Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER II
SESSION 2017/2018

COURSE NAME	:	STATISTICAL MODELLING FOR BIOLOGY
COURSE CODE	:	BWB 42703
PROGRAMME	:	BWQ
EXAMINATION DATE	:	JUNE/JULY 2018
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **ALL** QUESTIONS

THIS QUESTION PAPER CONSISTS OF **EIGHT(8)** PAGES

Q1 (a) Define the regression analysis.

(2 marks)

(b) Describe **THREE (3)** details used of regression analysis and explain **ONE (1)** example of using it.

(5 marks)

(c) Determine **THREE (3)** benefits of using regression analysis.

(6 marks)

(d) Differentiate **THREE (3)** points between linear regression and logistic regression.

(12 marks)

Q2 (a) Interpret the meaning of model specification.

(2 marks)

(b) The need for model selection often begins when a researcher wants to mathematically define the relationship between independent variables and the dependent variable. However, investigators measure many variables but include only some in the model. Analysts try to exclude independent variables that are not related and include only those that have an actual relationship with the dependent variable. Conclude **THREE (3)** effect on the including of the correct number of independent variables in the regression equation if they are:

(i) too few

(2 marks)

(ii) too many

(2 marks)

(iii) just right

(2 marks)

(c) There are **THREE (3)** standard approaches to model specification. Outline and explain each of standard clearly.

(9 marks)

(d) Regression model specification is as much a science as it is an art even statistical method could help, but we will need to place a high weight on theory, specification, simplicity and residual plots. Discuss **TWO (2)** out of four practical recommendations for model specification.

(8 marks)

Q3 (a) (i) Define the ANOVA test and explain **TWO (2)** example of test.
(6 marks)

(ii) Classify the meaning of one-way and two-way ANOVA.
(4 marks)

(iii) Explain the used of both one-way and two-way ANOVA.
(4 marks)

(b) (i) List **THREE (3)** assumptions of two-way ANOVA.
(3 marks)

(ii) The data sets available in R relate to an experiment into plant growth. The purpose of the experiment was to compare the yields on the plants for a control group and two treatments of interest. The response variable was a measurement taken on the dried weight of the plants. The analysis used the factor function to re-define the labels of the group variables that will appear in the output and graphs. All R-codes and output could be seen in **Table Q3(b)(ii)** in **Appendix A**. Explain all the conclusions that could be drawn from the output given.
(8 marks)

Q4 Mixcell Company trainers collects the `sat` from 50 states of the USA to study the possible relationship between SAT test results of pupils and expenditure on public education as well as other variables. The variables `sat` as in **Table 4(a)**. Answer following questions based on R-codes and output in **Table Q4(b)** in **Appendix B.**

Table Q4(a): SAT data of 50 states of the USA

`expend`	Average Expenditure per pupil 1994-95
`ratio`	Pupil/teacher ratio, 1994
`takers`	Average % of all eligible students taking the SAT, 1994-95
`math`	Average math SAT score, 1994-95

(a) Based on scatter plot of `math` against `expend` and a density estimate for `math`, explain on any possible relationship and shape of the distribution.
(4 marks)

(b) Interpret the `summary(mymodel)` and extract the estimates of the parameters β_0, β_1 and σ.
(3 marks)

(c) Perform a two-sided *t*-test of the hypothesis that $\beta_1 = 0$ with the associated *p-value* and explain the value.

(6 marks)

(d) Based on the ANOVA table, test for the overall significance of the regression using an appropriate *F* test, report the *p-value* and interpret that value. Explaining the similar of *p-value* obtained as in **Q4(c)**.

(6 marks)

(e) Determine the value of R^2 and interpret the value imply to the fit of the model

(2 marks)

(f) Reproduce scatter plot of `math` against `expend` in **Q4(a)** superimpose the estimated regression line. Examine this plot in terms of relationship.

(4 marks)

- END OF QUESTIONS -

FINAL EXAMINATION

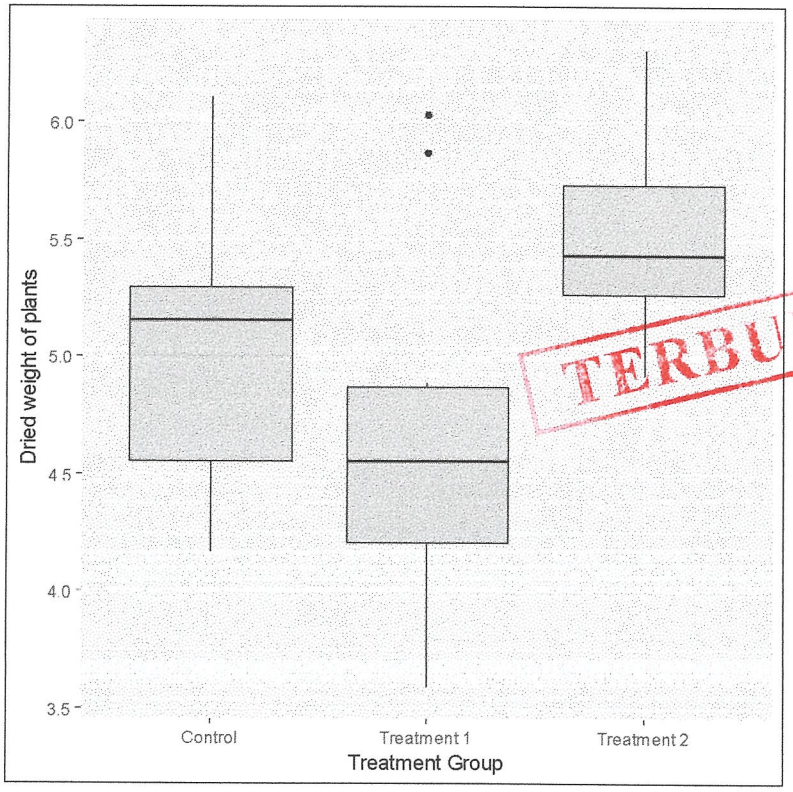
SEMESTER / SESSION : SEM II / 2017/2018
COURSE NAME: STATISTICAL MODELLING FOR BIOLOGY

PROGRAMME CODE : BWQ
COURSE CODE : BWB 42703

APPENDIX A: Q3(b)(ii)

Table Q3(b)(ii): Experiment Plant Growth

```
> plant.df = PlantGrowth
> plant.df$group = factor(plant.df$group,
+   labels = c("Control", "Treatment 1", "Treatment 2"))
> require(ggplot2)
> ggplot(plant.df, aes(x = group, y = weight)) +
+   geom_boxplot(fill = "grey80", colour = "black") +
+   scale_x_discrete() + xlab("Treatment Group") +
+   ylab("Dried weight of plants")
```



FINAL EXAMINATION

SEMESTER / SESSION : SEM II / 2017/2018
COURSE NAME: STATISTICAL MODELLING FOR BIOLOGY

PROGRAMME CODE : BWQ
COURSE CODE : BWB 42703

APPENDIX A: Q3(b)(ii)

```
> plant.mod1 = lm(weight ~ group, data = plant.df)
> summary(plant.mod1)

Call:
lm(formula = weight ~ group, data = plant.df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4180 -0.0060  0.2627  1.3690

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.0320     0.1971  25.527   <2e-16 ***
groupTreatment 1   -0.3710     0.2788  -1.331   0.1944
groupTreatment 2    0.4940     0.2788   1.772   0.0877 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared:  0.2641,    Adjusted R-squared:  0.2096
F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
> anova(plant.mod1)
Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value  Pr(>F)
group      2  3.7663  1.8832  4.8461 0.01591 *
Residuals 27 10.4921  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

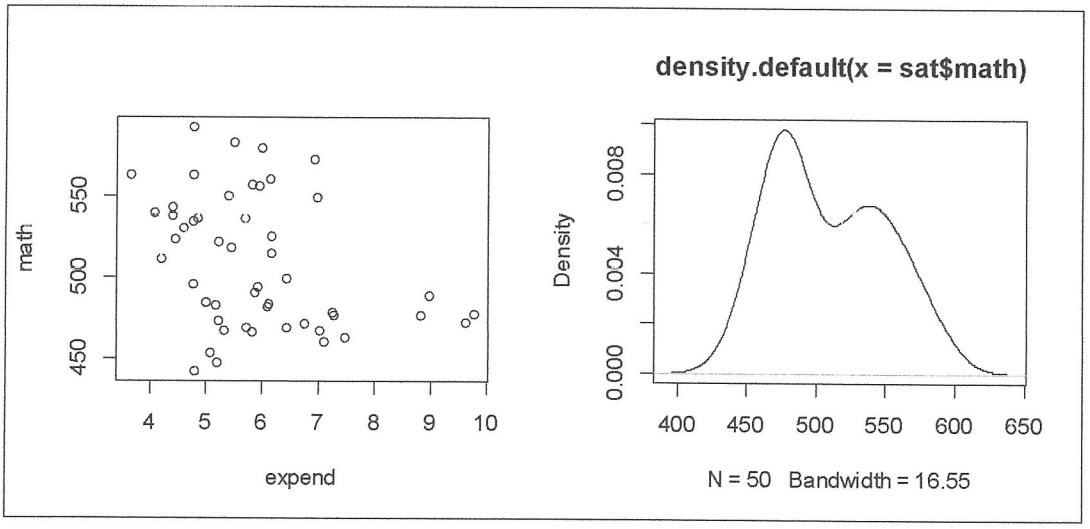
FINAL EXAMINATION

SEMESTER / SESSION : SEM II / 2017/2018
COURSE NAME: STATISTICAL MODELLING FOR BIOLOGY

PROGRAMME CODE : BWQ
COURSE CODE : BWB 42703

APPENDIX B: Q4(b)

```
> par(mfrow=c(1,2))
> plot(math~expend,data=sat)
> plot(density(sat$math))
```




```
> mymodel=lm(math~expend,data=sat)
> summary(mymodel)

Call:
lm(formula = math ~ expend, data = sat)

Residuals:
    Min      1Q  Median      3Q     Max
-77.204 -28.908   0.927  18.779  73.783

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   569.65      24.17  23.571   <2e-16 ***
expend        -10.31       3.99  -2.584   0.0129 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.06 on 48 degrees of freedom
Multiple R-squared:  0.1221,    Adjusted R-squared:  0.1038
F-statistic: 6.675 on 1 and 48 DF,  p-value: 0.01288
```

FINAL EXAMINATION

SEMESTER / SESSION : SEM II / 2017/2018
COURSE NAME: STATISTICAL MODELLING FOR BIOLOGY

PROGRAMME CODE : BWQ
COURSE CODE : BWB 42703

APPENDIX B: Q4(b)

```
> anova(mymodel)
Analysis of Variance Table

Response: math
           Df Sum Sq Mean Sq F value  Pr(>F)
Expend(X)   1   9670  9670.1  6.6753 0.01288 *
Residuals 48  69534  1448.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> plot(math~expend,data=sat)
> abline(mymodel)
> abline(lm(math~expend,data=sat)$coefficients)
```