INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2019 Examinations

Subject CS1 – Actuarial Statistics Core Principles (Part B)

Introduction

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

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

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

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

Mike Hammer
Chair of the Board of Examiners
July 2019

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

1. The aim of the Actuarial Statistics 1 subject is to provide a grounding in mathematical and statistical techniques that are of particular relevance to actuarial work.

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

3. 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.

4. 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.

5. Annotated plots and relevant comments should be provided when instructed to do so in the question.

6. Some of the questions in the examination paper admit alternative solutions from these 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, that are valid and can produce the correct output. All mathematically and computationally valid solutions or answers received credit as appropriate.

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

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

9. In cases where a question is based on simulations, 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 *student performance in this diet of the examination.*

1. Overall performance in CS1B varied considerably among candidates, and was less consistent compared to that in the theoretical part of the subject (CS1A).

2. Candidates demonstrated a good knowledge of the key R commands required for the application of the statistical techniques involved in this subject.

3. In general, candidates showed good understanding of certain Core Reading topics examined in this paper, for example hypothesis testing and general linear models. However, topics including Bayesian statistics were less well addressed.

4. The quality of analysis and commentary given alongside the R output varied significantly among candidates.

5. Generally there was inconsistency with answers being documented – for example R code, output and plots were not always submitted.

C. Pass Mark

The combined pass mark for CS1 in this exam diet was 58.

Solutions Subject CS1 – B

Q1

(i)

```
y <- c(87, 53, 72, 90, 78, 85, 83)
c(mean=mean(y), variance=var(y))

mean     variance
78.29      159.90
```

[1]
[1]
[1]

(ii)

```
xbar = s2 = numeric(10)
for (j in 1:10){
     x <- rpois(7, 78.29)
     xbar[j] = mean(x)
     s2[j] = var(x)
}
```

[3]

```
xbar

#[1] 77.85714 79.71429 68.71429 82.14286 69.71429 84.57143 77.28571 83.0000
# 76.85714 79.28571
```

[½]

```
s2
#[1] 104.80952 127.23810 136.23810  42.47619  51.90476 103.28571  83.90476
# 107.33333  49.80952  90.57143
```

[½]

It is unusual to get as large a difference between the mean and the variance

as that observed for these data, [1]

making it doubtful that these data are from a Poisson distribution. [1]

Part (i) was very well answered. However, a number of candidates showed only the R code and not the numerical answers. Answers in part (ii) were problematic with candidates making various errors in the computations and failing to provide a clear comment or conclusion at the end.

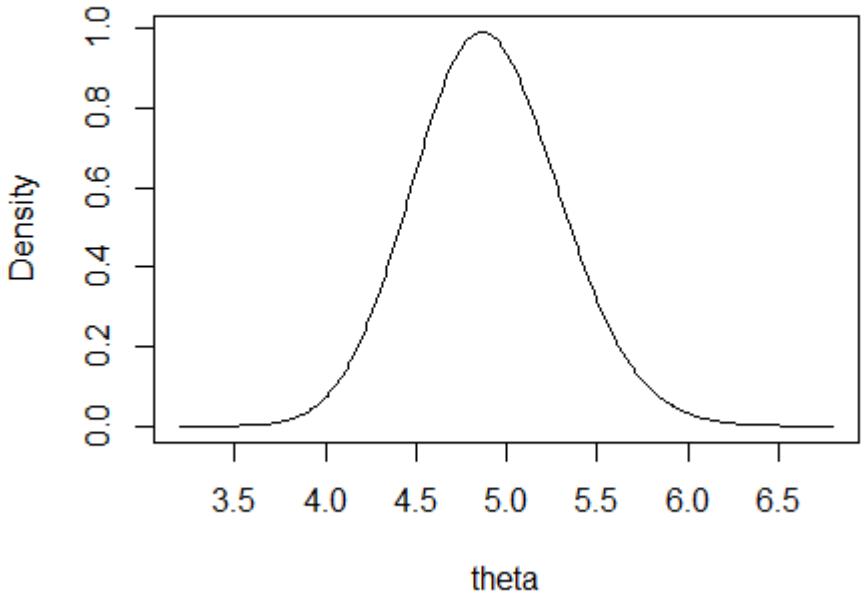
Q2

(i)

(a)

```
## Data entry

y = c (5, 5, 6, 2, 4, 10, 2, 5, 5, 2, 5, 3, 7, 4, 4, 5, 4, 6,
7, 2, 8, 4, 6, 4, 3, 6, 6, 6, 5, 7)

## plot the posterior pdf of theta

theta = seq(3.2, 6.8, by = 0.01)

plot(theta, dgamma(theta, sum(y)-1, length(y) + 0.01), ylab =
"Density", type = "l")
```

[2]



[2]

(b)

The posterior samples are

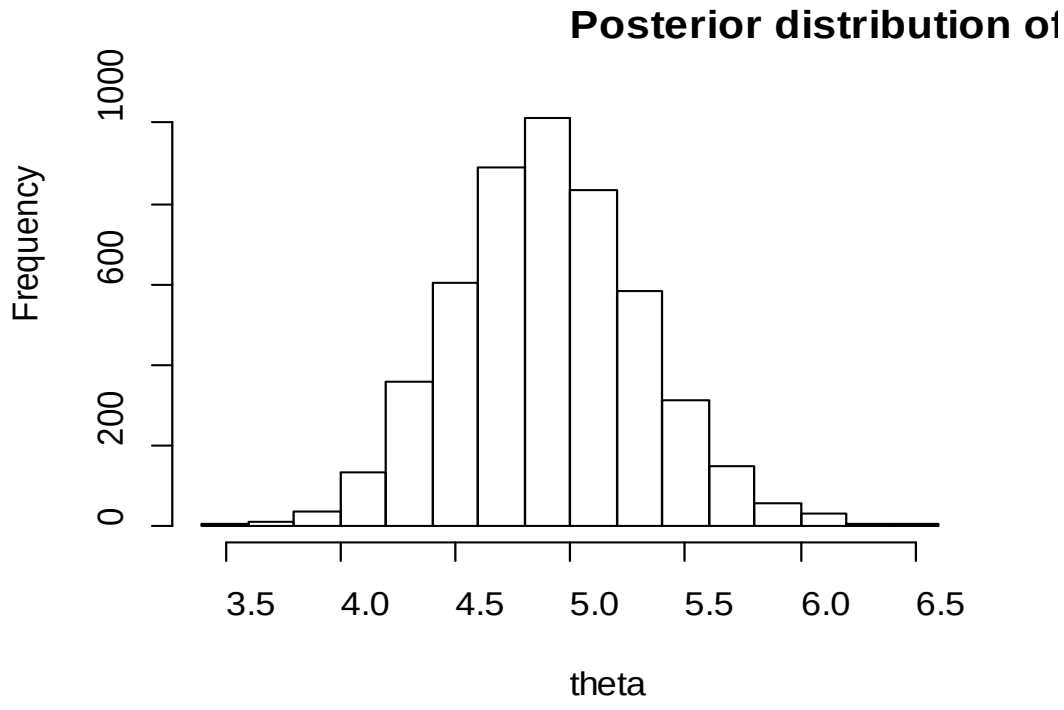
```
x = rgamma(5000, sum(y)-1, 30 + 0.01)
```

[4]

(ii) We can plot the histogram using

```
hist(x, main="Posterior distribution of theta",xlab="theta")
```

[1]



[1]

(iii)

```
mean(x)
# 5.003996
```

[1]

```
median(x)
# 4.997373
```

[1]

```
sd(x)
# 0.4117624
```

[1]

(iv) 15 is quite far away from the range of samples obtained for the posterior distribution of θ.

[1]

On the other hand 5 is more likely to be the true value.

[1]

15 is very unlikely to be the case if there is no calculation error.

5 fits well within the distribution and the values of the mean and the median are very close to it.

[1]

Parts (i), (ii) and (iii) were generally well answered. A common error was to use sum(y-1) instead of sum(y)-1 in the computations. In part (iii) a number of candidates provided the answers using the theoretical results based on the given posterior distribution – this was given full credit. Note that displaying the plots is required for full marks. Part (iv) was not particularly well answered, with many candidates not attempting this part despite having given answers in previous parts.

Note that the parameter of the gamma distribution given as $\sum_{i-1}^{n} y_i - 1$ *in the preamble of the question is theoretically incorrect and should have been* $\sum_{i-1}^{n} y_i + 1$. *The error did not affect the remainder of the question and the required answer, as the candidates were explicitly asked to work with the* $\sum_{i-1}^{n} y_i - 1$ *quantity as given in the question. The examiners did not find evidence of this error having a negative impact on candidates' performance.*

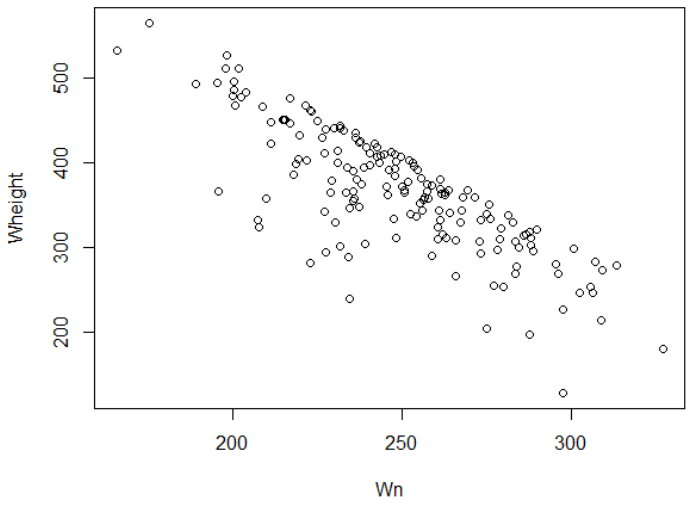
Q3

```
# load data

load("CS1waves.Rdata")
```

(i) #plot data

```
plot(Wn, Wheight)
```

[1]



[1]

(ii) There seems to be a linear relationship between wave height and number

of waves. [1]
The more waves per hour, the smaller the waves (negative association). [1]

(iii)

```
cor(Wn, Wheight,method = "pearson")
```
[1.5]

```
-0.8055382
```
[½]

(iv)

```
cor(Wn, Wheight, method = "spearman")
```
[1.5]

```
-0.7688486
```
[½]

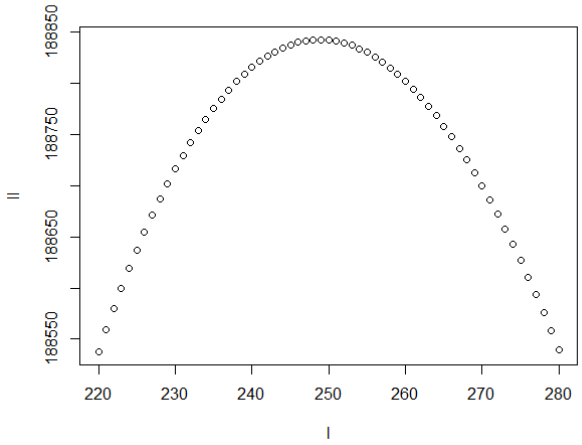
(v) Both correlation coefficients confirm the negative relationship that is already obvious in the plot. [1]
The rank correlation is lower than the Pearson correlation, [1]
indicating that the relationship is stronger when we take the magnitude of observations into account rather than just their relative rank. In other words, for observations with similar magnitude, the ranks are not always ordered. [1]

(vi)

```
l = 220:280
```
[1]

```
ll = log(l)*sum(Wn)-168*l
```
[2]

```
plot(l,ll)
```
[2]



(vii) By inspection, the maximum of $l(\lambda)$ is at about 250. [1]
The maximum likelihood estimate is $\hat{\lambda} \approx 250$ waves per hour. [2]

(viii) The exact MLE is the mean, that is, $\hat{\lambda} = \frac{1}{168}\sum_{i=1}^{168} X_i$ [2]

```
mean(Wn)
```
[1]

$\hat{\lambda}$ =248.8579

Most candidates performed very strongly in this question. Comments in part (v) were of mixed quality – credit was given to valid comments that may be different from those presented here. A common error in part (vii) was to provide the value of the log likelihood instead of lambda.

Q4

(i)
Data entry

```
sample1 <- c(21, 22, 28, 27, 20, 23, 26, 32, 25, 21, 30)
sample2 <- c(19, 18, 38, 33, 24, 39, 22, 20, 28, 26, 30)
```
[1]

(ii)
```
var.test(x = sample1, y = sample2, conf.level = 0.95)
```
[2]

Result
```
# F test to compare two variances
# data: sample1 and sample2
# F = 0.29259, num df = 10, denom df = 10, p-value = 0.06553
# alternative hypothesis: true ratio of variances is not equal to 1
# 95 percent confidence interval:
# 0.07872181 1.08750577
# sample estimates:
# ratio of variances
# 0.2925926
```

The p-value is 0.06553 > 0.05, so we have insufficient evidence to reject the assumption of equal variance.

[2]

(iii) (a)
```
## code
t.test(x = sample1, y = sample2, var.equal = TRUE, conf.level
= 0.95)
```
[2]

```
## Result
# Two Sample t-test
# data: sample1 and sample2
# t = -0.79396, df = 20, p-value = 0.4365
# alternative hypothesis: true difference in means is not
equal to 0
# 95 percent confidence interval:
# -7.254581 3.254581
```

```
# sample estimates:
# mean of x mean of y
#        25        27
```

The p-value of 0.4365 is much larger than the 5% significance level, therefore we have no evidence to suggest that the means are different between the two samples. [2]

(b)

The confidence interval can be read from the result above or extracted using

```
t.test(x = sample1, y = sample2, var.equal = TRUE, conf.level
= 0.95)$conf.int
```

to obtain

```
# -7.254581 3.254581
# attr(,"conf.level")
# 0.95
```

i.e. 95% CI is (-7.25, 3.25). [2]

(c)

It is clear that the confidence interval (-7.25, 3.25) contains 0, [1]
therefore the assumption of equal means holds. [1]

(iv) (a)

```
s1 = abs(sample1 - mean(sample1))
```
[1]
```
# s1
# 4 3 3 2 5 2 1 7 0 4 5
```

```
s2 = abs(sample2 - mean(sample2))
```
[1]
```
# s2
# 8 9 11 6 3 12 5 7 1 1 3
```

(b)
```
t.test(x = s1, y = s2, var.equal = TRUE, conf.level = 0.95)
```
[2]
```
# Two Sample t-test
# data: s1 and s2
# t = -2.1077, df = 20, p-value = 0.04788
# alternative hypothesis: true difference in means is not
equal to 0
# 95 percent confidence interval:
# -5.42646442 -0.02808103
```

```
# sample estimates:
# mean of x mean of y
#  3.272727  6.000000
```

The p-value 0.04788 is slightly less than 5%, so we reject the hypothesis of equal mean of the absolute deviations and therefore the equal variance assumption in the original data. [2]

(v) The tests in (ii) and (iv)(b) give different results [1.5]
but the *p*-values are quite similar. [1.5]

(vi) We would need to find an appropriate test that allows for the variances to be different and compare with the tests carried in (iii)(a). [2]

Parts (i) and (ii) were very well answered. In parts (iii)-(vi) there was wide variation in the quality of the answers. In part (iii) some candidates failed to specify 'var.equal=T' in the test – this error was only penalised in the first occurrence if also repeated later. A number of candidates did not attempt parts (v) and (vi), while some gave incomplete comments.

Q5

```
# read the data
cables_data<-read.csv("Cables_dataset.csv")
cables_data
```

```
    X Failure.Time Material.Type  Rainfall
1   1  0.093496420             1 0.2049910
2   2  0.064299790             2 0.2459758
3   3  0.037432940             3 0.1037756
4   4  0.036485400             4 0.3138880
5   5  0.080959110             1 0.2020806
6   6  0.002198732             2 0.2545738
7   7  0.028674680             3 0.2701437
8   8  0.032782080             4 0.3307581
9   9  0.074037110             1 0.2911203
10 10  0.059623200             2 0.2500904
11 11  0.030189960             3 0.1641475
12 12  0.030789370             4 0.4265417
13 13  0.071302530             1 0.1216770
14 14  0.046903810             2 0.3412739
15 15  0.033226010             3 0.3929279
16 16  0.038243150             4 0.3974997
17 17  0.064787050             1 0.2993994
18 18  0.042787140             2 0.3332971
19 19  0.033838870             3 0.2108672
20 20  0.025070570             4 0.2722977
```

(i)

```
# fit the linear model

linear_model<-lm(Failure.Time~Material.Type + Rainfall,
cables_data)
```

[3]

```
# output the linear model fit summary statistics

summary(linear_model)
```

[1]

```
#Call:
#lm(formula = Failure.Time ~ Material.Type + Rainfall, data =
#cables_data)
#
#Residuals:
#      Min        1Q    Median        3Q       Max
#-0.051313 -0.006694  0.002246  0.008590  0.025763
#
#Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
#(Intercept)    0.081086   0.012659   6.406 6.52e-06 ***
#Material.Type -0.014499   0.003575  -4.055 0.000822 ***
#Rainfall       0.005591   0.046745   0.120 0.906200
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
#1
#
#Residual standard error: 0.01627 on 17 degrees of freedom
#Multiple R-squared:  0.5326,    Adjusted R-squared:  0.4776
#F-statistic: 9.687 on 2 and 17 DF,  p-value: 0.001556
```

[1]

(ii) (a) From analysing the R output, we see that the fitted linear model is:

$$\hat{y} = 0.081086 - 0.014499x_1 + 0.005591x_2$$

[2]

where x_1 is the 'material type' variable, and x_2 is the 'rainfall' variable.

[1]

(ii) (b)

The R output shows that the 'material type' parameter is significantly different to zero (at the 0.1% level), [1]

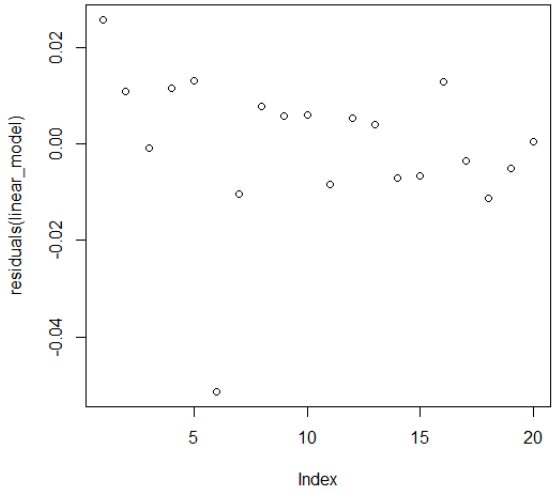
but the 'rainfall' parameter is not significantly different to zero – this is indicated by the t-tests shown in the fourth column of the R output. [1]

The intercept is also significantly different to zero [1]

(iii) (a)

```
# plot the residuals
plot(residuals(linear_model))
```

[1]



[1]

(iii) (b) The residuals exhibit a fairly random scatter around zero [1]
apart from the 6th point. [1]

(iv) (a)

```
# remove data point 6 and redefine the dataset
cables_data2 <- cables_data[-6,]
```

[2]

```
cables_data2
```

```
    X Failure.Time Material.Type  Rainfall
1   1   0.09349642             1 0.2049910
2   2   0.06429979             2 0.2459758
3   3   0.03743294             3 0.1037756
4   4   0.03648540             4 0.3138880
5   5   0.08095911             1 0.2020806
6   7   0.02867468             3 0.2701437
7   8   0.03278208             4 0.3307581
8   9   0.07403711             1 0.2911203
9  10   0.05962320             2 0.2500904
10 11   0.03018996             3 0.1641475
11 12   0.03078937             4 0.4265417
12 13   0.07130253             1 0.1216770
13 14   0.04690381             2 0.3412739
14 15   0.03322601             3 0.3929279
15 16   0.03824315             4 0.3974997
16 17   0.06478705             1 0.2993994
```

```
17 18   0.04278714             2 0.3332971
18 19   0.03383887             3 0.2108672
19 20   0.02507057             4 0.2722977
```

(iv) (b) The residuals plot indicates that data point 6 is an outlier. [2]

(v) (a)

```
# refit the linear model
linear_model2<-lm(Failure.Time~Material.Type + Rainfall,
cables_data2)
```

[1]

(v) (b)

```
summary(linear_model2)

#Call:
#lm(formula = Failure.Time ~ Material.Type + Rainfall,
#data = cables_data2)
#
#Residuals:
#       Min         1Q     Median         3Q        Max
#-0.0144060 -0.0082529 -0.0003768  0.0071557  0.0213878
#
#Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
#(Intercept)    0.086629   0.008094  10.703 1.06e-08 ***
#Material.Type -0.015576   0.002275  -6.846 3.93e-06 ***
#Rainfall       0.005152   0.029621   0.174    0.864
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
#' ' 1
#
#Residual standard error: 0.01031 on 16 degrees of
#freedom
#Multiple R-squared:  0.7756,    Adjusted R-squared:
#0.7475
#F-statistic: 27.64 on 2 and 16 DF,  p-value: 6.438e-06
```

[1]

The adjusted R^2 statistic for the model fitted to the data with the outlier removed is 0.7475. This shows an improved fit relative to the model fitted to all 20 data points, which had an adjusted R^2 statistic of 0.4776.

[2]

(vi) (a)

```
# fit a Gamma GLM
gfit<-glm(Failure.Time~Material.Type+Rainfall,
family=Gamma(link=inverse), cables_data2)
```

[2]

(vi) (b) The fitted model is:

```
# extract the coefficients
coef(gfit)

#  (Intercept) Material.Type      Rainfall
#    6.1920558     6.7286452     0.4814427
```

$$\hat{\eta} = \frac{1}{\mu} = 6.1920558 + 6.7286452x_1 + 0.4814427x_2$$

where x_1 is the 'material type' variable, and x_2 is the 'rainfall' variable.

[2]

(vi) (c)

```
# review the model fit
summary(gfit)

#Call:
#glm(formula = Failure.Time ~ Material.Type + Rainfall,
#family = Gamma(link = inverse),
#    data = cables_data)
#
#Deviance Residuals:
#     Min        1Q    Median        3Q       Max
#-0.26231  -0.14156  -0.03338   0.12850   0.25185
#
#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)     6.1921     2.6184   2.365    0.031 *
#Material.Type   6.7286     0.8359   8.050 5.12e-07 ***
#Rainfall        0.4814    10.6244   0.045    0.964
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
#' ' 1
#
#(Dispersion parameter for Gamma family taken to be
#0.03085425)
#
#    Null deviance: 2.99847  on 18  degrees of freedom
#Residual deviance: 0.48503  on 16  degrees of freedom
#AIC: -125.54
#
#Number of Fisher Scoring iterations: 4
```

Reviewing the model fit output from R, the 'rainfall' parameter is not significantly different to zero, whereas the 'material type' parameter is significant at the 0.1% level. [2]

Most candidates performed strongly in this question. Candidates answered well the parts of the question where various GLMs were fitted, e.g. parts (i), (v)(a) and (vi)(a). Answers regarding removing the outlier point and the fit of the resulting model (parts(iv), (v)), were mixed with a number of candidates failing to demonstrate understanding of the need for this and its practical importance.

END OF EXAMINERS' REPORT