I_{mat2,i} * I_{smoke,i} + \beta_5 I_{mat3,i} * I_{smoke,i} + \epsilon_i$$

where
Y is baby birth weight in ounces,
I_{mat2} is 1 if baby attained normal maturity, and 0 otherwise,
I_{mat3} is 1 if baby was postmature, and 0 otherwise,
I_{smoke} is 1 if baby's mother was a smoker, and 0 if non-smoker, and
ϵ is random error.

(1 for correct model and 2 for definition of variables)

(c) (1 mark) What practical quantity, if any, is being estimated by the estimate of the intercept?

The mean birth weight of premature babies whose mother was a non-smoker.

(d) For the test with p-value=0.09317,

i. (1 mark) What are the null and alternative hypotheses?

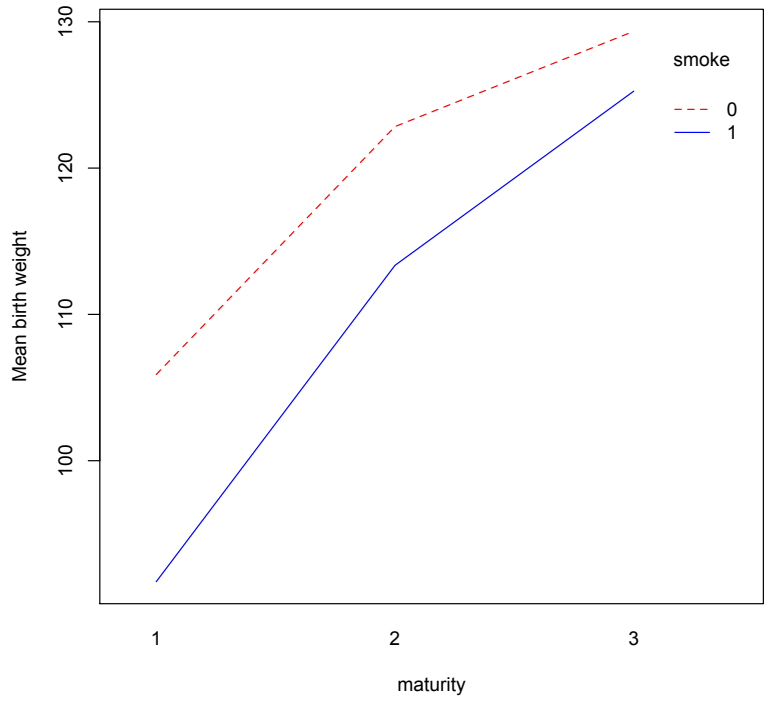
$H_0 : \beta_4 = \beta_5 = 0$,
H_a: at least one of β_4, β_5 is not zero

(0.5 marks for each hypothesis)

ii. (2 marks) Explain in practical terms, what you conclude from the test.

The effect of maternal smoking (Factor 1) on baby's birth weight does not change significantly with maturity level (Factor 2).

(e) (4 marks) Explain how the interaction plot below is consistent with the conclusions that can be drawn from inferences about the fitted model. Where possible, support your answer with relevant numbers from the R output.



Inference	Plot	Relevant p-value
There is weak evidence of interaction.	The lines are close to parallel.	0.09317
There is strong evidence of differences due to maturity level.	The lines are not horizontal.	< 0.0001 (optional)
There is evidence of a difference due to maternal smoking status.	The line for non-smoker is above the line for smoker.	

(2 marks for first line; 1 mark each for other lines.)

3. (20 marks) A health survey in Hague, Netherlands discovered an association between keeping pet birds and increased risk of lung cancer (`lc`). To investigate bird-keeping as a risk factor researchers conducted a case-control study of patients at four hospitals in the Hague. They identified 49 cases of lung cancer among patients who were registered with a general practice and they selected 98 controls from the population of residents having the same general age structure. In this question, we will investigate how well lung cancer incidence can be predicted from bird-keeping status (`keep`) and **age**.

   Relevant R codes and output are below and on the next page. Some numbers have been replaced by X's.

```
> str(lc)
 Factor w/ 2 levels "NoCancer","LungCancer": 2 2 2 2 2 2 2 2 2 2 ...
> str(age)   int [1:147] 37 41 43 46 49 51 52 53 56 56 ...
> str(keep)
 Factor w/ 2 levels "NoBird","Bird": 2 2 1 2 2 1 2 1 2 1 ...

-------------------------------------------------
MODEL 1
-------------------------------------------------
> fit1<-glm(lc~age+keep, family=binomial, data=lung)
> summary(fit1)
Call:
glm(formula = lc ~ age + keep, family = binomial, data = lung)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2371  -0.7031  -0.6661   1.1493   1.9043

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.32950    1.53189  -1.521 0.128342
age          0.01615    0.02569   XXXX 0.529594
keepBird     1.40271    0.38004   XXXX  0.000223 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 172.53  on 144  degrees of freedom
AIC: 178.53

Number of Fisher Scoring iterations: 4
```

```
-------------------------------------------------
MODEL 2
-------------------------------------------------
> fit2<-glm(lc~age+keep+age:keep, family=binomial, data=lung)
> summary(fit2)

Call:
glm(formula = lc ~ age + keep + age:keep, family = binomial,
    data = lung)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2794  -0.6688  -0.6680   1.1261   1.7964

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4058938  2.4052306  -0.585    0.559
age           0.0003369  0.0410580   0.008    0.993
keepBird     -0.0402338  3.0120224  -0.013    0.989
age:keepBird  0.0251607  0.0522527   XXXX    XXXX

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 172.30  on 143  degrees of freedom
AIC: 180.3

Number of Fisher Scoring iterations: 4
```

(a) (3 marks) Based on MODEL 1, write the likelihood and log-likelihood functions in terms of y_i and the model parameters to be estimated.

(1 mark) Likelihood function: $\mathcal{L} = \prod_{i=1}^{147} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

(1 mark) Log-likelihood function: $log\mathcal{L} = \sum_{i=1}^{147}\{y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\}$

(1 mark) where $\pi_i = \dfrac{\exp(\beta_0 + \beta_1 \mathtt{age}_i + \beta_2 I_{keepBird,i})}{1+\exp(\beta_0 + \beta_1 \mathtt{age}_i + \beta_2 I_{keepBird,i})}$

(b) (2 marks) What is the fitted equation according to MODEL 1? Define all variables.

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -2.33 + 0.02\texttt{age}_i + 1.40 I_{keepBird,i}$$

where
$\hat{\pi}_i$ is the estimated probability that the ith individual develops lung cancer,
$\texttt{age}_i$ is the age of the ith individual, and
$I_{keepBird,i}$ is 1 if the ith individual keeps a bird and 0 otherwise.

(1 mark for correct model and 1 mark for definition of variables.)

(c) (3 marks) Based on MODEL 1, does age have any effect on lung cancer incidence? Your answer should include the appropriate null and alternative hypothesis, test statistic and it's distribution under the null hypothesis, p-value and conclusion.

i. Null and Alternative hypotheses:

$H_0 : \beta_1 = 0,$
$H_a : \beta_1 \neq 0$

ii. Test Statistic: $z = \dfrac{0.01615}{0.02569} = 0.63$ or $z^2 = 0.63^2 = 0.40$

iii. Distribution of test statistic under H_0: $z \sim \mathcal{N}(0,1)$ **or** $z^2 \sim \chi^2_1$

iv. P-value (or range of p-value): **0.53**

v. Conclusion: **Since the p-value is large, we fail to reject the null hypothesis and conclude that there is no evidence that age has an effect on the odds of developing lung cancer over and above the bird keeping status of an individual.**

(1 mark each for parts (i, ii), (iii, iv), and v).

(d) (3 marks) Based on MODEL 1, what is the effect of bird-keeping on the odds of developing lung cancer? Explain. (*Hint: Your answer should include a 95% confidence interval.*)

(1 mark) **The estimated coefficient of keepBird, β_2 is 1.403. Hence, the odds ratio of bird-keeping versus no bird-keeping is** $\exp(1.403) = 4.07$.

(1 mark) **A 95% CI for β_2 is calculated using the formula $\hat{\beta}_j \pm z_{\frac{0.05}{2}} SE(\hat{\beta}_j)$. That is:**

$$1.403 \pm 1.96 * 0.38 = (0.66, 2.15)$$

(1 mark) **Finally, the odds of developing lung cancer for a subject who has a bird is 4.07 times the odds of a subject who is not a bird keeper. With 95% confidence, the related odds ratio is between** $\exp(0.66) = 1.93$ **and** $\exp(2.15) = 8.56$**.**

(e) (2 marks) From MODEL 1, what is the estimate of the probability of developing lung cancer for a 40-year old individual who keeps a bird?

(1 mark) The estimated probability of an event (developing lung cancer) is calculated using the formula: $\pi_i = \dfrac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \texttt{age}_i + \hat{\beta}_2 I_{keepBird,i})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \texttt{age}_i + \hat{\beta}_2 I_{keepBird,i})}$

(1 mark) Plugging in the estimated coefficients and values of the explanatory variables we get, $\pi_i = \dfrac{\exp(-2.33 + 0.0162(40) + 1.403(1))}{1 + \exp(-2.33 + 0.0162(40) + 1.403(1))} = \dfrac{0.757}{1 + 0.757} = 0.43$

(f) (3 marks) For MODEL 1, conduct a Global Likelihood Ratio test. *Your answer should include the appropriate null and alternative hypothesis, test statistic and it's distribution under the null hypothesis, p-value and conclusion.*
(0.5 marks) Null and Alternative hypotheses:
$H_0 : \beta_1 = \beta_2 = 0$, H_a : at least one of β_1, β_2 is not zero

(0.5 marks) Test Statistic: $G^2 = 187.14 - 172.53 = 14.61$
(0.5 marks) Distribution of test statistic under H_0: $G^2 \sim \chi^2_2$ (Chi-square with $df = 2$)

(0.5 marks) P-value (or range of p-value): $P(\chi^2_2 > 14.16) < 0.005$ (from tables)
(1 mark) Conclusion: **Since the p-value is very small, we have strong evidence against the null hypothesis. We conclude that the cited model is adequate, that is, at least one of the variables- age and bird-keeping status is useful in estimating the odds of developing lung cancer.**

(g) (3 marks) MODEL 2 includes the interaction of age and bird-keeping. Carry out an appropriate test for whether or not the interaction contributes in a statistically significant way to the explanation of the odds of lung cancer. *Your answer should include the appropriate null and alternative hypothesis, test statistic and it's distribution under the null hypothesis, p-value (or its range) and conclusion.*

	Wald approach OR	LRT procedure
H_0	($\gamma_3 = 0$ γ_3 represents the coefficient of the interaction term	Additive model is adequate
H_a	$\gamma_3 \neq 0$	Interaction model is better
Test statistic	$z = \frac{0.02516}{0.05225} = 0.48$	$G^2 = 172.53 - 172.30 = 0.23$
Distri. of TS under H_0	$z \sim \mathcal{N}(0,1)$	$G^2 \sim \chi^2_1$
range of p-value (using tables)	$2P(Z > 0.48) > 0.10$	$P(\chi^2_1 > 0.23) > 0.10$

Conclusion: **Since the p-value is large, we fail to reject the null hypothesis and conclude that the data are consistent with the coefficient of the interaction term being 0. Therefore, the interaction does not contributes in a statistically significant way to the explanation of the odds of developing lung cancer.**

(0.5 marks each for hypotheses, test statistic, distribution, p-value; 1 mark for conclusion)

(h) (1 mark) MODEL 2 estimates the log-odds of an individual having lung cancer. What would the fitted model be if it was estimating the log-odds of not developing lung cancer? Round your values to 4 decimal places.

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.4059 - 0.0003\texttt{age}_i + 0.0402 I_{keepBird,i} - 0.0252\texttt{age}_i * I_{keepBird,i}$$

(0.5 marks for correct sign of coefficients and 0.5 marks for correct absolute value of coefficients)

4. (6 marks) Fill out the table below to compare and contrast features of the models used in questions 1 and 3. *(1 mark for each cell)*

	Model in Question 1	First Model in Question 3
Underlying probability distribution of response (You do not need to specify) the parameters of the distribution.)	Normal	Bernoulli
Condition that must hold regarding the variance for inferences to be valid	Same for all observations	Variance changes for each observation; variance is $\pi_i(1-\pi_i)$
Probability distribution used to calculate the p-value for the test with null hypothesis that the coefficients of all parameters are 0, except for the intercept. (You do not need to specify) the parameters (or df) of the distribution.)	F	Chi-square, χ^2

END OF TEST