Institute
and Faculty
of Actuaries

EXAMINERS’ REPORT

CS2B - Risk Modelling and Survival Analysis

Core Principles

Paper B

September 2022

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.

Sarah Hutchinson
Chair of the Board of Examiners
December 2022

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

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.

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.

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.

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

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 *candidate performance in this diet of the examination.*

Candidates typically demonstrated their ability to use R to perform analysis but did not fully demonstrate their ability to interpret the results or to apply some of the techniques to unfamiliar situations. As with Paper A, the syllabus and Core Reading for Risk Modelling and Survival Analysis Core Principles covers multiple statistical techniques and modelling approaches.

Performance was quite uneven across areas of the syllabus with Question 3 on Ridge Regression as a means of Machine Learning typically receiving lower marks than the other two questions. Candidates are reminded that they need to prepare thoroughly across the entire syllabus and Core Reading. It is also important to remember that the primary aim of this examination is to test understanding of modelling approaches rather than of particular R packages.

To assist candidates in future preparation for this paper the examiners would suggest that study of R programming techniques is undertaken and practice questions attempted on a topic by topic basis alongside study for paper CS2A rather than afterwards as a separate exercise. In particular candidates may benefit from considering problem questions in Risk Modelling, Survival Analysis, Stochastic Processes and Time Series both from a traditional 'pen and paper' approach and also in R so as to build experience across the CS2 syllabus in both A and B paper question styles.

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 questions requiring charts, appropriate titles, axis labels and legends are necessary, and in questions requiring a specific numerical answer, this must be stated separately from the R output. Candidates are advised to take careful note of all instructions that are provided with the exam in order to maximise their performance in CS2B examinations.

Application skills questions were not well answered. Candidates should recognise that these are generally the questions which differentiate those candidates with a good grasp and understanding of the subject.

C. Pass Mark

The Pass Mark for this exam was 55
987 presented themselves and 195 passed.

Solutions for Subject CS2B - September 2022

Q1

(i)

```
> set.seed(912)

y=arima.sim(list(order=c(0,1,0)),n=400)
> fit=arima(y,order=c(1,0,0))
> fit
Call:
arima(x = y, order = c(1, 0, 0))

Coefficients:
ar1  intercept
      0.9978   -13.1872
s.e.  0.0024    12.6255

sigma^2 estimated as 1.018:  log likelihood = -575.34,  aic = 
1156.68
```

Setting the seed [½]
Simulate 400 realisations from the ARIMA(0,1,0) model and save them as vector y [½]
Fit to these data the model ARIMA(1,0,0) [½]
Display the fitted model fit [½]

(ii)

```
> fit$coef[1]-qnorm(0.975)*sqrt(fit$var.coef[1,1])
       ar1 
0.9931066 
> fit$coef[1]+qnorm(0.975)*sqrt(fit$var.coef[1,1])
     ar1 
1.002402
```
[1]

The standard error is 0.0024 and the 95% CI is (0.9931066, 1.002402) [1]

(iii)
The CI contains values >= 1 which indicate non-stationarity. [1]
This is not surprising as the data was generated from ARIMA(0,1,0) [1]

(iv)

```
> predict(fit, n.ahead = 10)                                  [1]
$pred
Time Series:
Start = 401
End = 410
Frequency = 1


[1] -30.92040 -30.88058 -30.84085 -30.80120 -30.76165
[6] -30.72219 -30.68281 -30.64352 -30.60433 -30.56521

$se
Time Series:
Start = 401
End = 410
Frequency = 1
[1] 1.009127 1.425520 1.743941 2.011473 2.246377 2.458030
[7] 2.652008 2.831948 3.000382 3.159152                       [1]
```

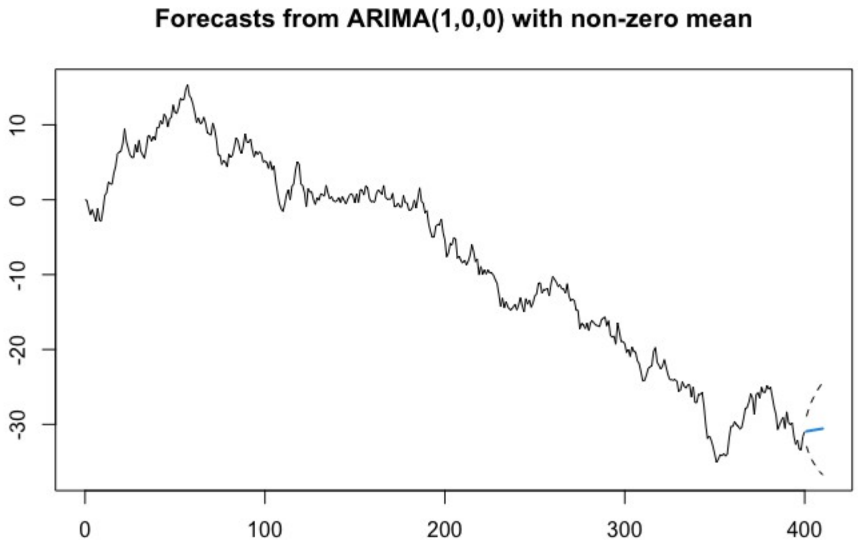
(v)

```
> A=cbind(predict(fit, n.ahead = 10)$pred,predict(fit
, n.ahead = 10)$se).                                          [1]
> A                                                           [½]
Time Series:
Start = 401
End = 410
Frequency = 1
    predict(fit, n.ahead = 10)$pred predict(fit, n.ahead =
10)$se

          401          -30.92040          1.009127
          402          -30.88058          1.425520
          403          -30.84085          1.743941
          404          -30.80120          2.011473
          405          -30.76165          2.246377
          406          -30.72219          2.458030
          407          -30.68281          2.652008
          408          -30.64352          2.831948
          409          -30.60433          3.000382
          410              -30.56521          3.159152            [½]
```

(vi)

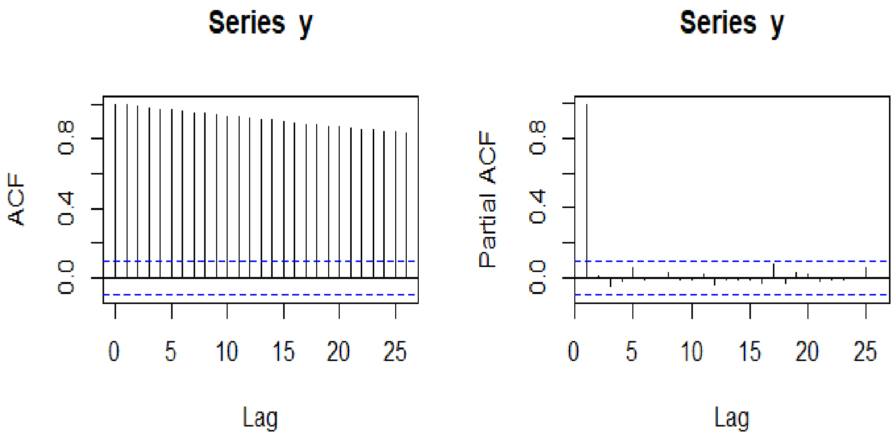
```
predV <- forecast(fit, h=10, level=c(95))
plot(predV,shaded=F)                                        [3]
```



[1]

(vii)

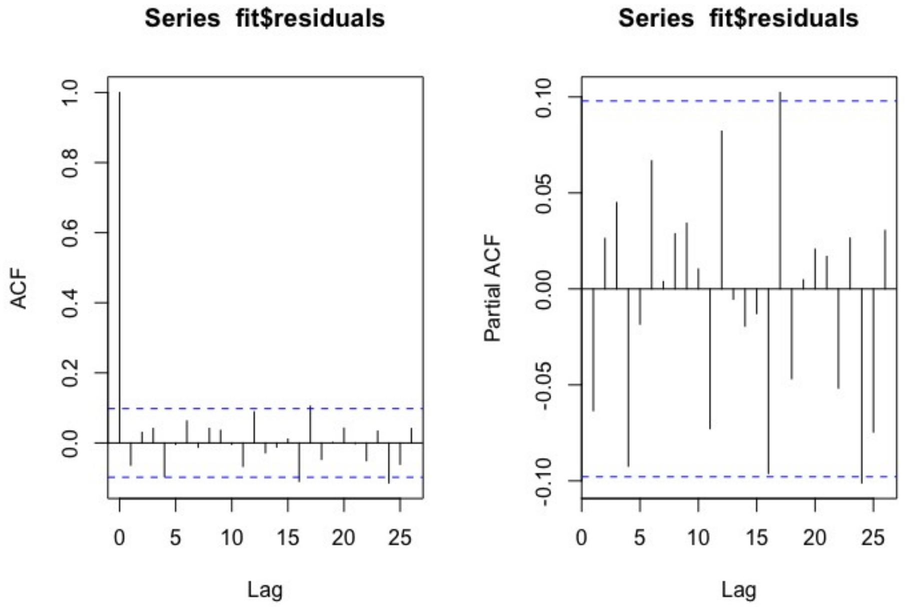
```
par(mfrow=c(1,2))
     acf(y)                                                 [½]
     pacf(y)                                                [½]
```



[1]

(viii)

```
par(mfrow=c(1,2))
     acf(fit$residuals)                                     [½]
     pacf(fit$residuals)                                    [½]
```



[1]

(ix)

The plots for y suggest that that the ACF is not decaying exponentially fast [½]

In fact the linear rate of decay suggests a unit root [½]

This is consistent with the ARIMA(0,1,0) behaviour [1]

The plots for the residuals of the model `fit`, however, generally lie within the confidence intervals [1]

This is consistent with the residuals forming a white noise process [1]

(x)

```
 > Box.test(fit$residuals,type="Ljung",fitdf = 1,lag=4)      [1½]
     Box-Ljung test
data:  fit$residuals
X-squared = 6.4628, df = 3, p-value = 0.09114                 [½]
>
> Box.test(fit$residuals,type="Ljung",fitdf = 1,lag=6)        [½]
     Box-Ljung test

data:  fit$residuals
X-squared = 8.0806, df = 5, p-value = 0.1519                  [½]

>
> Box.test(fit$residuals,type="Ljung",fitdf = 1,lag=12)       [½]
     Box-Ljung test

data:  fit$residuals
X-squared = 14.498, df = 11, p-value = 0.2067                 [½]
```

(xi)
From part (ix), the ACF and PACF plots of the residuals are consistent with an ARIMA(1,0,0) model. [1]
The three tests in part (x) are also consistent with an ARIMA(1,0,0) model at the 5% significance level, since the p-values are greater than 0.05 [1]
However, this is not sufficient to establish that the ARIMA(1,0,0) model is correct
We have simply not found evidence to conclude that it is incorrect [1]
We would expect the ARIMA(0,1,0) model that was used to generate the data to satisfy the tests as well [1]
Model ARIMA(0,1,0) can be shown to be also a good fit and lower AIC

[Total 30]

This question was generally well answered.

In part (ii) the confidence interval needed to be calculated using R not otherwise. Part (iii) was less well answered with many candidates failing to make the link to stationarity.

Parts (iv) and (v) were well answered. In part (vi) there are a number of different ways in R to calculate the forecast and generate the plot. It was pleasing to note that many candidates included proper titles, axis labels and a legend with their plot.

The ACF and PACF plots in parts (vii) and (viii) were generally properly constructed as well. In part (ix) there are a variety of points that could be made to secure the marks. With four plots to comment on, candidates are reminded to include a mention of each plot in their discussion.

Part (x) was generally well answered. Where an error was made it was most often with respect to the `fitdf` function in R and the resulting degrees of freedom.

Q2
(i)
Informative censoring is likely to be present [½]
The deaths for an unknown reason may or may not have been from blood clots [1]
Even if the deaths for an unknown reason were not from blood clots, they are still likely to constitute informative censoring [½]
This is because, had these lives not died, they were likely to have been in poorer health, and hence more likely to suffer from blood clots, than those remaining [1]

(ii)

```
data = read.csv(file=" CS2B_S22_Qu_2_Data.csv ")        [1½]
ST<-ifelse(data$Status==2,1,0)                           [2]
data_main<- cbind(data, ST)                              [½]
tail(data_main,20)                                       [1½]
```

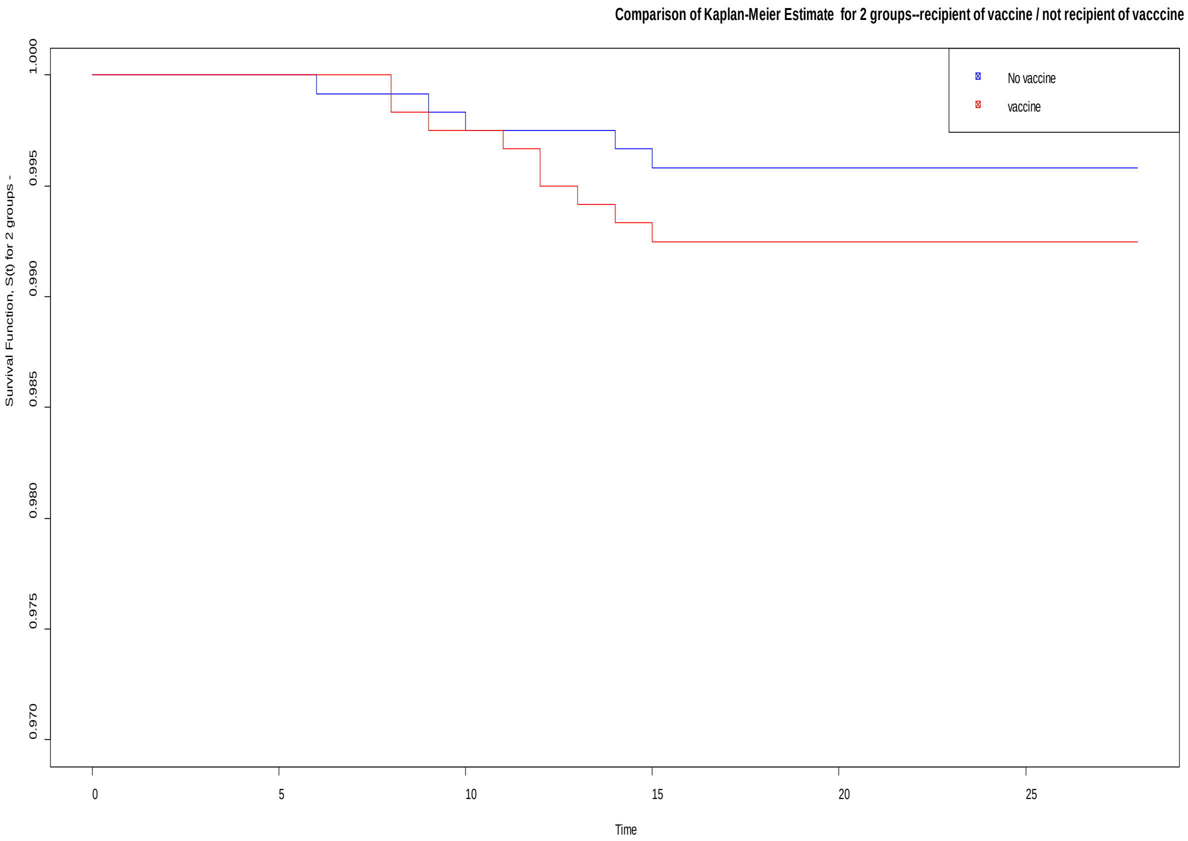
```
Life Drug Age co_morbidity already_infected Status Time ST
2381 2381    1   5          0              1      0   28  0
2382 2382    1   5          0              1      0   28  0
2383 2383    1   5          0              1      0   28  0
2384 2384    1   5          0              1      0   28  0
2385 2385    1   5          0              1      2   12  1
2386 2386    1   5          0              1      0   28  0
2387 2387    1   5          1              1      0   28  0
2388 2388    1   5          1              1      0   28  0
2389 2389    1   5          0              1      0   28  0
2390 2390    1   5          0              1      0   28  0
2391 2391    1   5          0              1      0   28  0
2392 2392    1   5          1              1      0   28  0
2393 2393    1   5          0              1      0   28  0
2394 2394    1   5          0              1      0   28  0
2395 2395    1   5          0              1      0   28  0
2396 2396    1   5          1              1      2    8  1
2397 2397    1   5          1              1      0   28  0
2398 2398    1   5          0              1      0   28  0
2399 2399    1   5          0              1      0   28  0
2400 2400    1   5          0              1      0   28  0
```

[½]

(iii)

```
S = survfit(                                                  [1]
Surv(data_main$Time, data_main$ST)                            [1]
~data_main$Drug)                                              [1]

plot(                                                         [½]
S,                                                            [½]
xlab = "Time",                                                [½]
ylab = "Survival Function, S(t)",                             [½]
ylim=c(.97,1),                                                [1]
col = c("blue", "red")  ,                                     [1]
main = "Comparison of Kaplan-Meier Estimate for 2 groups--
recipient of vaccine / not recipient of vacccine")            [½]
legend("topright", legend = c("No vaccine", "vaccine") ,      [½]
col = c("blue", "red")                                        [½]
, pch =7)
```



[½]

(iv)

Individuals who have not been administered vaccines have a lower possibility of blood clots within a 28 day period than vaccinated Individuals [1]

Any effect of vaccination occurs, if at all, occurs within the first 15 days [1]

The survival curves cross each other early in the curve [½]

In general, lines crossing each other may mean violation of proportionality in hazard rate. Here, it may be insignificant due to small sample size and/or small number of events occurring [½]

However, analyses does not consider possibility of other factors affecting the results [½]

[Marks available 3½, maximum 2]

(v)

H0: co-morbidity has no significant impact on blood clots along with vaccine indicator and age [½]

H1: co-morbidity has significant impact on blood clots along with vaccine indicator and age [½]

```
> cox_1<-                                        [½]
coxph(                                           [1]
Surv(data_main$Time,data_main$ST)                [1]
~ data_main$Drug*data_main$Age,                  [1]
ties = "breslow")                                [1]
```

```
>  cox_2<-coxph(Surv(data_main$Time,
data_main$ST)~data_main$Drug*data_main$Age*data_main$co
_morbidity, ties = "breslow")                                   [1]
```

Likelihood statistic follows Chi squared distribution with [½]
(7 - 3) degrees of freedom [½]
i.e. 4 degrees of freedom [½]

```
> L1<-cox_1$loglik[2]                                           [1]
>
> L2<-cox_2$loglik[2]                                           [1]
>
> 2 *( L2- L1)                                                  [1]
[1] 16.08793                                                    [½]

qchisq(0.95, 4)                                                 [½]
[1] 9.487729                                                    [½]
```

Conclusion: We reject Ho as the effect is statistically significant [½]
and conclude that co morbidity along with vaccine and age has an impact on blood clots [1]

[Total 34]

This question was generally well answered.

Part (i) on censoring was one of the parts where answers were weaker. Candidates need to apply definitions of censoring types from the Core Reading to the scenario and data set in the question.

In part (ii) candidates are reminded that where the question specifies the name to be given to a dataframe (as is the case here) then answers should use that name not another of the candidate's choosing.

Part (iii) was generally well answered. Some R functions default to plotting the 95% confidence interval around the Kaplan Meier estimate and whilst that was not asked for, candidates who did so were not penalised. Candidates should however be able to generate both Kaplan Meier estimates from the same `survfit` function rather than calculate them separately. Once again it was pleasing to see good titles, labels and legends in the majority of answer scripts.

In part (iv) it is also possible to obtain the likelihood statistics from the `annova()` function in R and marks were awarded if this route is followed. To obtain full marks from this route, the two likelihoods need to be produced within the `annova` output and then the likelihood ratio statistic still needs to be calculated from them and tested against the correct number of chi-squared degrees of freedom. Candidates who simply produced the `annova()` output with no further analysis in their script did not receive full marks.

Q3

(i)

```
data1 = read.csv("CS2B_S22_Qu_3_Data.csv")                    [1]
data1$One = 1                                                  [1]
X=as.matrix(cbind(data1$One, data1$mpg, data1$disp,
data1$qsec))                                                   [1½]
colnames(X) = c("One", "mpg", "disp", "qsec")                  [½]
head(X)                                                        [½]

            One mpg disp qsec
          [1,] 1 21.0 160 16.46
          [2,] 1 21.0 160 17.02
          [3,] 1 22.8 108 18.61
          [4,] 1 21.4 258 19.44
          [5,] 1 18.7 360 17.02
          [6,] 1 18.1 225 20.22                                [½]
```

(ii)

Ridge regression [1]

(iii)

```
ridge_fit =                                                    [½]
     function(lambda, y, X){                                   [1]
     I <- diag(ncol(X))                                        [1]
     beta_lambda <- solve( t(X)%*%X + lambda *I) %*% t(X)%*%y
                                                               [3]
     beta_lambda                                               [½]
}
```

(iv)

```
y <- data1$hp
ridge_fit(2, y, X)                                             [1½]
                  [,1]
   One 31.5838754
   mpg 1.7404198
   disp 0.5248482
   qsec -2.4052353                                             [½]
```

(v)

```
matrix_LAMBDA <- matrix(NA, 10001, 4)                          [1]
for(i in 0:10000){                                             [1]
     lambda <- i/10                                            [1]
     matrix_LAMBDA[i+1, ] <- ridge_fit(lambda, y, X)           [2]
}
```

(vi)

```
head(matrix_LAMBDA)                                            [½]
```

```
             [,1]       [,2]      [,3]       [,4]
[1,] 464.9608 -3.3834403 0.1946797 -16.539799
[2,] 275.8257 -1.1400920 0.3390069 -10.382313
[3,] 196.0660 -0.1945998 0.3998530  -7.784810
[4,] 152.0854  0.3263364 0.4333906  -6.351856
[5,] 124.2195  0.6560558 0.4546286  -5.443409
[6,] 104.9828  0.8833806 0.4692804  -4.815833
```
[½]

(vii)

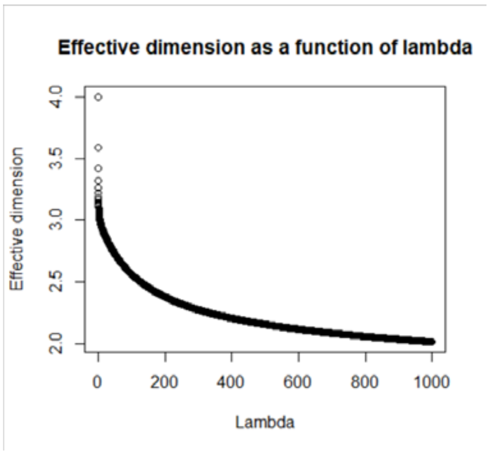
```
dim_fit =                                                     [½]
     function(lambda, X){                                     [½]
     I <- diag(ncol(X))                                       [½]
     H <- X%*% solve( t(X)%*%X + lambda * I )%*%t(X)          [2½]
     edim <- sum(diag(H))                                     [1½]
     edim                                                     [½]
```

(viii)

```
vector_dim <- numeric(10001)                                  [½]
for(i in 0:10000){                                            [1]
     lambda <- i/10                                           [½]
     vector_dim[i+1] <- dim_fit(lambda, X)                    [2]
```

(ix)

```
x <- c(0:10000)/10                                            [½]
plot(                                                         [½]
     x,                                                       [½]
     vector_dim,                                              [½]
xlab= "Lambda",                                               [½]
     ylab="Effective dimension",                              [½]
     main="Effective dimension as a function of lambda")      [½]
```



[½]

(x)

The effective dimension is positive [½]

The maximum dimension is 4 [½]

which corresponds to the number of parameters in the model [½]

The dimension reduces consistently as the value of lambda increases [½] which is as expected since increasing the penalty progressively reduces the effect of the covariates [½]

[Marks available 2½, maximum 2]

[Total 36]

This question was not very well answered with marks awarded generally much lower than those for the other two questions.

The main issue was candidates failing to read the question carefully and starting off with an incorrect model from part (i). The question clearly states that this is a four-parameter model and gives four beta values in the model specification. However because the data set provided had data for three explanatory variables ("mpg, disp and qsec") a large number of candidates simply proceeded to construct a three-parameter model.

In part (i) they omitted the code necessary for the beta_0 parameter. Those candidates were awarded partial marks for part (i) and then were not further penalised for this error in later parts, but candidates are reminded of the importance of reading the question and ensuring that results reflect the model specified not the model that seems to most simply reflect the dataset.

In part (iii) it is acceptable to assign the numerical value to I rather than use the `diag()` *function in R.*

[Paper Total 100]

END OF EXAMINERS' REPORT

Institute
and Faculty
of Actuaries

Beijing

14F China World Office 1 · 1 Jianwai Avenue · Beijing · China 100004
Tel: +86 (10) 6535 0248

Edinburgh

Level 2 · Exchange Crescent · 7 Conference Square · Edinburgh · EH3 8RA
Tel: +44 (0) 131 240 1300

Hong Kong

1803 Tower One · Lippo Centre · 89 Queensway · Hong Kong
Tel: +852 2147 9418

London (registered office)

7th Floor · Holborn Gate · 326-330 High Holborn · London · WC1V 7PP
Tel: +44 (0) 20 7632 2100

Oxford

1st Floor · Belsyre Court · 57 Woodstock Road · Oxford · OX2 6HJ
Tel: +44 (0) 1865 268 200

Singapore

5 Shenton Way · UIC Building · #10-01 · Singapore 068808
Tel: +65 8778 1784

www.actuaries.org.uk