Institute and Faculty of Actuaries

EXAMINERS' REPORT

CS2 - Risk Modelling and Survival Analysis

Core Principles

Paper B

September 2023

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.

For some candidates, this may be their first attempt at answering an examination using open books and online. The Examiners expect all candidates to have a good level of knowledge and understanding of the topics and therefore candidates should not be overly dependent on open book materials. In our experience, candidates that spend too long researching answers in their materials will not be successful either because of time management issues or because they do not properly answer the questions.

Many candidates rely on past exam papers and examiner reports. Great caution must be exercised in doing so because each exam question is unique. As with all professional examinations, it is insufficient to repeat points of principle, formula or other text book works. The examinations are designed to test "higher order" thinking including candidates' ability to apply their knowledge to the facts presented in detail, synthesise and analyse their findings, and present conclusions or advice. Successful candidates concentrate on answering the questions asked rather than repeating their knowledge without application.

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
November 2023

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 are reminded that preparation for R programming assessments is key. The assessment seeks to examine understanding of the syllabus, the use of basic functionality in R programming and an ability to problem solve using this understanding.

The syllabus for CS2 is extensive and candidates are reminded to ensure that they have prepared across the entire syllabus and not just the Survival Models and simple Loss Distributions part. Performance in Time Series was improved this session but average marks in questions on the introduction to Machine Learning techniques continue to be disappointing.

The examiners recommend combining the application of R and problem solving using R into the study of each section of the syllabus rather than considering R programming as a separate element after the 'theory' sections of the syllabus are completed. Practice in problem questions using R across all areas of the CS2 syllabus should be an integral part of examination preparation.

C. Pass Mark

The Pass Mark for this exam was 55.
993 presented themselves and 331 passed.

Solutions for Subject CS2B - September 2023

Q1

(i)

```
Deaths <-                                                    [½]
as.matrix(
read.csv("CS2B_S23_Q1_Deaths.csv"))                          [1]
Deaths <- Deaths[, -1]                                       [1]

Exposures <- as.matrix(read.csv("CS2B_S23_Q1_Exposures.csv"))
Exposures <- Exposures[, -1]                                 [1½]
```

(ii)

```
m_xt <-                                                      [½]
Deaths / Exposures                                           [1½]
```

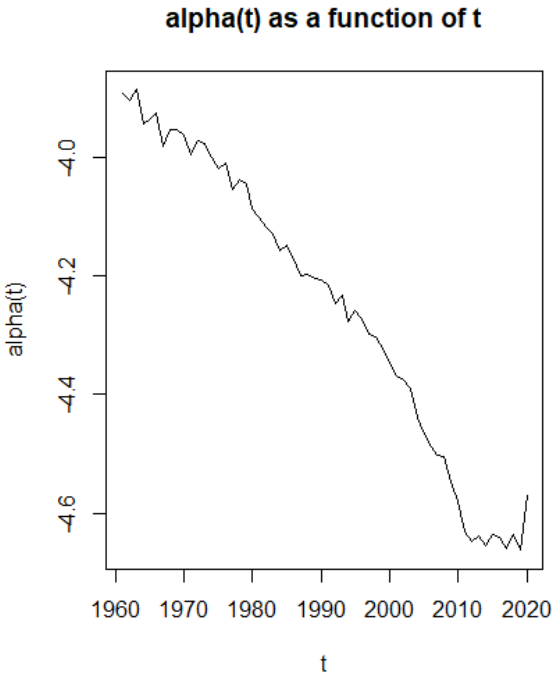
(iii)

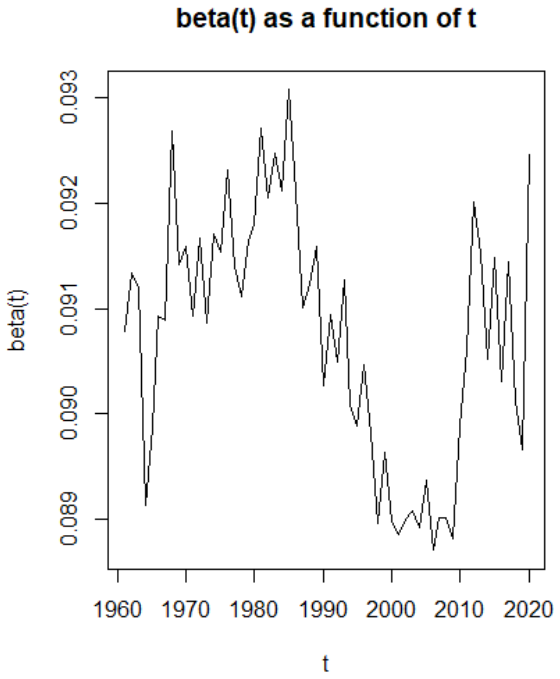
```
Gompertz <-                                                  [½]
matrix(nrow = 60, ncol = 2)                                  [½]

x <- -40:40                                                  [1]
for(j in 1:60)                                               [1]
{Gompertz[j,] <-                                             [½]
lm(                                                          [1]
log(m_xt[,j])                                                [1]
~ x)                                                         [½]
$coefficients}                                               [1]

head(Gompertz)                                               [½]
          [,1]       [,2]
[1,] -3.893014 0.09078393
[2,] -3.904192 0.09133845
[3,] -3.886557 0.09119119
[4,] -3.944754 0.08913961
[5,] -3.935357 0.08987763
[6,] -3.924543 0.09092293                                    [½]
```

(iv)

```
t = 1961:2020                                                [½]
par(mfrow = c(1,2))

plot(                                                        [½]
t,                                                           [½]
Gompertz[,1],                                                [1]
type = "l",                                                  [1]
main = "alpha(t) as a function of t",                        [½]
xlab = "t",                                                  [½]
ylab = "alpha(t)")                                           [½]
```

```
plot(t, Gompertz[,2], type = "l", main = "beta(t) as a
function of t", xlab = "t", ylab = "beta(t)")
```
[1½]





[½]

(v)

alpha(t) exhibits a downward trend as t increases [½]
indicating a trend of improving mortality over time [½]
beta(t) is positive throughout [½]
indicating that mortality increases with age [½]
beta(t) exhibits no consistent trend as t increases [½]
indicating no consistent trend in the rate of increase of mortality with age [½]
beta(t) exhibits greater volatility from year to year than alpha(t) [½]
Both alpha(t) and beta(t) show sharp increases in 2020 due to the COVID-19 pandemic [½]

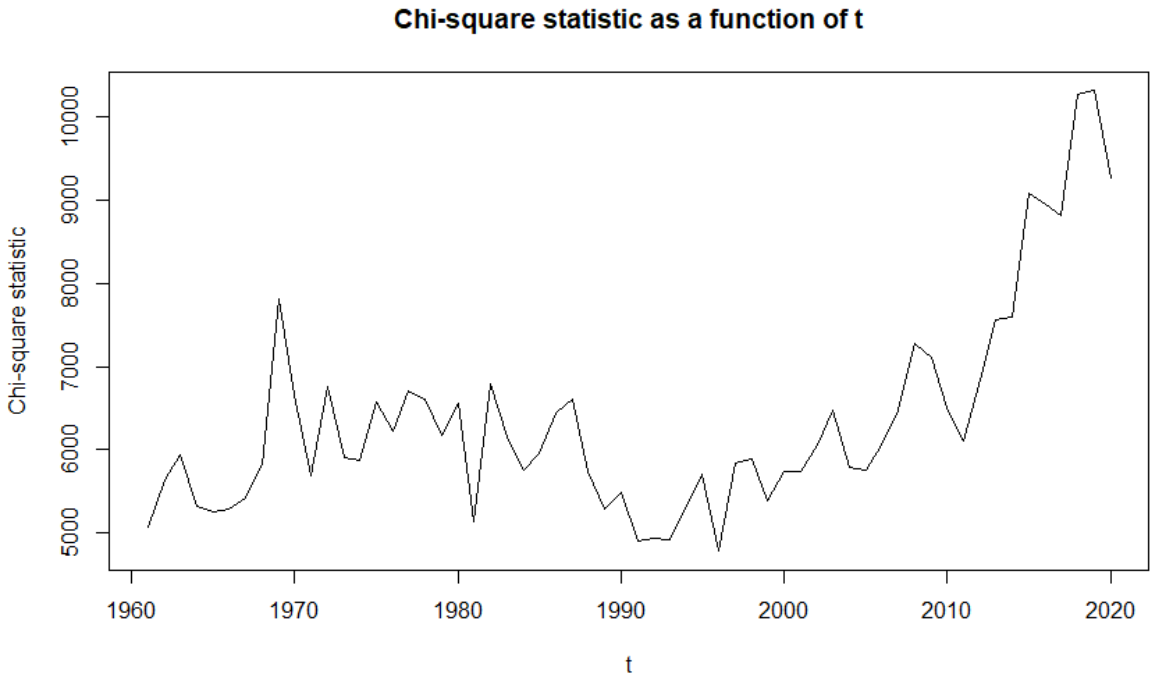
[Marks available 4, maximum 3]

(vi)

```
Expected <- matrix(nrow = 81, ncol = 60)                    [½]
for(i in 1:81)                                              [½]
{for(j in 1:60)                                             [½]
{Expected[i, j] <- Exposures[i, j] * exp(Gompertz[j, 1]
+ Gompertz[j, 2] * (i - mean(1:81)))}}                      [2]

chisq <- numeric(60)                                        [½]
for(i in 1:60)                                              [½]
{chisq[i] <- sum((Deaths[, i] - Expected[, i]) ^ 2
/ Expected[, i])}                                           [2]
```



```
plot(t, chisq, type = "l", main = "Chi-square statistic
as a function of t", xlab = "t", ylab = "Chi-square
statistic")
```
[1]



[½]

(vii)

The number of degrees of freedom to use in the chi-square test is 79 (= 81 - 2) [½]
Calculation of critical value using R or interpolation from Tables [½]
The chi-square statistics in the graph are therefore all highly significant [½]
Alternative models should therefore be investigated to improve the goodness of fit [½]
The chi-square statistic increases in the most recent years [½]
suggesting that there may be a particular need to improve the goodness of fit at the oldest ages [½]
since there will be more lives at those ages in the most recent years [½]
Maximum likelihood estimation should be investigated in place of linear regression [½]
This is equivalent to linear regression with the data points weighted by expected deaths [½]
This will improve the fit at the ages with the highest numbers of deaths [½]
Alternatively, linear regression with weighting by actual deaths could be used [½]

[Marks available 5½, maximum 3]

[Total 35]

This question was reasonably well answered.

In part (iii) setting x as age minus mean(age) works as well as manually setting the range -40 to 40.

Candidates are reminded that relevant graph titles and axis labels are required for full marks in plots like those in parts (iv) and (vi). Where the x axis is age or time this is particularly important.

A number of candidates took a different approach to part (vii) taking the material from the Core Reading on the limitations of the Chi-squared test and suggesting additional statistical tests, and this gained full marks.

With data to 2020 this question includes the period of the COVID-19 pandemic and as in the solution to (v) above the examiners would expect candidates to be able to relate survival modelling results to the pandemic where appropriate.

Q2

(i)

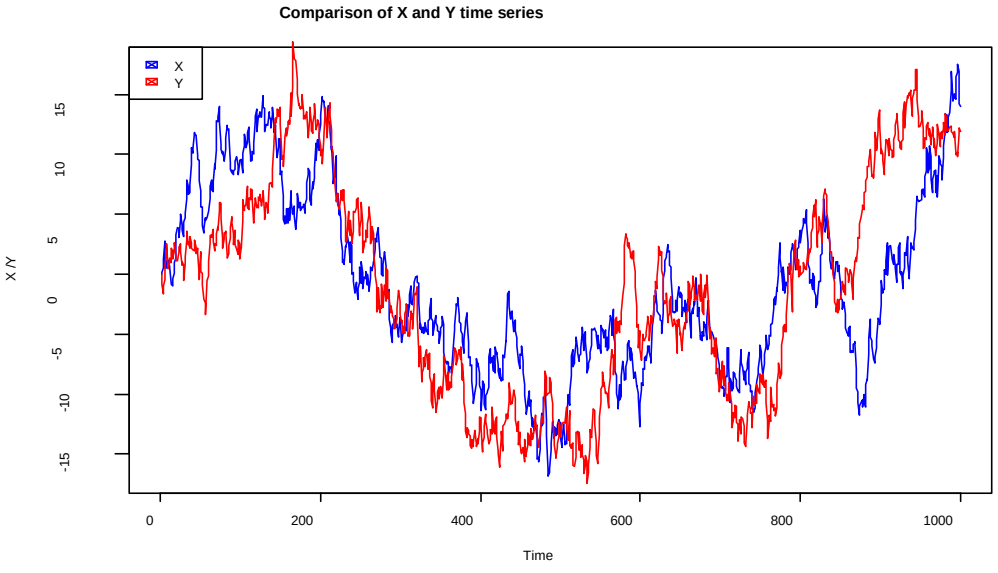
```
data = read.csv(file="CS2B_S23_Q2_Data.csv", head=TRUE)
X<-data[,2]
Y<-data[,3]
head(X,5)
[1] 0.0000000 0.1836433 0.5050501 1.2480126 2.7618762
```

[1] [1] [½] [½]

(ii)

```
plot(data$Time,X, xlab = "Time",  type= "l",  ylab = "X /Y", col = c("blue"), main =
"Comparison of X and Y time series " )
lines(Y, col = c("red" ))
legend("topleft", col = c("blue","red"), legend = c("X", "Y"),pch =7  )
```

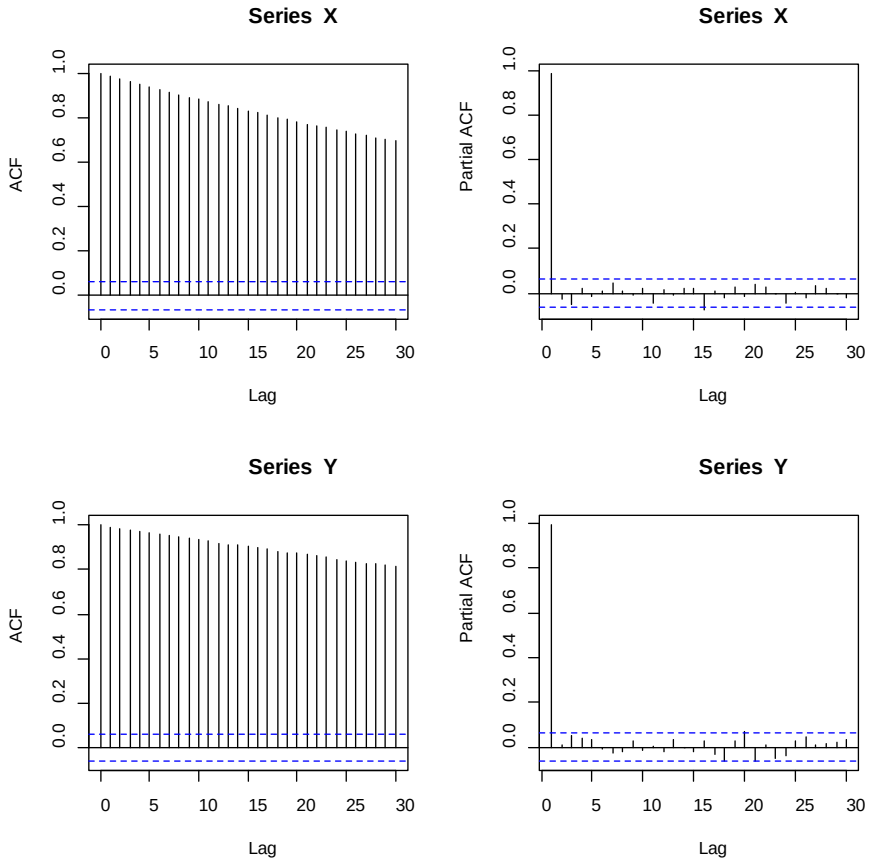
[2½] [1] [1]



[½]

(iii)

```
par(mfrow=c(2,2))
acf(X)
```
[½]
```
pacf(X)
```
[½]
```
acf(Y)
```
[½]
```
pacf(Y)
```
[½]



[1]

(iv)

In the case of a stationary time series, SACFs ultimately must converge towards zero exponentially fast which is not the case for SACFs of both X and Y. [½]

Hence X and Y are not stationary [½]

(v)

```
n = length(X)
X1= rep(NA,n-1)
X2= rep(NA,n-2)
X3= rep(NA,n-3)
```
[2]
```
for(i in 1 :length(X1)){ X1[i] = X[i+1] - X[i]}
```
[2]
```
i= 1
for(j in 1:length(X2))  {X2[j] = X1[j+1] - X1[j]}
```
[1]
```
i= 1
for(i in 1 :length(X3)){ X3[i] = X2[i+1] - X2[i]}
```
[1]
```
var_sr<- c(var(X),var(X1),var(X2),var(X3))
```
[2]

var_sr
1] 54.376104 1.055890 2.072943 6.206855 [½]
n = length(Y)
Y1= rep(NA,n-1)
Y2= rep(NA,n-2)
Y3= rep(NA,n-3) [½]
for(i in 1 :length(Y1)){ Y1[i] = Y[i+1] - Y[i]} [½]
i= 1
for(j in 1:length(Y2)) {Y2[j] = Y1[j+1] - Y1[j]} [½]
i= 1
for(i in 1 :length(Y3)){ Y3[i] = Y2[i+1] - Y2[i]} [½]
var_LT<- c(var(Y),var(Y1),var(Y2),var(Y3)) [1]
var_LT
[1] 79.979362 1.155266 2.357361 6.988118 [½]

It is normally the case that sample variance first decreases with 'd' until stationarity is achieved and then starts to increase. [1]
Sample variance is lowest at d = 1 under X and Y. [½]
Hence the X and Y have to be differenced once. [½]

```
> var(X)
[1] 54.37610384
> var(diff(X))
[1] 1.055890354
> var(diff(X, differences = 2))
[1] 2.072942932
> var(diff(X, differences = 3))
[1] 6.20685507
> var(Y)
[1] 79.97936178
> var(diff(Y))
[1] 1.155266085
> var(diff(Y, differences = 2))
[1] 2.357361092
> var(diff(Y, differences = 3))
[1] 6.988117921
```

(vi)
M<- lm(Y ~X) [1½]
summary(M) [1]
or
summary(M)$coefficients[1]; summary(M)$coefficients[2] [1]

Call:
lm(formula = Y ~ X)

Residuals:
Min 1Q Median 3Q Max
-11.689 -4.428 -1.084 3.414 18.335


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.27658    0.19059  -1.451    0.147
X            0.90145    0.02568  35.101   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.986 on 998 degrees of freedom
Multiple R-squared:  0.5525,    Adjusted R-squared:  0.552
F-statistic:  1232 on 1 and 998 DF,  p-value: < 2.2e-16
```

Therefore, a= -0.27658 , b = 0.90145 [½]

[Marks available 4, maximum 3]

(vii)

Y and X are I (1) time series.	[1]
Substituting Y into Z_t gives Z_t = a+ et	[1]
Zt becomes stationary as 'et' is the simplest case of stationarity	[1]
Hence Zt is co-integrated	[1]
with co-integration vector [1, -0.90145]	

(viii)

Residuals process 'et' could still be stationary even though it is not a white noise	[1]
Further analysis has to be done to check whether Zt is stationary	[1]
'et' not being a white noise would mean that the regression fit is not good	[1]
meaning relation between Y and X may not be linear	[1]
requiring a revised co-integrating relation	[1]
resulting in changing the co-integrating vector	[1]
It is also possible that there is no co-integrating relation between X and Y	[1]
Therefore, we cannot say anything on the inference	[1]

[Marks available 8, maximum 3]

[Total 35]

This question was well answered.

A number of candidates produced a scatter plot in part (ii) and whilst a line graph is more natural for time series data, a fully labelled scatter plot gained full marks.

In part (v) setting up a "for-loop" in R, whilst perhaps the most efficient method, was not required here given the relatively small number of variance calculations needed and candidates that repeated their calculation five times without the loop gained full credit. Similarly, alternative methods that used the diff() function were acceptable.

The part where answers were weakest was (viii) with many candidates assuming that because stationarity followed from the scenario in (vii) it did not in part (viii) whereas the correct answer is more nuanced.

Q3

(i)

dataset1 = read.csv("CS2B_S23_Q3_Pensioners.csv") [½]

tail(dataset1, 8) [1]

	Age	PopulationSize	DeathCount
29	83	407	98
30	84	388	107
31	85	356	111
32	86	325	114
33	87	298	118
34	88	269	122
35	89	247	124
36	90	234	128

[½]

(ii)

dataset1$logRate = log(dataset1$DeathCount/dataset1$PopulationSize) [1]

(iii)

model = lm(logRate ~ Age, data=dataset1) [1½]

round(coef(model), 5) [1]

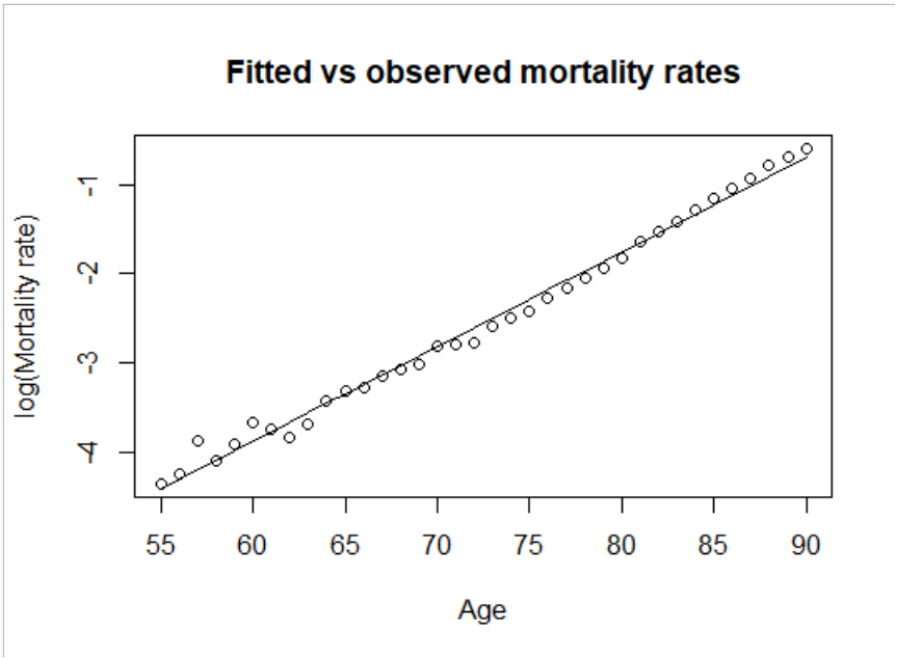
```
(Intercept)         Age
  -10.24444     0.10604
```

The intercept is -10.24444 and the slope is 0.10604 [½]

(Alternative with use of the `nlm()` *function is also acceptable)*

(iv)

```
plot(                                                    [½]
  dataset1$logRate~dataset1$Age,                         [1]
  xlab = "Age",                                          [½]
  ylab="log(Mortality rate)",                            [½]
  main= "Fitted vs observed mortality rates"             [½]
lines( model$fitted.values ~ dataset1$Age    )           [1½]
```

[½]

[Marks available 5, maximum 4]

(v)

Comment on this graph:

Mortality rate increases exponentially with age [½]
Log mortality rates increase linearly with age [½]
There is greater variation at the youngest ages [½]
The model seems to fit the data well overall [½]
but there is some systematic bias for higher ages [1]
mortality is underestimated above age 80 [½]
there may not be one linear model suitable for the full age range [½]

[Marks available 4, maximum 1]

(vi)

output_rss = [½]
data.frame(p = 1:5, rss = rep(NA, 5)) [1]

```
poly_data  <- data.frame(  logRate = dataset1$logRate,
x1 = dataset1$Age,
x2 = dataset1$Age^2,
x3 = dataset1$Age^3,
x4 = dataset1$Age^4,
x5 = dataset1$Age^5 )
```

[½] [½] [½] [½]

```
for(p in 1:5){
 M_p  = lm( logRate ~ ., data= poly_data[ ,c(1:(p+1))] )
  output_rss[p, 2] <- sum( (poly_data$logRate - M_p$fitted)^2 )
```

[½] [1½] [1½]

```
output_rss
```

[½]

```
p              rss
```


1	1	0.3801852
2	2	0.1776206
3	3	0.1773628
4	4	0.1648706
5	5	0.1645165 [½]

(vii)
The best model is M_5 [½]
Because it has the lowest RSS [½]

(viii)
```
p   <- 5
M_5 <- lm( logRate ~ ., data= poly_data[ ,c(1:(p+1))] )
```
[½]

```
Ages_for_prediction <- 91:110
```
[½]
```
data_for_prediction  <- data.frame(
x1 = Ages_for_prediction,
x2 = Ages_for_prediction^2,
```
[½]
```
x3 = Ages_for_prediction^3,
```
[½]
```
x4 = Ages_for_prediction^4,
```
[½]
```
x5 = Ages_for_prediction^5 )
```
[½]

```
forecast_91_110 <- predict( M_5, data_for_prediction )
```
[3]
forecast_91_110 [½]
-0.4935734 -0.4079749 -0.3346362 -0.2756383 -0.2332278 -0.2098221 -0.2080146 -0.2305803 -0.2804808 -0.3608699 -0.4750987 -0.6267210 -0.8194986 -1.0574067 -1.3446389 -1.6856130 -2.0849760 -2.5476094 -3.0786347 -3.6834185 [½]

(ix)
This model fits the data better compared to M_1 [½]
However, RSS has selected the most complex model among the five [½]
Because RSS does not have a penalty terms against model complexity [½]
The forecast trend is counter-intuitive [1]
Because the risk of dying increases as people age [½]
Using a penalise criteria such as AIC will probably lead to better result [½]
higher order polynomial are often unstable and unsuitable for forecasting [½]
this might be an example of model over-fitting [½]

[Marks available 4½, maximum 2]

[Total 29]

This question was not very well answered.

Parts (i) to (iv) are the application of simple linear regression modelling in R to a survival analysis scenario and are all quite straightforward.

The "machine learning" or "data analytics" component to this question comes in part (vi) where a number of approaches secured full marks including the one presented above, a variation that fitted regression models sequentially without the loop, or use of the I() or poly() functions in R.

Parts (vi), (viii) and (ix) were generally poorly answered and candidates are reminded of the importance of being able to apply regression modelling techniques to problem solving 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