THE UNIVERSITY OF MELBOURNE

Student Number

Semester 1 Assessment, 2018

School of Mathematics and Statistics

MAST90044 Thinking and Reasoning with Data

Writing time: 2 hours

Reading time: 15 minutes

This is NOT an open book exam

This paper consists of 11 pages (including this page)

Authorised Materials

- Mobile phones, smart watches and internet or communication devices are forbidden.

Instructions to Students

- You must NOT remove this question paper at the conclusion of the examination.
- Hand-held electronic calculators may be used.
- A single A4 sheet of notes (handwritten or typed, both sides) may be used.
- You should attempt all questions.
- For Question 2 onwards, show all your work, as marks will be awarded for correct work as well as correct answer.
- There are 6 questions with marks as shown. The total number of marks available is 82.

Instructions to Invigilators

- Students must NOT remove this question paper at the conclusion of the examination.
- Confiscate all magic wands.

Blank page (ignored in page numbering)

Question 1 (20 marks)

Multiple choice: In your script book give answers (as a letter A, B, C, D or E only) to each of the 10 questions. There is no need to show any working.

(a) In a statistics course, a linear regression equation was computed to predict the final exam score (y) from the score on an assignment (x). The equation of the least-squares regression line was

$$y = 5 + .9\,x$$

Suppose that $(x_0, 50)$ is a point on this line. What is the value of x_0?

A. 50

B. 45

C. 55

D. 60

E. None of the above.

A

(b) For the same linear regression problem in (a): Suppose Jane scored 75 on the assignment and 80 on the final exam. What would be the residual corresponding to this value?

A. 5

B. 0

C. -2.5

D. 7.5

E. None of the above

D

(c) A symmetric 95% confidence interval for the mean reading achievement score for a population of third grade students is $(41.1, 53.1)$. Suppose you compute a symmetric 99% confidence interval. Which of the following statements is correct?

A. the intervals are the same

B. the 95% interval is a subset of the 99% interval

C. the 99% interval is a subset of the 95% interval

D. neither interval need contain the other

E. you cannot determine the answer without knowing n.

B

(d) Which **one** of the following is likely to have a binomial distribution?

A. the number of accidents in a large factory during one 8-hour shift

B. the number of spades in a bridge hand (i.e. in a random selection of 13 cards from a pack of 52 cards)

C. the number of tosses of a fair coin until the 10th head is obtained

D. the number of years between floods at a certain location

E. the number of beetles that are killed when a random sample of 40 beetles is subjected to a specified dose of an insecticide.

E

(e) Suppose that the heights $X_1, \ldots, X_{20}$ of 20 random men and the heights $X'_1, \ldots, X'_{20}$ of 20 random women are both normally distributed, and denote the sample means by $\bar{X}_{20}$ and $\bar{X}'_{20}$, and the sample variances by $\hat{\sigma}^2$, and $\hat{\sigma}'^2$. Which of the following has a t distribution?

A. $\bar{X}_{20}$

B. $\hat{\sigma}^2$

C. $\bar{X}_{20} - \bar{X}'_{20}$

D. $\hat{\sigma}^2/\hat{\sigma}'^2$

E. None of the above.

E

(f) A random variable X has mean μ and standard deviation σ . Suppose n independent observations $X_1, \ldots, X_n$ with the same distribution are taken and the mean $\bar{X}_n$ of these n observations is calculated. We can assert that, if n is very large, $\bar{X}_n$ is approximately normal. This assertion follows from:

A. the law of large numbers

B. the law of rare events

C. the definition of sampling distribution

D. the central limit theorem

E. Pythagoras's Theorem

D

(g) In an opinion poll, 15% of 200 people sampled at random from a large population said that they were strongly opposed to euthanasia. A 95% Wald confidence interval for the population proportion who are opposed to euthanasia is:

A. $(0.025, 0.975)$

B. $(0.1, 0.2)$

C. $(0.25, 0.35)$

D. $(0.081, 0.219)$

E. $(\text{Wald} - .95, \text{Wald} + .95)$.

B

(h) Of the following statements about P-values, which one is *false*?

A. The P-value is the probability of a Type I error.

B. A large P-value does not prove that the null hypothesis is true.

C. In general, a small P-value is evidence against the null hypothesis.

D. The P-value is the probability of observing a value of the test statistic at least as extreme as the one observed, assuming that the null hypothesis is true.

E. The smaller the P-value, the more statistically significant the result.

A

(i) Suppose that we fit a linear regression model with 2 explanatory variables x_1 and x_2 (each of which is a factor with 4 levels) and the interaction between them. This is equivalent to fitting a linear regression model with a single explanatory variable (a factor) with how many levels?

A. 1

B. 2

C. 4

D. 8

E. 16

E

(j) What is the value of the 2nd quartile of a normal distribution having mean 0 and variance 1?

A. 0

B. 1

C. 0.25

D. 0.5

E. 1.96

A

Question 2 (9 marks) Choose *three* of the following five concepts, and explain the meaning of each. For each concept you choose, write a few sentences. Use a diagram or plot if it helps the explanation.

(a) P-value;

(b) Correlation coefficient;

(c) The binomial distribution;

(d) Testing the assumptions of linear models;

(e) Using AIC in model selection.

Question 3 (20 marks) Suppose that researchers follow 760 people for 1 week to see if coffee drinking increases the risk of insomnia (sleeping difficulties). Of the 760 participants, 540 are regular coffee drinkers and 220 do not drink coffee. Participants were asked whether they experienced difficulty sleeping during that week. 81 of the regular coffee drinkers and 27 of those who do not drink coffee said that they experienced difficulty sleeping.

(a) Describe some possible problems with this study. **Different types of coffee (decaf?), is regular defined properly, self reporting of sleeping difficulty (subjective)**

(b) Construct a contingency table for the relationship between coffee drinking and insomnia

	diff	**no diff**	
coffee	**81**	**459**	**540**
no coffee	**27**	**193**	**220**
	108	**652**	**760**

(c) Among those who did not drink coffee, what percentage had difficulty sleeping? **12.3%**

(d) Among those who regularly drink coffee, what percentage had difficulty sleeping? **15%**

(e) Find the odds ratio for the odds of insomnia during the week for regular coffee drinkers versus non-drinkers. **1.29**

(f) What is (in words) the sensible null hypothesis for this study? **There is no difference in sleep difficulty between coffee drinkers and no coffee drinkers.**

(g) Verify that the test statistic for the association between regular coffee drinking and insomnia is 0.95.

	diff	**no diff**	
coffee	**76.73684**	**463.2632**	**540**
no coffee	**31.26316**	**188.7368**	**220**
	108	**652**	**760**

$X^2 = \frac{(76.73684-81)^2}{76.73684} + \frac{(463.2632-459)^2}{463.2632} + \frac{(31.26316-27)^2}{31.26316} + \frac{(188.7368-193)^2}{188.7368} = 0.9537123$

(h) What is the approximate distribution of the test statistic in (g)? χ^2 **with 1 degree of freedom.**

(i) Use the output below to test the hypothesis from (f), and make a conclusion. Note that the `df` argument is blank: what should it have been?

```
> 1 - pchisq(0.9537089, df =  )

[1] 0.328777
```

Df = 1 and the difference is not statistically significant.

(j) Based on these results, can we conclude that regular coffee drinking causes insomnia? Why or why not? **No. No evidence of a relationship, but even if it were one you can't conclude causality.**

Question 4 (14 marks) To test the effectiveness of certain treatments on a variable y (the value of y is expected to be higher among diabetics), the following experiment was conducted with 120 participants of whom 65 were diabetic:

One third of all the participants were chosen at random to receive treatment 1, half of those who did not receive treatment 1 were chosen at random to receive treatment 2, and the remaining 40 participants received a control treatment. Variable y was measured for each participant 2 hours after receiving their treatment. Upon entering the data into R and using standard functions, the following R output is obtained:

```
> test=lm(y~treatment+diabetic)
> summary(test)
Call:  lm(formula = y ~ treatment + diabetic)

Residuals:
  Min      1Q  Median      3Q     Max
-41.085 -15.108   0.263  13.097  48.117

Coefficients:
              Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)   108.102    3.536        30.574    <2e-16 ***
treatment1    -7.174     4.376        -1.639     0.1039
treatment2    3.407      --A--         0.773      0.4409
diabeticYes   --B--       3.611        1.960      0.0524 .
---
  Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 19.55 on --C-- degrees of freedom
Multiple R-squared:  0.08376,Adjusted R-squared:  0.06006
F-statistic: 3.535 on 3 and --C-- DF,  p-value: 0.01701

> anova(test)
Analysis of Variance Table

Response: y
            Df Sum Sq  Mean Sq  F value  Pr(>F)
treatment   2   2586  1292.89   3.3813  0.03739 *
diabetic    1   1469  1468.99   3.8418  0.05239 .
Residuals 116   44355  382.37
---
  Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

(a) Describe the form of the linear model being considered here. $y_{i,j,k} = t_j + d_k + e_i$.

(b) What proportion of the variability in the response variable is explained by this model? Give a possible reason for this number. **R-squared, 0.084.**

(c) Find the values of A, B, C.
A=4.407, B=7.077, C=116

(d) Interpret the p-values for treatments in the linear regression model above. **There is little to no evidence of a difference between treatment1 and the baseline treatment. Likewise for treatment2 and the baseline.**

(e) Interpret the p-value for treatments in the ANOVA above. **There is evidence of a difference between treatments.**

(f) Compare the above two answers, explaining any similarities/differences. **These two answers are not contradictory because the estimated difference of largest magnitude is between treatment 2 and treatment 1.**

(g) Was this a sensible experimental design (in order to identify the effect of treatment on y)? If yes, explain why. If no, describe a better design for this experiment. **It makes sense to do blocking for diabetes. Even better would be to also measure the change in y after and before treatment (this essentially takes out individual effects) and you might want to include interaction effects.**

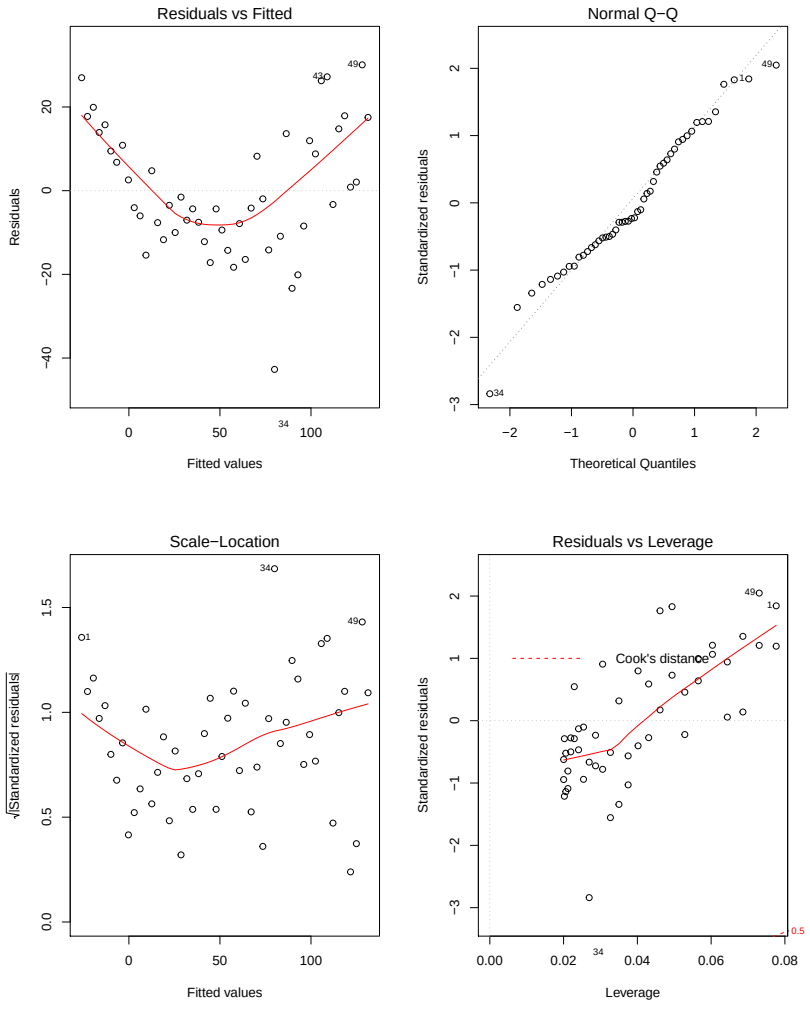
Question 5 (3 marks) A linear regression model was fitted to try to explain the effect of a variable x on a variable y. Based on the R output below: Is this model a good fit to the data? Why or why not?

```
lm(formula = y ~ x)

Residuals:
  Min      1Q  Median      3Q     Max
-42.716  -9.856  -3.404  11.668  30.069

Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   -29.150      4.380    -6.656    2.48e-08 ***
  x            32.100      1.495     21.476   < 2e-16 ***
  ---
  Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 15.25 on 48 degrees of freedom
Multiple R-squared:  0.9057,Adjusted R-squared:  0.9038
F-statistic: 461.2 on 1 and 48 DF,  p-value: < 2.2e-16
```



It is NOT. The R-squared looks good but the diagnostic plots clearly show some relationship was missed by the model. The Residuals vs. Fitted values plot shows it very very clearly.

Question 6 (16 marks) Recall the coronary heart disease data from assignment 3, where we wish to model coronary incidents for 200 males (26 of whom had coronary incidents) of varying ages (between 23 and 70) via logistic regression, and the variables in the data frame (called coronary) are:

age: age in years

sys_BP: systolic blood pressure, mm of mercury

dia_BP: diastolic blood pressure, mm of mercury

chol: cholesterol level in mg per DL

ht.in: height in inches

wt.lb: weight in pounds

incident: 0 for no coronary, 1 for coronary.

A stepwise variable selection method was performed using *step* in R, and gave the following output (some of which has been deleted).

```
Start:  AIC=148.85
incident ~ age + sys_BP + dia_BP + chol + ht.in + wt.lb

         Df Deviance    AIC
- dia_BP  1   134.88 146.88
- sys_BP  1   134.97 146.97
- ht.in   1   135.34 147.34
<none>        134.85 148.85
- chol    1   137.84 149.84
- age     1   138.73 150.73
- wt.lb   1   139.00 151.00

Step:  AIC=146.88
incident ~ age + sys_BP + chol + ht.in + wt.lb

         Df Deviance    AIC
- sys_BP  1   134.98 144.98
- ht.in   1   135.35 145.35
<none>        134.88 146.88
- chol    1   137.84 147.84
- age     1   138.77 148.77
- wt.lb   1   139.29 149.29

Step:  AIC=144.98
incident ~ age + chol + ht.in + wt.lb

        Df Deviance    AIC
- ht.in  1   135.52 143.52
<none>       134.98 144.98
- chol   1   138.05 146.05
- wt.lb  1   139.91 147.91
- age    1   140.30 148.30

Step:  AIC=143.52
incident ~ age + chol + wt.lb

        Df Deviance    AIC
<none>       135.52 143.52
- chol   1   138.77 144.77
- wt.lb  1   139.93 145.93
- age    1   142.30 148.30
```

```
Call:  .........

Coefficients:
(Intercept)          age         chol        wt.lb
  -9.255892     0.053004     0.006518     0.017539

Degrees of Freedom: 199 Total (i.e. Null);  196 Residual
Null Deviance:     154.6
Residual Deviance: 135.5  AIC: 143.5
```

Further output arising from the chosen model is as follows:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1049  -0.5541  -0.3777  -0.2510   2.7009

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.255892   2.071678  -4.468  7.9e-06 ***
age          0.053004   0.020827   2.545   0.0109 *
chol         0.006518   0.003589   1.816   0.0693 .
wt.lb        0.017539   0.008272   2.120   0.0340 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 154.55  on 199  degrees of freedom
Residual deviance: 135.52  on 196  degrees of freedom
AIC: 143.52

Number of Fisher Scoring iterations: 5



[1] "Accuracy Matrix for Logistic Regression, cutoff=0.5"

    FALSE TRUE
  0   174    0
  1    24    2
```

(a) What is the basic *R*-code required to fit a logistic regression model including all of the explanatory variables above?

glm(incident $\sim$., family=binomial, data=coronary)

(b) Which of the models appearing in the steps above would have the smallest residual deviance?

The full model should have the smallest residual deviance.

(c) What is the model chosen by the stepwise variable selection algorithm? Include in your answer what the response variable for this logistic model is.

$\log(p/(1-p)) = (-9.255892)+(0.053004)*age+(0.006518)*chol+(0.017539)*wt.lb$, **where p is the probability of coronary incident.**

For the chosen model:

(d) Interpret the meaning of each of the coefficients.

(i) Intercept β_0: this gives the value of $\log(p/(1-p))$ for a male of age=chol=wt.lb=0 (equivalently the probability for such a person is $e^{\beta_0}/(1+e^{\beta_0})$). This is not meaningful in this case.

(ii) Age coefficient β_1: all else held constant, adding 1 year of age multiplies the odds of incident by e^{β_1}.

(iii) Cholesterol coefficient β_2: all else held constant, increasing the cholesterol level by 1 multiplies the odds of incident by e^{β_2}.

(iv) Weight coefficient β_3: all else held constant, increasing the weight by 1 pound multiplies the odds of incident by e^{β_3}.

(e) Estimate the odds that a male of 50 years of age, with cholesterol level of 250 and weight 150 pounds has a coronary incident.

$\exp((-9.255892)+(0.053004)*50+(0.006518)*250+(0.017539)*150) = 0.09581443$

(f) Estimate the probability that a male of 50 years of age, with cholesterol level of 250 and weight 150 pounds has a coronary incident.

$0.09581443/(1 + 0.09581443) = 0.08743673$**.**

(g) When using the chosen model to predict whether participants had a coronary incident (predicting an incident if the probability of an incident is estimated as greater than 0.5), what is the proportion of correct predictions? Is this good or bad?

$(2 + 174)/200 = 0.88$**. If you were to say nobody will have a coronary incident your accuracy will be** 0.87**. This model increases your accuracy by** 0.01**. Not great.**

End of Exam—Total Available Marks = 82