THE UNIVERSITY of ADELAIDE

Examination in School of Mathematical Sciences
Semester 2, 2020

104843 STATS 2107 Statistical Modelling & Inference II

Total Duration: 190 mins

NUMBER OF QUESTIONS: 6 TOTAL MARKS: 100

Instructions

- Computations must be uploaded to MyUni as high quality images – take a photo and email the scan to your computer and upload the file into the MyUni examination (use of CamScanner and Canvas Student on your mobile device is recommended)
- Show all calculations and assumptions
- If you believe that a parameter or an important piece of information has been inadvertently omitted by the examiner, assume a suitable value, clearly stating it, and continue with the solution.
- It is your responsibility to ensure plenty of time for file scanning and uploading at end. You should upload early and then if you have an updated version of your exam with remaining time, upload again. All uploads are retained, your latest entry will be marked.
- You must review your submissions on MyUni before the end of the exam to check that they have uploaded, are legible, in the correct order, and correctly oriented. If they are not, then you must resubmit.

Materials

- This is an open-book, open-internet examination. You are permitted to access your notes and any material on MyUni. You are also permitted to access websites like Wolfram Alpha, Symbolab, Desmos, etc
- No discussion or consultation about the exam with any third party is permitted during the exam period.
- You may write your answers on paper and submit a scanned PDF. If you have a tablet and stylus, you may write your answers electronically and submit an exported PDF. No extra time allowance will be given for typesetting answers. If using paper:
  - Make sure you have plenty of blank A4 paper to write your answers on
  - Use a dark pen to write (not a pencil)

YOU MAY COMMENCE WRITING ONCE YOU ENTER THE EXAMINATION

1. Suppose $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed (i.i.d.) random variables with probability density function (PDF)

$$f_{Y_i}(y) = \frac{1}{\theta^2} y e^{-\frac{y}{\theta}}, \quad \text{for } y > 0,$$

where $\theta > 0$ is an unknown parameter. Let $\hat{\theta}_1 = Y_n$ and $\hat{\theta}_2 = \frac{1}{2}\bar{Y}$ be estimators of θ.

(a) Find the bias of $\hat{\theta}_1$ and $\hat{\theta}_2$.
Hint: Identify the distribution of Y_i.

(b) Find the mean square error (MSE) of $\hat{\theta}_1$ and $\hat{\theta}_2$.

(c) Which estimator, $\hat{\theta}_1$ or $\hat{\theta}_2$, is preferred by MSE? Justify your answer.

[10 marks]

2. Car crash tests are performed to assess the safety ratings of vehicles. Manufacturers of new cars are to send a few samples to an authorised agent for testing. Let p be the proportion of a certain model of car failing the crash test. Porsche had designed a new model and decided that in order to proceed with production, it needs p to be less than 0.05 for this car model. To test the hypothesis $p = 0.05$ versus $H_a : p > 0.05$, the following procedure was used.

i) Two cars of the new model were randomly selected from the production line.

ii) If both of them failed the crash test, then the null hypothesis is rejected.

iii) If only one of them failed the test, then a third car is randomly selected. If this third car failed the crash test, then the null hypothesis rejected.

iv) In all other cases, we do not reject the null hypothesis.

(a) Give the definition of Type I and Type II errors. Explain what is meant by committing a Type I error and committing a Type II error in the context of this problem.

(b) Find the probability of Type I error.
Hint: What is the distribution of the number of cars in the sample failing the crash test?

(c) Suppose the true proportion of cars that fails the crash test is p^*. Find the probability of Type II error.

(d) Find the power function of this test.

(e) Show that the value of the power function at $p = 0.05$ is the same as the value found in part (b).

[20 marks]

Please turn over for page 3

3. Suppose Y is a random variable with PDF

$$f_Y(y) = 2\left(\frac{\theta - y}{\theta^2}\right) \quad \text{for } 0 < y < \theta.$$

where θ is an unknown parameter.

(a) Find the distribution of $X = \frac{Y}{\theta}$ and explain why it is a pivotal quantity for θ. *Hint:* Use the change of variable formula.

(b) Find the 90% upper confidence interval for θ based on X, where the upper limit U is such that $P(\theta \leq U) = 0.90$.

(c) Give the 90% upper confidence interval for θ when $y = 0.6$.

[10 marks]

4. An analysis was performed to investigate whether ice-cream consumption is related to the price, the temperature of the day, and the income of the consumer. A multiple linear regression model was fitted to the data using R and the following output was obtained. Note that some of the entries in the R output are missing and you will be asked to calculate these in the question.

```
##
## Call:
## lm(formula = Consumption ~ Price + Temperature + Income, data = ic2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.115062 -0.025694  0.003299  0.022995  0.128025
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3462222  0.1166343   2.968  0.00341 **
## Price        ?????????  0.3533218  -4.840 2.83e-06 ***
## Temperature  0.0033775  0.0002239  15.083  < 2e-16 ***
## Income       0.0037759  0.0005619   ????? 2.43e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04151 on 176 degrees of freedom
## Multiple R-squared:  0.652,  Adjusted R-squared:  0.646
## F-statistic: 109.9 on 3 and 176 DF,  p-value: < 2.2e-16
```

(a) What was the size of the sample?

(b) What is the estimate of the coefficient of Price, β_1?

(c) What is the value of the t-statistic for $\hat{\beta}_3$?

(d) Perform a statistical test to determine if there is a significant (linear) relationship between Consumption and the three predictors (Price, Temperature, and Income) at the 5% significance level. State the null and alternative hypotheses, the value of the test statistic, the P-value, and your conclusion.

(e) Calculate the 95% confidence interval for β_2, the coefficient for Temperature, taking into account the other predictor variables. You may use R to calculate the appropriate critical value.

[15 marks]

5. A study was conducted to investigate the growth characteristics of blue mussel in New South Wales. Samples were taken from two distinct locations. Their age (in weeks) and weight (in grams) were recorded. An analysis was performed using R. The commands and outputs are given in Appendix A.

(a) Consider the scatterplot of weight against age given in Figure 1. Describe the relationship.

(b) Four observations from the data are:

`Weight`	Age	Location
0.44	3	1
0.50	3	1
1.76	3	2
2.38	4	2

Write down the design matrix for fitting a separate regression model of `Weight` on `Age` and `Location`, for these these four observations only.

(c) Consider the separate regression model. Write down the two lines of best fit for the relationship between weight and age: one for Location 1 and one for Location 2.

(d) Based on the separate regression model, estimate the weight of an 8 weeks old mussel at Location 2.

(e) Test for a statistically significant interaction term in the separate regression model at the 5% significance level. Remember to include the null and alternative hypotheses, the value of the test statistic, the P-value, and your conclusion.

(f) Using the Bayesian Information Criterion which model fits the data best? Justify your answer.

(g) Assess the assumptions of the linear model used in the parallel model. The plots given in Figure 2 may be used where appropriate.

[20 marks]

Please turn over for page 5

6. Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. random variables with PDF

$$f_{Y_i}(y;\theta) = \frac{2y}{\theta} e^{-\frac{1}{\theta}y^2}, \quad y > 0.$$

where $\theta > 0$ is an unknown parameter. The mean and variance of Y_i is given by

$$E[Y_i] = \sqrt{\frac{\pi\theta}{4}} \quad \text{and} \quad \text{var}(Y_i) = \frac{\theta(4-\pi)}{4},$$

respectively, for $i = 1, 2, \ldots, n$.

(a) Write down the likelihood function.

(b) Write down the log-likelihood function.

(c) Write down the score function.

(d) Find the maximum likelihood estimate $\hat{\theta}$ of θ.

(e) Find the Fisher information. You may assume that the usual regularity conditions hold.

(f) State the asymptotic distribution of $\hat{\theta}$ and hence give a $100(1-\alpha)\%$ confidence interval for θ.

(g) Suppose we want to test $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$.

(i) State the score test statistic and its asymptotic distribution.

(ii) State the test statistic and the critical region of the likelihood ratio test.

(h) An alternative form of the distribution is

$$f_{Y_i}(y;\sigma^2) = \frac{y}{\sigma^2} e^{-\frac{y^2}{2\sigma^2}}, \quad y > 0.$$

Give an expression for the maximum likelihood estimate of σ^2.

[25 marks]

Appendix A

Load the data

```
library(tidyverse)
mussel <- read.csv("mussel.csv", header=T)
mussel$Location <- factor(mussel$Location)
```

Visualise data

```
    ggplot(mussel, aes(x = Age, y=Weight, col = Location,
    shape = Location, fill=Location)) +
    geom_point(size=3, alpha=0.5) +
    geom_smooth(method = "lm", col = "black",aes(linetype=Location)) +
    scale_linetype_manual(values=c("1"="solid", "2"="dashed")) +
    labs(x = "Age (weeks)", y = "Weight (grams)", col="Location",
        shape="Location", fill="Location", linetype="Location") +
    theme(legend.position = "top")
```

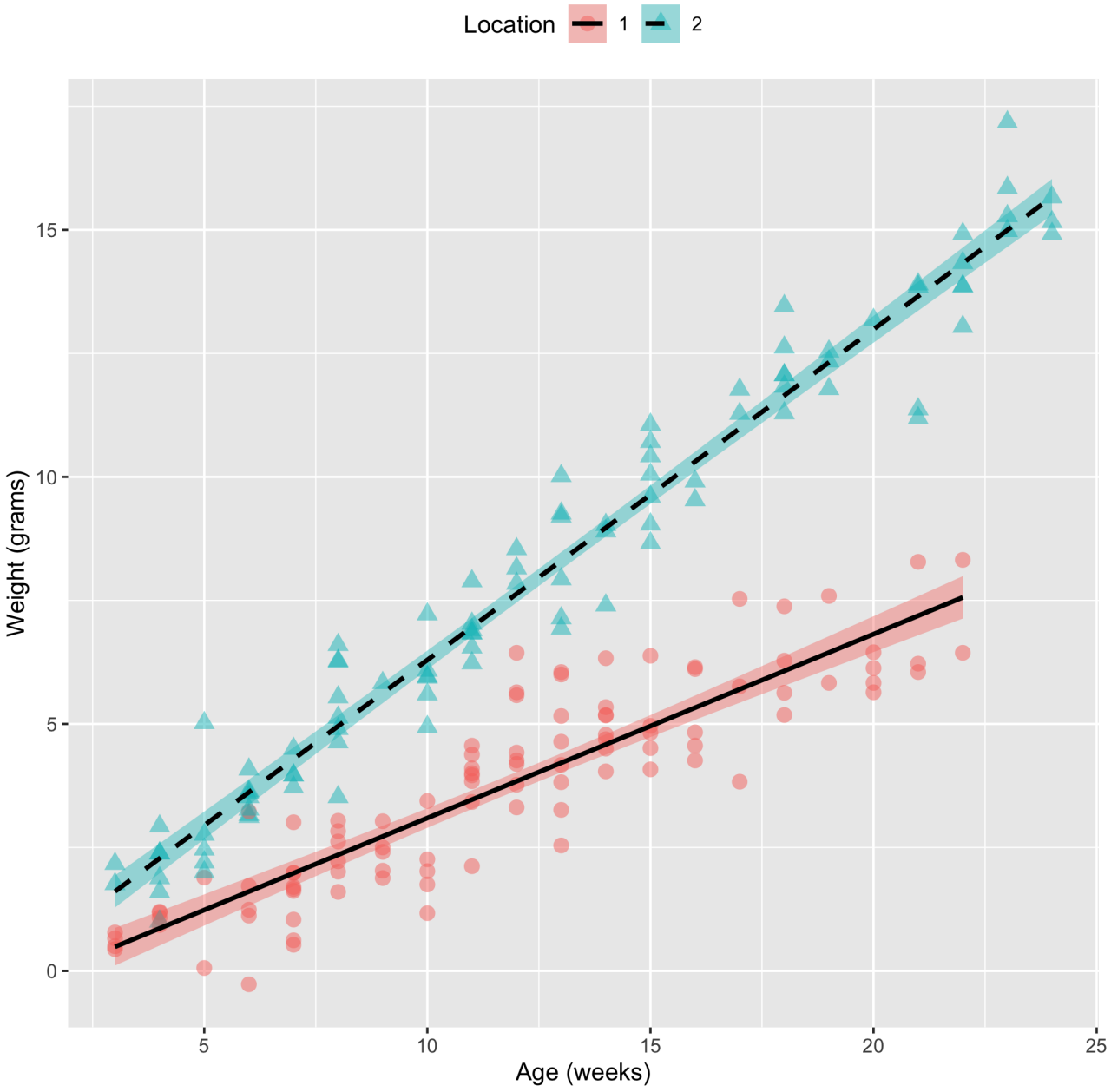

Figure 1: Scatterplot of Weight against Age for the mussel dataset. Colour and shape of points indicates Location.

Please turn over for page 8

Fit models

```
## Identical regression model
identical.model <- lm(Weight ~ Age, data = mussel)
summary(identical.model)
```

```
##
## Call:
## lm(formula = Weight ~ Age, data = mussel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0800 -1.7364  0.1562  1.8067  5.0823
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.19040    0.39124  -3.043  0.00267 **
## Age          0.57775    0.02914  19.830  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.269 on 193 degrees of freedom
## Multiple R-squared:  0.6708, Adjusted R-squared:  0.6691
## F-statistic: 393.2 on 1 and 193 DF,  p-value: < 2.2e-16
```

```
## Parallel regression model ----
parallel.model <- lm(Weight ~ Age + Location, data = mussel)
summary(parallel.model)
```

```
##
## Call:
## lm(formula = Weight ~ Age + Location, data = mussel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9364 -0.8978  0.0312  0.8541  3.4195
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.71453    0.22019  -12.33   <2e-16 ***
## Age          0.54959    0.01562   35.19   <2e-16 ***
## Location2    3.83452    0.17426   22.00   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.212 on 192 degrees of freedom
## Multiple R-squared:  0.9065, Adjusted R-squared:  0.9055
## F-statistic:   931 on 2 and 192 DF,  p-value: < 2.2e-16
```

```
## Separate regression model ----
separate.model <- lm(Weight ~ Age * Location, data = mussel)
summary(separate.model)
```

```
##
## Call:
## lm(formula = Weight ~ Age * Location, data = mussel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.46706 -0.51737 -0.00505  0.52470  2.60030
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.62660    0.23265  -2.693   0.0077 **
## Age            0.37219    0.01823  20.413   <2e-16 ***
## Location2      0.22773    0.31421   0.725   0.4695
## Age:Location2  0.29714    0.02360  12.592   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8984 on 191 degrees of freedom
## Multiple R-squared:  0.9489, Adjusted R-squared:  0.9481
## F-statistic:  1183 on 3 and 191 DF,  p-value: < 2.2e-16
```

Model Selection

```
BIC(identical.model, parallel.model, separate.model)
```

```
##                 df      BIC
## identical.model  3 886.7381
## parallel.model   4 646.5005
## separate.model   5 533.9220
```

Please turn over for page 10

Assumption checking

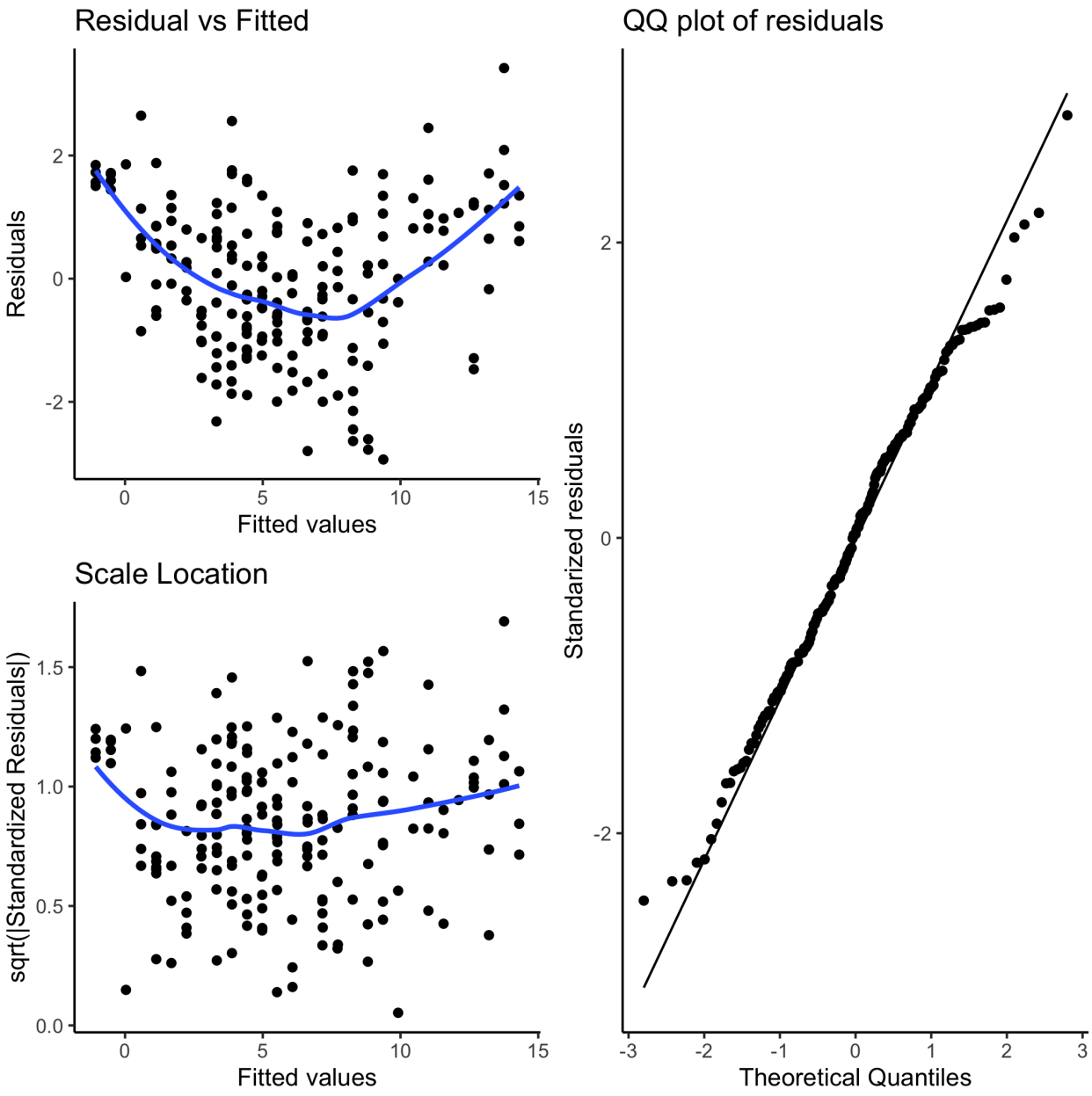

Figure 2: Assumption plots of the separate model for the mussel dataset.

Please turn over for page 12

Appendix B

Distribution	Probability mass function / probability density function	Expectation	Variance
Binomial	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$	np	$np(1-p)$
Geometric	$p(x) = p(1-p)^{x-1}$ for $x = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$	λ	λ
Uniform	$f(x) = \frac{1}{b-a}$ for $a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$f(x) = \lambda e^{-\lambda x}$ for $x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ for $x > 0$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Normal	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2\sigma^2)(x-\mu)^2}$ for $-\infty < x < \infty$	μ	σ^2
Beta	$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $0 < x < 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$