INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2020

Subject CS2B –Risk Modelling and Survival Analysis Core Principles

Introduction

The Examiners' Report is written by the Chief Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and using past papers as a revision aid and also those who have previously failed the subject.

The Examiners are charged by Council with examining the published syllabus. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

The report is written based on the legislative and regulatory context pertaining to the date that the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision.

Mike Hammer
Chair of the Board of Examiners
December 2020

A. General comments on the *aims of this subject and how it is marked*

1. The aim of the Risk Modelling and Survival Analysis subject is to provide a grounding in mathematical and statistical modelling techniques that are of particular relevance to actuarial work, including stochastic processes and survival models.

2. Candidates are reminded of the need to include the R code, that they have used to generate their solutions, together with the main R output produced, in their answer script. Where the R code was missing from a particular question part, no marks were awarded even if the output (e.g. a graph) was included. Partial credit was awarded in the cases where the R code was included but the R output was not.

3. The marking schedule below sets out potential R code solutions for each question. Other appropriate R code solutions gained full credit unless one specific approach had been explicitly requested in the question paper.

4. In cases where the same error was carried forward to later parts of the answer, candidates were given full credit for the later parts.

5. In higher order skills questions, where comments were required, well-reasoned comments that differed from those provided in the solutions also received credit as appropriate.

B. Comments on *candidates' performance in this diet of the examination.*

1. On the whole, performance was less than satisfactory. Candidates generally demonstrated their ability to use R to perform analysis but did not fully demonstrate their ability to interpret the results. As the first two questions covered topics that were tested in previous CS2B diets, a higher level of performance was expected.

2. Question 3 was poorly answered with many candidates failing to proceed beyond part (iv). Candidates are reminded that, in such circumstances, the best approach is to provide a "dummy" answer and carry on with the remaining parts of the question to receive carry forward credit.

3. It is important that appropriate commentary is provided alongside the R code and R output in the answer script, where relevant, to fully demonstrate sufficient understanding. For example, in Q1(v), it was important to clearly specify the sample ACF values separately from the R output and, in Q1(iii), to clearly label the different graphs. Instructions to this effect were communicated to candidates at the time of the exam. Candidates are advised to take careful note of all instructions that are provided with the exam in order to maximise their performance in future CS2B examinations. The instructions applicable to this diet can be found at the beginning of the solutions contained within this document.

4. Higher order skills questions were generally answered poorly. Candidates should recognise that these are generally the questions which differentiate those candidates with a good grasp and understanding of the subject.

5. The comments that follow the questions in the marking schedule below, concentrate on areas where candidates could have improved their performance. Candidates approaching the subject for the first time are advised to concentrate their revision in these areas.

C. Pass Mark

The combined Pass Mark for the CS2 exam was 56.

1,363 presented themselves and 476 passed.

Solutions for Subject CS2 Paper B September 2020

Please note that the following principles apply to the CS2B solutions. These principles were set out in the instructions provided to candidates along with the examination paper:

- Candidates **MUST** include the R code used to obtain their answers, together with the main R output produced, in the Word document in order to obtain marks. Please note that failure to include the R code used will result in **ZERO MARKS** for that particular question.

- Candidates **MUST** include appropriate titles, axes labels, and where relevant, legends in all graphical output that is generated in R for inclusion in the Word document. Please note that failure to include appropriate annotations will result in full credit not being given.

- When a question requires a particular numerical answer or conclusion, candidates **MUST** explicitly and clearly state this in the Word document, separately from, and in addition to the R output that contains the relevant numerical information. Please note that failure to include a separate answer or conclusion will result in full credit not being given.

- Candidates **MUST** type all their non-R code workings and answers into the Word document using standard keyboard typing. Candidates **DO NOT** need to use notation that requires specialised equation editing e.g. the "Equation Editor" functionality in Word. Calculations pasted in from another non-R application (e.g. Excel) will receive **ZERO MARKS**.

1

(i)

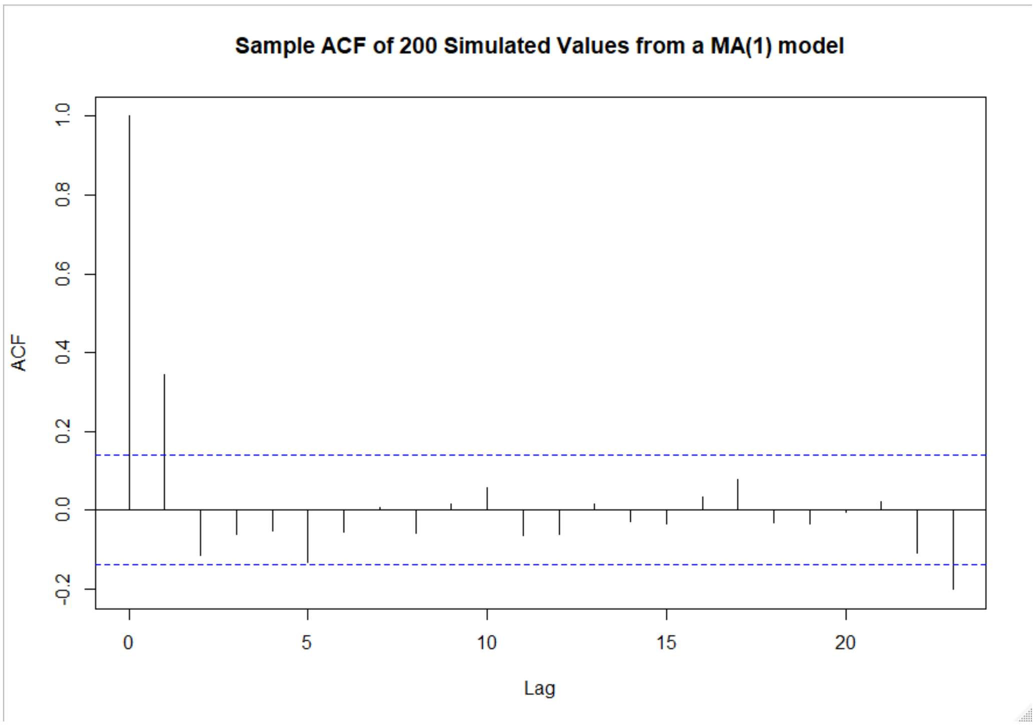
```
set.seed(967)                                          [½]
YMA = arima.sim(n=200, model=list(ma=c(0.4)))          [1½]
```

(ii)

```
set.seed(967)                                          [½]
YAR = arima.sim(n=200, model=list(ar=c(0.45)))         [1½]
```

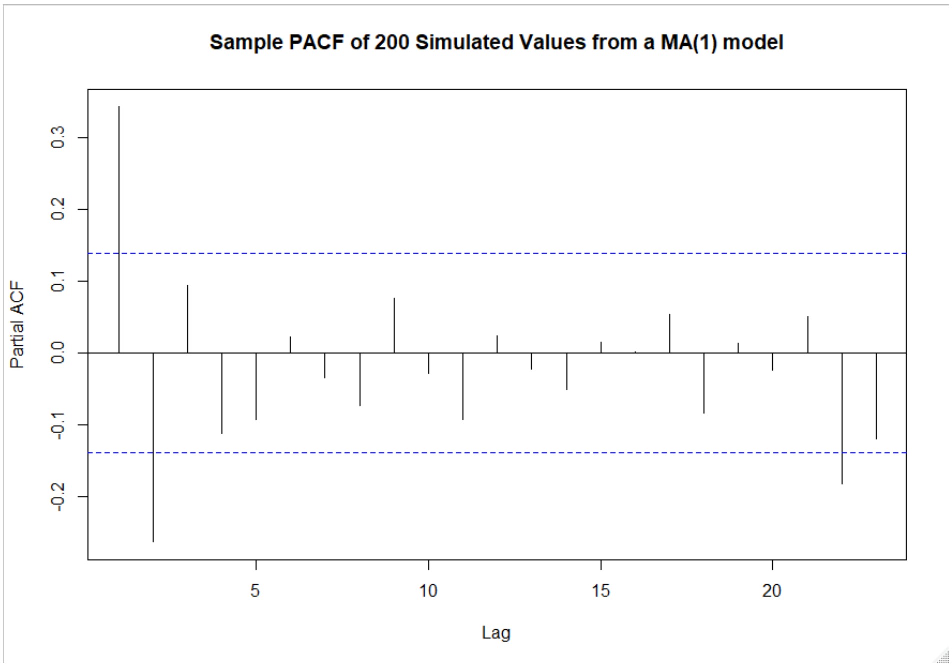
(iii)

```
acf(YMA, main ="Sample ACF of 200 Simulated Values from a
MA(1) model")
```



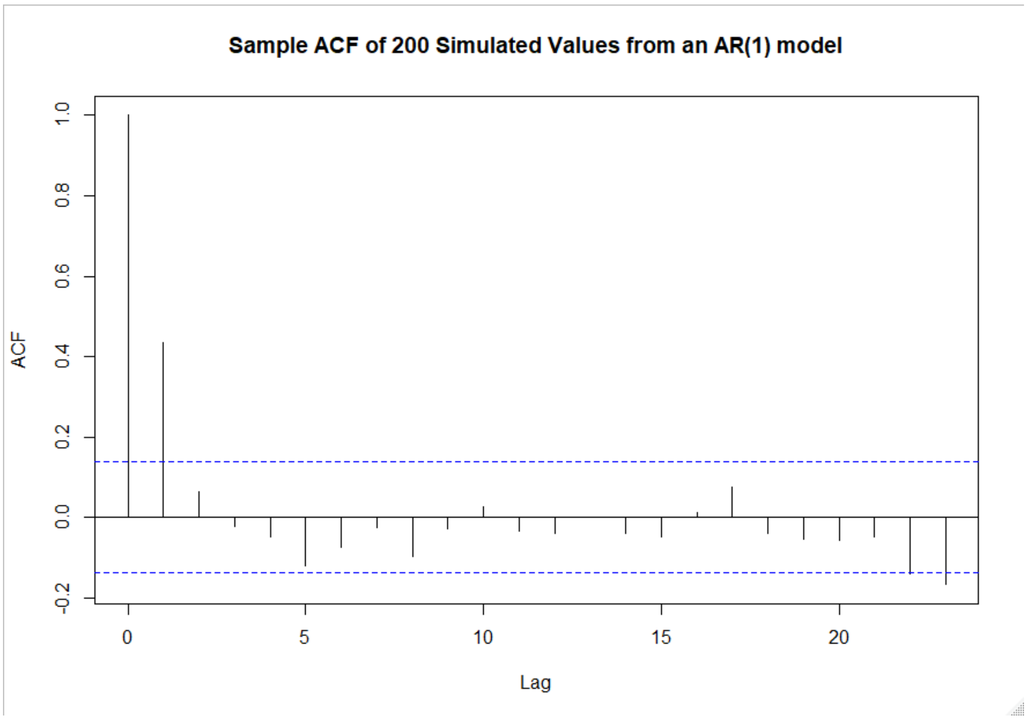
[1½]

```
pacf(YMA, main ="Sample PACF of 200 Simulated Values from a MA(1) model")
```



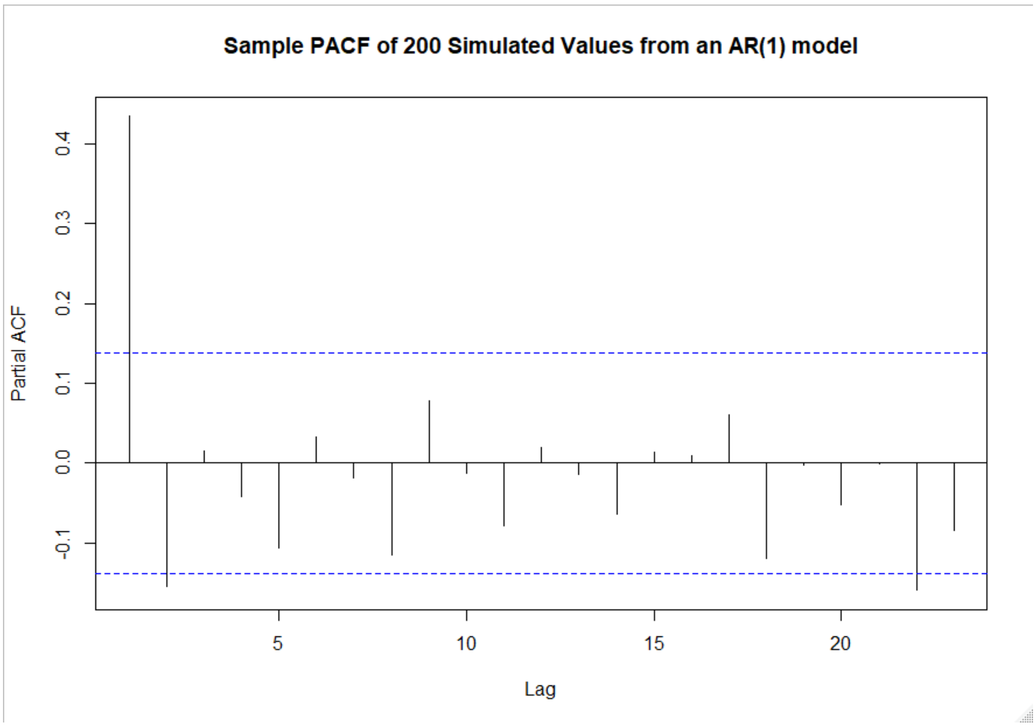
[1½]

```
acf(YAR, main ="Sample ACF of 200 Simulated Values from an AR(1) model")
```



[1½]

```
pacf(YAR, main ="Sample PACF of 200 Simulated Values from
an AR(1) model")
```



[1½]

(iv) For a MA(1) model we would expect…

…the PACF to decrease exponentially to 0… [½]

… and the ACF to have clear spikes up to lag 1 with insignificant spikes for all lags > 1. [1]

For an AR(1) model we would expect…

… the ACF to decrease exponentially to 0 [½]

… and the PACF to have one clear spike at lag 1 with insignificant spikes for all lags > 1. [1]

These features are broadly in line with the graphs above. [1]

There appear to be non-zero values at lag 22/23 in the graphs which may need to be investigated… [½]

… however, this may not be an issue as we would expect some random spikes as the sample ACFs won't conform perfectly with the theoretical behaviour. [½]

[Marks available 5, maximum 4]

(v)

```
ACF2MA = acf(YMA,plot = FALSE)$acf[3];ACF2MA          [½]
[1] -0.1146628

ACF2AR = acf(YAR,plot = FALSE)$acf[3];ACF2AR          [½]
[1] 0.06402174
```

The value of the sample ACF at lag 2 for YMA is -0.1146628. [½]

The value of the sample ACF at lag 2 for YAR is 0.06402174. [½]

(vi)

```
set.seed(967)                                         [1]
ACF2MA=1:1000                                         [1]
ACF2AR=1:1000                                         [1]

for (i in 1:1000){                                    [1]
YMA= arima.sim(n=200, model=list(ma=c(0.4)))          [1]
YAR= arima.sim(n=200, model=list(ar=c(0.45)))         [1]
ACF2MA[i]=acf(YMA,plot = FALSE)$acf[3]                [1]
ACF2AR[i]=acf(YAR,plot = FALSE)$acf[3]                [1]
}
```

(vii)

```
 mean(ACF2MA)
[1] -0.0108415                                        [½]

mean(ACF2AR)
[1] 0.188152                                          [½]

var(ACF2MA)
[1] 0.006459321                                       [½]

var(ACF2AR)
[1] 0.006076729                                       [½]
```

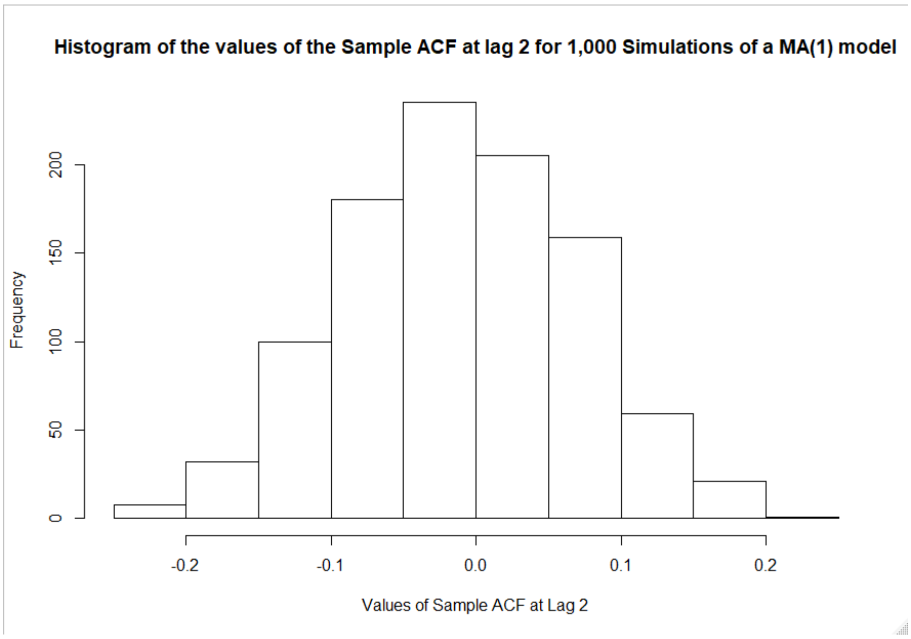
The mean of ACF2MA is -0.0108415.

The mean of ACF2AR is 0.188152. [½]

The variance of ACF2MA is 0.006459321.

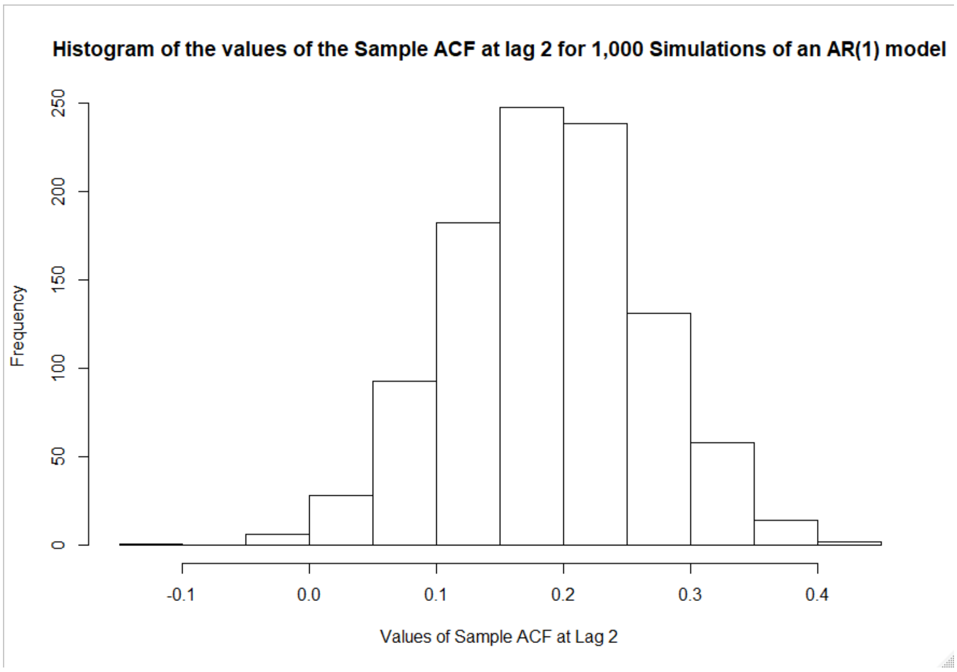
The variance of ACF2AR is 0.006076729. [½]

(viii)
```
hist(ACF2MA, xlab = "Values of Sample ACF at Lag 2", main
= "Histogram of the values of the Sample ACF at lag 2 for
1,000 Simulations of a MA(1) model")
```



[2]

```
hist(ACF2AR, xlab = "Values of Sample ACF at Lag 2", main
= "Histogram of the values of the Sample ACF at lag 2 for
1,000 Simulations of an AR(1) model")
```



[2]

(ix) The theoretical mean of ACF2MA = 0… [1]

…and the theoretical mean of ACF2AR = 0.45^2 = 0.2025. [2]

The asymptotic variance of ACF2MA

= (1 + [(0.4)/(1+0.4^2)]^2) / 200 = 0.005595 (1)

OR:

= (1 + 2 * ([(0.4)/(1+0.4^2)]^2)) / 200 = 0.006189 (2)

[2]

The histograms in part (viii) confirm the asymptotic normal behaviour of the estimates as they indicate normal distributions... [2]

… and the values of the mean and variance of ACF2MA calculated in part (vii) are similar to the theoretical asymptotic values… [1]

… and the value of the mean of ACF2AR calculated in part (vii) is similar to the theoretical asymptotic value… [1]

[Total 40]

Parts (i) and (ii) were very well-answered.

Part (iii) was very well-answered but some candidates lost marks for not including appropriate titles in the graphs. The minimum requirements for an appropriate title were that it must include the appropriate series name and also indicate whether the function plotted was a sample ACF or a sample PACF. The default axes labels were deemed appropriate in this case. Additionally, some candidates lost marks for not including either the R code or the graphs in their answer scripts.

Part (iv) was reasonably well-answered but some candidates either mixed up the expected behaviour of MA(1) and AR(1) processes or mixed up the expected behaviour of the ACF and PACF within each process.

Answers to part (v) were mixed with a number of candidates losing marks because they included the output of the ACF for a number of lags without making any attempt to pick out lag 2. Some candidates also lost marks because they picked out the value of the wrong lag. Additionally, some candidates lost marks for not including the R output and/or not separately stating the lag 2 values in their answer scripts.

Answers to part (vi) were also mixed with many candidates getting stuck here and not proceeding with later parts of the question.

Part (vii) was surprisingly poorly answered. This was mainly due to many candidates getting stuck in part (vi). Candidates are reminded that, in such circumstances, the best approach is to provide a "dummy" answer and carry on with the remaining parts of the question to receive carry forward credit. Additionally, some candidates lost marks for not including the R output and/or not separately stating the calculated values in their answer scripts.

Part (viii) was again poorly answered. Many candidates lost marks for not appropriately labelling the graphs. The minimum requirements for appropriate labelling were that either the x-axis and/or the title needed to clearly indicate which vector is being plotted and also needed to reference lag 2. The default y-axis labels were deemed appropriate in this case. Additionally, some candidates lost marks for not including either the R code or the graphs in their answer scripts.

Candidates are reminded to take careful note of all instructions that are provided with the exam in order to maximise their performance in future CS2B exams.

Part (ix) was extremely poorly answered. Very few candidates demonstrated an understanding of the expected asymptotic behaviour of these processes. Note that the formula in the Core Reading for the asymptotic variance of ACF2MA, denoted by equation (1), is incorrect. The correct formula for the asymptotic variance is given by equation (2).

2

(i)

```
length(X[X<=400])/1000
 [1] 0.987
```

OR:

```
sum(X<=400)/1000
[1] 0.987
```

[1½]

Therefore, the proportion of claims that are fully covered by the insurer = 98.7% [½]

(ii)

```
M = 400
Y = pmin(X, M)
```

[1]

(iii)

```
Z = pmax(0, X-M)
```

OR:

```
Z = X-Y
```

[1]

(iv)

```
S = sum(X[X<=400])
logLik = function(lambda){
     -5200 * lambda
     + 987 * log(lambda)
     - lambda * S
}
```

OR:

```
S = sum(Y)
logLik = function(lambda){
     987 * log(lambda)
     - lambda * S
}
```

OR:

```
nz = length(Y[Y==M])
Y_exc_M = Y[Y<M]
flnL = function(parameter){
        nz*pexp(M, rate=parameter, lower.tail=FALSE,
        log.p=TRUE)+
        sum(dexp(Y_exc_M, rate=parameter, log=TRUE))
}
```

[10]

(v)

```
987 / (5200 + sum(X[X<=400]))
 [1] 0.01023176
```

[5½]

Hence, the maximum likelihood estimate of *lambda* = 0.01023176. [½]

OR:

```
nlm(f = function(x) - logLik(x), p = 0.01)$estimate
 [1] 0.01023209
```

[5½]

Hence, the maximum likelihood estimate of *lambda* = 0.01023209. [½]

OR:

```
nlm(f = function(x) - flnL(x), p = 0.01)$estimate
 [1] 0.01023209
```

[5½]

Hence, the maximum likelihood estimate of *lambda* = 0.01023209. [½]

[Total 20]

Part (i) was well-answered. However, some candidates lost marks because they either calculated the number of claims that were covered by the insurer or because they calculated the proportion of claims covered by the reinsurer instead. Additionally, some candidates lost marks for not including the R output and/or not separately stating the proportion of claims covered in their answer scripts.

Parts (ii) and (iii) were very well-answered. However some candidates lost marks because they did not define M in the R code that they provided in their answer scripts.

Answers to part (iv) were mixed. Candidates who constructed a negative log-likelihood function did not lose marks provided that they allowed for the correct treatment of this function in part (v). A very common mistake was for candidates to construct the uncensored likelihood function rather than the censored likelihood function e.g. candidates used Y in the dexp function in the third alternative solution rather than Y_exc_M and omitted the first term in the flnL function.

Part (v) was well-answered. Candidates did not lose marks for using different starting estimates of p in the second and third alternative solutions provided that the estimate was reasonable. Candidates who constructed a negative log-likelihood function in part (iv) did not lose marks provided that they treated this correctly here. Some candidates lost marks for not including the R output and/or not separately stating the maximum likelihood estimate in their answer scripts.

3

(i)

```
portfolio$group_label_stage1 <- c(rep("A", length=100),
rep("B", length=100))
```

[4]

(ii)

```
x1_A <- mean(portfolio$x1[portfolio$group_label_stage1 ==
"A"]); x1_A
[1] 2.926666
x2_A <- mean(portfolio$x2[portfolio$group_label_stage1 ==
"A"]); x2_A
[1] -0.7054048
x1_B <- mean(portfolio$x1[portfolio$group_label_stage1 ==
"B"]); x1_B
[1] 2.829781
x2_B <- mean(portfolio$x2[portfolio$group_label_stage1 ==
"B"]); x2_B
[1] 0.7054048
```

OR:

```
model1 = kmeans(portfolio[1:100,1:2],1)
model1$centers
        x1         x2
1 2.926666 -0.7054048

x1_A <- model1$centers[1,1]; x1_A
[1] 2.926666

x2_A <- model1$centers[1,2]; x2_A
[1] -0.7054048

model2 = kmeans(portfolio[101:200,1:2],1)
model2$centers
        x1        x2
1 2.829781 0.7054048

x1_B <- model2$centers[1,1]; x1_B
[1] 2.829781

x2_B <- model2$centers[1,2]; x2_B
[1] 0.7054048
```

[5]

Therefore:
the coordinates of the centre of cluster "A", (*x*1_*A*, *x*2_*A*) = (2.926666, -0.7054048)
[½]

and:
the coordinates of the centre of cluster "B", (*x*1_*B*, *x*2_*B*) = (2.829781, 0.7054048)
[½]

(iii)

```
portfolio$dist_A <- sqrt((portfolio$x1 - x1_A)^2 +
(portfolio$x2 - x2_A)^2)                                  [4]
```

(iv)

```
portfolio$dist_B <- sqrt((portfolio$x1 - x1_B)^2 +
(portfolio$x2 - x2_B)^2)                                  [4]
```

(v)

```
portfolio$group_label_stage2 <- rep("A", 200)
portfolio$group_label_stage2[portfolio$dist_B <
portfolio$dist_A] <- "B"                                  [4]
```

(vi)

```
table(portfolio$group_label_stage1,
portfolio$group_label_stage2)
```

```
   A  B
A 70 30
B  3 97
```

[2]

(vii) There are 70 policyholders whose cluster labels were originally assigned to "A" in the first stage of the investigation which remained unchanged after the update. [1]

There are 97 policyholders whose cluster labels were originally assigned to "B" in the first stage of the investigation which remained unchanged after the update. [1]

However, the labels of 30 policyholders were updated from "A" to "B"… [1]

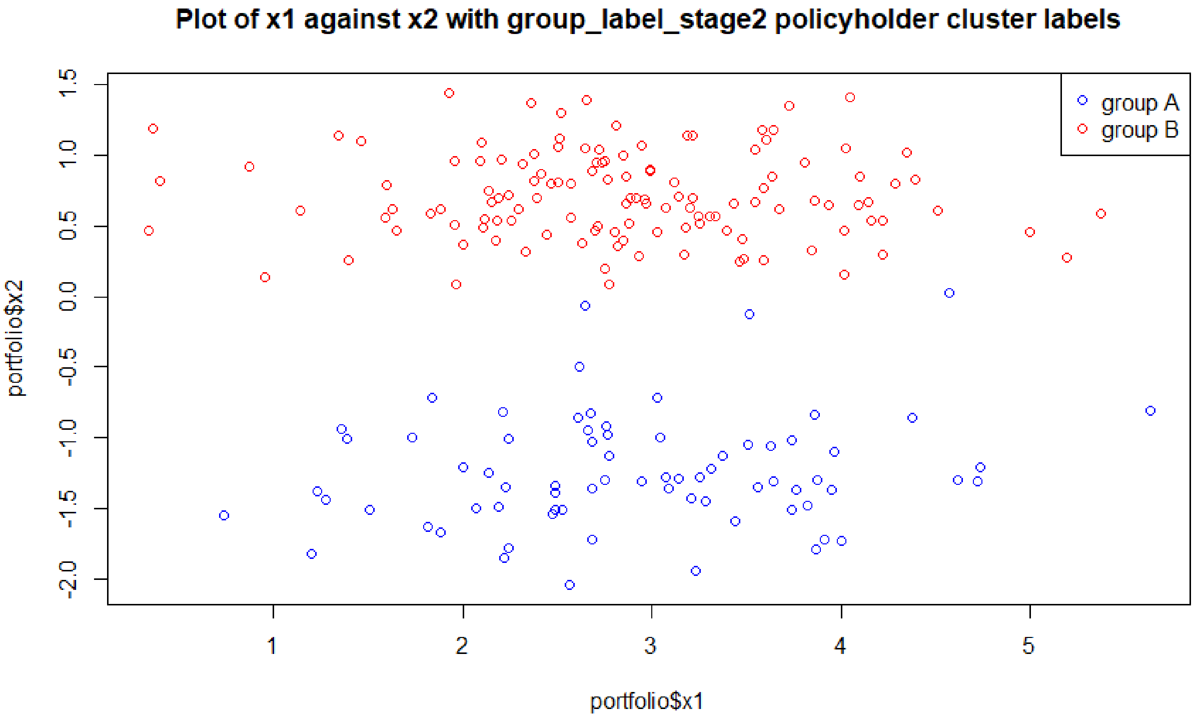
…whereas, the labels of 3 policyholders were updated from "B" to "A". [1]

(viii)

```
col_vec <- rep("blue", 200)                               [1]
col_vec[portfolio$group_label_stage2=="B"] <- "red"       [1]

plot(                                                     [½]
     portfolio$x1,                                        [½]
     portfolio$x2,                                        [½]
     main="Plot of x1 against x2 with group_label_stage2
     policyholder cluster labels",                        [½]
     col=col_vec)                                         [½]

legend("topright", legend=c("group A", "group B"),
col=c("blue", "red"), pch=1)                              [1]
```



[½]

(ix) The graph in part (viii) shows that the analyst was able to reasonably identify two sets of clusters. [1]

However, the analyst did not check the convergence of the clustering algorithm… [2]

… although, inspection of the table might suggest that the clusters are unlikely to change much. [1]

The top three group A policyholders might possibly be better assigned to group B. [1]

The analyst could potentially improve the results by updating the centres of the clusters, re-calculating the distances, updating the labels and repeating this process until convergence (i.e. until the labels remain constant).

OR:

The analyst could implement the full kmeans algorithm to ensure convergence of the final clusters.

[3]

The analyst may want to apply feature scaling / data normalisation to the values of x_1 and x_2 so that each of them contributes approximately proportionately to the Euclidean distances and then re-run the analysis. [2]

[Marks available 10, maximum 6]

[Total 40]

Part (i) was well answered. Candidates used a variety of different methods to populate the new group_label_stage1 column and all valid methods received full marks. Some candidates lost marks for not using the correct dataframe name or column name and also for not using A and B as cluster labels. Candidates are reminded of the need to read the question carefully.

Part (ii) was poorly answered with many candidates losing marks for using the kmeans function on all 200 policyholders together to derive the coordinates of the centres of the converged set of clusters A and B. Additionally, some candidates lost marks for not including the R output and/or not separately stating the cluster centre coordinates in their answer scripts.

Answers to parts (iii) and (iv) were mixed. A common mistake for candidates who had used the kmeans function in part (ii) to derive the converged set of clusters was to misinterpret the kmeans output and populate the coordinates of the cluster centres incorrectly. Additionally, some candidates lost marks for not using the correct column names specified in the question.

Part (v) was poorly answered with many candidates getting stuck here and not proceeding with later parts of the question. Some candidates lost marks for not using the correct column name and also for not using A and B as cluster labels. In many cases, candidates lost marks for using the value of the shortest Euclidean distance to populate the group_label_stage2 column.

Part (vi) was very poorly answered. This was mainly due to many candidates getting stuck in part (v). Candidates are reminded that, in such circumstances, the best approach is to provide a "dummy" answer and carry on with the remaining parts of the question to receive carry forward credit. Candidates did not lose marks for not including the R output as the command verb in this question was "Generate". Candidates also did not lose marks for generating the transpose of the table. Additionally, candidates who generated a matrix object with the correct entries did not lose marks.

Again part (vii) was extremely poorly answered. Some candidates lost marks for misinterpreting the table generated in part (vi).

Part (viii) was also very poorly answered. Candidates lost marks for not adding an appropriate title to the graph. The minimum requirement for an appropriate title was that it needed to refer to the stage 2 cluster labels. The default axes labels were deemed appropriate in this case. Candidates also lost marks for not adding an appropriate legend to the graph although candidates who manually added an appropriate legend next to the graph did not lose marks. Additionally, some candidates lost marks for not including either the R code or the graph in their answer scripts.

Again part (ix) was extremely poorly answered. Candidates who commented on exploring an alternative number of clusters did not receive credit as the question is specifically about dividing the policyholders into two clusters. Candidates who used the kmeans function to derive the converged set of clusters in part (ii) and who stated that the analyst's decision is reasonable because the use of the kmeans algorithm has resulted in a converged set of clusters, were awarded significant credit.

END OF EXAMINERS' REPORT