INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2021

Subject CS2A – Risk Modelling and Survival Analysis Core Principles Paper A

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.

Paul Nicholas
Chair of the Board of Examiners
July 2021

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

1. The aim of the Risk Modelling and Survival Analysis Core Principles 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. Some of the questions in this paper admit alternative solutions from those presented in this report, or different ways in which the provided answer can be determined. All mathematically correct and valid alternative solutions received credit as appropriate.

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

4. 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.

5. Candidates are advised to take careful note of all instructions that are provided with the exam in order to maximise their performance in future CS2A examinations. The instructions applicable to this diet can be found at the beginning of the solutions contained within this document.

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

1. Performance was generally satisfactory. Most candidates demonstrated a reasonable understanding and application of core topics in mathematical and statistical modelling techniques.

2. Topics that were not particularly well-answered in this paper include Stochastic Processes (e.g. Q3), Markov Jump Processes (e.g. Q4) and Machine Learning (e.g. Q8), despite these being examined through reasonably straightforward application questions. Candidates are reminded that it is very important to be familiar with all aspects of the syllabus.

3. It is important that candidates heed all of the instructions provided with the examination paper. A number of candidates lost marks because they did not include workings for numerical questions despite being forewarned about this in the instructions.

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 Pass Mark was 55.
1,422 candidates presented themselves and 480 passed.

Solutions for Subject CS2 Paper A April 2021

Please note the following principles apply to the CS2A solutions. These principles were set out in the instructions provided to candidates along with the examination paper:

1. Candidates **MUST** include typed workings, in addition to their typed answers, in the Word document for all numerical questions. Candidates using another software package to aid with calculations **MUST** ensure that all calculations appear in full in the Word document to ensure that they receive full marks. If full workings are not displayed then examiners will not be able to assess how the answer was determined, and full marks may not be awarded.

2. Candidates should type their 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.

3. Your Word document **MUST NOT** contain links to any other documents.

Q1

(i)

The required probability is

*C*_F(0.5, 0.5)
= - ln (1 + [(e^-0.5 - 1)^2]/[e^-1 - 1]) [1½]
= 0.280930 [½]

(ii)

The required probability is

*C*_F(0.5, 0.5)
= - (1/0.1) * ln (1 + [(e^-0.05 - 1)^2]/[e^-0.1 - 1]) [½]
= 0.253125 [½]

(iii)

The required probability is

*C*_Product (0.5, 0.5)
= *u* * *v* = 0.5 * 0.5 [½]
= 0.25 [½]

(iv)

The probabilities for the Frank copula in parts (i) and (ii) are higher than for the product copula in part (iii) [½]
because the Frank copula exhibits positive dependence [½]
As α approaches zero, the level of dependence approaches zero [½]
and so, the probability approaches that under the product copula. [½]

[Total 6]

Parts (i) and (ii) were well answered, although some candidates lost marks for not showing sufficient workings. To receive full marks candidates needed to include the correct formula, the correct value for u and v, the correct value of alpha and the numerical solution in their answer script.

Part (iii) was very well answered.

Part (iv) was very poorly answered. Many candidates included comments on tail dependency which were not relevant to the question. For candidates to receive full marks their answers needed to refer to ***both*** *the sign* ***and*** *the level of dependence, as specified in the question. Candidates are reminded to read the question carefully and ensure that their answer reflects what is being asked.*

Q2

(i)

E(X_t) = mu + E(e_t) + beta_1 * E(e_t-1) + beta_2 * E(e_t-2)
= mu + 0 + beta_1 * 0 + beta_2 * 0 [½]
= mu [½]
Var(X_t) = Var(e_t) + ((beta_1)^2) * Var(e_t-1) + ((beta_2)^2) * Var(e_t-2) [1]
= (1 + (beta_1)^2 + (beta_2)^2) * sigma^2 [1]

(ii)

gamma_0 = Var(X_t) = (1 + (beta_1)^2 + (beta_2)^2) * sigma^2 [½]
gamma_1 = Cov(beta_1 * e_t-1, e_t-1) + Cov(beta_2 * e_t-2, beta_1 * e_t-2) [1]

as covariances of all other cross-multiplied terms are 0
= (beta_1 + beta_1 * beta_2) * sigma^2 [½]
gamma_2 = Cov(beta_2 * e_t-2, e_t-2) [1]

as covariances of all other cross-multiplied terms are 0
= beta_2 * sigma^2 [½]
gamma_k = 0 for k > 2. [½]

[Total 7]

.

Parts (i) and (ii) were well answered, although some candidates lost marks for not showing sufficient workings.

To receive full marks candidates needed to be clear that the expected value and variance in part (i) had been calculated as the sum of the expected values and variances of the individual terms, and similarly for the covariances in part (ii).

A common alternative answer to part (i) was to derive the value of Var(X_t) *using covariances. This was awarded full credit provided sufficient workings were shown.*

Q3

(i)

EITHER:

Y_t has independent increments [1]
(because the random variables X_i are independent)
Y_t is therefore a Markov process [½]
because a process with independent increments possesses the Markov property [½]

OR:

Y_t = Y_(t-1) + X_t [½]
Therefore, the future development of the process Y_t, can be predicted from its present state alone without any reference to its history [1]
therefore, by definition, Y_t is a Markov process [½]

(ii)

Corr(Y_t, Y_(t+20)) = Cov(Y_t, Y_(t+20)) / sqrt (Var(Y_t) * Var(Y_(t+20))) [1]
Now,
Cov(Y_t, Y_(t+20))
= Cov(X_1 + … + X_t, X_1 + … + X_t + … + X_(t+20)) [½]
= Cov(X_1, X_1) + … + Cov (X_t, X_t) [1]
Since Cov(X_j, X_k) = 0 for j not= k as X_i are independent
= t * sigma^2 [½]
since X_i are identically distributed
Var(Y_t) = t * sigma^2 [½]
Var(Y_(t+20)) = (t + 20) * sigma^2 [½]

Therefore,
Corr(Y_t, Y_(t+20)) = (t * sigma^2) / sqrt (t * (t + 20) * sigma^4) [½]
= sqrt (t / (t + 20)) [½]

(iii)

Since Corr(Y_t, Y_(t+20)) = sqrt (t / (t + 20)) = sqrt (1 / (1 + 20 / t))
we can see that the value of the correlation coefficient increases as t increases [1]
and tends to 1 as t tends to infinity. [1]

[Total 9]

Part (i) was well answered.

Part (ii) was poorly answered with many candidates failing to quote the correct general formula for Corr(Y_t, Y_(t+20)). *Additionally, a number of candidates incorrectly thought that the first term in the stochastic process Y_(t+20) was X*_21 *rather than X*_1 *and therefore derived the wrong values for* Cov(Y_t, Y_(t+20)) and Var(Y_(t+20)).

Part (iii) was surprisingly poorly answered. Where candidates derived the wrong value in part (ii), credit was still awarded in part (iii) for a correct interpretation of how the derived correlation coefficient behaved as t increased.

Q4

(i)

The formula for the rate of transition to the recovered state produces recovery rates that decrease with increasing sickness duration [½]

which is usually observed in practice [½]

On the other hand, for some sicknesses the recovery rate may not decrease with sickness duration and so in this respect the model may not be reasonable [½]

The formula for the rate of transition to the dead state produces mortality rates that increase with increasing sickness duration [½]

which is consistent with a condition that progressively deteriorates over time [½]

On the other hand, for some sicknesses the mortality rate may not increase with sickness duration and so in this respect the model may not be reasonable [½]

It is necessary to include the parameter c to ensure that there is some positive level of mortality in the early stages of sickness, which will include mortality from causes other than the condition under consideration [1]

For a very long-term condition the formula for the rate of transition to the dead state would not be suitable since the mortality rates are bounded above by $a + c$, whereas in practice mortality increases without limit with age [1]

The model does not explicitly allow for other covariates (e.g. gender, smoker status etc.) in the modelling of transition rates [½]

although the use of the a, b and c parameters perhaps allows for some distinction between individuals [½]

The Markov property may not be a reasonable assumption as in reality the transition rates may depend on the previous history of the process rather than just the present state alone (e.g. the number of times the individual has been in the sick state may impact the transition rates) [½]

Depending on the condition, the fact that the model does not appear to allow for transition back to the sick state after reaching the recovered state may or may not be appropriate. [½]

Depending on the condition, duration dependent transition rates may or may not be appropriate. [½]

[Marks available 7½, maximum 2]

(ii)

Prob(Recovery)

= INT(0, infinity):[Prob (Remains in Sick State from time = 0 to t)
* (Transition Rate from Sick to Recovery at time t)] dt [1]

Now,

Prob (Remains in Sick State from time = 0 to t) = exp (-(a + c)t) [1]

since the sum of the recovery and death transition rates at time t is $a + c$. [1]

Therefore,

Prob (Recovery) = INT(0, infinity):[exp (-(a + c)t) * $a$exp (-bt)] [1]

= INT(0, infinity):[$a$exp (-(a + b + c)t)] [1]

= [-$a$exp (-(a + b + c)t) / (a + b + c)](0:infinity) [1]

= a / (a + b + c). [1]

[Total 9]

Part (i) was poorly answered, with many candidates commenting on only a single aspect of the reasonableness of the model. Alternative comments that were clear, distinct and relevant to the context of the question were also awarded credit.

Part (ii) was very poorly answered. Candidates are reminded that where the command verb is 'Demonstrate', it is particularly important to show sufficiently clear workings to indicate that a valid method has been used.

Q5

(i)

For Town A we have, using the census formula, the force of mortality is:

= 63 / (½(3,000 + 3,300)) = 0.02 [1]

For Town B we have:

= 26 / (½(1,770 + 1,674)) = 0.0151 [1]

(ii)

<u>ASSUMPTIONS (Maximum = 1 mark)</u>

Assume that the ratio of the smoker force of mortality to that of the non-smoker in Towns A and B is the same as the national average. [1]

Assume that the ratio of the smoker force of mortality to that of the non-smoker in Towns A and B remains constant throughout the year. [½]

Assume that the policies in force of both towns varies linearly between census dates. [½]

[Marks available 2, maximum 1]

<u>THEN EITHER:</u>

Let the force of mortality for smokers in Town A be mu_s(A), and that for non-smokers be mu_n(A).

We therefore have 0.5 * mu_s(A) + 0.5 * mu_n(A) = 0.02 [1]

mu_s(A) = 1.5 * mu_n(A) [½]

and hence:

0.5 * 1.5 * mu_n(A) + 0.5 * mu_n(A) = 0.02

mu_n(A) = 0.02 / 1.25 = 0.016 [½]

mu_s(A) = 1.5 * 0.016 = 0.024 [½]

Let the force of mortality for smokers in Town B be mu_s(B), and that for non-smokers be mu_n(B).

We therefore have 0.2 * mu_s(B) + 0.8 * mu_n(B) = 0.0151 [1]

mu_s(B) = 1.5 * mu_n(B) [½]

and hence:

0.2 * 1.5 * mu_n(B) + 0.8 * mu_n(B) = 0.0151

mu_n(B) = 0.0151 / 1.1 = 0.01373 [½]

mu_s(B) = 1.5 * 0.01373 = 0.02059 [½]

<u>OR:</u>

Let the force of mortality for smokers in Town A be mu_s(A), and that for non-smokers be mu_n(A).

Let the number of deaths for smokers in Town A be *d_s(A)* and that for non-smokers be *d_n(A)*.

Let the exposed to risk for smokers in Town A be *E_s(A)* and that for non-smokers be *E_n(A)*.

Then,

E_n(A) = ½(3,000 + 3,300) * 50% = 1,575 [½]

E_s(A) = ½(3,000 + 3,300) * 50% = 1,575 [½]

d_n(A) = 63 * (50% / (50% + 50%*1.5)) = 25.2 [½]

d_s(A) = 63 * ((50% * 1.5) / (50% + 50%*1.5)) = 37.8 [½]

Therefore,

mu_n(A) = 25.2 / 1,575 = 0.016

mu_s(A) = 37.8 / 1,575 = 0.024 [½]

Similarly for Town B,

E_n(B) = ½(1,770 + 1,674) * 80% = 1,377.6 [½]

E_s(B) = ½(1,770 + 1,674) * 20% = 344.4 [½]

d_n(B) = 26 * (80% / (80% + 20%*1.5)) = 18.9091 [½]

d_s(B) = 26 * ((20% * 1.5) / (80% + 20%*1.5)) = 7.0909 [½]

Therefore,

mu_n(B) = 18.9091 / 1,377.6 = 0.01373

mu_s(B) = 7.0909 / 344.4 = 0.02059 [½]

(iii)

The company would do better to vary the premiums on the basis of geographical area [½]

as it is clear that death rates in Town A for both smokers and non-smokers are higher than those in Town B [½]

If the company does not differentiate its prices on the basis of geographical area, it may lose business in Town B to a rival company which does differentiate [½]

This could explain why the company has fewer policies in Town B [1]

Also, in Town A it may attract new business from rival companies, but will under-price the product and hence risk becoming insolvent [½]

There is relatively little data, so it could adopt a "wait and see" approach [½]

The higher premium for smokers is appropriate [½]

1.5 times the death rate will not translate as 1.5 times the premium due to e.g., company renewal expenses being similar for smokers and non-smokers. The difference may be relatively small, although it is a 25-year term assurance so it may be reasonably significant [1]

The ratio of the mortality rates between smokers and non-smokers may not follow the national average and so a revised pricing structure may be more appropriate [1]

The company make take the view that the differences in mortality rates between Towns A and B are not statistically significant and hence that their pricing approach is justified. [½]

There may be other practical or commercial reasons why the company may not want to differentiate in price between towns (e.g. administrative ease) [½]

[Marks available 7, maximum 3]

[Total 11]

Part (i) was the best answered question in the whole paper.

Part (ii) was well answered. The most common error in part (ii) was to assume that the number of deaths, rather than the force of mortality, was 50% higher for smokers than for non-smokers. This approach produced the correct results for Town A but not for Town B. Many candidates also lost marks in part (ii) for not clearly stating their assumptions. Candidates are reminded to read the question carefully.

Part (iii) was poorly answered. Better performing candidates tended to comment both on the potential need to vary the premiums by geographical area and on the appropriateness of the higher premium for smokers, since both are relevant to the question. Alternative comments that were clear, distinct and relevant to the context of the question were also awarded credit.

Q6

Let the individual total claim costs be denoted by X.
Then $X = Y + Z$ where Y is the cost of the claim and Z is the claim handling expense.

$E(X) = E(Y) + E(Z) = 120 + 0.2 * 30 = 126$ [1]

<u>THEN EITHER:</u>

$E(X^2) = E(Y^2) + 2E(YZ) + E(Z^2)$ [½]

Using the independence of Y and Z we have that $E(YZ) = E(Y)E(Z)$

$E(X^2) = E(Y^2) + 2E(Y)E(Z) + E(Z^2)$ [½]

Now:

$E(Y^2) = \mathrm{Var}(Y) + (E(Y))^2 = 2 * E^2(Y)$

$= 2 * 120^2 = 28{,}800$ [1]

$E(Z^2) = 0.2 * 30^2 = 180$ [1]

So $E(X^2) = 28{,}800 + 2 * 120 * 6 + 180 = 30{,}420 = 174.41^2$ [1]

Using the Normal approximation for the total claim amounts S we have for n policies

$E(S) = 0.4 * 126 * n = 50.4 * n$. [1]

$\mathrm{Var}(S) = 0.4 * n * 174.41^2 = 12{,}168 * n = n * 110.31^2$. [1]

<u>OR:</u>

$\mathrm{Var}(Y) = 120^2 = 14{,}400$ [1]

$\mathrm{Var}(Z) = 0.2 * 30^2 - 6^2 = 144$ [1]

$\mathrm{Var}(X) = \mathrm{Var}(Y) + \mathrm{Var}(Z)$ [1]

(using the independence of Y and Z)

$= 14{,}400 + 144$

$= 14{,}544$ [1]

$E(S) = 0.4 * 126 * n = 50.4 * n$. [1]

$\mathrm{Var}(S) = 0.4 * n * 14{,}544 + 0.4 * n * 126^2 = 12{,}168 * n = n * 110.31^2$. [1]

The corresponding premium income is $60 * n$. [½]

We need to find the smallest n for which:

$P(N(50.4 * n, n * 110.31^2) < 60 * n) >= 0.99$ [1]

i.e. $P(N(0,1) < (9.6 * n) / (110.31 * \mathrm{sqrt}(n))) >= 0.99$

$(9.6 * n) / (110.31 * \mathrm{sqrt}(n)) >= 2.326348$ [1]

sqrt(*n*) >= 2.326348 * (110.31 / 9.6) = 26.73087
n >= 26.73087^2 = 714.5393 [½]
So, the minimum number is *n* = 715 policies. [1]

[Total 11]

This question was well answered in general. The most common errors were:

Treating expenses as a deterministic addition to the claim amounts, rather than 30 with a probability of 20% and zero with a probability of 80% as per the question
Omitting expenses entirely
Omitting the factor of n in E(S) and Var(S).

Q7

(i)

10_p_65 = 1_p_65 * 7_p_66 * 2_p_73
= exp(-INT(0,1): 0.040 *dt*) * exp(-INT(0,7): 0.005 *dt*)
* exp(-INT(0,2): 0.080 *dt*) [1]
= exp(-(0.040 + 7 * 0.005 + 2 * 0.080)) [1½]
= 0.790571 (to 6 dp's) [½]

(ii)

EITHER:

The suggested formula follows Makeham's Law of mortality
*mu_x = A + B * c^x*
where:
A = 0.0020291
B = 0.0001000
c = 1.0793496

Using the formulae on page 32 of the Tables,
10_p_65 = (*s*^10) * *g*^((*c*^65)*(*c*^10-1)) [1]
where:
s = exp(-0.0020291) [1]
g = exp(-0.0001000 / log (1.0793496)) [1]
c = 1.0793496 [½]
Hence,
10_p_65 = 0.790571 (to 6 dp's) [½]

OR:

10_p_65
= exp(-INT(65,75): (0.0020291 + 0.0001 * 1.0793496^*x*) *dx*) [1]
= exp(-[0.0020291 * *x* + (0.0001 / ln(1.0793496)) * 1.0793496^*x*]:(65,75)) [2]
= exp(-(0.0020291 * 10 + 0.0013096 * (1.0793496^75 - 1.0793496^65)))
= exp(-0.235000) [½]
= 0.790571 (to 6 dp's) [½]

(iii)

The matching probabilities show that the average mortality rate over the age range 65-75 is approximately the same [½]

A number of tests would need to be performed (e.g. a chi-square goodness-of-fit test, a signs test, a grouping of signs test etc.) before a firm conclusion could be reached about whether the suggested model was a good fit for the observed mortality rates. [1]

However, from inspection, the individual mortality rates at each age differ quite significantly over the period: [½]

Age	*Force of mortality (p.a.)* *Observed*	*Force of mortality at start of 'age' year* *Graduated*
65	0.040	0.016
66	0.005	0.017
67	0.005	0.019
68	0.005	0.020
69	0.005	0.021
70	0.005	0.023
71	0.005	0.025
72	0.005	0.026
73	0.080	0.028
74	0.080	0.030

[1]

In particular, the mortality rates derived from the suggested model increase smoothly with age [½]

and this is in line with what we expect and desire in practice [½]

In contrast, the mortality rates derived from the observed lives do not increase with age - they decrease from age 65 to age 66, remain flat until age 73 where they then increase again and remain flat [½]

As the study only lasted for 10 years, it does not produce any reliable mortality rates beyond age 75 and so it is very difficult to determine how good a fit the suggested model would be beyond age 75 [½]

The observed rates appear to be atypical, perhaps suggesting that it might be difficult to find a suitable standard table to graduate against if the suggested model does not produce a good fit as expected [½]

It may be better to try to use another parametric formula to fit to the observed rates if the suggested model proves to be inadequate [½]

The atypical observed rates perhaps suggest that the sample size for the study was too small to provide reliable observed rates [½]

The observed rate of 0.005 for durations 1-8 may be incorrectly calculated and should be checked for data / calculation errors (e.g. a value of 0.05 may be more reasonable). [½]

The analyst could look to other comparable studies to obtain more data [½]

[Marks available 7½, maximum 4]

[Total 11]

Part (i) was very well answered, although, here and in part (ii), some candidates did not quote their answers to six decimal places. Candidates are reminded to read the question carefully.

Answers to part (ii) were satisfactory. A common mistake was to multiply probabilities of surviving successive age ranges [x,x+1] calculated as exp(-mu_x). Candidates using this approach received very few marks since the definition of mu_x treats x as a continuous variable. A smaller number of candidates multiplied probabilities calculated as exp(-½(mu_x + mu_x+1)). This is more appropriate but did not gain full marks since it introduces an approximation error compared with the exact calculation.

Part (iii) was very poorly answered. Marks were not awarded for comments on the advantages and disadvantages of Makeham's Law in general, without reference to the specific scenario in the question. Alternative comments that were clear, distinct and relevant to the context of the question were also awarded credit.

Q8

(i)

Let *Y_i* represent the actual medical costs of the *i*th policyholder.
Then, the residual sum of squares (RSS) in region *S_k* is given by:
RSS (*alpha_k*) = sum (all *i* that are members of set *S_k*):[(*Y_i* - *alpha_k*)^2] [1]
Setting *d* RSS/*d alpha_k* = 0 gives: [1]
(*alpha_k*)^hat = sum (all *i* that are members of set *S_k*):[*Y_i*] / *n_k* [½]
where *n_k* = number of policyholders in region *S_k.* [½]
and *k* = 1, …, 5.

(ii)

The regions assigned to each of the nine policyholders in the training data set are as follows:

Age	*BMI*	*Actual Medical Costs (£)*	*Region Assigned*
50	26.3	27,809	*S*_3
48	28.0	23,568	*S*_3
28	24.0	17,663	*S*_2
45	22.9	21,099	*S*_1
59	29.8	30,185	*S*_5
56	20.0	22,413	*S*_4
38	19.3	15,821	*S*_1
61	29.9	30,942	*S*_5
34	25.3	18,972	*S*_2

[3]

Therefore,

(*alpha*_1)^hat = (£21,099 + £15,821) / 2 = £18,460.00
(*alpha*_2)^hat = (£17,663 + £18,972) / 2 = £18,317.50
(*alpha*_3)^hat = (£27,809 + £23,568) / 2 = £25,688.50
(*alpha*_4)^hat = £22,413.00
(*alpha*_5)^hat = (£30,185 + £30,942) / 2 = £30,563.50

[1]

(iii)
Predicted medical costs for each of the three policyholders in the test data set, using the recursive binary decision tree model, are as follows:

Policyholder Reference Number	*Age*	*BMI*	*Region Assigned*	*Predicted Medical Costs (£)*
1	57	27.9	*S*_4	22,413.00
2	60	28.1	*S*_5	30,563.50
3	40	21.1	*S*_1	18,460.00

[2]

(iv)
Predicted medical costs for each of the three policyholders in the test data set, using the linear regression model, are as follows:

Policyholder Reference Number	*Age*	*BMI*	*Predicted Medical Costs (£)*
1	57	27.9	-8500 + 304 * 57 + 698 * 27.9 = 28,302.20
2	60	28.1	-8500 + 304 * 60 + 698 * 28.1 = 29,353.80
3	40	21.1	-8500 + 304 * 40 + 698 * 21.1 = 18,387.80

[2]

(v)
A comparison of the predicted results of the two machine learning models is set out in the table below:

Policyholder Reference Number	*Actual Medical Costs (£)*	*Predicted Medical Costs using Recursive Binary Decision Tree model (£)*	*Predicted Medical Costs using Linear Regression model (£)*
1	27,768.00	22,413.00	28,302.20
2	30,023.00	30,563.50	29,353.80
3	18,524.00	18,460.00	18,387.80

The recursive binary decision tree model provided reasonable predictions for policyholders 2 and 3 in the test data set but it was reasonably far out for policyholder 1 [½]
This may be due, in part, to the fact that there was only one policyholder in region *S*_4 in the original training data set [½]

In comparison, the linear regression model provided reasonable predictions for all three policyholders in the test data set [½]

Even though, the recursive binary decision tree model was a closer predictor for policyholders 2 and 3 it would appear that the linear regression model is a better overall predictor for the data in the test data set [½]

Different values of the a_1, a_2, b_1, and b_2 constants in the recursive binary decision tree model should be considered in order to improve the model's performance [½]

Extra nodes in the recursive binary decision tree model should also be considered in order to try to improve the model's performance [½]

Both the training and test data sets are small and so further testing on larger data sets would be required before more robust conclusions could be drawn [1]

Additionally, the linear regression model produces smooth results [½]

whereas, the recursive binary decision tree model exhibits jumps from one region to another [½]

The linear regression model assumes that the dependence of medical costs on age and BMI is linear, whereas the recursive binary decision tree model does not make this assumption [½]

The recursive binary decision tree model allows for interaction between age and BMI, whereas the linear regression model does not [½]

[Marks available 6, maximum 4]

[Total 15]

Most candidates did not attempt this question and therefore this was the least well-answered question on the whole paper. Candidates are reminded of the need to be familiar with all aspects of the syllabus.

Parts (i), (ii), (iii) and (v) were very poorly answered. Answers to part (iv) were more satisfactory.

To receive full marks in part (i), candidates needed to derive the least squares estimators in the specific scenario in the question and to define the notation they used.

Many candidates who attempted part (v) received few marks because they only commented that the linear regression model was more accurate overall, or that it was more accurate for policyholder 1, without drawing any conclusions or suggesting any further investigations that should be performed. Alternative comments that were clear, distinct and relevant to the context of the question were also awarded credit.

Q9

(i)

The characteristic polynomial here is:

1 - 0.3*B* - 0.1(*B*^2) [½]

= (1 - 0.5*B*)(1+ 0.2*B*) [½]

with roots 2 and -5. [½]

Therefore, *Y_t* is stationary as both roots are greater than 1 in magnitude. [½]

Hence, *Y_t* is an ARIMA(2,0,0) process. [1]

(ii)

From the Yule-Walker equations for autocorrelation values we have:

rho_k = 0.3 * *rho*_(*k*-1) + 0.1 * *rho*_(*k*-2) for *k* >= 1 [1]

For *k* = 1,

*rho*_1 = 0.3 * *rho*_0 + 0.1 * *rho*_1 = 0.3 + 0.1 * *rho*_1

*rho*_1 = 0.3 / 0.9 = 1/3 [1]

For *k* = 2,

*rho*_2 = 0.3 * *rho*_1 + 0.1 * *rho*_0 = 0.3 * *rho*_1 + 0.1

*rho*_2 = 0.3 * (1/3) + 0.1 = 0.2 = 1/5 [1]

For *k* = 3,

*rho*_3 = 0.3 * *rho*_2 + 0.1 * *rho*_1

*rho*_3 = 0.3 * (1/5) + 0.1 * (1/3) = 0.09333 = 7/75 [1]

(iii)

For the partial autocorrelation values we have:

*phi*_1 = *rho*_1 = 1/3 [1]

*phi*_2 = (*rho*_2 - (*rho*_1)^2) / (1 - (*rho*_1)^2)

= (1/5 - 1/9) / (1 - 1/9)= (4/45) / (8/9) = 1/10 = 0.1 [1]

*phi*_3 = 0 since *Y_t* is AR(2). [1]

(iv)

Using the formula on page 42 of the Tables, the test statistic is:

n(*n* + 2) * (sum (*k*=1, *m*):[((*r_k*)^2) / (*n* - *k*)])

= *n*(*n* + 2) * (((1/3)^2)/(*n* - 1) + ((1/5)^2)/(*n* - 2) + ((7/75)^2)/(*n* - 3)) [1]

Under the null hypothesis of a white noise process, the test statistic follows a chi-squared distribution with (3 - 0 - 0 =) 3 degrees of freedom [1]

According to page 169 of the Tables, the critical value of the chi-squared distribution with 3 degrees of freedom at the 5% level is 7.815 [1]

We therefore need to find the least positive integer *n* such that:

n(*n* + 2) * (((1/3)^2)/(*n* - 1) + ((1/5)^2)/(*n* - 2) + ((7/75)^2)/(*n* - 3)) >= 7.815 [1]

The LHS is less than 7.815 for *n* = 45 but greater than 7.815 for *n* = 46 [1]

Hence, the required minimum sample size is *n* = 46. [1]

(v)

Sample Autocorrelation Values Equal to Theoretical Values

(Maximum = 2½ marks)

The absolute value of the test statistic increases with the number of lags m. [½]

As the autocorrelation values, *rho_k*, generally decrease with increasing *k* [½]

adding further lags to the Ljung and Box "portmanteau" test tends to increase the test statistic more slowly than the corresponding chi-squared critical value [½]
and hence would tend to increase the value of *n* in part (iv) [½]
This would tend to support the use of a small number of lags [½]
to maximise the power of the test for a fixed value of *n* [½]
However, if too few lags are used then we may miss meaningful ACF values present in the model [½]
which may lead to a larger value of *n* being required in part (iv) [½]
This would tend to support the use of a larger number of lags [½]
to maximise the power of the test for a fixed value of *n* [½]
A balance will therefore need to be found between these two potential concerns [½]
[Marks available 5½, maximum 2½]

Sample Autocorrelation Values ***Not*** *Equal to Theoretical Values*
(Maximum = 2½ marks)
<u>EITHER:</u>
The absolute value of the test statistic increases as the entries (r_k)^2 increase [½]
If the sample autocorrelation values were not equal to the theoretical values but instead the first few autocorrelation values happen to be small [½]
using a small number of lags will result in a relatively small value of the test statistic under the alternative hypothesis [½]
A large value of *n* would then be required to reject the null hypothesis [½]
This would tend to support the use of a larger number of lags [½]
to maximise the power of the test for a fixed value of *n* [½]
[Marks available 3, maximum 2½]

<u>OR:</u>
The absolute value of the test statistic increases as the entries (r_k)^2 increase [½]
If the sample autocorrelation values were not equal to the theoretical values but instead the first few autocorrelation values happen to be large [½]
using a small number of lags will result in a relatively large value of the test statistic compared to the chi-squared critical value [½]
A smaller value of *n* would then be required to reject the null hypothesis [½]
This would tend to support the use of a smaller number of lags [½]
… to maximise the power of the test for a fixed value of *n* [½]
[Marks available 3, maximum 2½]
[Total 21]

Parts (i), (ii) and (iii) were very well answered.

*In part (iii), some candidates lost marks because they did not explain why phi_*3 = 0.

The most common error in part (iv) was to use one degree of freedom. This is incorrect since the number of degrees of freedom in the Ljung and Box 'portmanteau' test is the number of lags less p + q, where the ***null*** *hypothesis is that Y_t is an ARMA(p, q) process. In this case p = q = 0 since the null hypothesis is that Y_t is a white noise process.*

Part (v) was very poorly answered. Many candidates implied either that a small number of lags is always preferable or that a large number of lags is always preferable. Candidates instead needed to discuss the ***circumstances*** *in which a large or a small number of lags would result in a more powerful test. Alternative comments that were clear, distinct and relevant to the context of the question were also awarded credit.*

[Paper Total 100]

END OF EXAMINERS' REPORT