MATH97082

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May-June 2021

This paper is also taken for the relevant examination for the
Associateship of the Royal College of Science

Statistical Modelling 2

Date: Monday, 17 May 2021

Time: 09:00 to 11:30

Time Allowed: 2.5 hours

Upload Time Allowed: 30 minutes

This paper has 5 Questions.

Candidates should start their solutions to each question on a new sheet of paper.

Each sheet of paper should have your CID, Question Number and Page Number on the top.

Only use 1 side of the paper.

Allow margins for marking.

Any required additional material(s) will be provided.

Credit will be given for all questions attempted.

Each question carries equal weight.

SUBMIT YOUR ANSWERS AS SEPARATE PDFs TO THE RELEVANT DROPBOXES ON BLACKBOARD INCLUDING A COMPLETED COVERSHEET WITH YOUR CID NUMBER, QUESTION NUMBERS ANSWERED AND PAGE NUMBERS PER QUESTION.

1. (a) This question part concerns the normal linear model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$, with parameters estimated by least squares. Determine whether or not each of the following statements is true or false in general. Justify your answers briefly.

(i) The residuals all have identical variance.

(1 mark)

(ii) The errors all have identical variance.

(1 mark)

(iii) If the model fits well, then a plot of residuals $\widehat{\boldsymbol{\epsilon}}$ against the fitted values $\widehat{\boldsymbol{y}}$ should show no pattern.

(2 marks)

(iv) If the model fits well, then a plot of residuals $\widehat{\boldsymbol{\epsilon}}$ against the response values $\boldsymbol{y}$ should show no pattern.

(2 marks)

(b) A particular mouse gene has two variant forms, a and A. A study investigates the relationship between genetic type and the blood concentration of a particular protein. Mice have two copies of the relevant gene, so there are three types: aa, AA and Aa. These are coded in R as factors with levels `a`, `b` and `c`, respectively. Note that R considers levels of a factor in alphabetical order.

A linear model of blood concentration `y` is proposed with `genotype` as a categorical covariate, using Helmert contrasts. Let μ_{aa}, μ_{AA} and μ_{Aa} denote the mean blood concentrations in the three categories. R output for this model can be found on the following page. In what follows, let the parameters of the model be β_0, β_1 and β_2, denoting the `intercept`, `genotype1` and `genotype2` coefficients, respectively.

(i) Write down the form of a row of the design matrix for an observation in each of the three categories.

(3 marks)

(ii) State an estimate of the grand mean $\frac{\mu_{aa}+\mu_{AA}+\mu_{Aa}}{3}$.

(1 mark)

(iii) Test the hypothesis $\mu_{aa} = \mu_{AA}$.

(2 marks)

(iv) State in the context of the data the null hypothesis that corresponds to $\beta_2 = 0$. Perform the test and explain its conclusion in plain language.

(3 marks)

(v) Write out the variance-covariance matrix of the regression parameter estimators in terms of σ^2.

(3 marks)

(vi) State the number of degrees of freedom of the F-statistic, labelled ####, and write down the null hypothesis corresponding to this F-statistic.

(2 marks)

Question 1 continues on the following page

Continuation of Question 1

```
lm(formula = y ~ genotype, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.26428 -0.41579  0.08087  0.56774  1.19535 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.7310     0.2218  30.342 1.03e-12 ***
genotype1    -1.7377     0.2717  -6.396 3.42e-05 ***
genotype2     0.3813     0.1569   2.431   0.0317 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8592 on 12 degrees of freedom
Multiple R-squared:  0.796,Adjusted R-squared:  0.762 
F-statistic: 23.41 on #### DF,  p-value: 7.213e-05
```

(Total: 20 marks)

2. The R output at the end of the question shows the result of fitting a binomial regression model. The dataset consists of n observations in total, planted in different conditions to determine the effects of rainfall and soil nitrogen content on the germination of seeds. For each observation $i = 1, 2, \ldots, n$, the response variable y_i is the number of seeds that germinated out of $m = 90$ that were planted, x_i is the amount of rainfall, and z_i is the soil nitrogen content at time of planting. Both covariates are measured in arbitrary units.

(a) Explain why a binomial GLM is reasonable here. Comment on what you would expect to see in the residual plots if a normal linear model were fit to this data. (4 marks)

(b) State the link function that was used in this model.

(1 mark)

(c) Determine n, the number of observations.

(1 mark)

(d) Under the assumption that the model fits well, and that asymptotic results are valid, perform a test of the null hypothesis that nitrogen content has no effect on the probability of germination. State clearly the asymptotic results used.

(2 marks)

(e) Continuing to assume the validity of the model, and any necessary results, show how to use the values given to form an approximate 95% confidence interval for the multiplicative effect of a unit increase in rainfall on the odds of germination. You need not simplify your answer.

(2 marks)

(f) The seeds for the experiment are sold in packets of 15, and each observation used the all the seeds from six different packets. Suggest how this might violate the assumptions of the model, and identify values from the R output that support your suggestion.

(4 marks)

(g) Suggest how the model violations in (f) might be accommodated within the model. State the effect on your conclusions in parts (d) and (e).

(4 marks)

(h) Suggest how the biologist who collected the data might improve their experimental design for future studies.

(2 marks)

Question 2 continues on the following page

Continuation of Question 2

```
Call:
glm(formula = y ~ x + z, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.4917  -1.6280   0.1752   1.9098   5.1646

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.13614    0.06484   2.100 0.035762 *
x            0.44715    0.08191   5.459 4.79e-08 ***
z            0.17656    0.05174   3.412 0.000644 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 598.04  on 92  degrees of freedom
Residual deviance: 552.41  on 90  degrees of freedom
AIC: 1004.8
```

(Total: 20 marks)

3. This question concerns count data from an experiment studying the folding of artificially designed proteins. The folding of individual protein molecules is observed, and the experiment records counts of transitions between states, stored in the variable `counts`. The experimenters are interested in the relationship between `counts` and a predicted quality score from another model, stored in the variable `score`.

Initially, the experimenters fit a normal linear model. A scale-location plot of its residuals is shown in Figure 1. After consulting a statistician, they fit a Poisson GLM using its canonical link function. A scale location plot of its residuals is shown in Figure 2. The data were collected in two different experimental runs, which are labelled in Figures 1 and 2. The relationship between the variables of interest is not believed to be different for the different runs.

Where questions refer to completing R code, full credit will be given for mathematically correct responses. Syntax errors will not be penalized.

(a) Evaluate the fit of the linear model based on the scale-location plot given in Figure 1.

(3 marks)

(b) Explain the properties of the Poisson GLM (in contrast to the normal linear model) that might allow for a better fit.

(2 marks)

(c) Give the missing expressions in the code at the end of the question for:

(i) The deviance function `D` labelled ##1##

(2 marks)

(ii) The inverse link function `inv.link` labelled ##2##

(1 mark)

(iii) Initial values of β labelled ##3##

(1 mark)

(iv) The adjusted response labelled ##4##

(2 marks)

(v) Weights labelled ##5##

(2 marks)

(d) Contrast the stopping criteria used in the implementation below with those in the `glm` function that is supplied in R.

(3 marks)

(e) Comment on the scale-location plot for the Poisson model given in Figure 2.

(2 marks)

(f) The true response distribution in this example is better modelled by a negative binomial distribution. The preliminary consultation that resulted in a Poisson GLM recommendation was based only on the first run labelled with triangles in the residual plots. Suggest why that analysis found the Poisson model to be adequate.

(2 marks)

Question 3 continues on the following page

Continuation of Question 3

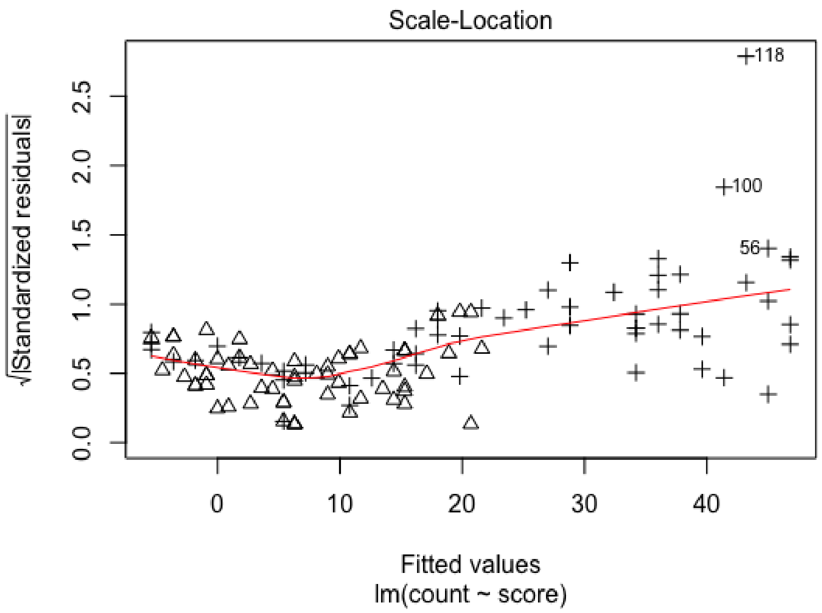

Figure 1: Scale-location plot for normal linear model. The different point shapes correspond to two different experimental runs.

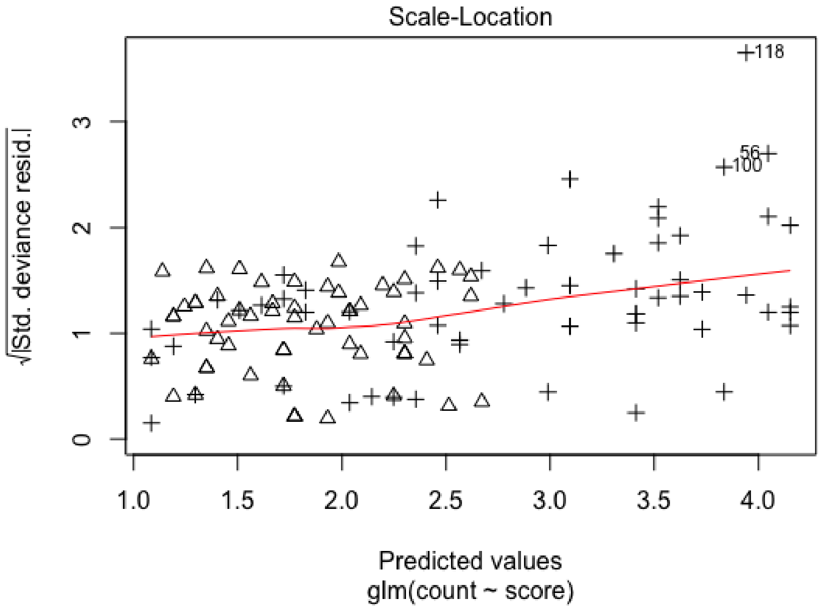

Figure 2: Scale-location plot for Poisson GLM.

Question 3 continues on the following page

Continuation of Question 3

```
D <- function(y,mu){
  ##1##
}
inv.link<-function(eta){
  ##2##
}


X<-cbind(1,score)
y<-count

jj <- 0

# get initial estimate
fit0<-lm(!##3##)
beta<-as.numeric(fit0$coefficients)
oldD <- D(y,inv.link(X%*%beta))

while(jj==0){
  eta <- X%*%beta
  mu <- inv.link(eta)
  z <- eta + ##4##
  w <- ##5##
  lmod <- lm(z~0+X, weights=w)
  beta <- as.numeric(lmod$coeff)
  newD <- D(y,inv.link(X%*%beta))
  control <- abs(newD-oldD)/(abs(newD)+0.1)
  if(control<1e-8)
    jj <- 1
    oldD <- newD
}
```

(Total: 20 marks)

4. A normal linear mixed model is written as $\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\nu} \sim N(0, \sigma_\nu^2 I_m)$ is a vector of length m, $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 I_n)$ is a vector of length n, independent of $\boldsymbol{\nu}$, X is the design matrix for the fixed effects, and Z is the model matrix for the random effects.

(a) Write down the distribution of the random variable $\boldsymbol{Y}$.

(3 marks)

(b) Suppose data are collected on the relationship between mathematics scores in national tests and a numerical measure of socio-economic status. Students are sampled from a large number of schools. Around 20 students are sampled from each school, although from some schools as few as 5 students were sampled. A linear mixed model is used to model the relationship between mathematics score and socio-economic status.

Explain the advantages of the linear mixed model, in comparison with:

(i) A *complete pooling* model, in which the data are combined to produce a single regression line.

(ii) A *no pooling* model in which a separate regression line is fitted for each school.

(4 marks)

(c) Two different linear mixed models are fit to data as described in (b) using the code given at the end of the question.

(i) State the code that should replace ##1## so that the models can be compared. Justify your answer.

(2 marks)

(ii) Determine the number of residual degrees of freedom for each of the models, and use your answers to verify the AIC for each of the two models.

(2 marks)

(iii) Justifying your answer briefly, determine which of the two models is to be preferred.

(2 marks)

(iv) Use the output to write down an expression for the intra-school correlation coefficient for model 2.

(2 marks)

(v) Considering the data context, comment on the difference in intra-school correlation coefficient estimated by the two models.

(3 marks)

(vi) Comment on the QQ-plot of the residuals given in the figure.

(2 marks)

Question 4 continues on the following page

Continuation of Question 4

```
mylme <- lmer(score~1+(1|school),data=mathres,REML=##1## )

summary(mylme)
     AIC      BIC   logLik deviance df.resid
  2119.5   2130.9  -1056.8   2113.5      ##2##

Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept)  2.058   1.435
 Residual             37.411   6.116
Number of obs: 326, groups:  school, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)  13.8023     0.7255   19.02

mylme2 <- lmer(score~SES+(1|school),data=mathres,REML=##1## )

     AIC      BIC   logLik deviance df.resid
  2110.7   2125.8  -1051.3   2102.7      =##3##

Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept)  0.04746 0.2179
 Residual             36.99420 6.0823
Number of obs: 326, groups:  school, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)  13.7119     0.3511  39.053
SES           1.9514     0.4345   4.491
```

Question 4 continues on the following page

Continuation of Question 4

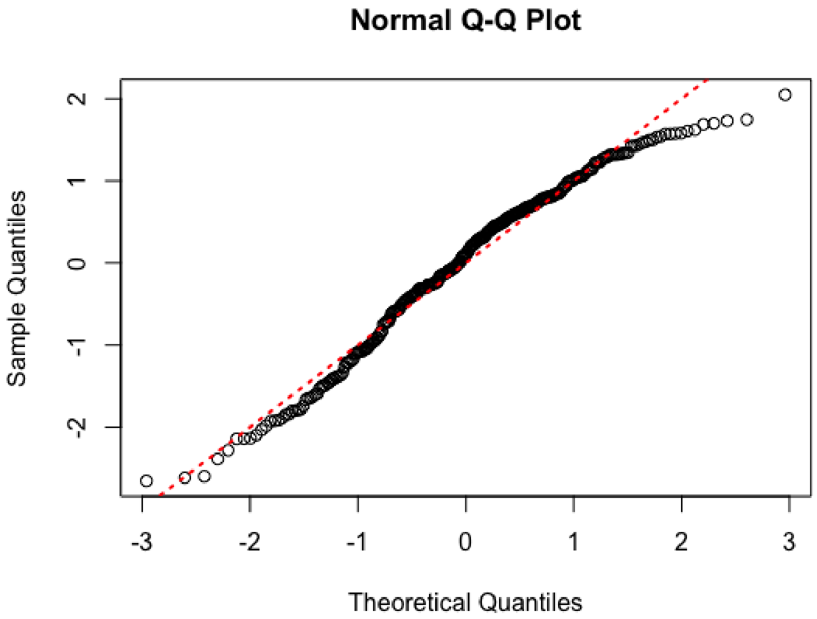

Figure 3: QQ plots of residuals standardized to unit variance for mylme2

(Total: 20 marks)

5. This question refers to the article *Overdispersion and Poisson Regression* by Berk and MacDonald.

(a) Explain what is meant by overdispersion in a generalized linear model, and give examples of when it arises.

(3 marks)

(b) Give a brief summary of the argument in the article, as to why the negative binomial is not usually the way to improve the models commonly used in criminology applications.

(4 marks)

(c) Stating clearly any results that you need, and any necessary conditions, justify the deduction in the third sentence on page 272,

It follows that for real data generated by a Poisson process, the mean should be approximately the same as the variance.

(2 marks)

(d) Write the probability mass function of the negative binomial in exponential family form. Identify the canonical parameter.

(3 marks)

(e) Give a brief derivation of equation (6).

(3 marks)

(f) Derive the expression for the variance of the negative binomial.

(2 marks)

(g) Justify the assertion below from page 281, noting any limitations.

The residual deviance is 511.6, approximately equal to the residual degrees of freedom, which is one indication that the model is performing as it should.

(3 marks)

(Total: 20 marks)

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2021

This paper is also taken for the relevant examination for the Associateship.

MATH65051

Statistical Modelling 2 (Solutions)

Setter's signature	Checker's signature	Editor's signature
....	....	....

1. (a) In what follows, let $P = X(X^TX)^{-1}X^T$ be the hat matrix, such that $\widehat{\boldsymbol{y}} = P\boldsymbol{y}$ and $\widehat{\boldsymbol{\epsilon}} = (I - P)\boldsymbol{y}$.

seen ⇓

1, A

(i)

$$\text{Cov}(\widehat{\boldsymbol{\epsilon}}) = \text{E}\left(\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}^T\right),$$

since $\text{E}(\widehat{\boldsymbol{\epsilon}}) = 0$.

Now

$$\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}^T = (I - P)\boldsymbol{y}\boldsymbol{y}^T(I - P),$$

so that

$$\text{Cov}(\widehat{\boldsymbol{\epsilon}}) = (I - P)\text{E}\left(\boldsymbol{y}\boldsymbol{y}^T\right)(I - P) = (I - P)\sigma^2 I(I - P) = \sigma^2(I - P),$$

since $(I - P)^2 = I - P$.
In general, $I - P$ need not have diagonal entries - different residuals will have different variance.

seen ⇓

1, A

(ii) We are given that $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$ - the errors do indeed have constant variance.

seen ⇓

2, A

(iii) By standard properties of projection matrices, $P^2 = P = P^T$, so that $\widehat{\boldsymbol{y}}^T\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y}P(I - P)\boldsymbol{y}$ and $P(I - P) = P - P^2 = 0$, so that $\widehat{\boldsymbol{\epsilon}}$ and $\widehat{\boldsymbol{y}}$ are uncorrelated normal random variables. A plot should show no pattern.

unseen ⇓

2, B

(iv) With the notation above, we see that

$$\boldsymbol{y}^T\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y}^T(I - P)\boldsymbol{y} = \boldsymbol{y}^T(I - P)^2\boldsymbol{y} = \boldsymbol{y}^T(I - P)^T(I - P)\boldsymbol{y} = RSS.$$

We see that the vectors $\widehat{\boldsymbol{\epsilon}}$ and $\boldsymbol{y}$ are in general positively correlated (unless the data fits the model perfectly). So a plot of residuals against observed response values will show a positive linear relationship.

seen/sim.seen ⇓

3, B

(b) (i) The design matrix of a factor with three levels, with Helmert contrasts has rows of the form
level a: $1, -1, -1$
level b: $1, 1, -1$
level c: $1, 0, 2$.

seen/sim.seen ⇓

1, A

(ii) The grand mean is estimated by the intercept of the linear model, which is $6.7(\pm 0.2)$.

unseen ⇓

2, C

(iii) Considering the design matrix, the coefficient β_1 labelled `genotype1` is an estimate of $\frac{1}{2}(\mu_{AA} - \mu_{aa})$. So a test of the null hypothesis $\beta_1 = 0$ is required. From the output given, the p-value for this test is extremely small, suggesting strong evidence against the null hypothesis.

unseen ⇓

3, D

(iv) For this design matrix, the coefficient β_2 labelled `genotype2` is an estimate of $\frac{1}{3}(\mu_{Aa} - \frac{\mu_{AA}+\mu_{aa}}{2})$. If $\beta_2 = 0$, then we have an additive dose response relationship between the number of copies of the A variant and the concentration of the protein: the average difference in concentration between individuals with zero copies and with one copy is the same as the average difference between individuals with one copy and with two copies.

If the model fit is reasonable, this analysis gives evidence against the null hypothesis that $\beta_2 = 0$. Hence , we conclude that

$$\mu_{Aa} > \frac{1}{2}\left(\mu_{AA} + \mu_{aa}\right).$$

Rearranging,

$$0 > \mu_{AA} - 2\mu_{Aa} + \mu_{aa} = \mu_{AA} - \mu_{Aa} - (\mu_{Aa} - \mu_{aa}),$$

so that

$$\mu_{Aa} - \mu_{aa} > \mu_{AA} - \mu_{Aa}$$

The increase in concentration of the protein with the second copy of A is smaller than the increase from the first copy.

unseen ⇓

3, C

(v) Note that since we have an equal number m of subjects in each group, the columns of the design matrix are orthogonal. Hence the matrix X^TX is diagonal. The ith diagonal entry is given by the inner product of column i with itself.

For the first column, $1 \times 1 + \ldots + 1 \times 1 = 3m$.

For the second column, $(-1) \times (-1) + \ldots (-1) \times (-1) + 1 \times 1 + \ldots + 1 \times 1 + 0 + \ldots + 0 = 2m$.

For the third column, $(-1) \times (-1) + \ldots (-1) \times (-1) + (-1) \times (-1) + \ldots + (-1) \times (-1) + 2 \times 2 + \ldots + 2 \times 2 = 6m$.
hence

$$X^TX = \begin{pmatrix} 3m & 0 & 0 \\ 0 & 2m & 0 \\ 0 & 0 & 6m \end{pmatrix}$$

The variance-covariance matrix of the estimators is then given by

$$\sigma^2(X^TX)^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{3m} & 0 & 0 \\ 0 & \frac{1}{2m} & 0 \\ 0 & 0 & \frac{1}{6m} \end{pmatrix}$$

sim. seen ⇓

2, C

(vi) The model has $p = 3$ parameters and $n = 15$ observations. The F statistic in the R output corresponds to a test of the null hypothesis that $\beta_1 = \beta_2 = 0$. This null model has one degree of freedom. Hence the F statistic has an $F(3-1, 15-3) = F(2, 12)$ distribution.

sim. seen ⇓

4, A

2. (a)
   - $*$ Response variable is the number of successes in a fixed number of binary trials.
   - $*$ With a suitable choice of link function, Binomial GLM allows a real-valued linear predictor to be smoothly mapped to the probability scale.
   - $*$ Accounts for the non-trivial mean-variance relationship for a binomial proportion.
   - $*$ Scale-location plot would show non-constant variance of standardized residuals
   - $*$ Normal QQ-plot of residuals may show pronounced non-normality (depending on m, n and p)

sim. seen ⇓

1, A

(b) Logistic link function (since it is the default for binomial GLMs).

sim. seen ⇓

1, A

(c) From the output, the model has $n - p = 90$ residual degrees of freedom and $p = 3$ parameters, so $n = 93$.

sim. seen ⇓

2, B

(d) Assuming that n and m are sufficiently large, $\widehat{\boldsymbol{\beta}}$ should be normally distributed around its populaton value. The row for z in the table gives an estimate of the regression coefficient for nitrogen content, and a test of the null hypothesis of zero effect. We see from the p-value that there is strong evidence against the null hypothesis. The observed relationship is stronger than the associations that might plausibly arise in random samples of size n if there were no effect. We conclude that in this experiment, higher nitrogen content is positively associated with the odds of germination,

sim. seen ⇓

2, B

(e) Continuing to assume that $\widehat{\boldsymbol{\beta}}$ is normally distributed around its true value, an approximate 95% confidence interval for β_1, the regression coefficient for rainfall, is $(0.45 \pm 1.96 \times 0.08)$. This is a confidence interval on the linear predictor scale. Since $\eta = \log(\frac{p}{1-p})$, exponentiation of the endpoints of the interval gives an approximate 95% confidence interval for the multiplicative effect on the odds of a unit change in x. Hence a suitable interval is

$$(\exp(0.45 - 1.96 \times 0.08), \exp(0.45 + 1.96 \times 0.08)) = (1.3, 1.8)$$

Final numerical values not required.

unseen ⇓

4, D

(f)
   - $*$ Seeds from the same packet may be correlated (storage conditions, genetics).
   - $*$ So m correlated trials rather than m independent trials.
   - $*$ Correlated observations give rise to over-dispersion.
   - $*$ Assuming asymptotic results hold, scaled deviance should be $\chi^2(n-p)$ distributed.
   - $*$ This means $D/\phi \approx n - p$. But for the binomial $\phi = 1$.
   - $*$ The R output gives 552.41 and $n - p = 90$. This suggests $\phi \approx 6 \gg 1$.

unseen ⇓

(g)
- $*$ Use quasi-binomial model: estimate ϕ from the data. 2, B
- $*$ This leaves estimates of the regression coefficients unchanged.
- $*$ but inflates their standard errors by $\sqrt{\phi}$. 2, C
- $*$ Effect of nitrogen content no longer significantly different from zero.
- $*$ Effect of rainfall is still significantly different from zero, but confidence intervals now wider.

sim. seen ⇓

(h) Randomize seeds from different packets across conditions. If packet-to-packet variability is substantial, could include a random effect corresponding to packet. 2, C

3. (a) Clear evidence of heterogeneity of variance in the scale-location plot. This contradicts the constant variance assumption underlying the linear model. Observation 118 is very unusual $2.5^2 = 6.25$, extremely atypical for a standard normal variable. Clear difference in mean level between the two groups - because of the heterogeneity of variance, this means that may be exerting undue influence over the model. This could be checked with a residuals vs leverage plot.

sim. seen ⇓

3, A

(b) A Poisson GLM accommodates a non-constant mean-variance relationship. Specifically, in a Poisson GLM, the (conditional) variance of an observation with fitted value μ is also μ. This means that the model recognises that larger fitted values will correspond to more variable observations. The effect in this example (if the relationship is correct) would be to reduce the extreme influence of the higher-variance points on the right hand side of the plot.

sim. seen ⇓

1, A

1, B

(c) (i)

sim. seen ⇓

2, A

```
D <- function(y,mu){
  a <- y*log(y/mu)
  b <- -y + mu
  a[y==0] <- 0
  2*sum(a+b)
}
```

(ii)

sim. seen ⇓

1, A

```
inv.link<-function(eta){
  exp(eta)
}
```

(iii)

sim. seen ⇓

1, B

```
fit0<-lm(log(y+1e-6)~0+X)
```

- Need to cope with the possibility of zeros.
- May use weighted regression.

(iv)

sim. seen ⇓

2, A

```
z <- eta+((y-mu)/mu)
```

(v)

sim. seen ⇓

2, A

```
w <- mu
```

(d)

- · glm also uses a stopping criterion based on the change in deviance, as here.
- · In glm the absolute and relative tolerances can be specified as parameters.
- · In addition, glm will perform a fixed maximum number of iterations (25 by default).
- · Whereas the loop considered here might, in principle never terminate.
- · Since for Poisson regression the MLE may not exist (likelihood surface can have a monotonic increasing direction), it is good practice to have a loop that terminates.

unseen ⇓

1, C

2, D

sim. seen ⇓

2, C

(e)

- ∗ Still evidence of unmodelled mean-variance relationship, particularly for points in the triangle group.
- ∗ Relatively flat smoothing line in the left-hand side of the plot.
- ∗ Extreme points have, if anything, become more extreme.
- ∗ Conclude that the mean-variance relationship is still not correctly specified.
- ∗ Should investigate including an additional parameter for group.
- ∗ Should check the extent of the influence of the most extreme points on the overall model fit.

sim. seen ⇓

2, D

(f)

- ∗ Can see from the Poisson GLM that the smoother line is relatively flat for predicted values 1.0 to 2.0 on the linear predictor scale.
- ∗ Essentially all the triangular points lie in the region for which the smoother is roughly flat.
- ∗ The variance function for the negative binomial is $V(\mu) = \mu(1 + \frac{\mu}{r}) \approx \mu$ when $\mu \ll r$. For a dataset wholly in this range, a Poisson fit should be close to that obtained using the true negative binomial model.

4. (a) As a linear combination of multivariate normal variables, Y must be multivariate normal. It suffices to compute its mean vector and variance-covariance matrix. For the mean

$\boxed{\text{seen} \Downarrow}$

$\boxed{3, \text{A}}$

$$
\begin{aligned}
\mathrm{E}(\boldsymbol{Y}) &= \mathrm{E}(X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon}) \\
&= X\boldsymbol{\beta} + Z\underbrace{\mathrm{E}(\boldsymbol{\nu})}_{=0} + \underbrace{\mathrm{E}(\boldsymbol{\epsilon})}_{=0} \\
&= X\boldsymbol{\beta}
\end{aligned}
$$

And for the variance-covariance matrix,

$$
\begin{aligned}
\mathrm{Cov}(\boldsymbol{Y}) &= \mathrm{Cov}(X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon}) \\
&= Z\mathrm{Cov}(\boldsymbol{\nu})Z^T + \mathrm{Cov}(\boldsymbol{\epsilon}) \quad \text{(since } \boldsymbol{\epsilon}\text{, } \boldsymbol{\nu} \text{ indep.)} \\
&= ZI_m\sigma_\nu^2 Z^T + I_n\sigma_\epsilon^2 \\
&= \sigma_\epsilon^2 \left(ZI_m \frac{\sigma_\nu^2}{\sigma_\epsilon^2} Z^T + I_n \right) \\
&= \sigma_\epsilon^2 \left(I_n + Z\Psi Z^T + I_m \right),
\end{aligned}
$$

where $\Psi = \frac{\sigma_\nu^2}{\sigma_\epsilon^2} I_n$.

$\boxed{\text{seen} \Downarrow}$

$\boxed{4, \text{A}}$

(b) A linear mixed model respects the structure of the data, with students nested within schools. A *complete pooling* analysis ignores the fact that observations within a single school are likely to be correlated. A complete pooling model assumes observations are independent, so that even if the resulting parameter estimates are reasonable, the uncertainty estimates will be too small. Equivalently, complete pooling ignores school-to-school variability.

A *no pooling* analysis treats schools as entirely separate entities, and so makes it difficult to generalize conclusions to unseen schools. By fitting independent models, we ignore potentially relevant information - the resulting coefficients will have large standard errors, particularly for schools with as few as five observations. The linear mixed model should shrink extreme regression coefficients from schools with very few observations back to more reasonable values.

$\boxed{\text{seen} \Downarrow}$

$\boxed{2, \text{B}}$

(c) (i) `REML = FALSE`. REML computes the likelihood on a transformed data set LY, where L is a linear transformation that eliminates the fixed effects. For two different fixed effect structures, the transformations L would in general be different, leading to likelihoods that are not computed on the same data. (For nested fixed effects, the restricted likelihoods would have different dimensions as probability densities).

seen ⇓

2, A

(ii) For the model mylme, there is one fixed effect parameter and two variance components, so $p = 3$, and so

$$AIC = -1056.8 \times (-2) + 2 \times 3 = 2119.5$$

For the model mylme2, there are two fixed effect parameters and two variance components, so $p = 4$, and so

$$AIC = -1056.8 \times (-2) + 2 \times 4 = 2110.6$$

sim. seen ⇓

2, B

(iii) Asymptotically, choosing the model with the smaller AIC should choose the model with the smaller leave-one-out cross-validation error. mylme2 has the smaller AIC, so it is to be preferred.

sim. seen ⇓

2, A

(iv)

$$\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\epsilon^2} = \frac{0.04746}{0.04746 + 36.99420} \approx 0.13\%$$

sim. seen ⇓

3, D

(v)
- Intra-class correlation in the smaller model is substantially larger $\frac{2}{2+37}$.
- In the smaller model, maths scores within schools appear correlated.
- But much of this correlation appears to be due to socio-economic status.

unseen ⇓

2, D

(vi)
- Appears OK in the very centre of the distribution
- Somewhat asymmetric in the tails: upper tail has only one observation greater than 2; lower tail has substantially more.
- Tails lighter than the normal.
- Consistent with finite min and max score of the test. May be worth transforming data before fitting the model.

5. (a) Overdispersion is the presence of excess variation in the fitted values of a model, beyond what would be expected from the mean-variance relationship induced by the GLM. Dispersion is the manifestation of unmodelled components of variabilty, e.g. dependence between observations or excess variability in subjects with the same value of the linear predictor. $\boxed{3,\ \mathrm{M}}$

$\boxed{4,\ \mathrm{M}}$

(b)

- ∗ Negative binomial models are have been used in quantitative criminology work to account for overdispersion in Poisson GLMs.
- ∗ But the negative binomial model only captures a specific kind of overdispersion relative to the Poisson.
- ∗ Where overdispersion in the Poisson GLM results from a misspecified linear predictor, i.e. omitted variable bias, using a a negative binomial model will not change the (biased) coefficient estimates.

$\boxed{2,\ \mathrm{M}}$

(c) By the weak law of large numbers, for any $k > 0$ for which the expectation exists, $\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{\mathcal{P}} \mathrm{E}(X^k)$, so in the setting of interest here, moment estimators are consistent estimators of the moments of the Poisson distribution.

$\boxed{3,\ \mathrm{M}}$

(d)

$$
\begin{aligned}
f(y;\lambda,\theta) &= \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)}\left(\frac{\lambda}{\lambda+\theta}\right)^{y}\left(\frac{\theta}{\lambda+\theta}\right)^{\theta} \\
&= \exp\left(\log\Gamma(y+\theta) - \log\Gamma(y+1) - \log\Gamma(\theta) + y\log\left(\frac{\lambda}{\lambda+\theta}\right) + \theta\log\left(\frac{\theta}{\lambda+\theta}\right)\right)
\end{aligned}
$$

For fixed, known θ, this is in exponential family form, and the canonical parameter can be identified as the factor multiplying y, which is $\log\frac{\lambda}{\lambda+\theta}$.

$\boxed{3,\ \mathrm{M}}$

(e)

$$
\begin{aligned}
f(y;\lambda,\theta) &= \int_0^\infty \exp(-\lambda t)\frac{(\lambda t)^y}{y!}\frac{\theta^\theta}{\Gamma(\theta)}t^{\theta-1}\exp(-\theta t)\ dt = \int_0^\infty t^{y+\theta-1}\exp(-(\lambda+\theta)t)\frac{\lambda^y}{y!}\frac{\theta^\theta}{\Gamma(\theta)}\ dt \\
&= \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}\frac{\lambda^y\theta^\theta}{(\lambda+\theta)^{y+\theta}}\int_0^\infty(\lambda+\theta)^{y+\theta}\frac{t^{y+\theta-1}}{\Gamma(y+\theta)}\exp(-(\lambda+\theta)t)\ dt \\
&= \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}\frac{\lambda^y\theta^\theta}{(\lambda+\theta)^{y+\theta}},
\end{aligned}
$$

since the final integrand is a probability density, and so the integral is 1.

(f) Consider the formulation above where $Y|T \sim \textsc{Poisson}(\lambda T)$ and $T \sim \Gamma(\theta, 1)$. Note that $\mathrm{E}(T) = 1$ and $\mathrm{Var}(t) = \frac{1}{\theta}$. Then, by the law of total variance $\mathrm{Var}(Y) = \mathrm{E}\left[\mathrm{Var}(Y|T)\right] + \mathrm{Var}\left[\mathrm{E}(Y|T)\right] = \mathrm{E}(\lambda T) + \mathrm{Var}(\lambda T) = \lambda + \frac{\lambda^2}{\theta} = \lambda\left(1 + \frac{\lambda}{\theta}\right)$ as given in the article. $\boxed{2, \mathrm{M}}$

$\boxed{3, \mathrm{M}}$

(g) The result being used here is the approximate asymptotic chi-square distribution of the scaled deviance of a GLM. For a Poisson GLM, which has dispersion parameter 1, the scaled deviance is just the deviance. Under the assumption that the model is fitting well, the deviance should be (very roughly) $\chi^2(n-p)$, where $n-p$ is the number of residual degrees of freedom once p parameters have been estimated using n observations. This distribution has mean $n-p$, so the observed value of the deviance is consistent with what would be expected if the model were fitting well.

Review of mark distribution:

Total A marks: 33 of 32 marks
Total B marks: 17 of 20 marks
Total C marks: 14 of 12 marks
Total D marks: 16 of 16 marks
Total marks: 100 of 80 marks
Total Mastery marks: 20 of 20 marks

If your module is taught across multiple year levels, you might have received this form for each level of the module. You are only required to fill this out once for each question.

Please record below, some brief but non-trivial comments for students about how well (or otherwise) the questions were answered. For example, you may wish to comment on common errors and misconceptions, or areas where students have done well. These comments should note any errors in and corrections to the paper. These comments will be made available to students via the MathsCentral Blackboard site and should not contain any information which identifies individual candidates. Any comments which should be kept confidential should be included as confidential comments for the Exam Board and Externals. If you would like to add formulas, please include a sperate pdf file with your email.

ExamModuleCode	QuestionNumber	Comments for Students
MATH96051MATH97082	1	This question attracted many good responses. Part b) iv was generally found difficult, particularly with interpretation. In b) vi, there was some confusion about the null model. Some candidates found b) difficult overall, because they failed to use Helmert contrasts
MATH96051MATH97082	2	Generally all parts were well done - though sometimes in part f) more clarity was required on the discussion of the issue of correlation of seeds from the same packet and what modelling assumptions were being violated.
MATH96051MATH97082	3 a)	Sometimes well completed - but generally additional issues in the plots (other than heterogenity of variance) such as outliers, difference in patterns between the runs etc were missed
MATH96051MATH97082	3 b)	The Poisson GLM and its accommodation of a non-constant mean/variance relationship was often not mentioned
MATH96051MATH97082	3 c)	Generally well completed - but some lack of attention to detail present.
MATH96051MATH97082	3 d)	The main stopping criteria was generally noted - but often the additional stopping criteria of ' number of iterations' and the implications of having this or not was missed.
MATH96051MATH97082	3 e)	Sometimes completed well - but often the 'flat smoother region' for one of the runs was missed

MATH96051MATH97082	3 f)	Often good note was made of the difference in the runs - and sometimes (though not often) a good explanation of why this might be was also provided.
MATH96051MATH97082	4 a)	Almost always correct - but often literally just a statement with no justification/explanation for the terms
MATH96051MATH97082	4 b)	Sometimes completed well - but generally more depth was required.
MATH96051MATH97082	4 c) i) - iv)	These were mainly calcuations (or a definition) and were generally completed accurately - so an opportunity to gain marks
MATH96051MATH97082	4 v) & vi)	Sometimes completed well - but generally more depth was required.
MATH96051MATH97082	5	Overall, this question was well done. I was pleased with how well candidates had understood the article and its shortcomings. Derivations from earlier statistics work were well understood