Institute
and Faculty
of Actuaries

EXAMINERS' REPORT

CS1B - Actuarial Statistics

Core Principles

Paper B

April 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
July 2023

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

The aim of the Actuarial Statistics subject is to provide a grounding in mathematical and statistical techniques that are of particular relevance to actuarial work.

In particular, the CS1B paper is a problem-based examination and focuses on the assessment of computer-based data analysis and statistical modelling skills.

For the CS1B exam candidates are expected to include the R code that they have used to obtain the answers, together with the main R output produced, such as charts or tables.

When a question requires a particular numerical answer or conclusion, this should be explicitly and clearly stated, separately from, and in addition to the R output that may contain the relevant numerical information.

Some of the questions in the examination paper accept alternative solutions from those presented in this report, or different ways in which the provided answer can be determined. In particular, there are variations of the R code presented here, which are valid and can produce the correct output. All mathematically and computationally valid solutions or answers received credit as appropriate.

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 questions where comments were required, valid comments that were different from those provided in the solutions also received full credit where appropriate.

In cases where a question is based on simulations, and no seed was specified, all numerical answers provided in this document are examples of possible results. The numerical values presented here will be different if the simulations are repeated.

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

Well prepared candidates were able to score highly. However,

Most candidates demonstrated sufficient knowledge of the key R commands required for the application of the statistical techniques involved in this subject.

In some occasions candidates failed to provide full R code, output and/or appropriate graphs (e.g. parts of Question 2, Question 3). Candidates must include the R code used to obtain their answers, together with the main R output produced in their answers.

In some cases, the layout of the provided answers was not satisfactory, with input and output being separated and presented with large gaps, and with significant duplications.

C. Pass Mark

The Pass Mark for this exam was 62.
1460 presented themselves and 746 passed.

Solutions for Subject CS1B - April 2023

Q1

(i)
R code:
norm.val = rnorm(10000) [1]

(ii)
R code:
chisq4.val = rchisq(10000,4) [2]

(iii)
The main disadvantage of the inverse transform method is the necessity to have an explicit expression for the inverse of the distribution function, which is not available for the chi-square distribution. [2]
The use of the method would require numerical methods for inverting the cdf. [1]

(iv)
A random variable having the *t_n* distribution can be simulated as
Z / sqrt(Y/n) [1]
where Z ~ N(0,1) and $Y \sim \chi_n^2$ [1]
independently. [1]

R code:
t4.val = norm.val/sqrt(chisq4.val/4) [2]

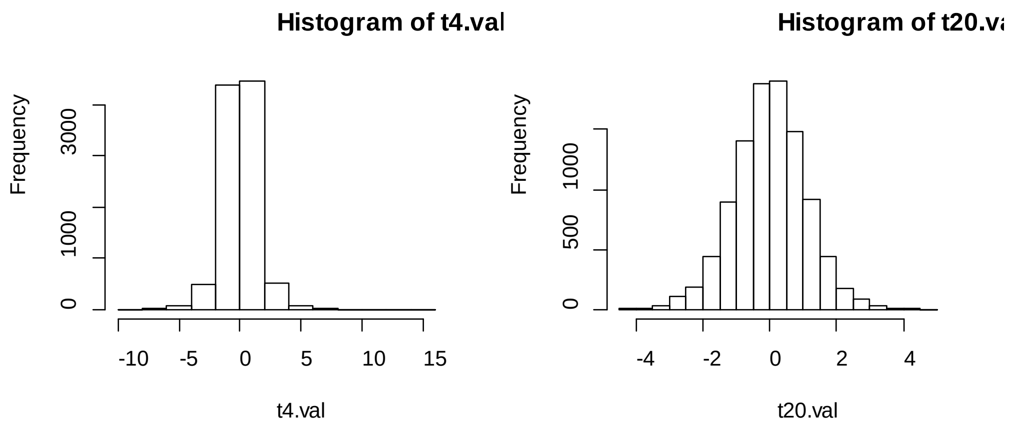
(v)
R code:
norm.val = rnorm(10000) [1]
[Or, use same values from (i)]

chisq20.val = rchisq(10000,20) [2]

t20.val = norm.val/sqrt(chisq20.val/20) [2]

(vi)
R code:
(vii)
par(mfrow=c(1,2))
hist(t4.val) [1]
hist(t20.val) [1]

Histograms are shown below: [2]



The histograms illustrate that the *t* distribution approaches the standard normal distribution as the number of the degrees of freedom increases. [2]

[Total 22]

Parts (i), (ii): Very well answered. A small number of candidates used runif() in (i).

Part (iii): Candidates' comments were often unclear. Many candidates failed to mention the need to use numerical methods.

Part (iv): Mostly well answered. A common error was failing to mention the required independence.

Part (v): Mostly well answered. There were some errors concerning the degrees of freedom of the involved chi square distribution.

Part (vi): The answers here were of mixed quality. A range of different plots were presented and received credit as appropriate.

Q2

(i)

We need to assume that the 1,000 policies are independent [1]

and that the probability to claim in a six-month period is the same for all policyholders [1]

(ii)

```
model1 = glm(claimed ~ age, data = CS1policies, family =
binomial)                                                    [2]
#OR
#model1 <- glm(claimed~age, family = "binomial" (link =
#"logit"))
summary(model1)                                              [1]
```

intercept = -2.336974, [1]

slope (age) = 0.024023 [1]

(iii)
```
model2 = glm(claimed ~ duration, data = CS1policies, family =
binomial)                                                    [2]
summary(model2 )                                             [1]
```

intercept = 0.390062, [1]
slope (duration) = -0.027608 [1]

(iv)
Residual deviance(model1) = 1051.8 , deviance(model2) = 948.98 [1]
AIC(model1) = 1055.8 , AIC(model2) = 952.98 [1]

The residual deviance and the AIC are both smaller for the model in part (iii). [1]
So duration seems to be a better predictor than age, and we would choose the model in part (iii).

Alternative answer:
The accuracy of the AIC depends on which R command is used: model1, summary(model1) or AIC(model1), which give 1056, 1055.8 and 1055.78, respectively. For model2 they give 953, 952.98 and 952.9805, respectively.

(v)
```
model3 = glm(claimed ~ age + duration, data = CS1policies,
family = binomial)                                           [2]
summary(model3)                                              [1]
```

```
Intercept = -0.909944
Slope(age) = 0.029434
Slope(duration) = -0.028040                                  [2]
```

(vi)
The AIC (AIC (model3)= 945.51) has decreased compared to the model in part (iii) (which is already better than the model in part (ii)) [1]

We would therefore use both, age and duration, in a linear predictor, and choose the model in part (v). [1]

Alternative answer:
The accuracy of the AIC depends on which R command is used: model3, summary(model3) or AIC(model3), which give 945.5, 945.51 and 945.5106, respectively.

[Total 22]

Parts (i)-(iii) were well answered generally. In part (ii) (and later parts) a number of candidates failed to report clearly the estimated parameter values.

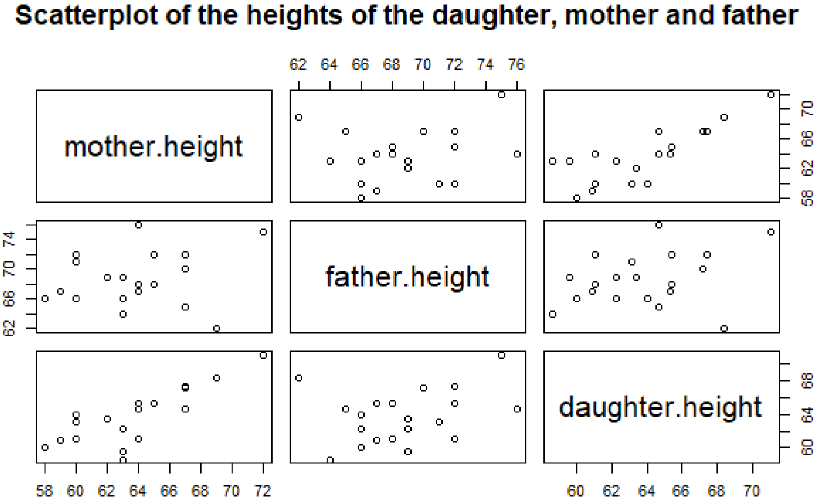
Part (iv): Well answered. A fair number of candidates failed to state the statistics (or the conclusion) explicitly. Some candidates considered the AIC only and not the deviance.

Part (v): Candidates performed well here and consistently with their answers in earlier parts (ii) and (iii). Answers including the interaction term received full marks as appropriate (and later).

Part (vi): Well answered generally. Some candidates used an alternative answer involving the deviances and anova() to compare models, and received credit as appropriate.

Q3
(i)
R code:

```
plot(heights, main = "Scatterplot of the heights of the daughter, mother
and father")
```
[1]



[2]

(ii)
The scatterplots suggest a (strong) positive and roughly linear relationship between mothers' and daughters' height. [2]

The relationship between fathers' and daughters' height seems weaker [1]

There is no association between mothers' and fathers' height. [1]

(iii)
R code:

```
mod1 = lm(daughter.height ~ mother.height + father.height)
summary(mod1)
```
[2]

```
#Residuals:
#    Min      1Q  Median      3Q     Max
#-3.8805 -0.6942  0.5915  0.8651  3.3138
#
#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)     7.4543    10.8804   0.685    0.503
#mother.height   0.7072     0.1289   5.488    4e-05 ***
#father.height   0.1636     0.1266   1.293    0.213
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#Residual standard error: 1.94 on 17 degrees of freedom
#Multiple R-squared:  0.6752,    Adjusted R-squared:  0.637
#F-statistic: 17.67 on 2 and 17 DF,  p-value: 7.057e-05
```
[1]

The fitted equation is

$\hat{y} = 7.454 + 0.707\ Mother's\ height + 0.164\ Father's\ height$ [2]

(iv)

Mother's height has a positive [½]
significant positive effect on daughter's height (p-value = 4e-05). [1½]
On the other hand, although daughters' height also increases with father's height, [½]
the effect is not significant (p-value = 0.213). [1½]

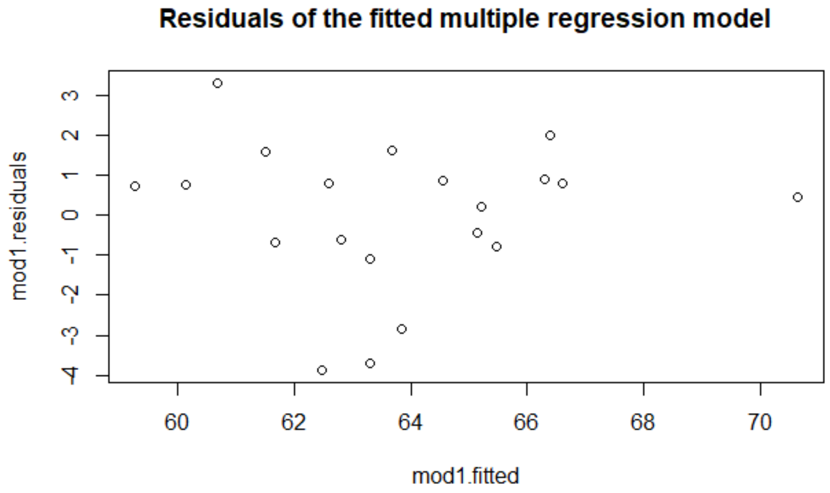
(v)

R code:

```
mod1.residuals = residuals(mod1)
# Or
# mod1.residuals = resid(mod1)

mod1.fitted = fitted(mod1)
#Or
# mod1.fitted = fitted.values(mod1)
plot(mod1.fitted,mod1.residuals, main = "Residuals of the fitted multiple regression model")
```
[3]

Alternative code: plot(mod1,1)



[2]

(vi)
There seem to be doubts about the assumption of constant variance, with residuals around fitted values 62-64 showing higher variation. [1]

Also, although there seems to be random variation around zero, there is a larger number of positive residuals, indicating possible underestimation. [1]

(vii)
R code:
```
newdata = data.frame(mother.height=61, father.height=63)
predict(mod1,newdata,interval="predict",level=0.95)
```
[2]
The 95% prediction interval (56.4, 65.4) [1]
So, a daughter's height of 67.5 inches seems unlikely, based on the fitted model. [1]

[Total 27]

Part (i): Well answered. A common error was not including annotations on the plot.

Part (ii): Overall well answered. Some candidates failed to identify all 3 relevant points of the relationships shown in the plots.

Part (iii): Well answered generally. A number of candidates failed to state the fitted equation.

Part (iv): Mixed answers. A common error was not referring explicitly to the p-values or the positive effect.

Part (v): Overall very well answered. A common error was plotting the residuals against an index, without reference to the explanatory variables or fitted values.

Part (vi): Mixed answers. In a number of cases comments were unclear, but most candidates provided some valid comments.

Part (vii): Well answered generally.

Q4

Note the answers shown here are in units of 1,000s (or 1,000,000s for variances).

> n = ncol(claims)

(i)(a)

> m <- mean(rowMeans(claims)) [1]

[1] 3,383.2 [1]

(i)(b)

> v <- var(rowMeans(claims)) - mean(apply(claims,1,var))/n [2½]

[1] 3,232,667 [½]

(i)(c)

> s <- mean(apply(claims,1,var)) [1½]

[1] 1,887,450 [½]

(ii)

> Z <- n/(n+s/v) [1]

[1] 0.8954367 or 89.54%

> Z * rowMeans(claims)+(1-Z)*m [1]

[1] 3960.578 4270.399 1202.633 5391.485 2090.906

Estimated claims amount for company E is 2,090.91. [1]

(iii)(a)

> n <- ncol(claims)

> N <- nrow(claims) [½]

> X <- claims/volumes [½]

> Xi_bar <- rowSums(claims)/rowSums(volumes) [1]

> Pi <- rowSums(volumes) [½]

> P <- sum(Pi) [½]

> P_star <- sum(Pi*(1-Pi/P))/(N*n-1) [2]

> m2 <- sum(claims)/P [1]

[1] 7.319775 [1]

(b)

> v2 <- (sum(rowSums(volumes*(X-m2)^2))/(N*n-1) - mean(rowSums(volumes*(X-Xi_bar)^2)/(n-1)))/P_star [3]

[1] 3.151102 [1]

(iii)(c)

> s2 <- mean(rowSums(volumes*(X-Xi_bar)^2)/(n-1)) [2]


```
[1] 4356.445
```
[1]

(iv)

```
> Zi <- Pi/(Pi+s2/v2)
[1] 0.6662620 0.6097364 0.3807366 0.7799425 0.3902952
```

or 39.03% credibility factor for company E.

```
> Zi*Xi_bar + (1-Zi)*m2
```
[½]

```
[1] 7.304668 9.030223 6.656035 6.088278 8.740714
```

Therefore the premium (i.e. expected claims) for company E in the coming year will be: 8.740714 * 130 = 1,136.29 [1½]

(v)

The expected claim amount for company E is lower using EBCT model 2. [1]
EBCT model 2 allows for the number of claims (volumes), [½]
which has been declining [½]

[Total 29]

Part (i): Very well answered overall. In (b) and (c) a small number of candidates gave confused answers between v and s.

Part (ii): Very well answered overall. A number of candidates did not mention explicitly the figure for company E.

Part (iii): Very well answered overall. As in part (i), a small number of candidates showed confused answers between v and s.

Part (iv): Well answered generally, with some small calculation errors. As in part (ii), some candidates did not mention explicitly the figure for company E.

Part (v): This part was not well answered. A number of candidates did not attempt it, but most candidates received partial credit for providing some valid comments.

[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