# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINERS' REPORT

April 2021

## Subject CS1 – Actuarial Statistics
## Core Principles
## Paper 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.

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 Actuarial Statistics 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. 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.

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

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

8. 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.*

1. Overall performance in CS1B was satisfactory. Well prepared candidates were able to score highly.

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

3. The quality of the commentary given alongside the R output was not always strong and varied significantly among candidates.

4. In some occasions candidates failed to provide R code, output and/or appropriate graphs. Candidates must include the R code used to obtain their answers, together with the main R output produced in their answers.

5. Questions corresponding to parts of the syllabus that are not frequently examined

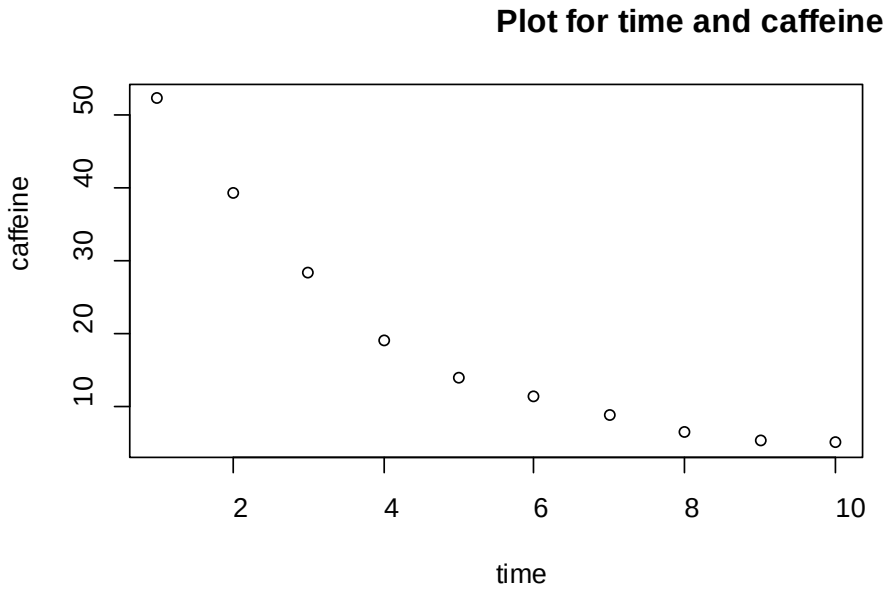
6. were generally poorly answered (e.g. parts of Q3, Q4). This highlights the need for candidates to cover the whole syllabus when they revise for the exam and not only rely on themes appearing in past papers.

### C. Pass Mark

The Pass Mark for this exam was 56.
1,482 candidates presented themselves and 779 passed.

**Solutions for Subject CS1 Paper B April 2021**

**Q1**
(i)




[1]

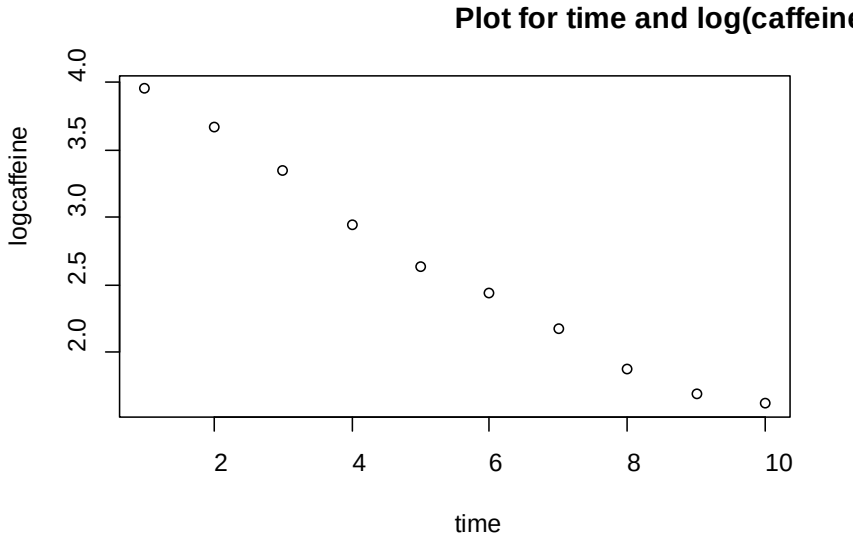
(ii)
The plot shows a non-linear and inverse relationship between C and T (i.e. the level of caffeine in the blood does not reduce linearly over time - instead it appears to decay at an exponential rate) [3]

(iii)
Given the shape of the graph in (i), a log transformation should be used on the data (i.e. transform C to log(C)) [2]
This is because the original plot from part (i) has an exponential shape [1]

(iv) > logcaffeine <- log(caffeine) [½]
> logcaffeine
3.955082 3.671225 3.342862 2.946017 2.636196 2.438863 2.172476 1.879465
1.691939 1.615420 [½]

(v)(a)
> plot(time,logcaffeine,main = "Plot for time and log(caffeine)") [1]




[1]

(b)

> cor(time,logcaffeine,method="pearson") [1]

#[1] -0.9919573 [1]

(c)

We can see via the plot in part (v)(a) and the calculated Pearson coefficient in part that a strong negative linear relationship exists between the transformed variable and time [2]

**[Total 15]**

*Generally very well answered. A common error in part (ii) was failing to refer to the strength of the relationship between caffeine and time. Answers in part (v)(c) were varied, with candidates often giving partial comments.*

**Q2**

(i)(a)

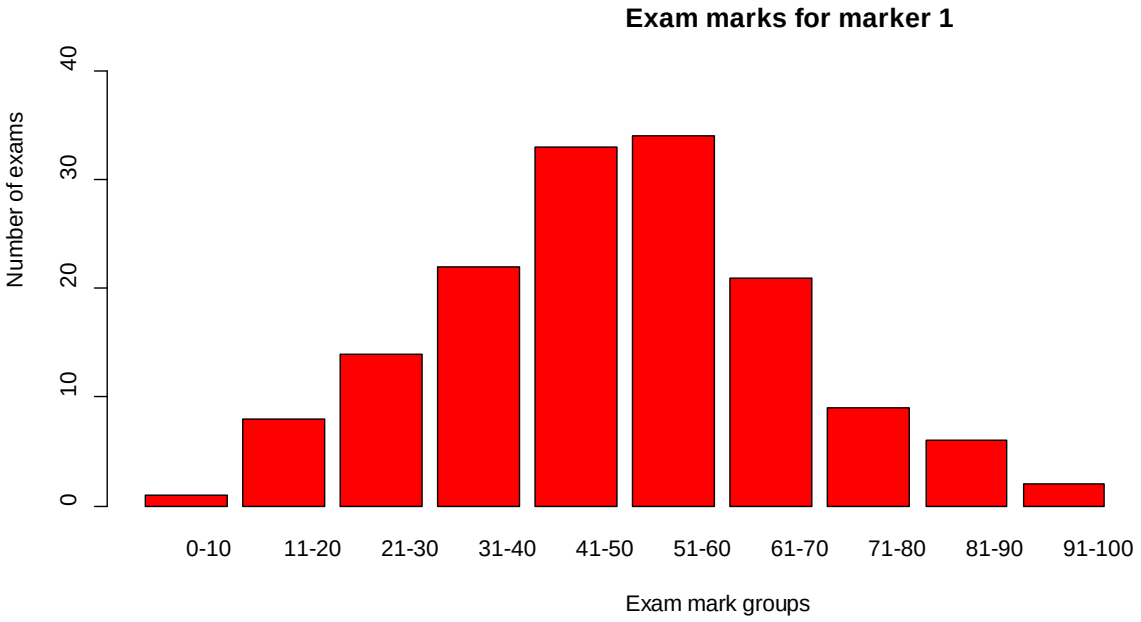
```
> axis = c("0-10","11-20","21-30","31-40","41-50","51-60","61-70","71-80","81-90","91-100")

> barplot(marks_1,xlab = "Exam mark groups", ylab = "Number of exams", main = "Exam marks for marker 1",col = "Red",names = axis, ylim = c(0,40))
```

[1]

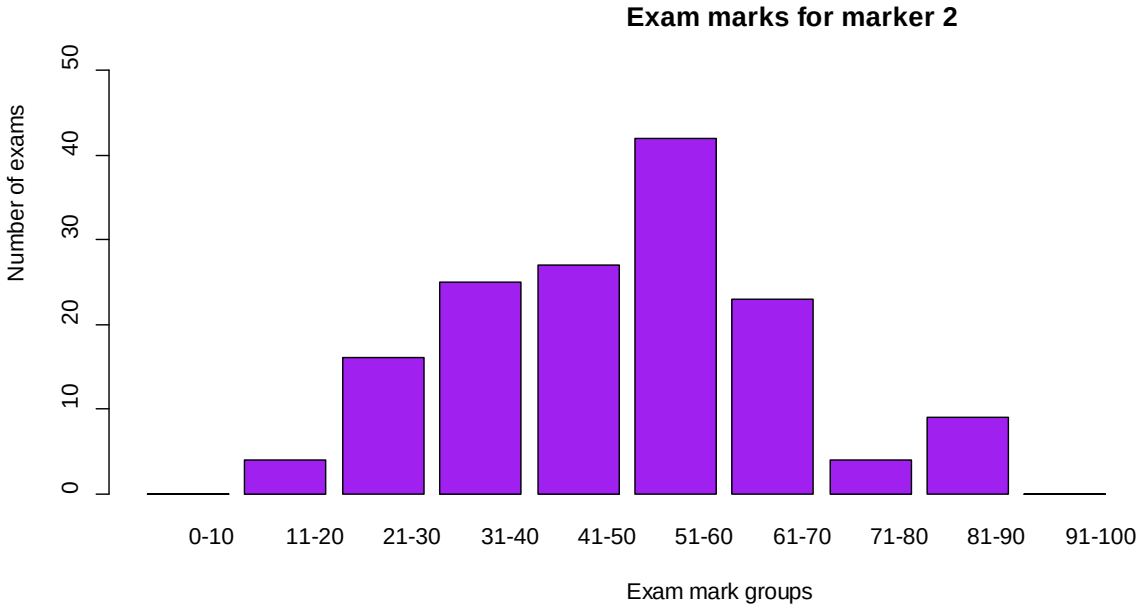
```
> barplot(marks_2,xlab = "Exam mark groups", ylab = "Number of exams", main = "Exam marks for marker 2",col = "Purple",names = axis, ylim = c(0,50))
```

[1]




[1]




[1]

(b)

The distributions of marks look similar, especially for middle scores [1]

However, there appears to be some differences in marking for low and high scoring exams [1]

The plot for marker 1 resembles a Normal shape (but it is not as clear for maker 2, where there appears to be some skewness) [1]

Overall, the plots suggest that the two markers are generally consistent [1]

[Marks available 4, maximum 3]

(ii)

$H_0$ : difference in means is zero $vs$ $H_A$ : difference in means is not zero. [½]

```
> t.test(marker_1,marker_2,paired=TRUE)
Paired t-test
data:  marker_1 and marker_2
t = 2.862, df = 9, p-value = 0.01872
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1886284 1.6113716
sample estimates:
mean of the differences
0.9
```

[2] [½]

P-value is equal to 0.01872 [½]
which is less than the significance level (5%) [½]
Therefore reject the null hypothesis [½]
There appears to be difference in the mean scores between the two markers [½]

(iii)

$H_0$ : difference in means is zero $vs$ $H_A$ : difference in means is not zero [½]

```
> t.test(marker_1,marker_2)
Welch Two Sample t-test
data:  marker_1 and marker_2
t = 1.1968, df = 17.675, p-value = 0.2472
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6820491  2.4820491
sample estimates:
mean of x mean of y
      3.9       3.0
```

[1] [½]

P-value is equal to 0.2472 [½]
which is more than the significance level (5%) [½]
Therefore do not reject the null hypothesis [½]
There appears to be no difference between the mean scores of the two markers [½]

(iv)

Parts (ii) and (iii) lead to different conclusions [½]
We would expect the data to not be independent of one another since both markers have marked the same exam papers [½]
Therefore the paired test in (ii) gives smaller variation for the test statistic and leads to rejecting the hypothesis of equal means, at the 5% level [1]

(v)

If a paired problem is analysed as though it involved independent samples, then the results would be invalid because the assumption of independence is violated [2]
Alternatively, if independent samples are analysed as though they were paired, then the results would be valid although they would be making inefficient use of the data due to the discarding of possible information about the means and variances of the two separate populations [2]

**[Total 22]**

*Parts (i)-(iii) were generally well answered. Plots in part (i) were varied, with a number of candidates not using reasonable axis labels (or no axis at all), while comments were often missing in (i)(b). A common error in part (ii) was to omit the PAIRED = TRUE parameter, as the question requested. Answers in (iv), (v) were mixed with a range of comments.*

**Q3**

(i)(a)

```
> p = c(0.1, 0.2, 0.3, 0.4, 0.5)
```
[½]

```
> gp=c(0.5, 0.2, 0.2, 0.05, 0.05)
```
[½]

```
> mean_A=sum(p*gp)
> mean_A
#[1] 0.195
```
[1]

```
>sd_A = sqrt(sum(gp*p^2)-mean_A^2)
>sd_A
#[1] 0.1160819
```
[2]

(b)

```
> samples_beta = rbeta(10000, 3, 12)
```
[1]

(c)

```
> mean_N = mean(samples_beta)
> mean_N
#[1] 0.2012128
```
[1]

```
> sd_N = sd(samples_beta)
> sd_N
#[1] 0.09913621
```
[1]

(d)

The first and second moments of the two priors are very similar [1]

The two prior beliefs are similar despite one being based on a continuous distribution and the other on a discrete distribution [1]

(ii)

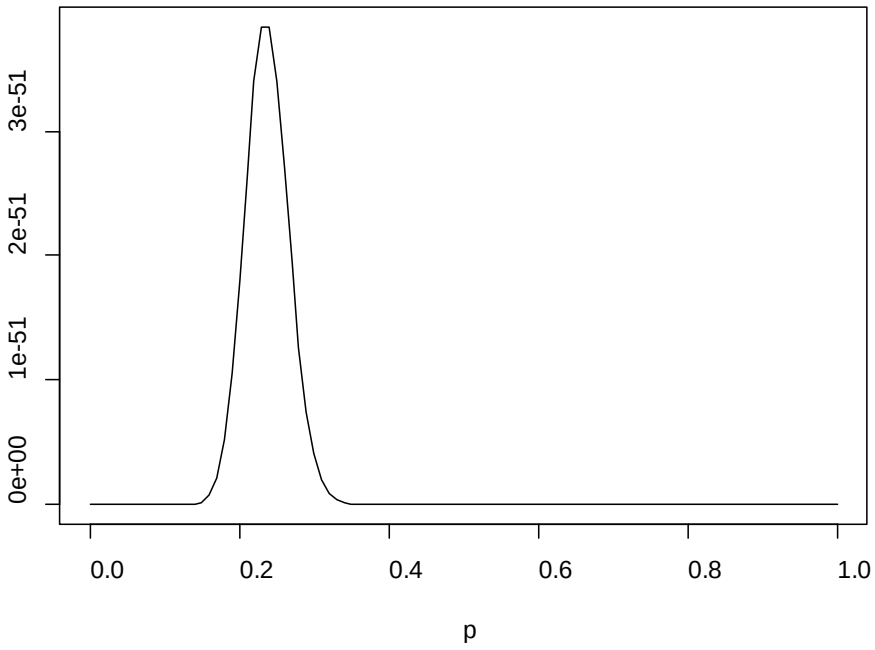
Read the data in:

```
> load("BinaryTrain.RData")

> p = seq(0, 1, by = 0.01)

> dens = p^(2+sum(y))*(1-p)^(11+length(y)-sum(y))

> plot(p, dens, type = "l", ylab ="",xlab="p")
```
[3]

Or,

```
> curve(x^(2+sum(y))*(1-x)^(11+length(y)-sum(y)))
```




[1]

(iii)

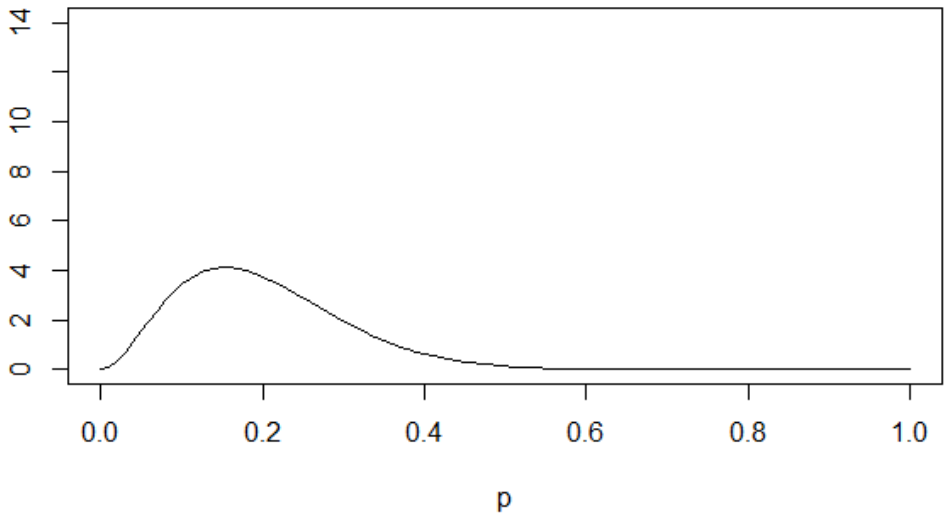
Prior distribution

```
> plot(p, dbeta(p, 3, 12), type = "l", ylim =c(0, 14), ylab ="")
```

[2]

Or,

```
> curve(dbeta(x, 3, 12), type="l",ylim=c(0, 14), ylab ="", xlab="p")
```




[1]

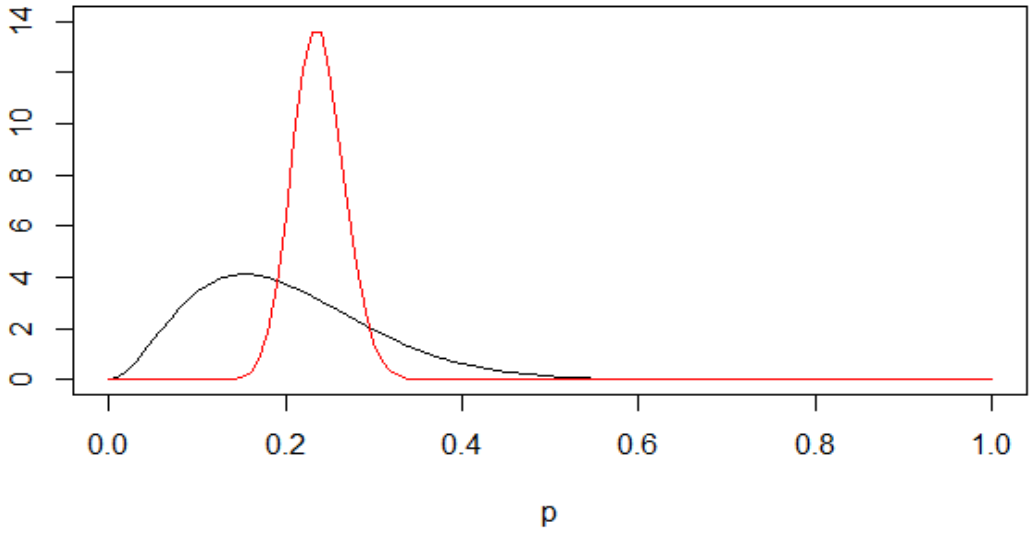
(iv)(a)

```
> x = p
> curve(dbeta(x, 3+sum(y), 12+length(y)-sum(y)), type = "l",
add=TRUE, col="red")
```

[2]

Or,

```
>lines(p, dbeta(p,3+sum(y),12+length(y)-sum(y)),col="red")
```




[1]

(b)

Clearly, the posterior is much narrower than the prior. [1]

(c)

The posterior distribution is more affected by the data than by the prior. [1]

(v)

90% interval for p

```
> qbeta(c(0.05, 0.95), 3+sum(y), 12+length(y)-sum(y))
#[1] 0.1910197 0.2861843
```

[1]

90% interval for p is (0.191, 0.286). [1]

(vi)

```
> 1 - pbeta(0.25, 3+sum(y), 12+length(y)-sum(y))
#[1] 0.3216195
```

[1]

[1]

(vii)(a)

```
> z = rbeta(10000, 3+sum(y), 12+length(y)-sum(y))
```

[1]

(b)

```
> sum(z>0.25)/10000
  # 0.3271
```

[1]

[½]

Or,

```
> length(z[z>0.25])/length(z)
```

(c) As expected the proportion is very similar to the answer in (vi). [½]

**[Total 28]**

*A number of candidates did not attempt this question, or only attempted part (i). For candidates that attempted it, the overall performance was mixed. In part (iii) many candidates plotted the graphs using the simulated values from part (i)(b) or using inappropriate plots, which led to difficulty later in the question.*

**Q4**

(i)

load("CS1passenger.RData")

Linear predictor for modelling:

(a) $\alpha_i + \beta \times temp$: where the intercept $\alpha_i, i = 1,2$ depends on the semester [2]

(b) $\alpha_i + \beta_i \times temp$: where $\alpha_i$ as above, $\beta_i, i = 1,2$ also depends on the semester [2]

(c) $\alpha_i + \beta_i \times temp + \gamma_j$ with $\alpha_i$, $\beta_i$ as above, $\gamma_j$, $j$ = 1,2 depends on the route. [2]

Alternative answer:

(a) $y = a + b_1 x_1 + b_2 x_2$ , where $x_1$ is temperature and $x_2 = 0$ (nonSemester), $x_2 = 1$ (Semester)

(b) $y = a + b_1 x_1 + b_2 x_2 + \gamma x_1 x_2$ , where $x_1$and $x_2$ as above

(c) $y = a + b_1 x_1 + b_2 x_2 + \gamma x_1 x_2 + b_3 x_3$ , where $x_3 = 0$ (8am), $x_3 = 1$ (9am)]

(ii)(a)

```
>Model1<- glm(Passengers~temp*semester + route, family=poisson(link="log"))
>summary(Model1)
```
[2]

```
Call:
glm(formula = Passengers ~ temp * semester + route, family = poisson(link = "log"))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.8128  -0.6263  -0.1566   0.5162   1.3991

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.40210    0.31155   1.291   0.1968
temp                -0.07878    0.03576  -2.203   0.0276 *
semestersemester     0.53514    0.46691   1.146   0.2517
route9am             0.17370    0.44520   0.390   0.6964
temp:semestersemester  0.10779    0.05741   1.878   0.0604 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
```
[1]

```
Null deviance: 30.406  on 19  degrees of freedom
Residual deviance: 13.833  on 15  degrees of freedom
AIC: 62.187
```

(b)

Temperature (temp) is significant [1]
Semester is not significant [½]
Route is not significant [½]
The interaction between temperature (temp) and semester is not significant at 5% significance level [½]
but it is close to being significant [½]

(iii)(a)

```
>Model2<- update(Model1,~.-route)
```
[2]

Or,

```
Model2 <- glm(Passengers~temp*semester,family="poisson" (link = "log"))
>summary(Model2)
Call:
glm(formula = Passengers ~ temp + semester + temp:semester, family = poisson(link
= "log"))
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.84542  -0.66323  -0.06209   0.43732   1.34790
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.44284    0.29121   1.521   0.1283
temp                -0.07452    0.03387  -2.200   0.0278 *
semestersemester     0.54602    0.46390   1.177   0.2392
temp:semestersemester  0.10012    0.05316   1.883   0.0597 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 (Dispersion parameter for poisson family taken to be 1)
Null deviance: 30.406  on 19  degrees of freedom
Residual deviance: 13.982  on 16  degrees of freedom
AIC: 60.336
```
[1]

(b)

The AIC has fallen from 62.187 to 60.336 - so new model has improved the initial model [1]

(iv)(a)

```
>Model3<- glm(Passengers~temp+temp:semester,family=poisson(link="log"))
>Modela<- glm(Passengers~temp+ semester,family=poisson(link="log"))
```
[2]

```
>Modelb<- glm(Passengers~temp*semester,family=poisson(link="log"))
```
[2]

```
>Model3$aic
 59.65976
```
[1]

```
>Modela$aic
 62.03591
```
[1]

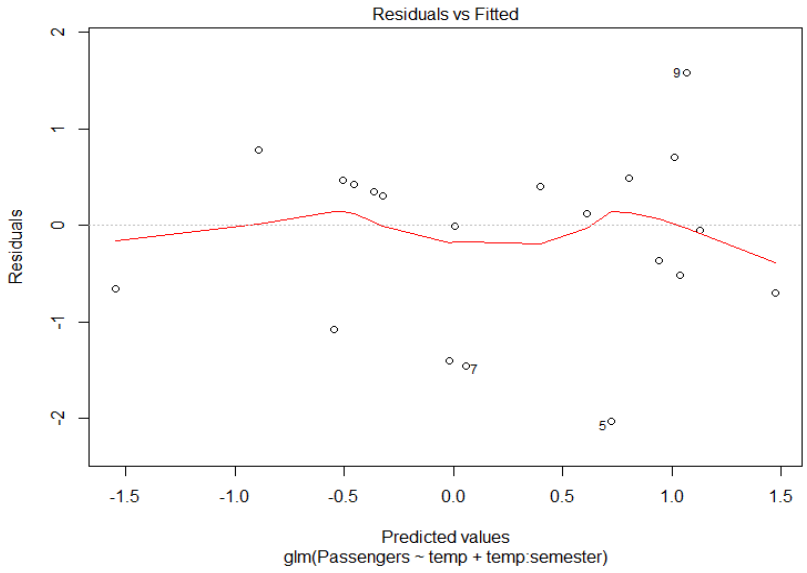
```
>Modelb$aic
 60.33588
```
[1]

Model3 has the lowest AIC compared with the other models. We conclude that Model3 outperforms the other models considered here [1]

(b)
Model3 doesn't include both of the main effects. Despite this, the model still suits the data well [1]

(v)(a)

```
> plot(Model3,1)
```
[2]




[1]

(b)
The residuals plot shows no patterns - exhibiting a fairly random scatter around zero with constant variance and no outliers [2]
The plot suggests that the model is appropriate [1]

(vi)

```
>predict(Model3, data.frame(temp=0,semester="semester",route="8am"),type =
        "response")
```
[3]
Predicted number is: 1.866568 [1]

**[Total 35]**

*A number of candidates did not attempt parts of his question.*
*Part (i) was answered poorly. Parts (ii) and (iii) were generally very well answered. However a number of candidates failed to comment on all aspects of the model in (ii). Answers in part (iv) were weak. Many candidates used deviance (anova) tests to answer this question. Note that the deviance test requires the compared models to be nested, and therefore cannot be used here on all three models. Part (v) was well answered. A common error here was plotting against the index, rather than the predicted values. Answers in part (vi) were mostly correct.*

**[Paper Total 100]**

# END OF EXAMINERS' REPORT