IMPERIAL

MATH70044

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May 2024

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

Applied Statistical Inference

Date: Friday, May 17, 2024

Time: 14:00 – 16:30 (BST)

Time Allowed: 2.5 hours

This paper has 5 Questions.

Please Answer Each Question in a Separate Answer Booklet

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

Supplementary books may only be used after the relevant main book(s) are full.

Any required additional material(s) will be provided.

Allow margins for marking.

Credit will be given for all questions attempted.

Each question carries equal weight.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO

Throughout this paper, no simplification of numerical answers is required.

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 using the method of maximum likelihood. X can be taken to be full rank.

(a) Write down the likelihood function and show that maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ satisfies the linear system

$$X^T X\widehat{\boldsymbol{\beta}} = X^T\boldsymbol{Y}.$$

(4 marks)

(b) In the simplest case of an intercept-only model with parameter β_0, for which $y_i = \beta_0 + \epsilon_i$, show that the form of the maximum likelihood estimator is $\widehat{\beta}_0 = \bar{Y}$, the sample mean, and state its standard error.

(2 marks)

(c) The code and output at the end of the question come from a randomized controlled trial with $2n$ subjects, of whom n receive a placebo and n receive a treatment. The most general model considered is given below:

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + \beta_3 x_i z_i + \epsilon_i, \qquad i = 1, \ldots, 2n.$$

(i) Using the output of `fit2`, comment on whether `fit1` is appropriate.

(2 marks)

(ii) Determine n, the number of individuals in each group.

(2 marks)

(iii) Supply the values in the ANOVA table labelled `##1##`, `##2##` and `##3##`, leaving your answer in terms of other values given in the output.

(3 marks)

(iv) Find the sample variance of $y_1, \ldots, y_n$, in terms of other values given in the output.

(4 marks)

(v) State a property of the experimental design that ensures that fitting the model `y ~ z` would still give an unbiased estimate of β_1, and explain the benefit that comes from adding the covariate `x` to this model.

(3 marks)

Question 1 continues on the following page

```
fit1 <- lm(y ~ z + x)

summary(fit1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1286     0.1772   0.726   0.4716
z             0.6702     0.2507   2.674   0.0103 *
x             2.0512     0.1686  12.165 3.98e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8857 on 47 degrees of freedom
Multiple R-squared:  0.7703,  Adjusted R-squared:  0.7605


fit2 <- lm(y ~ z * x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1256     0.1783   0.704   0.4847
z             0.6702     0.2521   2.659   0.0108 *
x             1.9351     0.2398   8.069 2.34e-10 ***
z:x           0.2322     0.3392   0.685   0.4971

anova(fit1, fit2)

  Res.Df      RSS    Df    Sum of Sq     F  Pr(>F)
1     47    ##1##
2     46    ##### ##2##      0.37177 ##3##  0.4971
```

(Total: 20 marks)

2. Consider a Poisson generalized linear model with the canonical link, i.e.

$$Y_i \sim \text{Poisson}(\mu_i), \qquad \mu_i = \exp(\eta_i), \qquad \eta_i = X_i\boldsymbol{\beta}, \qquad i = 1, \ldots, n.$$

(a) For the model above, with log likelihood l, derive,

(i) The score function $U(\boldsymbol{\beta}) = \nabla l$.

(ii) The Fisher Information $I(\boldsymbol{\beta}) = \mathrm{E}\left[-\nabla^2 l\right]$.

(iii) The deviance.

(6 marks)

(b) Explain how the quantities in Part (a) can be used within a numerical method to determine values of the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ given a sample of data $\boldsymbol{y} = (y_1, \ldots, y_n)$.

(4 marks)

(c) State the approximate asymptotic distribution of the scaled deviance and suggest how this result can be used as part of an approximate solution to the problem caused by overdispersion.

(2 marks)

(d) For each of the forms of model misspecification below, explain briefly how the mean-variance relationship would be affected, indicating in each case whether overdispersion or underdispersion would occur if a Poisson model were assumed.

(i) For $n \geq 1$ and $0 < \pi < 1$ such that $n\pi = \lambda$,

$$Y \sim \text{Binomial}(n, \pi).$$

(2 marks)

(ii) For $Z \sim \text{Bernoulli}(\pi)$, $0 < \pi < 1$ and $\lambda_0 \neq \lambda_1$,

$$Y|Z = 0 \sim \text{Poisson}(\lambda_0), \qquad Y|Z = 1 \sim \text{Poisson}(\lambda_1).$$

(3 marks)

(iii)

$$Y = X|X > 0, \text{ where } X \sim \text{Poisson}(\lambda).$$

(3 marks)

(Total: 20 marks)

3. This question concerns data on subjects from the Framingham Heart Study. We consider the following variables, defined for each subject,

   - PREVAP, binary response variable, taking values 0 (no angina) and 1 (angina).
   - SEX, binary covariate taking values 0 (male) and 1 (female).
   - TOTCHOL, numerical covariate giving total cholesterol measured in mg/dL.

   The joint distribution of PREVAP and SEX is computed below.

```
cont_table <- dat %>% group_by(PREVAP, SEX) %>% summarize(count = n())
```

```
  PREVAP   SEX count
1      0     0  1852
2      0     1  2435
3      1     0    92
4      1     1    55
```

   The mean total cholesterol in the sample is 237 mg/dL and the standard deviation is 45 mg/dL.

   A model to see whether the prevalence of angina is different by sex can be fitted as follows,

```
fit0 <- glm(PREVAP ~ SEX, family = binomial, data = dat)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.0022     0.1068  -28.11  < 2e-16 ***
SEX          -0.7881     0.1732   -4.55 5.36e-06 ***
```

   A model that allows for the effect of cholesterol is

```
fit1 <- glm(PREVAP ~ SEX + TOTCHOL, family = binomial, data = dat)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.036823   0.433144  -9.320  < 2e-16
SEX         -0.841818   0.175877  -4.786  1.7e-06
TOTCHOL      0.004375   0.001740   2.515   0.0119
```

Question 3 continues on the following page

(a) State the link function used in `fit1` above, and hence write out the relationship between the mean μ_i and the linear predictor η_i for an observation i from this model. (2 marks)

(b) In terms of ratios of values given in the table, write down the value of

(i) The estimated intercept in `fit0`. (2 marks)

(ii) The estimated `SEX` parameter in `fit0`. (2 marks)

(c) Use the summary of the total cholesterol distribution in the sample to give a plain language interpretation of the parameter for `TOTCHOL` estimated in `fit1`. You may find it helpful to note that $\exp(0.2) \approx 1.2$.

(4 marks)

(d) Using the information at the end of the question, comment on whether there is evidence that the effect of cholesterol on angina differs by sex. (2 marks)

(e) Explain the statistical purpose of the code below labelled `bootstrap routine` and state the distribution that `lrt` would be expected to follow. (4 marks)

(f) Comment on the output of `fit3`, in which some values have been replaced by #####. In particular, state with justification the two missing values labelled ##1## and ##2##. (4 marks)

```
n_boot <- 1000
lrt <- rep(0, times = n_boot)
coef2 <- matrix(0, nrow = n_boot, ncol = 4)
for(i in 1:n_boot){ # bootstrap routine
  bt_idx <- sample(n, replace = TRUE)
  bt_dat <- dat[bt_idx, ]
  fit1_bt <- glm(PREVAP ~ SEX + TOTCHOL, family = binomial, data = bt_dat)
  fit2_bt <- glm(PREVAP ~ SEX * TOTCHOL, family = binomial, data = bt_dat)
  coef2[i, ] <- coef(fit2_bt)
  lrt[i] <- 2*(as.numeric(logLik(fit2_bt)) - as.numeric(logLik(fit1_bt)))
}

quantile(coef2[,4], c(0.025, 0.975))
2.5%        97.5%
  -0.0069  0.0068

fit3 <- glm(count ~ PREVAP * SEX, family = poisson(), data = cont_table)
    Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    ##1##  0.02324  323.795  < 2e-16 ***
PREVAP      -3.00223  0.10682  -28.107  < 2e-16 ***
SEX          0.27368  0.03083    8.876  < 2e-16 ***
PREVAP:SEX     ##2##    #####    #####    ##### ***
```

(Total: 20 marks)

4. Consider a normal linear mixed model of the form

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

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) The mixed model at the end of the question relates to the `Ergostool` dataset seen in class. The data come from a designed experiment to study the effort required to get up from four different stools. Each subject tries each stool once.

(i) Explain why it is appropriate to use fixed effects for stool type, but random effects for subject.

(2 marks)

(ii) Explain why is it is important to randomize the order in which the subjects are tested on the different stools.

(2 marks)

(iii) Write down an estimate of the intra-class correlation coefficient.

(2 marks)

(iv) Interpret the estimated parameter `TypeT4`.

(2 marks)

(v) Explain the difficulty that would arise when fitting the model below.

(2 marks)

```
fit2_ergoStool <- lm(effort ~ Type * Subject,
                     data = ergoStool)
```

Question 4 continues on the following page

(c) Consider a situation where there are m groups with 2 observations per group,

$$y_{ij} = \mu_j + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2), \qquad i = 1, \ldots, m,$$

where $\mu_1, \ldots, \mu_m$ are fixed effect parameters to be estimated.

(i) Show that the maximum likelihood estimator of σ^2 is biased in this case.

(2 marks)

(ii) Show that an unbiased estimate of σ^2 can be made by applying a linear transformation to the data, and state explicitly a suitable linear transformation in this case.

(3 marks)

(iii) Explain how the approach in c(ii) can be extended to estimation of variance components in the more general mixed model considered in part (a).

(2 marks)

```
fit_ergoStool <- lmer(effort ~ Type + (1|Subject),
                      data = ergoStool)

 Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 1.775    1.332
 Residual             1.211    1.100

Fixed effects:
            Estimate Std. Error t value
(Intercept)   8.5556     0.5760  14.853
TypeT2        3.8889     0.5187   7.498
TypeT3        2.2222     0.5187   4.284
TypeT4        0.6667     0.5187   1.285
```

(Total: 20 marks)

5. This question concerns data from a study of the reflexes of different subjects under different experimental conditions. The n measurements y_i take the values 1 or 0, depending on whether or not the reflex was observed, and a statistical model is specified for each y_i as follows:

$$y_i|\boldsymbol{b} \sim \text{BERNOULLI}(1, \mu_i), \qquad \mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}},$$

where $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$. The columns of the matrix $\boldsymbol{X}$ specify values of covariates or consist of zeros and ones to encode factors such as sex or experimental conditions. The matrix $\boldsymbol{Z}$ specifies which measurements correspond to which individual. The vector $\boldsymbol{\beta}$ contains predictors for the covariates and factors and the vector $\boldsymbol{b}$ contains the subject-specific random effects, with $b_j \sim N(0, \sigma_b^2)$ for each of the n_b subjects.

(a) Write down explicitly the log of the joint density $f(\boldsymbol{y}, \boldsymbol{b}; \sigma_b)$ of the data and random effects. (2 marks)

(b) The R function `lfyb` given at the end of the question computes the log joint density above, as well as its gradient vector and Hessian matrix. Derive the form of the missing terms in the code labelled `##A`, where A is a number. Answers may be given in mathematical notation or R code. (5 marks)

(c) Explain the function of the code `lal`. A line-by-line description is not required, but you should comment in your answer on the purpose of the lines marked `##Comment A`, where `A` is a number. (6 marks)

(d) Give a clear derivation of the approximation used in `lal`. (3 marks)

(e) State the large sample distribution of the maximum likelihood estimators of the fixed effect coefficients and explain how to test whether these coefficients are significantly different from zero. (4 marks)

```
lfyb <- function(y, b, eta0, Z, sig.b) {
    ## log joint density of y and b, for model
    ##   eta0 <- X%*%beta, eta = eta0 + Zb, p = exp(eta)/(1+exp(eta)),
    ## y ~ bernoulli(p)
      eta <- as.numeric(eta0 + Z%*%b)
      lf <- sum(y*eta - log(1 + exp(eta)))  +  ##1
      mu <- exp(eta)/(1 + exp(eta))
      g <- ##2
      diagH <- colSums(Z*Z*(mu*(mu - 1))) - 1/sig.b^2
      list(lf = lf, g = g, diagH = diagH)
}
```

Question 5 continues on the following page

```
lal <- function(theta, y, X, Z) {
    beta <- theta[-1]
    sig.b <- exp(theta[1])
    eta0 <- X%*%beta
    b <- rep(0, ncol(Z))
    lf <- lfyb(y, b, eta0, Z, sig.b)
    for (i in 1:200) {
        if (max(abs(lf$g)) < abs(lf$lf)*1e-9) break ## Comment 1
        step <- -lf$g/lf$diagH ## Comment 2
        repeat {
            lf1 <- lfyb(y, b + step, eta0, Z, sig.b)
            if (lf1$lf<lf$lf) step <- step/2 else break ##Comment 3
        }
    ## now update...
    lf <- lf1
    b <- b + step
    }
    if (i==200) warning("failed") ## Comment 4

    la <- lf$lf + length(b)*log(2*pi)/2 - sum(log(abs(lf$diagH)))/2 ## Comment 5
    -la
}
```

(Total: 20 marks)

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2024

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

60044/70044

Applied Statistical Inference (Solutions)

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

1. (a)

seen ⇓

4, A

$$L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2).$$

Since $\boldsymbol{\epsilon} \sim N(0,\, \sigma^2 I_n)$, we see $\boldsymbol{Y} \sim N(X\boldsymbol{\beta},\, \sigma^2 I_n)$.

Recall that for the general multivariate normal vector $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \Sigma)$, we have

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right).$$

Taking $\boldsymbol{\mu} = X\boldsymbol{\beta}$ and $\Sigma = \sigma^2 I_n$ gives

$$L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}) \right),$$

as required.
Maximizing L is equivalent to maximising $l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = \log\big(L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})\big)$. This is just

$$l(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}).$$

For fixed σ^2, maximizing l is then minimising the sum of squares,

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}).$$

We now compute the gradient ∇S. Using the product rule for ∇,

$$\nabla S(\boldsymbol{\beta}) = -2X^T (\boldsymbol{y} - X\boldsymbol{\beta}).$$

It might be more transparent to use components,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - X_i\boldsymbol{\beta})^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2.$$

The kth component of the gradient is then

$$[\nabla S(\boldsymbol{\beta})]_k = \frac{\partial}{\partial \beta_k} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 = -\sum_{i=1}^{n} 2x_{ik} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right) = \left[-2X^T (\boldsymbol{y} - X\boldsymbol{\beta}) \right]_k.$$

The maximizer $\widehat{\boldsymbol{\beta}}$ satisfies $\nabla S(\boldsymbol{\beta}) = 0$, equivalently $X^T X \widehat{\boldsymbol{\beta}} = X^T \boldsymbol{y}$.

seen ⇓

(b) For the intercept-only model, $X^T X$ is simply the 1×1 matrix (n) and $X^T \boldsymbol{y} = \sum_{i=1}^{n} y_i$, and so

2, B

$$\widehat{\beta_0} = (X^T X)^{-1} X^T \boldsymbol{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y},$$

as required. As a linear combination of normal variables, $\widehat{\beta_0}$ is clearly normal. By linearity,

$$\mathrm{E}\left[\widehat{\beta_0}\right] = \beta_0.$$

Using independence, its variance is just

$$\mathrm{Var}\left[\bar{Y}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[Y_i\right] = \frac{\sigma^2}{n}.$$

The sampling distribution of our estimator is therefore

$$\widehat{\beta_0} \sim N\left(\beta_0, \frac{\sigma^2}{n}\right).$$

unseen ⇓

(c) (i) The interaction term is `z:x`, and the p-value shows that this estimate is not significantly different from zero. Hence, no reason to doubt the model excluding this term .

2, C

2, B

(ii) `fit1` has 47 residual degrees of freedom and 3 estimated parameters so there must be $2n = 47 + 3 = 50$ observations, hence $n = 25$ in each group.

2, C

(iii) `##1##` is the RSS of `fit1`, which is 47×0.8857^2.
`##2##` is the number of additional model degrees of freedom of `fit2`, which is just 1.
`##3##` is the F statistic, which for a model with 1 additional model degree of freedom is just the square of the corresponding t-statistic. From the output for `fit2`, this is 0.685^2.

1, D

4, D

(iv) In terms of quantities available in the output, the RSS is

$$(n-p)\widehat{\sigma^2} = 47 \times 0.8857^2.$$

The R^2 value gives the proportion of variance explained, so the sample variance is

$$\frac{RSS}{n(1-R^2)} = \frac{47 \times 0.8857^2}{50 \times (1 - 0.7703)}$$

sim. seen ⇓

(v)
- Subjects are *r*andomly assigned to treatment or control, hence the treatment groups should be balanced for any covariate that influences y.
- Controlling for `x` reduces the residual standard error, hence leads to an estimate of the treatment effect with a lower sampling variance.
- Hence greater power to detect a given size of treatment effect.

3, D

2. (a) (i) The score function is the gradient of the log likelihood $U(\boldsymbol{\beta}) = \nabla l(\beta)$. seen ⇓

The likelihood has the form 2, A

$$L(\beta) = \prod_{i=1}^{n} \exp\left(-\mu_i + y_i \log \mu_i - log y_i!\right) = \exp\left(\sum_{i=1}^{n} -\mu_i + y_i \log \mu_i - \log y_i!\right).$$

Then with $\eta_i = X_i\boldsymbol{\beta}$, we have

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\mu_i + y_i \log \mu_i - \log y_i! = \sum_{i=1}^{n} -\exp(\eta_i) + y_i\eta_i - \log y_i!,$$

so that for $j = 1, \ldots, p$,

$$[\nabla l(\beta)]_j = \frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{n} -\exp(\eta_i) + y_i\eta_i - \log y_i!\right) = \sum_{i=1}^{n} -X_{ij}\exp(\eta_i) + y_i X_{ij} = \sum_{i=1}^{n} X_{ij}(y_i - \mu_i).$$

(ii) The Fisher information is given by 2, A

$$I = \mathrm{E}\left(-\nabla^2 l(\beta)\right).$$

Then

$$\left[\nabla^2 l(\beta)\right]_{jk} = \frac{\partial}{\partial \beta_k} U(\boldsymbol{\beta}) = -\sum_{j=1}^{n} X_{ij} X_{ik} \mu_i.$$

We recognise this as the (j, k) entry of $X^T W X$, where the matrix W is diagonal with ith diagonal entry μ_i. This quantity is non-random, and so taking expectation with respect to $\boldsymbol{y}$ has no effect. Asymptotically, the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by $(X^T W X)^{-1}$.

(iii) The deviance of a generalized linear model is the expression 2, A

$$D = 2\phi\left(l(\boldsymbol{y}; \boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}; \boldsymbol{y})\right),$$

where the first argument of the log likelihood is the estimated mean and the second argument is the response variable $\boldsymbol{y}$.
In this case, the dispersion parameter $\phi = 1$ and we have

$$l(\boldsymbol{y}; \boldsymbol{y}) = \sum_{i=1}^{n} -y_i + y_i \log y_i - \log y_i!$$

so that the deviance is

$$2\left(\sum_{i=1}^{n} -y_i + y_i \log y_i - \log y_i! - \sum_{i=1}^{n} -\hat{\mu}_i + y_i \log \hat{\mu}_i - \log y_i!\right) = 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \mu_i).$$

(b) 4, A

1. Choose a starting estimate $\widehat{\boldsymbol{\beta}}_0$ e.g. using a simple moment estimator.
2. For a given estimate $\widehat{\boldsymbol{\beta}}_k$, use a variation on Newton's method to obtain an improved estimate.
3. Specifically, form the second order Taylor approximant to the log likelihood at $\widehat{\boldsymbol{\beta}}_k$, this is

$$l(\boldsymbol{\beta}) \approx l(\widehat{\boldsymbol{\beta}}_k) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_k)^T \boldsymbol{U}(\widehat{\boldsymbol{\beta}}_k) - \frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_k)^T I(\widehat{\boldsymbol{\beta}}_k)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_k).$$

   (Strictly, we will do Fisher scoring, since we have replaced the hessian of the log likelihood at $\widehat{\beta}$ by its expectation.)
4. Take the maximum point of this second order approximant as an improved estimate, i.e.

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + I(\widehat{\boldsymbol{\beta}}_k)^{-1} \boldsymbol{U}(\widehat{\boldsymbol{\beta}}_k).$$

5. A reasonable point to stop can be determined by considering how the deviance changes at each iteration. By default, the `glm` routine in R stops when the deviance change criterion

$$\frac{D_{k+1} - D_k}{D_{k+1} + 0.1} < \epsilon,$$

   where $\epsilon = 10^{-8}$ by default.

2, B

(c) Asymptotically, the scaled deviance has an approximate $\chi^2(n-p)$ distribution. (More precisely, it is asymptotically $\chi^2(n-p, \nu)$, where the non-centrality parameter ν is typically small).

We use moment estimation,

$$\mathrm{E}\left[\frac{D}{\phi}\right] = \mathrm{E}\left[\chi^2(n-p)\right] = n - p,$$

hence $\widehat{\phi} = \frac{D}{n-p}$.

Overdispersion is the existence of an unmodelled variance component. It does not affect parameter estimates, but instead increases their variance, so that the asymptotic confidence intervals would be too small. The estimator above can be used within a quasi-Poisson model to account for overdispersion: in such a model, the parameter estimates would be unchanged, but all standard errors would be adjusted by a multiplicative factor of $\sqrt{\widehat{\phi}}$.

sim. seen ⇓

(d) (i) In this case, we have $\mathrm{E}[Y] = n\pi = \lambda$ and $\mathrm{Var}[Y] = np(1-p) = \lambda(1-p) < \lambda$, so that we would see underdispersion if a Poisson model were used.

2, B

unseen ⇓

(ii) Here we have

$$\mathrm{E}[Y] = \mathrm{E}_Z[\lambda_1 Z + \lambda_0(1-Z)] = \lambda_1\pi + \lambda_0(1-\pi).$$

3, D

Now by the law of total variance, and the fact that conditional on Z we have a Poisson distribution,

$$\begin{aligned}\mathrm{Var}\left[Y\right] &= \mathrm{E}\left[\mathrm{Var}\left[Y|Z\right]\right] + \mathrm{Var}\left[\mathrm{E}\left[Y|Z\right]\right] \\ &= \mathrm{E}_Z\left[\lambda_1 Z + \lambda_0(1-Z)\right] + \mathrm{Var}_Z\left[(\lambda_1-\lambda_0)Z + \lambda_0\right] \\ &= \lambda_1\pi + \lambda_0(1-\pi) + (\lambda_1-\lambda_0)^2\pi(1-\pi) \\ &> \mathrm{E}\left[Y\right],\end{aligned}$$

so that in this case we have overdispersion relative to a Poisson model.

unseen ⇓

3, D

(iii) We compute the mean by conditional probability.

$$\lambda = \mathrm{E}\left[X\right] = \mathrm{E}\left[X|X=0\right]\Pr(X=0) = \mathrm{E}\left[X|X>0\right]\Pr(X>0) = \mathrm{E}\left[Y\right](1-\exp(-\lambda)),$$

so that

$$\mathrm{E}\left[Y\right] = \frac{\lambda}{1-\exp(-\lambda)} > \lambda.$$

We can also apply the same argument to X^2. First, we note that $\mathrm{Var}[X] = \lambda$, so that $E\left[X^2\right] = \lambda + \lambda^2$.

$$\lambda+\lambda^2 = \mathrm{E}\left[X^2\right] = \mathrm{E}\left[X^2|X=0\right]\Pr(X=0) = \mathrm{E}\left[X^2|X>0\right]\Pr(X>0) = \mathrm{E}\left[Y^2\right](1-\exp(-\lambda)),$$

so that

$$\mathrm{E}\left[Y^2\right] = \frac{\lambda+\lambda^2}{1-\exp(-\lambda)},$$

and hence

$$\begin{aligned}\mathrm{Var}[Y] &= \frac{\lambda+\lambda^2}{1-\exp(-\lambda)} - \frac{\lambda^2}{(1-\exp(-\lambda))^2} \\ &= \frac{\lambda}{1-\exp(-\lambda)}\left((1+\lambda) - \frac{\lambda}{1-\exp(-\lambda)}\right) \\ &= \mu(1+\lambda-\mu) < \mu,\end{aligned}$$

hence in this case we have underdispersion relative to a Poisson model.

3. (a) No link function is mentioned, hence R will use the canonical link by default, which in this case is the logit link.

meth seen ⇓

2, A

Hence

$$\log\left(\frac{\mu}{1-\mu}\right) = \eta.$$

(b) (i)

meth seen ⇓

2, A

$$\widehat{\beta}_0 = \log\left(\frac{92}{1852}\right).$$

2, A

(ii)

$$\widehat{\beta}_1 = \log\left(\frac{55}{2435}\right) - \log\left(\frac{92}{1852}\right).$$

(c) A change in cholesterol by one unit (here mg/dL) increases the log odds of angina by an amount $\widehat{\beta}_1 = 0.004375$. So a 1 standard deviation increase in total cholesterol adds $44.65 \times 0.004375 \approx 0.2$ to the log odds of having angina. In practical terms, it increases the odds of having angina by a multiplicative factor of $\exp(0.2) \sim 1.2$. In plain language, controlling for sex, a 1 standard deviation increase in total cholesterol is associated with a 20% increase in the odds of having angina.

meth seen ⇓

4, B

(d) Zero is contained in the bootstrap confidence interval, hence we have no reason to reject the null hypothesis that the interaction effect is zero. There is no evidence that the effect of cholesterol on angina differs by sex.

meth seen ⇓

2, C

(e)
- This is a non-parametric bootstrap routine.
- The aim is to compare a model that assumes a common effect of cholesterol for the two sexes with a more general model in which the effect of cholesterol is allowed to vary between sexes.
- The method approximates the sampling distribution of statistics of interest, e.g. the interaction coefficient in part(d) and the likelihood ratio test statistic, by their bootstrap distributions.
- The method also looks at the distribution of the likelihood ratio test statistic comparing the two models.
- Under the null hypothesis, assuming asymptotic results hold, we would expect this statistic to have a $\chi^2(1)$ distribution.

sim. seen ⇓

4, B

unseen ⇓

(f)
- This is an instance of the relationship between the Poisson and multinomial distributions.
- The number of observations in the four cells of the contingency table are multinomially distributed;
- Therefore, equivalently, we can model these as four Poisson counts (with different means), conditional on a fixed total.
- `fit3` is a Poisson GLM with the log link.
- `##1##` is the intercept of this log-linear model. Using the contingency table, this must be $\log 1852$.
- `##2##` is the log of the difference in the rates of presence/absence of angina in the two different sexes, hence must also be the coefficient of `SEX` in `fit0`.

2, C

2, D

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.

seen ⇓

3, A

For the mean

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

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

meth seen ⇓

2, B

(b) (i)
- The aim of the experiment is to estimate the amount of effort required to get up from different types of stool. Therefore, we need to estimate a parameter for each of the stool types.
- By contrast, we are not interested in estimating the overall ability of individual subjects; we just need to allow for the fact that there will be a component of variability due to subject, i.e. that different observations of the same subject will be correlated.
- Hence we model the different abilities of individual subjects as realisations of a random variable $\boldsymbol{\nu}$ with mean zero and variance estimated from the data.

unseen ⇓

2, C

(ii)
- In the context of the experiment, a "learning effect" is plausible: on average the amount of effort needed may not be constant, e.g. people find it easier after the first attempt.
- If all subjects tried the stools in the same order, the learning effect would be entirely confounded with stool type.
- Hence we randomize the order in which subjects try the different stools, so that stool type and order are unrelated.

seen ⇓

2, A

(iii)
$$\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\epsilon^2} = \frac{1.775}{1.775 + 1.211}$$

seen ⇓

2, B

(iv) This is an estimate of the mean difference in effort required to get up from a stool of type $T4$ and from a stool of type $T1$. Considering the t-statistic, there is no evidence for a difference in mean effort for these two stool types.

(v) Each subject tries each stool type once. This model therefore has the same number of parameters as observations. Hence it is saturated, and so perfectly reproduces the data it is trained on (overfit). There are no residual degrees of freedom with which to estimate sampling uncertainty, and so it is not possible to make inferences from the sample to the population.

seen ⇓

2, C

(c) (i) $Y_{ij} \sim N(\mu_j, \sigma^2)$, for $i = 1, 2$ and $j = 1, 2, \ldots, m$. In this case the maximum likelihood estimator of the mean is just the sample average for each pair $\widehat{\mu}_j = \frac{1}{2}(Y_{1j} + Y_{2j})$.

seen ⇓

2, A

The maximum likelihood estimator for σ^2 is then

$$\frac{1}{m}\sum_{j=1}^{m} S_i^2 = \frac{1}{m}\sum_{j=1}^{m} \frac{1}{2}\left((Y_{1j} - \widehat{\mu}_j)^2 + (Y_{2j} - \widehat{\mu}_j)^2\right).$$

Simplifying individual terms,

$$S_j^2 = (Y_{1j} - \widehat{\nu}_j)^2 + (Y_{2j} - \widehat{\nu}_j)^2 = \frac{(Y_{1j} - Y_{2j})^2}{4},$$

but clearly $Y_{1j} - Y_{2j} \sim N(0, 2\sigma^2)$, so that $\mathrm{E}\left[S_j^2\right] = \frac{1}{2}\sigma^2$.

(ii) Consider the transformation $B_j = Y_{1j} - Y_{2j}$. Then $B_1, \ldots, B_m$ are independent and $N(0, 2\sigma^2)$ distributed. This transformation leads to a distribution that does not involves the nuisance parameters $\mu_1, \ldots, \mu_m$.

seen ⇓

3, A

The joint pdf of $B_1, \ldots, B_m$ is then

$$f(b_1, \ldots, b_m) = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi(2\sigma^2)}} \exp\left\{-\frac{1}{2(2\sigma^2)} b_j^2\right\}.$$

Hence the log-likelihood is

$$-\frac{m}{2}\log(4\pi\sigma^2) - \frac{1}{4\sigma^2}\sum_{j=1}^{m} B_j^2,$$

and it follows that

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{m}{2\sigma^2} + \frac{1}{4(\sigma^2)^2}\sum_{j=1}^{m} B_j^2.$$

Therefore, the MLE of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{2m}\sum_{j=1}^{m} B_j^2.$$

Finally, we can compute the expectation of this MLE and find that is it indeed an unbiased estimator of σ^2:

$$\mathrm{E}(\widehat{\sigma}^2) = \frac{1}{2m} \cdot m \cdot 2\sigma^2 = \sigma^2.$$

seen ⇓

2, A

(iii) The method of restricted maximum likelihood, we apply a linear transformation L to the response $\boldsymbol{Y}$ that projects the data onto the space orthogonal to that spanned by the columns of the design matrix. $L\boldsymbol{Y}$ is then independent of the fixed effects $\boldsymbol{\beta}$, and we can maximize the resulting restricted likelihood for the parameters $(\sigma_\nu^2, \sigma_\epsilon^2)$, obtaining unbiased estimates of these variances. If needed, these estimates can then be plugged in to the likelihood, which can then be maximized over the fixed effects $\boldsymbol{\beta}$

5. (a) The joint log density is given by

2, M

$$\begin{aligned}\log f(\boldsymbol{y}, \boldsymbol{b}) &= \log f(\boldsymbol{y}|\boldsymbol{b}) + \log f(\boldsymbol{b}) \\ &= \sum_{i=1}^{n} \eta_i y_i - \sum_{i=1}^{n} \log(1 + e^{\eta_i}) - \frac{n_b}{2} \log(2\pi\sigma_b^2) - \frac{1}{2\sigma_b^2} \sum_{j=1}^{n_b} b_j^2.\end{aligned}$$

5, M

(b) The missing code should be

```
##1 sum(dnorm(b,0,sig.b,log=TRUE))
##*2 t(Z)%*%(y-mu)  - b/sig.b^2
```

The second line comes from calculating the gradient vector as follows

$$\begin{aligned}\eta_i &= (\eta_0)_i + \sum_{k=1}^{n_b} Z_{ik} b_k, \\ \frac{\partial \sum_i \log(1 + e^{\eta_i})}{\partial b_j} &= \sum_i \frac{e^{\eta_i}}{1 + e^{\eta_i}} \frac{\partial \eta_i}{\partial b_j} = \sum_i \mu_i Z_{ij}, \\ \frac{\partial \sum_i y_i \eta_i}{\partial b_j} &= \sum_i y_i Z_{ij},\end{aligned}$$

so that

$$\begin{aligned}\frac{\partial \log f(\boldsymbol{y}, \boldsymbol{b})}{\partial b_j} &= \sum_i (y_i - \mu_i) Z_{ij} - b_j/\sigma_b^2, \\ \frac{\partial \log f(\boldsymbol{y}, \boldsymbol{b})}{\partial \boldsymbol{b}} &= \boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{b}/\sigma_b^2\end{aligned}$$

6, M

(c) The function `lal` uses Newton's method find the mode of $\log f(\boldsymbol{y}, \boldsymbol{b})$ over the argument $\boldsymbol{b}$. It then approximates the log likelihood $\log f(\boldsymbol{y}; \boldsymbol{\theta})$ using the Laplace approximation.

`##Comment 1` This line checks whether Newton's method has converged to within the defined tolerance level, specified as the ratio of the largest entry of the log likelihood's gradient to the function value.

`##Comment 2` Specifies the descent direction according to Newton's method. $\boldsymbol{\Delta} = \nabla f H^{-1}$, although since H is a diagonal matrix here, the update is simply carried out by element-wise division.

`##Comment 3` Checks whether a step in the direction Δ has in fact led to a decrease in the function, and if not halves the length of Δ, while preserving its direction. Since Δ is a descent direction (by the positive definiteness of H), a sufficiently small step will reduce the value of the objective function.

`##Comment 4` Ensures the Newton routine terminates after finitely many iterations, and returns a warning if convergence to within the required tolerance has not been

achieved.

##Comment 5 Computes the approximate likelihood using Laplace's approximation (see below). Note the simplified form of the determinant for a diagonal matrix.

3, M

(d) The Laplace approximation proceeds via a Taylor expansion of $\log f(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\theta})$ to second order (considered as a function of $\boldsymbol{b}$) about the mode $\widehat{b}$.

$$\log f(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\theta}) \approx \log f(\boldsymbol{y}, \widehat{\boldsymbol{b}}; \boldsymbol{\theta}) - \frac{1}{2}(\widehat{\boldsymbol{b}} - \boldsymbol{b})^T H (\widehat{\boldsymbol{b}} - \boldsymbol{b})$$

where H is the negative matrix of second derivatives of $\log f$ with respect to the b_j. We can now approximate $f(\boldsymbol{y}; \boldsymbol{\theta})$ as follows,

$$\begin{aligned} f(\boldsymbol{y}; \boldsymbol{\theta}) &= \int f(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{b} \approx \int f(\boldsymbol{y}, \widehat{\boldsymbol{b}}; \boldsymbol{\theta}) \exp\left(-\frac{1}{2} (\boldsymbol{b} - \widehat{\boldsymbol{b}})^\top H (\boldsymbol{b} - \widehat{\boldsymbol{b}}) \right) \mathrm{d}\boldsymbol{b} \\ &= (2\pi)^{n/2} \det(H)^{-1/2} f(\boldsymbol{y}, \widehat{\boldsymbol{b}}; \boldsymbol{\theta}) \end{aligned}$$

using the fact that the integrand is merely an unnormalized multivariate Normal density function. Note that $\widehat{\boldsymbol{b}}$ and H depend on both $\boldsymbol{y}$ and $\boldsymbol{\theta}$.

4, M

(e) $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{I}_{\boldsymbol{\beta}}^{-1})$. Use the command `optim` with method BFGS to find the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$. The command `Hessian=T` ensures that the method will return an approximate Hessian when it has converged. The inverse of this matrix should then be close to the information matrix, the square roots of whose diagonal entries are standard errors for the parameters. This determines a z score from which the significance of each parameter can be tested.

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

MATH60044 Applied Statistical Inference

Question	Marker's comment
1	Parts (a) and (b) were generally well done.Part (c) (i) - (ii) were also well done.Part (c) (iii) was found difficult. Almost no-one used the fact that the F-statistic for comparing models 1 and 2 was the square of the t-statistic in model 2.Part (c) (iv) was a good differentiator, with some good attempts.Almost no-one gave a clear answer to Part (c) (v).
2	Parts (a) and (b) were done uniformly well, with a good understanding shown of the Poisson GLM. Part (c) had some good answers, although the quasi-Poisson model was not always described in detail. Part (d) was a good discriminator. (i) was answered well overall. (ii) was somewhat harder, with many candidates assuming properties of the Variance operator that do not hold, instead of using the Law of Total Variance. Part (iii) was challenging with relatively few complete answers.nbsp;
3	This question was based on a data set that was used in a tutorial class.(a) was well done by essentially all candidates(b) also done well by most; good understanding of contingency tables shown(c) some good answers, especially by those who were familiar with the relevant tutorial.(d) very well done overall. The most common incorrect response was to misunderstand the question as referring to the effect of sex on angina, rather than on whether there is evidence that sex modifies the effect of cholesterol on angina.(e) Many good attempts, but a good discriminator(f) Found challenging by most. Surprisingly, no-one referred to the relationship between the multinomial and Poisson distributions, which was the point of the question. A reasonable minority were able to calculate the parameters of the model, however.
4	This question was basically well done. Some marks were lost for failing to discuss fully aspects of the question. How do we interpret the significance of the coefficient of stool T4 (there is no significant difference from T1)? Having estimated the variance components by REML, what do we then do to infer the fixed effects?

MATH70044 Applied Statistical Inference

Question	Marker's comment
4	This question was basically well done. Some marks were lost for failing to discuss fully aspects of the question. How do we interpret the significance of the coefficient of stool T4 (there is no significant difference from T1)? Having estimated the variance components by REML, what do we then do to infer the fixed effects?
5	Almost all attempts showed serious engagement with the mastery material. This was an excellent capstone for the module, allowing candidates to show the insights they had developed over the term.nbsp;