Imperial College
London

MATH97082 MATH70044

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May-June 2022

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

Statistical Modelling 2

Date: 06 June 2022

Time: 09:00 – 11:30 (BST)

Time Allowed: 2:30 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 ONE PDF TO THE RELEVANT DROPBOX ON BLACKBOARD WITH COMPLETED COVERSHEETS WITH YOUR CID NUMBER, QUESTION NUMBERS ANSWERED AND PAGE NUMBERS PER QUESTION.

1. This question concerns normal linear models of the form

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$, with parameters estimated by least squares.

Below, six models of the form above are estimated in R. Assume that `x` is a continuous covariate and `z` is a categorical covariate, taking values labelled `0` and `1`. You may assume that the design matrix of each model has full rank.

```
fit1 <- lm(y ~ x)
fit2 <- lm(y ~ x + z)
fit3 <- lm(y ~ x * z)
fit4 <- lm(y ~ 1)
fit5 <- lm(y ~ 0 + x)
fit6 <- lm(y ~ 0 + z)
```

Abridged output for one of the models is given below.

```
summary(fit6)

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
z0   1.8713     0.4332    4.32 7.81e-05 ***
z1   2.2413     0.2166   10.35 8.17e-14 ***

Residual standard error: 1.37 on 48 degrees of freedom
Multiple R-squared:  0.7237,Adjusted R-squared:  0.7122
```

(a) State the number of regression parameters estimated in

(i) `fit1`

(ii) `fit3`

(iii) `fit5`

(3 marks)

(b) State two models that are not nested. (1 mark)

Question 1 continues on the following page

Continuation of Question 1

(c) State, giving brief reasons for your answers,

(i) the model with the highest value of R^2.

(2 marks)

(ii) the number of degrees of freedom of the t-statistic for testing the coefficient for `x` in `fit1`.

(2 marks)

(iii) the number of degrees of freedom of the F-distribution when performing the test `anova(fit1, fit3)`.

(2 marks)

(d) Justifying your answers carefully, identify all of the models above for which it is possible that the residuals do not sum to zero.

(4 marks)

(e) Give a condition on the vector `x` for the design matrix in `fit1` to have full rank.

(1 mark)

(f) Determine the number of observations in each of the categories defined by `z`.

(3 marks)

(g) Interpret the parameter estimates given in the output for `fit6`.

(2 marks)

(Total: 20 marks)

2. This question concerns defects observed in a type of industrial machine. The R output shows the result of fitting a binomial generalized linear model.

The data used to fit the model consists of n independent observations (n_i, x_i, y_i), where n_i is the number of machines surveyed, x_i is the number of hours of operation, common to all machines in observation i, and y_i is the number of machines that were found to be defective. In R, these variables are denoted `machines`, `hours` and `defects`, respectively.

```
response <- cbind(bin_dat$defects,
                  bin_dat$machines - bin_dat$defects)
fit0 <- glm(response ~ hours, family = binomial, data = bin_dat)

summary(fit0)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.9235966  0.3779589 -10.381   <2e-16 ***
hours        0.0009992  0.0001142   8.754   <2e-16 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 112.670  on 10  degrees of freedom
Residual deviance:  10.331  on  9  degrees of freedom
AIC: 49.808

d_t <- bin_dat$machines*predict(fit0, type = "response")
sum(d_t)

[1] 106
```

(a) Explain in detail the numerical procedure for fitting a binomial generalized model. Include in your answer

- How the numerical method works
- How to initialize the method
- A suitable stopping criterion.

(5 marks)

(b) Under the assumption that the model fits well, state the effect of an additional 1000 hours of running time on the odds that machine is defective.

(2 marks)

Question 2 continues on the following page

Continuation of Question 2

(c) Stating any necessary asymptotic results, explain how to produce a confidence interval for the effect on the odds found in (b).

(2 marks)

(d) Comment briefly on the value of the residual deviance, making any necessary distributional assumptions.

(2 marks)

(e) State what is stored in the variable `d_t`, giving the meaning of the integer output in the context of the data. (2 marks)

In an alternative model, a machine becomes defective when its stress S exceeds the threshold t, which is common to all machines. The stress of a machine cannot be observed directly, but whether or not a machine is defective is observed. S is assumed to be a random variable, independent of the stress of all other machines. Its distribution after x hours of operation is normal with constant variance σ^2 and a mean τ_x that changes linearly with the number of hours of operation,

$$\tau_x = \alpha_0 + \alpha_1 x.$$

Let Y_{ij} be a binary random variable indicating whether or not the jth machine in the ith sample is defective.

(f) Determine a function G such that

$$\Pr(Y_{ij} = 1|x_i) = G(\eta_i),$$

where $\eta_i = \beta_0 + \beta x_i$ and the parameters β_0 and β_1, which are functions of α_0, α_1, t and σ, could be estimated from the data.

(3 marks)

(g) By considering the first order Taylor expansion of $G(\eta)$ about a value η_0 such that the corresponding fitted value $\mu_0 \approx \frac{1}{2}$, determine the approximate relationship between the regression parameters estimated by the model in (f) and those esimated in `fit0`.

(4 marks)

(Total: 20 marks)

3. This question concerns a dataset collected by a researcher interested in the variola virus, which causes small lesions to develop in living tissues.

An experiment was conducted in which chicken eggs were exposed to different concentrations of the virus. After allowing time for the virus to develop, the lesions on each egg's membrane were counted. The code below fits two different models in R. The response variable `y` contains the number of lesions counted on each egg, and the covariate `x` is the $\log_2$ of the dilution factor of the virus concentration. This means that for an observation with `x = 1`, the virus concentration has been diluted by a factor of 2, i.e. it has half the virus concentration of an observation with `x = 0`.

```
fit0 <- lm (y ~ x, data = dat)
fit1 <- glm(y ~ x, family = "poisson", data = dat)

summary(fit1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.26793    0.02255  233.60   <2e-16 ***
x           -0.68094    0.01544  -44.09   <2e-16 ***

sum(residuals(fit1, type = "pearson")^2)
[1] 291.5915

nrow(dat)
[1] 48
```

(a) Considering the left plot in Figure 1, comment on the suitability of the linear model `fit0`.

(3 marks)

(b) State the link function that is used in the Poisson model. Justify the choice of link function with reference to the modelling context.

(3 marks)

(c) Assuming the Poisson model is adequate, give a plain language summary of the relationship between dilution factor and the mean number of lesions observed.

(3 marks)

Question 3 continues on the following page

Continuation of Question 3

(d) State the approximate sampling distribution of the maximum likelihood estimators $\hat{\beta}$ of the parameters of the model, in terms of the information matrix $\mathcal{J}(\beta)$. Hence explain how to construct an approximate 95% confidence interval for the mean response for an observation with a given value of the dilution factor, x_0. State any additional information needed, beyond that given in the R output.

(4 marks)

(e) Identify a feature of the output given that suggests poor model fit for the Poisson model.

(2 marks)

(f) Suggest how the output of the Poisson model can be modified, as an approximate solution to the problem in (e). State the effect of the change on the confidence interval in (d).

(3 marks)

(g) The right plot in Figure 1 shows the relationship between the log of the variance of the count at each dilution and the log of the mean count. Suggest how this information could be used to select an alternative model. (2 marks)

(Total: 20 marks)

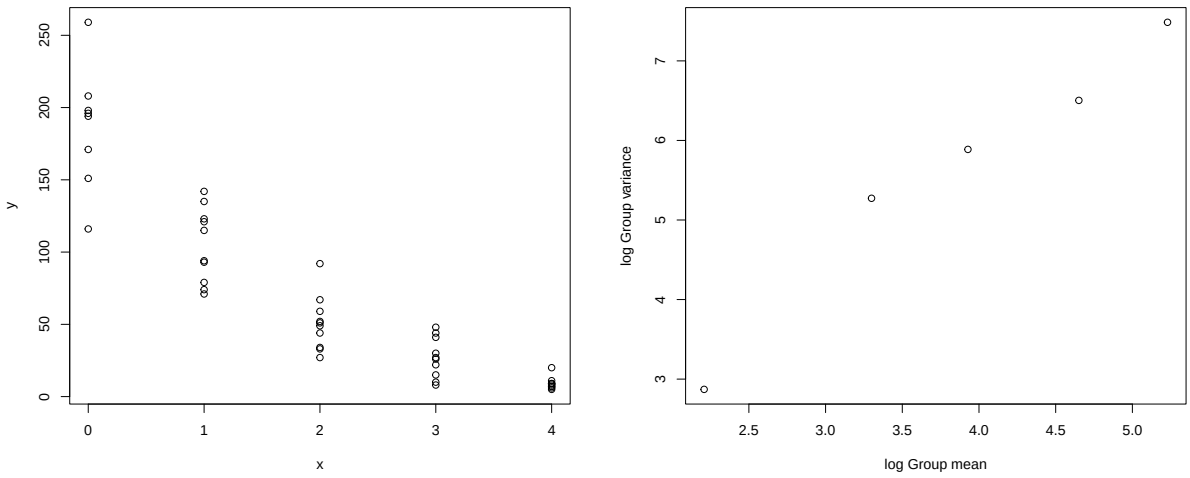


Figure 1: Left: Plot of the variables in Question 3. x is the $\log_2$ dilution factor and y is the lesion count. Right: Log of the count variance for each dilution level plotted against the log of the corresponding count mean.

4. (a) The data in this part of the question come from a market research study, which evaluated four different beauty products. The study asked 8 different members of the public (the subjects) to evaluate each of the products, giving a score to each.

Initially, the researchers fit a linear model to the data. Abridged R output is given below.

```
fit_lm <- lm( score ~ product, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.500      1.072   2.333 0.027079 *
productP2      6.458      1.516   4.261 0.000208 ***
productP3      3.958      1.516   2.612 0.014322 *
productP4      1.458      1.516   0.962 0.344199

Residual standard error: 3.031 on 28 degrees of freedom
Multiple R-squared:  0.4299,Adjusted R-squared:  0.3688
F-statistic: 7.038 on 3 and 28 DF,  p-value: 0.001135
```

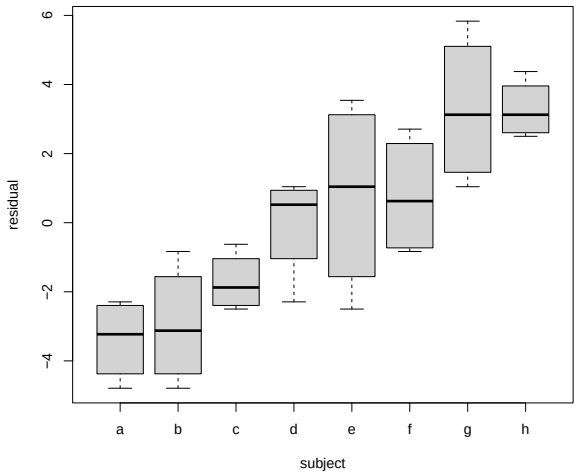


Figure 2: Box plot of residuals by participant for Question 4(a).

(i) Comment on the output of the model and the diagnostic plots in Figure 2. (3 marks)

(ii) Use the output to write down the mean score for the product P2. (1 mark)

Question 4 continues on the following page

Continuation of Question 4

(iii) A second model is fit below. Explain the assumptions behind this model, and suggest why it would be preferred to a model that includes a fixed effect for subject. State briefly the method used to fit the model.

(4 marks)

```
fit_re <- lmer( score ~ product + (1|subject), data = dat)
summary(fit_re)

Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 5.696    2.387
 Residual             3.493    1.869
Number of obs: 32, groups:  subject, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)   2.5000     1.0717   2.333
productP2     6.4583     0.9345   6.911
productP3     3.9583     0.9345   4.236
productP4     1.4583     0.9345   1.561
```

(iv) Calculate the intra-class correlation coefficient (numerical simplification not required).

(2 marks)

(v) Explain briefly how the experiment should be conducted, if the models are to produce unbiased estimates of the scores for each product. (2 marks)

(b) The following R code and output relates to a randomized controlled trial aiming to compare the effect of a new drug with the standard existing treatment. For each patient, observations are made of a severity score for the condition at six equally spaced timepoints. Lower scores indicate less severe disease, and so are better. The first measurement is at time 0, immediately before the drug is administered, and the first measurement is one day later. The treatment group is coded as a binary variable with 0 indicating the standard treatment group and 1 indicates the new drug.

Question 4 continues on the following page

Continuation of Question 4

```
fit0 <- lmer(score ~ treatment * time + (1| id), data = df, REML = FALSE)
summary(fit0)

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 1.387    1.178
 Residual             5.916    2.432
Number of obs: 84, groups:  id, 14

Fixed effects:
               Estimate Std. Error t value
(Intercept)    147.6357     0.8005 184.425
treatment       -0.8962     1.1321  -0.792
time             1.4384     0.2198   6.545
treatment:time  -1.7496     0.3108  -5.630
```

In parts (i) and (ii) below, you may assume that `fit0` provides a reasonable fit to the data, and that the sampling distribution of the fixed effect parameters is normal.

(i) Comment on whether there is any evidence for differences between the groups at time 0. (1 mark)

(ii) Explain what can be concluded about the effectiveness of the treatment compared with standard existing treatment. Sketch the expected score as a function of time for the two treatment groups. (3 marks)

(iii) Comment on the correlation assumed in this model between different measurements on the same individual. (1 mark)

(iv) An alternative model is now fit using the command below. Explain how this model differs from that in `fit0`, and suggest how the two models might be compared in practice. (3 marks)

```
fit1 <- lmer(score ~ treatment * time + (time | id), data = df,
 REML = FALSE)
```

(Total: 20 marks)

5. *Parts (a) to (e) refer to the faith vs gender example discussed in section 3.3.4 of the extract provided.*

(a) State the null hypothesis that is tested when the models `mod.0` and `mod.1` are compared. State the conclusion of the hypothesis test in plain language.

(3 marks)

(b) Explain why the parameters in the model at the top of page 139 are *"obviously not identifiable"*.

(2 marks)

(c) Tha aim of this part of the question is to verify a particular case of the statement on page 139, that *"when the total number of subjects in the table ... [is] fixed, then it can be shown that the correct likelihood can be written as a product of Poisson p.m.fs, conditional on various fixed quantities"* .

Suppose that $Y_1, \ldots, Y_m$ are independent Poisson random variables with $Y_j \sim \text{POISSON}(\lambda_j)$. For the natural number $n \geq 0$, show that the joint distribution of $(Y_1, Y_2, \ldots, Y_m)$ conditional on the event $\{\sum_{j=1}^{m} Y_j = n\}$ is multinomial, with parameters that you should determine.

(3 marks)

(d) Show explicitly how the third fitted value `377.901` can be written in terms of coefficients of `mod.0`

(2 marks)

(e) Explain from first principles why the sum of the fitted values for males is equal to the number of males in the observed data.

(3 marks)

Parts (f) to (h) refer to the sole eggs example in section 3.3.5 of the extract provided.

(f) Explain, in terms that would be accessible to an ecologist without much statistical training, why the quasi-Poisson model adopted in the text might be expected to produce a better-fitting model than a normal linear model.

(2 marks)

(g) Explain why the summary of the model `b` on page 144 has `AIC: NA`.

(1 mark)

(h) One problem with the models explored is the excess of zeros in the diagnostic plots. An alternative model might consider only the observations with positive observed egg density. Show that if $Y \sim \text{POISSON}(\lambda)$, the conditioned random variable $Y|Y > 0$ defines an exponential family. Identify the mean $\mu(\lambda)$ and variance function $V(\lambda)$ that should be used in the corresponding quasi-likelihood model.

(4 marks)

(Total: 20 marks)

```
       Df Deviance   AIC     LRT Pr(>Chi)
<none>      75.198 133.2
tf     14  261.702 291.7 186.504   <2e-16 ***
trt     1   76.799 132.8   1.601   0.2057
```

So although the effect of `auto` relative to `allo` is estimated to be 0.70, the effect is not significant, and its inclusion is only barely supported by AIC. The positive effect means that the hazard is estimated to be higher for the `auto` group. To visualize what is happening it can help to plot the survival curve estimates for the two groups. The following code does this by following section 3.1.10.

```
te <- sort(unique(bone$t[bone$d==1])) ## event times
## predict survivor function for "allo"...
pd <- data.frame(tf=factor(te),trt=bone$trt[1])
fv <- predict(b,pd)
H <- cumsum(exp(fv)) ## cumulative hazard
plot(stepfun(te,c(1,exp(-H))),do.points=FALSE,ylim=c(0,1),
     xlim=c(0,550),main="",ylab="S(t)",xlab="t (days)")
## add s.e. bands...
X <- model.matrix(~tf+trt-1,pd)
J <- apply(exp(fv)*X,2,cumsum)
se <- diag(J%*%vcov(b)%*%t(J))^.5
lines(stepfun(te,c(1,exp(-H+se))),do.points=FALSE,lty=2)
lines(stepfun(te,c(1,exp(-H-se))),do.points=FALSE,lty=2)
```

Adding similar code for the `auto` group produces figure 3.14. For much more on survival analysis see Collett (2015), Klein and Moeschberger (2003) and Therneau and Grambsch (2000).

3.3.4 Log-linear models for categorical data

The following table classifies a random sample of women and men according to their belief in the afterlife:

	Believer	Non-Believer
Female	435	147
Male	375	134

The data (reported in Agresti, 1996) come from the US General Social Survey (1991), and the 'non-believer' category includes 'undecideds'. Are there differences between males and females in the holding of this belief? We can address this question by using analysis of deviance to compare the fit of two competing models of these data: one in which belief is modelled as independent of gender, and a second in which there is some interaction between belief and gender. First consider the model of independence. If y_i is an observation of the counts in one of the cells of the table, then we could model the expected number of counts as

$$\mu_i \equiv \mathbb{E}(Y_i) = n\gamma_k\alpha_j \text{ if } y_i \text{ is data for gender } k, \text{ and faith } j,$$

where n is the total number of people surveyed, α_1 the proportion of believers, α_2 the proportion of non-believers and γ_1 and γ_2 the proportions of women and men,

respectively. Taking logs of this model yields

$$\eta_i \equiv \log(\mu_i) = \log(n) + \log(\gamma_k) + \log(\alpha_j).$$

So defining $\tilde{n} = \log(n)$, $\tilde{\gamma}_k = \log(\gamma_k)$ and $\tilde{\alpha}_j = \log(\alpha_j)$ the model can be written as

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{n} \\ \tilde{\gamma}_1 \\ \tilde{\gamma}_2 \\ \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{bmatrix}.$$

This is clearly a GLM structure, but is obviously not identifiable. Dropping $\tilde{\gamma}_1$ and $\tilde{\alpha}_1$ solves the identifiability problem yielding

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{n} \\ \tilde{\gamma}_2 \\ \tilde{\alpha}_2 \end{bmatrix}.$$

Note how gender and faith are both factor variables with two levels in this model.

If the counts in the contingency table occurred independently at random, then the obvious distribution to use would be Poisson. In fact even when the total number of subjects in the table, or even some other marginal totals, are fixed, then it can be shown that the correct likelihood can be written as a product of Poisson p.m.f.s, conditional on the various fixed quantities. Hence provided that the fitted model is forced to match the fixed total, and any fixed marginal totals, the Poisson is still the distribution to use. As was shown in section 3.1.6, forcing the model to match certain fixed totals in the data is simply a matter of insisting on certain terms being retained in the model.

The simple 'independence' model is easily estimated in R. First enter the data and check it:

```
> al <- data.frame(y=c(435,147,375,134),gender=
+   as.factor(c("F","F","M","M")),faith=as.factor(c(1,0,1,0)))
> al
    y gender faith
1 435      F     1
2 147      F     0
3 375      M     1
4 134      M     0
```

Since gender and faith are both factor variables, model specification is very easy. The following fits the model and checks that the model matrix is as expected:

```
> mod.0 <- glm(y ~ gender + faith, data=al, family=poisson)
> model.matrix(mod.0)
  (Intercept) genderM faith1
1           1       0      1
2           1       0      0
3           1       1      1
4           1       1      0
```

Now look at the fitted model object `mod.0`

```
> mod.0

Call:  glm(formula=y~gender+faith,family=poisson,data=al)

Coefficients:
(Intercept)      genderM       faith1
     5.0100      -0.1340       1.0587

Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
Null Deviance:      272.7
Residual Deviance: 0.162        AIC: 35.41

> fitted(mod.0)
      1       2       3       4
432.099 149.901 377.901 131.099
```

The fit appears to be quite close, and it would be somewhat surprising if a model with interactions between faith and gender did significantly better. Nevertheless such a model could be:

$$\eta_i \equiv \log(\mu_i) = \tilde{n} + \tilde{\gamma}_k + \tilde{\alpha}_j + \tilde{\zeta}_{kj} \text{ if } y_i \text{ is data for gender } k \text{ and faith } j,$$

where $\tilde{\zeta}_{kj}$ is an 'interaction parameter'. This model allows each combination of faith and gender to vary independently. As written, the model has rather a large number of un-identifiable terms.

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{n} \\ \tilde{\gamma}_1 \\ \tilde{\gamma}_2 \\ \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \tilde{\zeta}_{11} \\ \tilde{\zeta}_{12} \\ \tilde{\zeta}_{21} \\ \tilde{\zeta}_{22} \end{bmatrix}.$$

But this is easily reduced to something identifiable:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{n} \\ \tilde{\gamma}_2 \\ \tilde{\alpha}_2 \\ \tilde{\zeta}_{22} \end{bmatrix}.$$

The following fits the model, checks the model matrix and prints the fitted model object:

```
> mod.1 <- glm(y ~ gender*faith, data=al, family=poisson)
> model.matrix(mod.1)
```

```
  (Intercept) genderM faith1 genderM:faith1
1           1       0      1              0
2           1       0      0              0
3           1       1      1              1
4           1       1      0              0
> mod.1

Call:  glm(formula=y ~ gender*faith,family=poisson,data=al)

Coefficients:
   (Intercept)         genderM          faith1  genderM:faith1
       4.99043        -0.09259         1.08491        -0.05583

Degrees of Freedom: 3 Total (i.e. Null);  0 Residual
Null Deviance:      272.7
Residual Deviance: 9.659e-14    AIC: 37.25
```

To test whether there is evidence for an interaction between gender and faith the null hypothesis that `mod.0` is correct is tested against the more general alternative that `mod.1` is correct, using analysis of deviance.

```
> anova(mod.0,mod.1,test="Chisq")
Analysis of Deviance Table

Model 1: y ~ gender + faith
Model 2: y ~ gender * faith
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         1    0.16200
2         0  9.659e-14  1  0.16200   0.68733
```

A p-value of 0.69 suggests that there is no evidence to reject model 0 and the hypothesis of no association between gender and belief in the afterlife.

Notice that, in fact, the model with the interaction is the saturated model, which is why its deviance is numerically zero, and there was not really any need to fit it and compare it with the independence model explicitly — in this case we could just as well have examined the deviance of the independence model. However, the general approach taken for this simple 2-way contingency table can easily be generalized to multi-way tables and to arbitrary number of groups. In other words, the approach outlined here can be extended to produce a rather general approach for analyzing categorical data using log-linear GLMs.

Finally, note that the fitted values for `mod.0` had the odd property that although the fitted values and original data are different, the total number of men and women is conserved between data and fitted values, as is the total number of believers and non-believers. This results from the fact that the log link is canonical for the Poisson distribution, so by the results of section 3.1.6, $\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\hat{\boldsymbol{\mu}}$. The summations equated on the two sides of this last equation are the total number of subjects, the total number of males and the total number of believers: this explains the match between fitted values and data in respect of these totals.

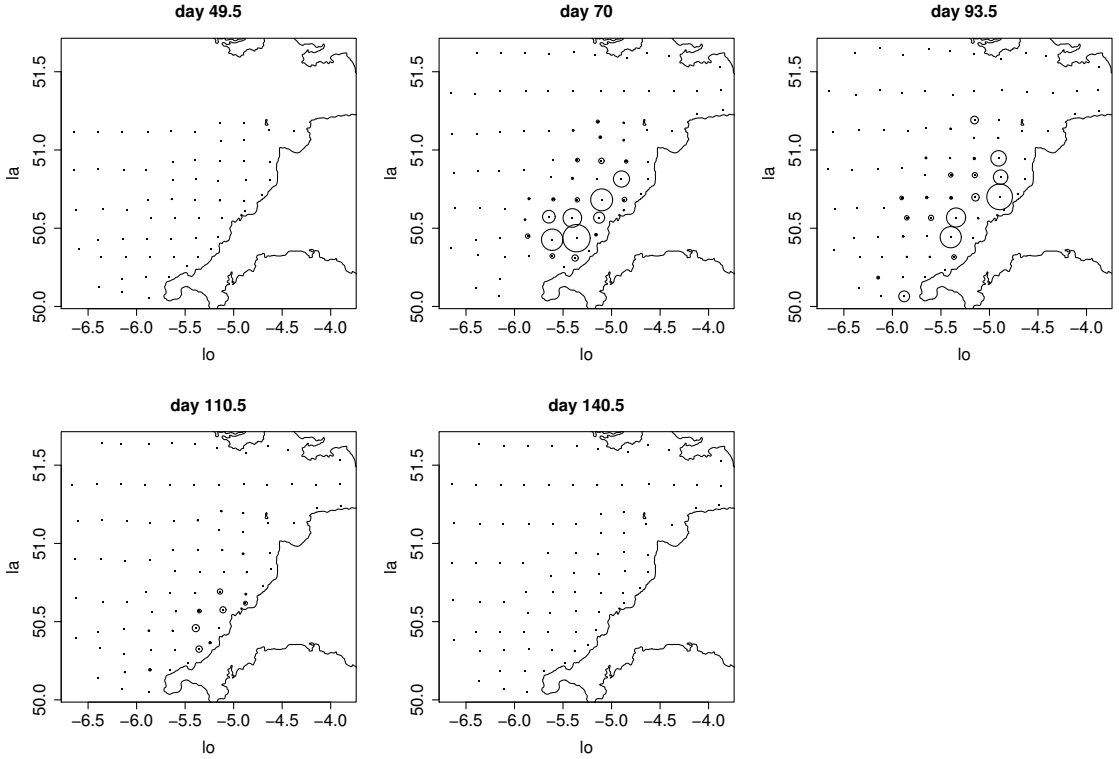


Figure 3.15 *Density per* m^2 *sea surface of stage I sole eggs in the Bristol channel. The days given are the Julian day of the survey midpoint (day 1 is January 1). The symbol sizes are proportional to egg density and a simple dot indicates a station where no eggs were found.*

3.3.5 Sole eggs in the Bristol channel

Fish stock assessment is difficult because adult fish are not easy to survey: they tend to actively avoid fishing gear, so that turning number caught into an assessment of the number in the sea is rather difficult. To get around this problem, fisheries biologists sometimes try and count fish eggs, and work back to the number of adult fish required to produce the estimated egg population. These 'egg production methods' are appealing because eggs are straightforward to sample. This section concerns a simple attempt to model data on sole eggs in the Bristol channel. The data (available in Dixon, 2003) are measurements of density of eggs per square metre of sea surface in each of 4 identifiable egg developmental stages, at each of a number of sampling stations in the Bristol channel on the west coast of England. The samples were taken during 5 cruises spaced out over the spawning season. Figure 3.15 shows the survey locations and egg densities for stage I eggs for each of the 5 surveys. Similar plots could be produced for stages II–IV. For further information on this stock, see Horwood (1993) and Horwood and Greer Walker (1990).

The biologists' chief interest is in estimating the *rate* at which eggs are spawned at any time and place within the survey arena, so this is the quantity that needs to be estimated from the data. To this end it helps that the durations of the egg stages

are known (they vary somewhat with temperature, but temperature is known for each sample). Basic demography suggests that a reasonable model for the density of eggs (per day per square metre of sea surface), at any age, a, and location-time with covariates $\mathbf{x}$, would be

$$d(a, \mathbf{x}) = S(\mathbf{x})e^{-\delta(\mathbf{x})a}.$$

That is, the density of eggs of age a is given by the product of the local spawning rate S and the local survival rate. δ is the per capita mortality rate, and, given this rate, we expect a proportion $\exp(-\delta a)$ of eggs to reach age a. Both S and δ are assumed to be functions of some covariates.

What we actually observe are not egg densities per unit age, per m^2 sea surface, but egg densities *in particular developmental stages* per m^2 sea surface: y_i, say. To relate the model to the data we need to integrate the model egg density over the age range of the developmental stage to which any particular datum relates. That is, if a_i^- and a_i^+ are the lower and upper age limits for the egg stage to which y_i relates, then the model should be

$$\mathbb{E}(y_i) \equiv \mu_i = \int_{a_i^-}^{a_i^+} d(z, \mathbf{x}_i)dz.$$

Evaluation of the integral would be straightforward, but does not enable the model to be expressed in the form of a GLM. However, if the integral is approximated so that the model becomes

$$\mu_i = \Delta_i d(\bar{a}_i, \mathbf{x}_i),$$

where $\Delta_i = a_i^+ - a_i^-$ and $\bar{a}_i = (a_i^+ + a_i^-)/2$, then progress can be made, since in that case

$$\log(\mu_i) = \log(\Delta_i) + \log\{S(\mathbf{x}_i)\} - \delta(\mathbf{x}_i)\bar{a}_i. \tag{3.15}$$

The right hand side of this model can be expressed as the linear predictor of a GLM, with terms representing $\log(S)$ and δ as functions of covariates and with $\log(\Delta)$ treated as an 'offset' term — essentially a column of the model matrix with associated parameter fixed at 1.

For the sole eggs, a reasonable starting model might represent $\log(S)$ as a cubic function of longitude, `lo`, latitude, `la`, and time, `t`. Mortality might be modelled by a simpler function — say a quadratic in `t`. It remains only to decide on a distributional assumption. The eggs are sampled by hauling a net vertically through the water and counting the number of eggs caught in it. This might suggest a Poisson model, but most such data display overdispersion relative to Poisson, and additionally, the data are not available as raw counts but rather as densities per m^2 sea surface. These considerations suggest using quasi-likelihood, with the variance proportional to the mean.

The following R code takes the `sole` data frame, calculates the mean ages and offset terms required and fits the suggested model. Since polynomial models can lead to numerical stability problems if not handled carefully, the covariates are all translated and scaled before fitting.

```
> sole$off <- log(sole$a.1-sole$a.0)# model offset term
> sole$a<-(sole$a.1+sole$a.0)/2     # mean stage age
```

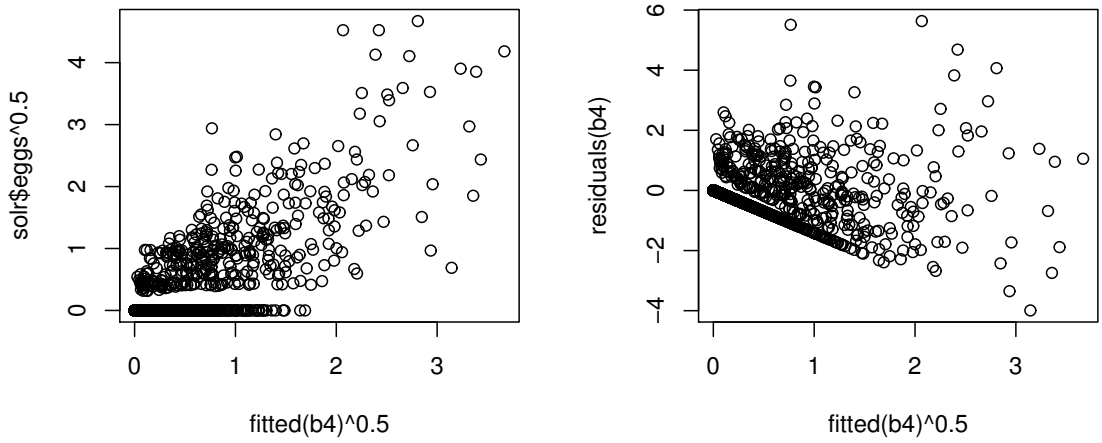


Figure 3.16 *Residual plots for the final sole egg model.*

```
> solr<-sole                              # make copy for rescaling
> solr$t<-solr$t-mean(sole$t)
> solr$t<-solr$t/var(sole$t)^0.5
> solr$la<-solr$la-mean(sole$la)
> solr$lo<-solr$lo-mean(sole$lo)
> b <- glm(eggs ~ offset(off)+lo+la+t+I(lo*la)+I(lo^2)+I(la^2)
+          +I(t^2)+I(lo*t)+I(la*t)+I(lo^3)+I(la^3)+I(t^3)+
+          I(lo*la*t)+I(lo^2*la)+I(lo*la^2)+I(lo^2*t)+
+          I(la^2*t)+I(la*t^2)+I(lo*t^2)+ a +I(a*t)+I(t^2*a),
+          family=quasi(link=log,variance="mu"),data=solr)
> summary(b)

Call:
glm(formula = eggs~offset(off)+lo+la+t+I(lo*la)+
    I(lo^2)+I(la^2)+I(t^2)+I(lo*t)+I(la*t)+I(lo^3)+
    I(la^3)+I(t^3)+I(lo*la*t)+I(lo^2*la)+I(lo*la^2)
    +I(lo^2*t)+I(la^2*t)+I(la*t^2)+I(lo*t^2)+
    a+I(a*t)+I(t^2*a),family = quasi(link=log,
    variance = "mu"), data = solr)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-4.10474  -0.35127  -0.10418  -0.01289   5.66956

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.03836    0.14560  -0.263 0.792202
lo              5.22548    0.39436  13.251  < 2e-16 ***
```

```
la              -5.94345    0.50135 -11.855  < 2e-16 ***
t               -2.43222    0.25761  -9.442  < 2e-16 ***
I(lo * la)       3.38576    0.61797   5.479 4.99e-08 ***
I(lo^2)         -3.98406    0.36744 -10.843  < 2e-16 ***
I(la^2)         -4.21517    0.56228  -7.497 1.10e-13 ***
I(t^2)          -1.77607    0.26279  -6.758 1.97e-11 ***
I(lo * t)        0.20029    0.35117   0.570 0.568518
I(la * t)        1.82637    0.47332   3.859 0.000119 ***
I(lo^3)         -3.46452    0.49554  -6.991 4.03e-12 ***
I(la^3)          8.53152    1.28587   6.635 4.48e-11 ***
I(t^3)           0.70085    0.12397   5.653 1.87e-08 ***
I(lo * la * t)  -1.10150    0.90738  -1.214 0.224959
I(lo^2 * la)     5.20779    0.88873   5.860 5.65e-09 ***
I(lo * la^2)   -12.87497    1.24298 -10.358  < 2e-16 ***
I(lo^2 * t)      0.79928    0.54238   1.474 0.140774
I(la^2 * t)      5.42159    1.08911   4.978 7.14e-07 ***
I(la * t^2)     -1.14220    0.46440  -2.459 0.014021 *
I(lo * t^2)      0.65862    0.36929   1.783 0.074705 .
a               -0.12285    0.02184  -5.624 2.21e-08 ***
I(a * t)         0.09456    0.04615   2.049 0.040635 *
I(t^2 * a)      -0.18310    0.05998  -3.053 0.002306 **
---
Sig. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasi family taken to be 1.051635)

    Null deviance: 3108.86  on 1574  degrees of freedom
Residual deviance:  913.75  on 1552  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

The summary information suggests dropping the `lo*t` term (it seems unreasonable to drop the constant altogether). Rather than re-type the whole `glm` command again, it is easier to use:

```
b1 <- update(b, ~ . - I(lo*t))
```

which re-fits the model, dropping the term specified. Repeating the process suggests dropping `lo*la*t`, `lo*t^2` and finally `lo^2*t`, after which all the remaining terms are significant at the 5% level. If `b4` is the final reduced model, then it can be tested against the full model:

```
> anova(b,b4,test="F")
Analysis of Deviance Table
[edited]
  Resid. Df Resid. Dev   Df Deviance      F Pr(>F)
1      1552     913.75
2      1556     919.28   -4    -5.54 1.3161 0.2618
```

which gives no reason not to accept the simplified model.

The default residual plots are unhelpful for this model, because of the large number of zeroes in the data, corresponding to areas where there really are no eggs. This

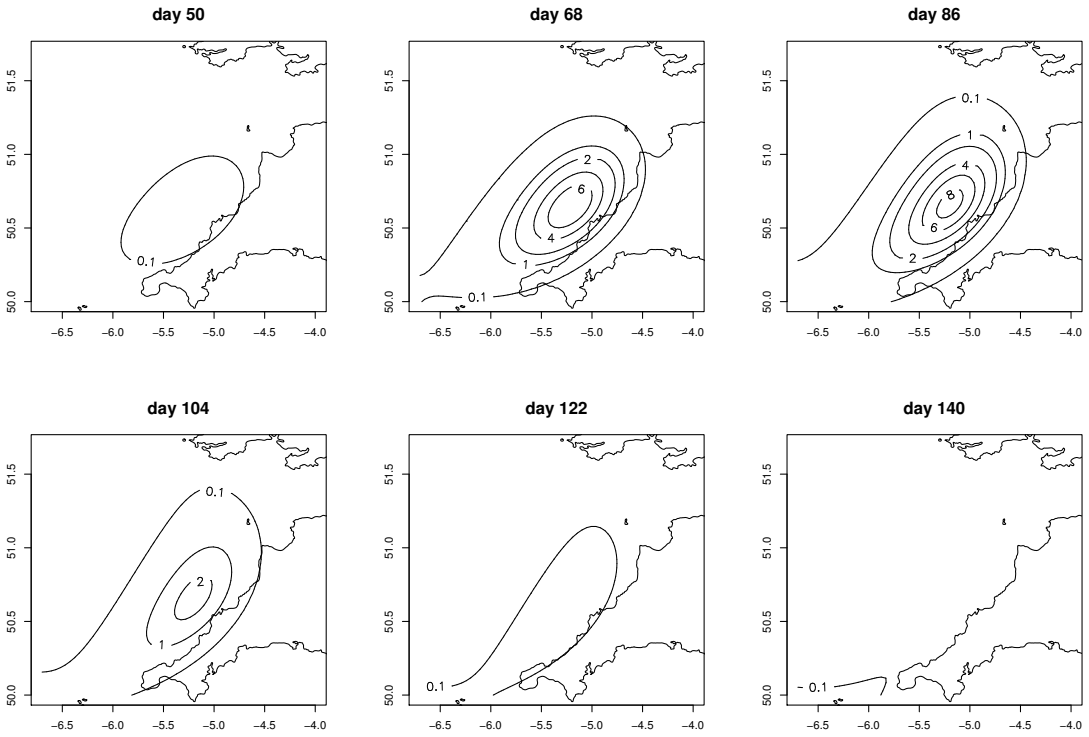


Figure 3.17 *Model predicted sole spawning rates in the Bristol channel at various times in the spawning season.*

tends to lead to some very small values for the linear predictor, corresponding to zero predictions, which in turn lead to rather distorted plots. The following residual plots are perhaps more useful.

```
par(mfrow=c(1,2)) # split graph window into 2 panels
plot(fitted(b4)^0.5,solr$eggs^0.5) # fitted vs. data plot
plot(fitted(b4)^0.5,residuals(b4)) # resids vs. sqrt(fitted)
```

The plots are shown in figure 3.16. The most noticeable features of both plots relate to the large number of zeros in the data, with the lower boundary line in the right hand plot corresponding entirely to zeros, for which the raw residual is simply the negative of the fitted value. The plots are clearly far from perfect, but it is unlikely that great improvements can be made to them with models of this general type.

The fitted model can be used for prediction of the spawning rate over the Bristol channel, by setting up a data frame containing the times and locations at which predictions are required, the age at which prediction is required — always zero — and the offset required — zero if spawning rate per square metre is the desired output. The time and location co-ordinates must be scaled in the same way as was done for fitting, of course. Figure 3.17 shows model predicted spawning rates produced in this way from model `b4`. It has been possible to get surprisingly far with the analysis of these data using a simple GLM approach, but the fitting did become somewhat unwieldy when it came to specifying that spawning rate should be a smooth function of

location and time. For a less convenient spatial spawning distribution, it is doubtful that a satisfactory model could have been produced in this manner. This is part of the motivation for seeking to extend the way in which GLMs are specified, to allow a more compact and flexible way of specifying smooth functional relationships within the models, i.e., part of the motivation for developing GAMs.

3.4 Generalized linear mixed models

Recall that a GLMM models an exponential family random variable, Y_i, with expected value μ_i using

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}, \quad \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\psi}_\theta).$$

That is, a GLMM is a GLM in which the linear predictor depends on some Gaussian random effects, $\mathbf{b}$, multiplied by a random effects model matrix $\mathbf{Z}$. The main difficulty in moving from linear mixed models to GLMMs is that it is no longer possible to evaluate the log likelihood exactly, since it is not generally analytically tractable to integrate $\mathbf{b}$ out of the joint density of $\mathbf{y}$ and $\mathbf{b}$ to obtain the likelihood.

One effective way to proceed is to follow the approach of section 2.4, Taylor expanding around, $\hat{\mathbf{b}}$, the mode of $f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})$, to get

$$f(\mathbf{y}|\boldsymbol{\beta}) \simeq \int \exp\left\{\log f(\mathbf{y}, \hat{\mathbf{b}}|\boldsymbol{\beta}) + \frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})^\mathsf{T}\frac{\partial^2 \log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})}{\partial \mathbf{b}\partial \mathbf{b}^\mathsf{T}}(\mathbf{b}-\hat{\mathbf{b}})\right\} d\mathbf{b}.$$

This differs from section 2.4 in being approximate, since we have neglected the higher order terms in the Taylor expansion, rather than those terms being identically zero. However, having accepted this approximation the rest of the evaluation of the integral follows section 2.4 exactly. In the current case the required Hessian is $-\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{Z}/\phi - \boldsymbol{\psi}^{-1}$, where $\mathbf{W}$ is the IRLS weight vector based on the $\boldsymbol{\mu}$ implied by $\hat{\mathbf{b}}$ and $\boldsymbol{\beta}$. So

$$f(\mathbf{y}|\boldsymbol{\beta}) \simeq f(\mathbf{y}, \hat{\mathbf{b}}|\boldsymbol{\beta})\frac{(2\pi)^{p/2}}{|\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{Z}/\phi + \boldsymbol{\psi}_\theta^{-1}|^{1/2}}.$$

This sort of integral approximation is known as 'Laplace approximation'. Substituting the explicit form for the random effects density and taking logs gives

$$l(\boldsymbol{\theta}, \boldsymbol{\beta}) \simeq \log f(\mathbf{y}|\hat{\mathbf{b}}, \boldsymbol{\beta}) - \hat{\mathbf{b}}^\mathsf{T}\boldsymbol{\psi}_\theta^{-1}\hat{\mathbf{b}}/2 - \log|\boldsymbol{\psi}_\theta|/2 - \log|\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{Z}/\phi + \boldsymbol{\psi}_\theta^{-1}|/2. \tag{3.16}$$

Notice that as well as direct dependencies, the right hand side depends on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ via $\hat{\mathbf{b}}$ and $\mathbf{W}$. The dependence of $\mathbf{W}$ on $\boldsymbol{\beta}$ means that the MAP estimate and MLE of $\boldsymbol{\beta}$ no longer correspond exactly, in contrast to the LMM case. Nonetheless it is convenient to use the MAP estimates, since they can be computed easily along with the corresponding $\hat{\mathbf{b}}$, using a penalized version of the IRLS method used for GLMs. If we do this then we can also define a Laplace approximate profile likelihood $l_p(\boldsymbol{\theta}) = l(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the MAP estimate given $\boldsymbol{\theta}$.

Laplace approximate REML also follows in a similar way, from the linear model

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2022

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

MATH60044/70044/97082

Statistical Modelling 2 (Solutions)

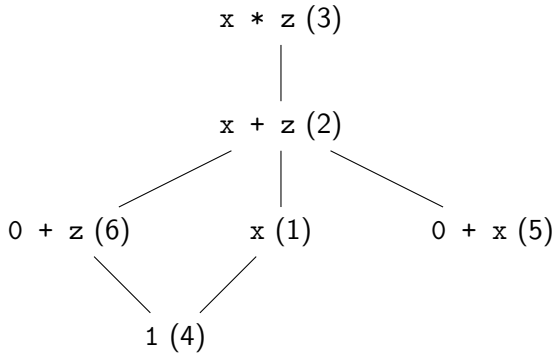
Setter's signature	Checker's signature	Editor's signature
.....	.....	.....

1. (a) (i) 2 — seen ⇓

(ii) 4 — 3, A

(iii) 1 — unseen ⇓

(b) The diagram below (not required) shows all nesting relationships. Allow any non-nested pair, e.g. `fit5` and `fit6`. — 1, A



(c) (i) `fit3`, because it is the model that has the most free parameters, and all other models are nested submodels of it. — sim. seen ⇓, 2, A

(ii) Since `fit6` has 48 degrees of freedom with two free parameters, there must be $n = 50$ observations. The t distribution for the coefficient of `x` in model 1 will therefore also have 48 degrees of freedom. — sim. seen ⇓, 2, B

(iii) The numerator degrees of freedom is given by the difference in the numbers of parameters estimated by the two models, which is $4-2 = 2$. The denominator degrees of freedom is the number of residual degrees of freedom of the larger model, which is $50 - 4 = 46$. Hence we have an $F(2, 46)$ distribution. — sim. seen ⇓, 2, C

(d) Residuals necessarily sum to zero in models with an intercept, as the residuals are orthogonal to each column of the design matrix. So need a model with no intercept. `fit5` is the only possibility. `fit6` has no intercept but the column vector of ones is a linear combination of its two columns, and the vector of residuals is orthogonal to both columns. — unseen ⇓, 2, A, 2, D

(e) For the design matrix to have full rank, require that the column corresponding to `x` not be a scalar multiple of the intercept column. Hence it is enough that not all of the entries of x are equal. — sim. seen ⇓, 1, C

(f) Conside the standard errors of the output in `fit6`. Their ratio is $0.4332/0.2166 = 2$. But for the categorical covariate `z`, this is just $\frac{\sqrt{n_1}}{\sqrt{n_0}}$. Hence $4n_0 = n_1$. Since $n_0 + n_1 = 50$, this gives $n_0 = 10$ and $n_1 = 40$. — 3, D, unseen ⇓

(g) 1.8713 is the mean of y for individuals with $z = 0$, and 2.2413 is the mean of y for individuals with $z = 1$. — 2, B, sim. seen ⇓

2. (a) Use iterated weighted least squares - a form of Newton's method - to maximize the log likelihood as a function of $\boldsymbol{\beta}$.

seen ⇓

5, A

Given an estimate $\boldsymbol{\beta}^m$, define $U^m = \nabla l(\boldsymbol{\beta}^m)$ and $\mathcal{J}^m = \nabla^2 l(\boldsymbol{\beta}^m)$. Then the standard Newton iteration is

$$\boldsymbol{\beta}^{m+1} = \boldsymbol{\beta}^m + (\mathcal{J}^m)^{-1} \boldsymbol{U}^m.$$

At each iteration, this chooses an updated value $\boldsymbol{\beta}^{m+1}$ to maximize the best quadratic approximation to the log likelihood at $\boldsymbol{\beta}^m$. Since the log likelihood is convex, the sequence $(\boldsymbol{\beta}^m)_{m\geq 0}$ converges to the maximizer of l.

The method needs to be initialized with an estimate $\boldsymbol{\beta}_0$, obtained e.g. from estimating $\pi_i = \frac{y_i}{n_i}$ and fitting the linear model $\log\left(\frac{\pi_i}{1-\pi_i}\right) \approx \beta_{01} + \beta_{02} x$.

Assess convergence using the reduction in deviance between successive iterations. Stop when the reduction is small, or after some maximum number of iterations (R defaults to 25).

(b) The logistic regression model considered here models p_i, the proportion of defective machines in observation i, as

2, B

meth seen ⇓

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i.$$

An additional 1000 hours of running time would increase the log odds above by $1000\beta_1 = 0.999 \approx 1$, so the odds would increase by a multiplicative factor of $\exp(1) \approx 2.7$.

meth seen ⇓

(c) Assuming that the number of machines in each observation is sufficiently large, the sampling distribution of the maximum likelihood estimates is roughly normal, so that an approximate 95% confidence interval for the additive effect on the log odds of each additional hour is given by

2, C

$$\beta_1 \pm 1.96\mathsf{SE}(\beta_1).$$

Then the confidence interval for the additive effect is $1 \pm 1.96 \times 0.1141$. On the odds scale, the confidence interval is

$$\left(\exp\left(1 - 1.96 \times 0.1141\right), \exp\left(1 + 1.96 \times 0.1141\right)\right) = (2.17, 3.40).$$

(Final numerical answer not needed.)

(d) Assuming that numbers of machines in each observation is sufficiently large, the scaled deviance would be expected to have a roughly chi-square distribution with 9 degrees of freedom. The observed value of the deviance, 10.331 is consistent with this. Equivalently, the estimated dispersion parameter for the binomial GLM, $\hat{\phi} = 10.331/9 \approx 1.15$ is close to its theoretical value of 1. This gives no grounds for concern about the model.

sim. seen ⇓

2, D

(e) `dt` contains the fitted values on the response scale for each observation. In context, this is the predicted number of defective machines for each observation.

unseen ⇓

2, D

Since this model uses the canonical link, the system of equations satisfied by the maximum likelihood estimates reduces to

$$X^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = 0.$$

Now, since the model includes an intercept, this says in particular that $\sum_i \hat{\mu}_i = \sum_i y_i$. In context, 106 is the total number of failed machines in the data.

(f) Note that $S = \sigma Z + \alpha_0 + \alpha_1 x$, where $Z \sim N(0,1)$ with CDF Φ. Then

sim. seen ⇓

3, C

$$\Pr(Y = 1|X = x) = \Pr(S > t|X = x) = \Pr(\sigma Z + \alpha_0 + \alpha_1 x > t) = \Phi\left(\frac{\alpha_0 + \alpha_1 x - t}{\sigma}\right),$$

using $\Phi(-z) = 1 - \Phi(z)$. This is the probit link function.

In terms of identifiable parameters, $\beta_0 = \frac{\alpha_0 - t}{\sigma}$ and $\beta_1 = \frac{\alpha_1}{\sigma}$.

(g) The link functions have similar shape, with a central linear section in the region where $\mu \sim \frac{1}{2}$. Let x_0 be such that $\mu(x_0) = g(\beta_0 + \beta_1 x_0) = g(0)$ is close to $\frac{1}{2}$, for both link functions. Then for x sufficiently close to x_0,

unseen ⇓

4, D

$$\mu(\eta) = \mu(\eta_0) + g'(\eta_0)(\eta - \eta_0).$$

holds for both link functions. Hence the regression coefficients for the logistic model are roughly given by multiplying those of the probit model by a factor $\frac{\Phi'(0)}{g'(0)}$.
To determine this factor, we differentiate,

$$\frac{d}{dx}\frac{1}{1+\exp(-x)}\bigg|_{x=0} = \frac{\exp(-x)}{(1+\exp(-x))^2}\bigg|_{x=0} = \frac{1}{4}$$

$$\frac{d}{dx}\Phi(x)\bigg|_{x=0} = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\bigg|_{x=0} = \frac{1}{\sqrt{2\pi}}$$

Hence the coefficients for the probit model are roughly given by multiplying those of the logistic model by $\frac{\sqrt{2\pi}}{4} \approx 0.63$.

3. (a)
* Substantial reduction in variance as the dilution level decreases. This violates an assumption of the normal linear model, which requires a constant variance.

sim. seen ⇓

3, A

* Relationship appears somewhat nonlinear, suggesting again that a normal linear model is inappropriate.

* Linear fit would imply negative values for x slightly greater than 4, which is physically unreasonable.

seen ⇓

(b) By default, R has used the log link function. The model assumes

1, A

2, D

$$\mu_x = \exp(\beta_0 + \beta_1 x).$$

This means that a unit increase in x, which corresponds to a multiplicative change in the concentration, also corresponds to a multiplicative change in the mean number of lesions. If we suppose that the number of lesions should be proportional to the number of viral particles, this is a reasonable relationship.

sim. seen ⇓

(c) A unit increase in x, i.e. a halving of the concentration, changes the mean number of lesions by a multiplicative factor of $\exp(\beta_1) = \exp(-0.68094) \approx 0.5$. This is consistent with the assumption that the number of lesions is roughly proportional to the amount of virus.

3, B

sim. seen ⇓

(d) To form an approximate confidence interval for the mean response at x_0, assume that the maximum likeihood estimator $\hat{\beta} \sim N(\beta, \mathcal{I}(\beta)^{-1})$. As a linear function of $\hat{\beta}$, we have that

4, B

$$\hat{\eta}(x_0) = (1, x_0)\hat{\beta} \sim N(\eta(x_0), (1, x)\mathcal{I}(\beta)^{-1}(1, x_0)^T).$$

To compute the distribution of $\hat{\eta}(x_0)$, we therefore also need the an estimate of the variance-covariance matrix of $\hat{\beta}$, which is not given in the available output.

An approximate 95% confidence interval for $\eta(x_0)$ is

$$\left(\hat{\eta}(x_0) - 1.96\mathsf{SE}\left[\hat{\eta}(x_0)\right], \hat{\eta}(x_0) + 1.96\mathsf{SE}\left[\hat{\eta}(x_0)\right]\right)$$

Then $\hat{y}(x_0) = \exp\left(\hat{\eta}(x_0)\right)$, and the corresponding confidence interval for the prediction is

$$\left(\exp\left(\hat{\eta}(x_0) - 1.96\mathsf{SE}\left[\hat{\eta}(x_0)\right]\right), \exp\left(\hat{\eta}(x_0) + 1.96\mathsf{SE}\left[\hat{\eta}(x_0)\right]\right)\right)$$

(e) The line of output immediately following the summary computes the Pearson chi-square statistic,

sim. seen ⇓

2, B

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

When normalized by the number of degrees of freedom, this gives an estimate of the dispersion parameter. (Could also use the deviance for this.) For the Poisson, we would expect $\phi \sim 1$, but in fact is it is substantially larger,

$$\hat{\phi} = \frac{292}{48} \approx 6.$$

This indicates overdispersion - for a given value μ_x of the mean, there is more variability than would be expected.

sim. seen ⇓

(f) Use a quasi-poisson model. Adopt the estimate $\hat{\phi}$ from the data. This results in no changes to the estimated model parameters, but increases the standard error for each parameter by a factor of $\sqrt{\hat{\phi}}$. This then propagates through to a wider confidence interval for $\hat{y}(x_0)$. In terms of the original standard errors above, the new confidence interval is

3, C

$$\left(\exp\left(\hat{\eta}(x_0) - 1.96\sqrt{\hat{\phi}}\mathsf{SE}\left[\hat{\eta}(x_0)\right]\right), \exp\left(\hat{\eta}(x_0) + 1.96\sqrt{\hat{\phi}}\mathsf{SE}\left[\hat{\eta}(x_0)\right]\right)\right).$$

unseen ⇓

(g) A roughly linear pattern is evident in the plot of $\log V(\mu)$ against $\log \mu$. The slope of the line is greater than 1, indicating a relationship $V(\mu) = \mu^\alpha$ for $\alpha > 1$. Could estimate a reasonable value of α and use this in a quasi-likelihood model.

1, B

1, D

4. (a) (i)
- Box plot of residuals show substantial between-subject variability, relative to the residual standard error. Some subjects have residuals all of the same sign.

sim. seen ⇓

3, A

- Need to account for this variability in order to obtain reasonable estimates of the sampling uncertainty.

- This means that standard errors, p-values etc. do not faithfully represent the uncertainty in parameter estimates.

(ii)

1, A

$$2.500 + 6.458 = 8.958.$$

sim. seen ⇓

(iii)

4, A

- A model that included a fixed effect for each subject would be treating individuals as entirely separate entities, rather than members of a population. Model would have no validity for subjects outside the study: it would be overfit to the specific individuals in the study.

sim. seen ⇓

- Instead the linear mixed model assumes zero mean, normally distributed subject-to-subject variation, $b_j \sim N(0, \sigma_b^2)$, where σ_b^2 is estimated from the data. The score for product i given by subject j takes the form

$$y_{ij} = \beta_0 + \beta_i + b_j + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ is an error term, independent of other random errors and of b_j.

- The model `fit_re` is estimated using restricted maximum likelihood, which produces unbiased estimators of the variance σ_b^2.

(iv)

sim. seen ⇓

2, A

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} = \frac{5.696}{3.493 + 5.696} \approx 0.62.$$

(v) Need to randomize the order in which subjects see the different products, to balance out exerimental fatigue or other confounding effects.

sim. seen ⇓

2, A

(b) (i) The coefficient for `treatment` is a measure of the difference between the groups at baseline. Its t-value is -0.792, which is small in absolute value. This suggests there is no material difference between the groups at baseline, under the assumption that the sampling distribution of the parameters is normal.

unseen ⇓

1, A

unseen ⇓

3, B

(ii) In the group with `treatment == 1`, the gradient of the line is negative, so that the score is decreasing with time, whereas in the group with `treatment == 0`, the gradient is positive, so that the score is increasing with time. Assuming that the sampling distribution of the estimates is normal, the $t-$statistic is far into the tails of the null distribution, so this difference is statistically significant. This suggests that there is evidence that the treatment is effective.
Mark for sketch requires the gradients and intercepts to be roughly correct.

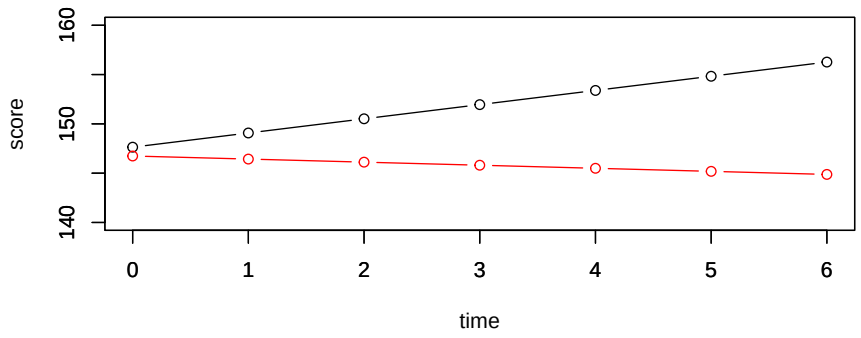


Figure 1: Sketch showing mean score for the two treatment groups. Standard care in black, new treatment in red.

unseen ⇓

1, D

(iii) Model assumes constant correlation between any two observations on the same subject. A more realistic model would allow observations closer in time to be more closely correlated than those farther apart, but this would require additional parameters to be estimated.

unseen ⇓

2, A

1, D

(iv)
- This new model allows for a random slope as well as a random intercept.
- It has been fit using maximum likelihood rather than restricted maximum likelihood.
- In order for the two models to be compared, the first model should be re-fit using maximum likelihood.
- Comparison can be made by considering the log likelihood ratio. As the null distribution of this test statistic may not take a standard asymptotic form (as the null hypothesis corresponds to a boundary of the parameter space), its sampling distribution should be computed by parametric bootstrap.

5. (a) The null hypothesis corresponds to independence of `gender` and `faith`. Expressing the form of an observation in terms of the model parameters, 3, M

$$\mathrm{E}(y_i j) = \beta_0 + \beta_M + \beta_1 + \beta_{1M},$$

and the null hypothesis corresponds to a test of the null hypothesis $\beta_{1M} = 0$.

The conclusion of the test is that there is no evidence for an association between faith and gender.

2, M

(b) The design matrix has more columns than rows. Hence, it clearly has linearly dependent columns, and parameters corresponding to such columns are not identifiable.

3, M

(c) Note first that $\sum_{j=1}^m Y_j \sim \textsc{Poisson}(\sum_{j=1}^m \lambda_j)$.

$$\begin{aligned}
\Pr(Y_1 = y_1, \ldots, Y_m = y_m | \sum_{j=1}^m Y_j = n) &= \frac{\Pr(Y_1 = y_1, \ldots, Y_m = y_m, \sum_{j=1}^m Y_j = n)}{\Pr(\sum_{j=1}^m Y_j = n)} \\
&= \frac{\prod_{j=1}^m \exp(-\lambda_j) \frac{\lambda_j^{y_j}}{y_j!}}{\exp(-\sum_{j=1}^m \lambda_j) \frac{\left(\sum_{j=1}^m \lambda_j\right)^n}{n!}} \qquad \text{if } \sum_{j=1}^m = n \\
&= \frac{n!}{\prod_{j=1}^m y_j!} \prod_{j=1}^m \left(\frac{\lambda_j}{\sum_{l=1}^m \lambda_m} \right)^{y_j},
\end{aligned}$$

This is indeed the probability mass function of a multinomial.

2, M

(d)

$$\eta_3 = 5.0100 - 0.1340 + 1.0587$$

Then

$$377.9266 = \exp(\eta_3).$$

(e) When using a GLM with the canonical link, the maximum likelihood equations for the parameters take on a simple form, when written in terms of the fitted values $\hat{mu}$, specifically

3, M

$$X^t(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = 0.$$

Taking each the column for males, this says that
$\sum_{i=1}^n x_i y_i = \sum_{i=1}^n x_i \hat{\mu}_i$,
where $x_i = 1$ for males and $x_i = 0$ otherwise. This then says that the sum of the fitted values for males will be the total of y_i over all male subjects.

2, M

(f)
* Accounts for differences in variance for different values of the mean. A normal linear model assumes constant variability for different mean levels.
* Respects the fact that count data cannot be negative.

1, M

(g) The quasi-poisson does not correspond to an underlying exponential family of probability distributions, and so there is no corresponding likelihood function. Hence no AIC, which is a function of the likelihood.

4, M

(h) The probability mass function for the truncated Poisson is

$$\Pr(Y = y|Y > 0) = \frac{\exp(-\lambda)\lambda^y}{y!(1 - \exp(-\lambda))}, \qquad y \geq 1.$$

Writing this in exponential family form

$$\Pr(Y = y|Y > 0) = \exp\left(-\lambda + y\log\lambda - \log(1 - \exp(-\lambda)) - \log y!\right).$$

This is an exponential family with canonical parameter $\theta = \log\lambda$.

Relate the mean $\mu(\lambda)$ to the mean of the corresponding Poisson random variable:

$$\lambda = \mathrm{E}(Y) = \mathrm{E}(Y|Y = 0)\Pr(Y = 0) + E(Y|Y > 0)\Pr(Y > 0) = 0 + \mu(\lambda)(1 - \exp(-\lambda)).$$

Hence $\mu(\lambda) = \frac{\lambda}{(1-\exp(-\lambda))}$.

Similarly

$$\mathrm{Var}(Y)+\mathrm{E}(Y)^2 = \lambda+\lambda^2 = \mathrm{E}(Y^2) = \mathrm{E}(Y^2|Y=0)\Pr(Y=0)+E(Y^2|Y>0)\Pr(Y>0),$$

giving

$$\lambda + \lambda^2 = \left(V(\lambda) + \mu(\lambda)^2\right)(1-\exp(-\lambda)).$$

Simplifying,

$$\begin{aligned} V(\lambda) &= \frac{\lambda+\lambda^2}{(1-\exp(-\lambda))} - \frac{\lambda^2}{(1-\exp(-\lambda))^2} \\ &= \frac{\lambda}{(1-\exp(-\lambda))^2}\left((1+\lambda)(1-\exp(-\lambda)) - \lambda\right) \\ &= \frac{\lambda}{(1-\exp(-\lambda))^2}\left(1-(1+\lambda)\exp(-\lambda)\right). \end{aligned}$$

Review of mark distribution:
Total A marks: 32 of 32 marks
Total B marks: 19 of 20 marks
Total C marks: 11 of 12 marks
Total D marks: 18 of 16 marks
Total marks: 100 of 80 marks
Total Mastery marks: 20 of 20 marks

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 separate pdf file with your email.

ExamModuleCode	QuestionNumber	Comments for Students
Statistical Modelling 2_MATH60044 MATH97082 MATH70044	1	Many good attempts overall, showing good conceptual understanding of the linear model.
Statistical Modelling 2_MATH60044 MATH97082 MATH70044	2	Responses to (a) were often too general, and did not address the question, which asked for "how the method works". Many gave an explicit algorithm, but very few stated clearly that iterated weighted least squares works by iteratively finding the maximum of second-order Taylor approximations to the log likelihood function.
Statistical Modelling 2_MATH60044 MATH97082 MATH70044	3	enerally a well-done question. Relatively few made good use of the plot to answer (g).

Statistical Modelling 2_MATH60044 MATH97082 MATH70044	4	Part (a) was generally well done, although there were very few good answers to part (v), which asked specifically for how the experiment should be performed. Many suggested an approach for unbiased estimation of the variance parameters, which is not what hte question asked. Part (b) was found difficult by many. Answers to (i) and (ii) were often quite confused, although these should really be core knowledge. No-one gave a convincing answer to (iii): many calculated an estimate of the correlation in the model, but the question asked for a comment on the correlation assumed: the point is that the model assumes equal correlation between all measurements on the same individual (more realistically, correlation between observations might be expected to decrease according to their separation in time)
Statistical Modelling 2_MATH60044 MATH97082 MATH70044	5	(a) - (e) were well done on the whole. (h) was found to be very difficult - in hindsight perhaps this was too long for the final part.