University of Toronto

Department of Statistical Sciences

STA 303H1S / 1002 HS – Winter 2015

Midterm Test- SOLUTIONS

February 24, 2015

Duration- 90 minutes

Last Name: ____

First Name:____

Student Number:___________________________________

Section enrolled in (Circle one): **STA303** **STA1002**

Aids allowed: Non-programmable calculator

Instructions:

- This test has 13 pages, including this page and 4 main questions. The last page includes some useful formulae and percentile points from various distributions. Please check that all pages are included.
- Show all your work and answer in the space provided, in ink. Pencil may be used, but then remarks will not be allowed. Use the back of pages for rough work.
- If you would like clarification of a question, or are having some other difficulty, please do not hesitate to seek assistance from your instructor or TA.
- Answer questions completely, using supporting statistical values where appropriate. Use a benchmark statistical significance level of 5% and 95% level for confidence intervals, unless stated otherwise.
- The maximum score is 50.
- Do your very best!

Question 1
An investigator was interested in uncovering the effect of genotype of mother (**`motgen`**) and genotype of litter (**`litter`**) on litter weight (**`weight`**). Hence, a foster feeding experiment was conducted with rat mothers and litters of four different genotypyes: A, B, I, and J. The measurement was the litter weight (in grams) after a trial feeding period.

For the purposes of this test, litters of genotypes A and B were combined to form litter 'AB' and litters of genotypes I and J were combined to form litter 'IJ'. Litters AB consisted of 31 pups, while litters IJ consisted of 29 pups. Here is some output from SAS to compare litters 'AB' to litters 'IJ'. Answer the following questions.

```
                              The GLM Procedure
                           Class Level Information

                        Class         Levels    Values
                        litter             2    AB IJ

                    Number of Observations Read          60
                    Number of Observations Used          60

Dependent Variable: weight
                                      Sum of
Source                      DF       Squares    Mean Square   F Value   Pr > F

Model                        1     45.831058      45.831058      (F)    0.4181
Error                       58   3996.657442      68.907887
Corrected Total             59   4042.488500

              R-Square     Coeff Var      Root MSE    weight Mean
              0.011337      15.41662      8.301077       53.84500

Source                      DF    Type III SS    Mean Square   F Value   Pr > F
litter                       1    45.83105840    45.83105840      (F)   0.4181

                                               Standard
    Parameter           Estimate              Error    t Value    Pr > |t|
    Intercept        52.94137931 B       1.54147139      34.34      <.0001
    litter    AB      1.74894327 B       2.14452071       (T)      0.4181
    litter    IJ      0.00000000 B        .                 .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.
```

(a) (2 marks) What are the mean weights of pups of litter genotypes AB and IJ? Show your work.

Mean of litterAB is $\hat{\beta}_0 + \hat{\beta}_1$=52.94+1.75=54.69
Mean of litterIJ is $\hat{\beta}_0$=52.94

1 mark for correct group mean

(b) (2 marks) Is there evidence of a difference in the mean weight between pups of litter genotypes AB and IJ? Explain.

No. We have evidence (p=0.4181) that the coefficient of the dummy variable of litter genotype (litter=1, when the litter has A or B genotype) is 0.

0.5 mark for saying yes, and 1.5 marks for good explanation (at least 0.5 for providing correct p-value)

(c) (2 marks) What are the 2 missing numbers **F** and **T**? Are they related? Explain.

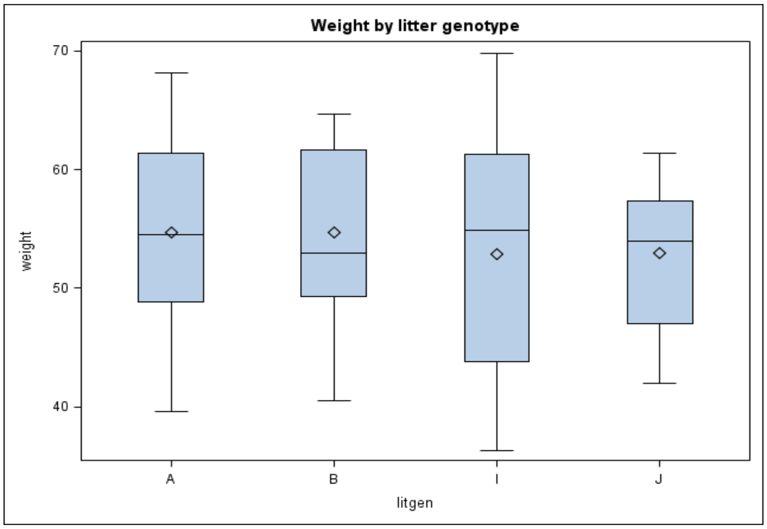
F=45.831/68.9079=0.67

T=1.7489/2.1445=0.82

Yes, $T_{58}^2=F_{1,58}$.

0.5 mark each for correct value of F and T. 0.5 mark for saying yes and 0.5 mark for correct explanation.

(d) (3 marks) Given the side-by-side boxplots and table of group sizes below of the four litter genotypes: A, B, I and J, discuss whether it was reasonable to combine litter genotypes A and B, and litter genotypes I and J. Do you have any concerns about these combinations, with respect to the assumptions of the general linear model?



Litter genotype	*n*
A	**16**
B	**15**
I	**14**
J	**15**

Yes, it seems reasonable since the group means of A and B are very similar and the group means of I and J are similar but both less than those of A and B.
However, we see that groups differ in size and variation. This is concerning since it may affect the assumption of equal group variance of the general linear model.

1 mark for comparison of means, 1 mark for comparison of variances and group sizes, and 1 mark for correct explanation.

Question 2

An alternative formulation of the model that could have been used in question 1 is

$$Y_{gi} = \theta_g + \epsilon_{gi,} \qquad g = 1, 2, 3, 4$$

where Y_{gi} is the weight of the *i*th pup with genotype *g* and ϵ_{gi} are random errors. By the method of least squares, the estimates of θ_g are found by minimizing

$$\sum_{g=1}^{4}\sum_{i=1}^{n_g}(Y_{gi} - \theta_g)^2$$

with respect to θ_1, θ_2, θ_3, *and* θ_4.

(a) (2 marks) Find the least squares estimates of θ_1, θ_2, θ_3, *and* θ_4.

Let *R* be the expression above that should be minimized:

$$\boldsymbol{\frac{\partial R}{\partial \theta_g} = -2\sum_{i=1}^{n_g}(Y_{gi} - \theta_g)}$$

Setting the above equal to 0 and solving gives

$$\boldsymbol{\hat{\theta}_g = \frac{\sum_{i=1}^{n_g} y_{gi}}{n_g} = \bar{y}_g, g = 1, 2, 3, 4}$$

1 mark for correct differentiation and 1 mark for solutions

(b) (3 marks) How are θ_1, θ_2, θ_3, *and* θ_4 and related to the parameters of the model fit in question 1?

The model fit in question 1 is

$$\boldsymbol{Y = \beta_0 + \beta_1 I_{litterAB} + \in}$$

where *I*$_{\text{litterAB}}$ is 1 if the litter genotype is A or B and 0 otherwise. Hence, equating the expectations of *Y*gi to the β's, we have

$$\boldsymbol{\beta_0 + \beta_1 = \theta_1 = \theta_2}$$
$$\boldsymbol{\beta_0 = \theta_3 = \theta_4}$$

1.5 mark each for each equation

Question 3

In this question, we will work with the same data as in question 1, keeping the classification of litter genotype as AB or IJ. Further, we will include genotype of the foster mother (`motgen`) with types A, B, I and J.

Some edited SAS output from 2 models is given below and on the next page. Some numbers have been replaced by X's.

MODEL 1

```
                              The GLM Procedure
                           Class Level Information
                       Class         Levels    Values
                       litter             2    AB IJ
                       motgen             4    A B I J

                    Number of Observations Read          60
                    Number of Observations Used          60

Dependent Variable: weight
                                           Sum of
Source                      DF          Squares    Mean Square   F Value   Pr > F
Model                        7       974.619452     139.231350      2.36   0.0359
Error                       52      3067.869048      58.997482
Corrected Total             59      4042.488500

              R-Square     Coeff Var      Root MSE    weight Mean
              0.241094      14.26499      7.680982       53.84500

Source                      DF      Type III SS    Mean Square   F Value   Pr > F
litter                       1       19.2961063     19.2961063      0.33   0.5699
motgen                       3      769.5066216    256.5022072      4.35   0.0083
litter*motgen                X      XXXXXXXXXXX     XXXXXXXXXX      XXXX   0.3483

                                                Standard
 Parameter                   Estimate              Error    t Value    Pr > |t|
 Intercept                49.20000000 B       2.71563716      18.12      <.0001
 litter        AB         -1.11428571 B       3.97528576      -0.28      0.7804
 litter        IJ          0.00000000 B        .               .          .
 motgen        A           2.04285714 B       3.97528576       0.51      0.6095
 motgen        B          11.03333333 B       4.14820429       2.66      0.0104
 motgen        I           3.50000000 B       3.84049091       0.91      0.3663
 motgen        J           0.00000000 B        .               .          .
 litter*motgen AB A        8.14642857 B       5.62190304       1.45      0.1533
 litter*motgen AB B       -1.56904762 B       5.74547611      -0.27      0.7859
 litter*motgen AB I        2.43928571 B       5.52741054       0.44      0.6608
 litter*motgen AB J        0.00000000 B        .               .          .
 litter*motgen IJ A        0.00000000 B        .               .          .
 litter*motgen IJ B        0.00000000 B        .               .          .
 litter*motgen IJ I        0.00000000 B        .               .          .
 litter*motgen IJ J        0.00000000 B        .               .          .
 NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.
```

(SAS output for this question continues on the next page)

MODEL 2

(Initial output that is the same as for MODEL 1 has been deleted)

```
Dependent Variable: weight
                                    Sum of
Source                     DF        Squares    Mean Square   F Value   Pr > F
Model                       4     775.879358     193.969839      3.27   0.0179
Error                      55    3266.609142      59.392893
Corrected Total            59    4042.488500

              R-Square     Coeff Var      Root MSE    weight Mean
              0.191931      14.31271      7.706678       53.84500

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
litter                      1     22.2216910     22.2216910      0.37   0.5433
motgen                      3    730.0482992    243.3494331      4.10   0.0107

                                              Standard
    Parameter            Estimate                Error    t Value    Pr > |t|
    Intercept         48.11000426 B       2.19724550      21.90      <.0001
    litter    AB       1.22141944 B       1.99684115       0.61      0.5433
    litter    IJ       0.00000000 B        .                .         .
    motgen    A        6.23190537 B       2.81722809       2.21      0.0311
    motgen    B        9.89204177 B       2.87152180       3.44      0.0011
    motgen    I        4.64178602 B       2.77056176       1.68      0.0995
    motgen    J        0.00000000 B        .                .         .
```

(Questions based on the foregoing output begin here and continue on the next pages)

(a) (4 marks) Write the model that is being estimated in the output labeled MODEL 1; clearly define all variables.

$$\boldsymbol{Y_i = \beta_0 + \beta_1 I_{litterAB,i} + \beta_2 I_{motgenA,i} + \beta_3 I_{motgenB,i} + \beta_4 I_{motgenI,i}}$$
$$\boldsymbol{+\beta_5 I_{litterAB,i} I_{motgenA,i} + \beta_6 I_{litterAB,i} I_{motgenB,i} + \beta_7 I_{litterAB,i} I_{motgenI,i} + \in_i}$$

where *Y* is the weight of litter,
$I_{litterAB}$=1 if litter has genotype A or B and 0 otherwise,
$I_{motgenA}$=1 if litter has foster mother with genotype A and 0 otherwise,
$I_{motgenB}$=1 if litter has foster mother with genotype B and 0 otherwise,
$I_{motgenI}$=1 if litter has foster mother with genotype I and 0 otherwise, and
ϵ is random error

1 mark for correct model and 0.5 mark for each correct definition of variable.

(b) For the test in MODEL 1 with *p*-value 0.3483,

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

Ho: β_5= β_6 =β_7=0
Ha: at least one of β_5, β_6, and β_7 is non-zero

0.5 mark each

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

There is no evidence that differences in mean weight among the 4 mother genotypes differ with litter genotype.

iii. (4 marks) What are the **first 3** missing numbers (replaced by X's in MODEL 1)?

DF= 3

Type III SS =974.619-775.879=198.74

Mean Square = 198.74/3=66.25

1 mark for DF and MS, and 2 marks for SS.

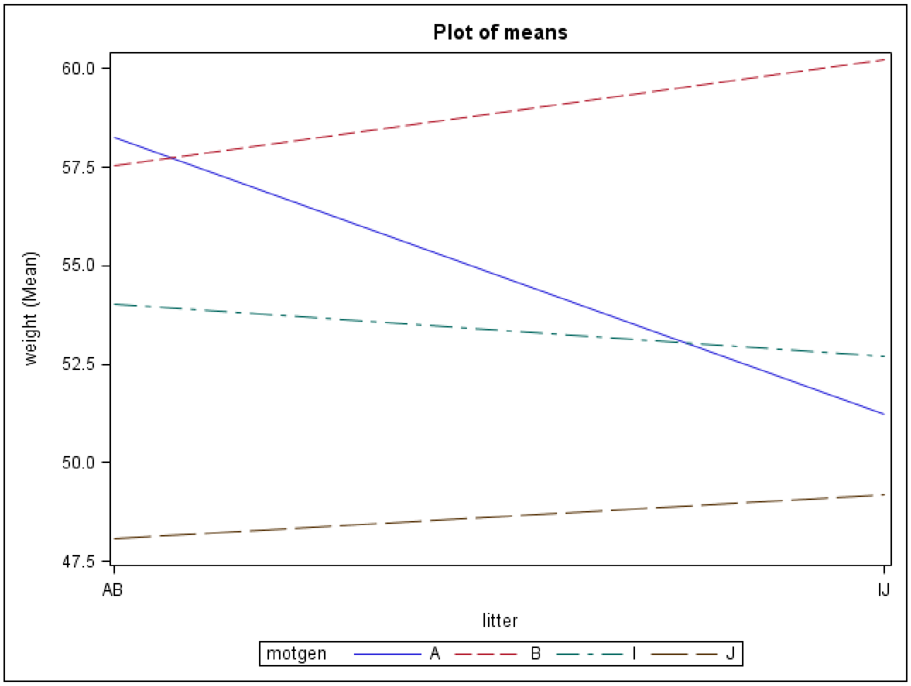
(c) (1 mark) For MODEL 2, what practical quantity, if any, is being estimated by the estimate of the intercept?

The mean weight of litter with genotype I or J and foster mother with genotype J.

(d) (1 mark) For MODEL 2, estimate the mean weight for a pup with litter genotype IJ and whose foster mother genotype is B.

Estimate is: $\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_3$ = 48.11+9.89 = 58 grams

(e) (6 marks) Here is a plot showing the mean weight of a pup for each litter genotype combination (AB are on the left and IJ are on the right) with separate lines for each mother genotype. The dashed top line is for B, the (mostly) second solid line is for A, the third dot-dashed line is for I and the bottom long dashed line is for J.



Explain how this interaction plot is consistent with the conclusions that can be drawn from inferences about the fitted models above. Support your answer with relevant numbers from the SAS output.

Inference	Plot	Relevant p-value
There is no evidence of interaction.	**Most lines are close to parallel**	**0.3483**
There is some evidence of differences based on mother genotype.	**Line for mother genotype J is lower than lines for others**	**0.0107**
There is no evidence of a difference between litter genotype groups	**Most lines are horizontal**	**0.5433**

2 marks for each correct row above

Question 4

In this question we will work with the same data as in question 1. However, here we will dichotomize weight as either above average of 54 grams or not; the new related variable is ‘above’=1 or ‘above’=0 respectively. Then we model how litter genotype and foster mother genotype can be used to predict the odds of a rat pup attaining an above average weight. Some edited SAS output from the model is given below. Some numbers have been replaced by X’s.

```
                  The LOGISTIC Procedure
                  Model Information

    Data Set                      WORK.RAT2
    Response Variable             above
    Number of Response Levels     2
    Model                         binary logit
    Optimization Technique        Fisher's scoring

        Number of Observations Read          60
        Number of Observations Used          60

                    Response Profile
           Ordered                     Total
             Value        above     Frequency
                 1            1            29
                 2            0            31

            Probability modeled is above=1.

                Class Level Information
         Class      Value     Design Variables
         motgen     A          1      0      0
                    B          0      1      0
                    I          0      0      1
                    J          0      0      0
         litter     AB         1
                    IJ         0

               Model Convergence Status
    Convergence criterion (GCONV=1E-8) satisfied.

                  Model Fit Statistics
                                     Intercept
                      Intercept            and
        Criterion          Only     Covariates
        AIC              85.111         79.156
        SC               87.205         89.628
        -2 Log L         83.111         69.156

          Testing Global Null Hypothesis: BETA=0
Test                     Chi-Square       DF     Pr > ChiSq
Likelihood Ratio            XXXXXXX        X         0.0074
Score                       12.7156        4         0.0128
Wald                        10.1531        4         0.0379
```

(SAS output for this question continues on the next page)

```
                   Type 3 Analysis of Effects
                                         Wald
              Effect      DF    Chi-Square    Pr > ChiSq
              litter       1        XXXXXX        0.7770
              motgen       3       10.1530        0.0173

               Analysis of Maximum Likelihood Estimates
                                     Standard         Wald
Parameter       DF    Estimate        Error    Chi-Square    Pr > ChiSq
Intercept        1     -1.7969       0.8017        5.0244        0.0250
litter    AB     1     -0.1658       0.5853        XXXXXX         0.7770
motgen    A      1      2.0191       0.9215        4.8014        0.0284
motgen    B      1      3.1929       1.0053       10.0870        0.0015
motgen    I      1      1.8798       0.9106        4.2617        0.0390

                         Odds Ratio Estimates
                                Point          95% Wald
          Effect             Estimate      Confidence Limits
          litter AB vs IJ       0.847       0.269       2.668
          motgen A vs J         7.532       XXXXX      XXXXXX
          motgen B vs J        24.359       3.396     174.734
          motgen I vs J         6.552       1.100      39.039
```

(a) (4 marks) Does litter genotype have any effect on the odds of being above average in weight? In your answer, include the appropriate null and alternative hypotheses, test statistic and its distribution under the null hypothesis, *p*-value and conclusion.

Hypotheses: *1 mark*

Ho: $\beta_1 = 0$
Ha: $\beta_1 \neq 0$

Test Statistic: **(Wald Chi-square with 1 df)** *1.5 marks*

$$\boldsymbol{Z^2 = \left(\frac{-0.1658}{0.5853}\right)^2 = 0.0802}$$

***p*-value** = 0.7770 *0.5 mark*

Conclusion: *1 mark*

Since the *p*-value is large, we fail to reject Ho and conclude that there is no evidence that litter genotype has an effect on the odds of being above average, over and above foster mother genotype of a litter.

(b) (3 marks) Compare the effect of mother genotype A to the effect of mother genotype J on the odds of being above average? Explain. *(Hint: your answer should include a 95% confidence interval.)*

The coefficient of mother genotype A, $\widehat{\beta}_2$ is 2.0191, hence the odds ratio of mother genotype A to J is exp(2.0191)=7.532.

1 mark for some version of these results

A 95% confidence interval for $\widehat{\beta}_2$ is calculated using the formula $\widehat{\beta}_2 \pm Z_{\propto/2} S.E.(\widehat{\beta}_2)$ which results in (0.213, 3.825).

1 mark for correct CI

Finally, the odds of being above average for a litter with foster mother whose genotype is A are 7.53 times the odds of a litter with foster mother of genotype J. The 95% CI is (exp(0.213), exp(3.825)) = (1.237, 45.83).

1 mark for correct final statement

(c) The log-odds of being above average were estimated to be

$$-1.8 - 0.17 I_{\text{litterAB}} + 2.02 I_{\text{motgenA}} + 3.2 I_{\text{motgenB}} + 1.88 I_{\text{motgenI}},$$

i. (2 marks) From the model above, what is the estimate of the probability of being above average in weight for a rat pup with type A genotype whose foster mother has type I genotype?

$$\hat{\pi} = \frac{\exp(-1.88 - 0.17 + 1.88)}{1 + \exp(-1.88 - 0.17 + 1.88)} = \frac{\exp(-0.09)}{1 + \exp(-0.09)} = 0.48$$

1 mark for correct formula and 1 mark for correct calculation

ii. (2 marks) What would be the estimated equation for the log-odds of attaining equal or less than average weight, if the indicator variable for litter were 1 for litters IJ and 0 for litters AB?

$$logit(\pi_i) = 1.97 + 0.17 I_{\text{litterAB}} - 2.02 I_{\text{motgenA}} - 3.2 I_{\text{motgenB}} - 1.88 I_{\text{motgenI}}$$

1 mark for correct form of equation and 1 mark for correct values and signs

(d) (4 marks) Is there evidence that this model is adequate? In your answer, include the appropriate null and alternative hypotheses, test statistic and its distribution under the null hypothesis, *p*-value and conclusion.

Hypotheses: *1 mark*

Ho: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

Ha: at least one of β_1, β_2, β_3, β_4 and β_5 is non-zero

Test Statistic: (follows Chi-square distribution with 3 df) *1.5 marks*

G^2=83.111-69.156 = 13.955

***p*-value** = 0.0074 *0.5 mark*

Conclusion: *1 mark*

Since the p-value is small, we have strong evidence against Ho. We conclude that the fitted model is adequate, that is, at least one of the factors- litter genotype and mother genotype is useful in predicting the odds of being above average.

(e) Compare the two analyses relating to the response variable weight, the first by the general linear additive model (MODEL 2) in Question 3 and the second by the logistic model in this question. State how they are similar or different with respect to,

i. (1 mark) A model assumption.

Similarity:
Both models assume that $g(E(Y))=f(X;\beta)$ is a linear function of the β's.

OR

Difference:
Underlying distribution of the response is Normal for the general linear model versus Bernoulli for the logistic model.
OR
The general linear model assumes that the group variances are the same while the logistic model assumes that the variance of *Y* follows the Bernoulli distribution.

ii. (1 mark) The statistical significance of litter genotype.

We can conclude from both models that litter genotype is not statistically relevant in predicting weight, over and above mother genotype.