Imperial College
London

MATH70044

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May 2023

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

Statistical Modelling 2

Date: 1 June 2023

Time: 10:00 – 12:30 (BST)

Time Allowed: 2.5hrs

This paper has 5 Questions.

Please Answer All Questions in 1 Answer Booklet

Candidates should start their answers 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

© 2023 Imperial College London

1. For the analysis of a random sample of data points $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, a Normal linear model is proposed

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n),$$

where X is an $n \times p$ matrix ($p < n$) of covariates, including an intercept, and X has full rank.

(a) Write down a system of linear equations satisfied by the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$. State the distribution of $\boldsymbol{Y}$ and use it to determine the distribution of $\widehat{\boldsymbol{\beta}}$.

(2 marks)

(b) Find an expression for the matrix P such that the fitted values $\widehat{\boldsymbol{Y}}$ can be written as $P\boldsymbol{Y}$. Show that P is a projection matrix.

(3 marks)

(c) Write down a projection matrix Q such that $\boldsymbol{e}$, the vector of residuals, can be written as $\boldsymbol{e} = Q\boldsymbol{Y}$.

(1 mark)

(d) Find the joint distribution of $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$.

(5 marks)

(e) Figure 1 shows two plots, labelled A and B. One of the plots displays the residuals plotted against the fitted values $\widehat{\boldsymbol{Y}}$, and the other shows the residuals plotted against the response variable $\boldsymbol{Y}$. Explain which plot is which, and state which of the two is more useful as a diagnostic plot. State a form of misspecification that your preferred plot can be used to identify. (5 marks)

(f) Normal QQ-plots are often used in model criticism. The usual diagnostic QQ plot in R shows the standardized residuals against the standard normal distribution. Explain the function of this plot, and suggest why it is more informative than a normal QQ plot of the response variable $\boldsymbol{Y}$ or of the raw residuals $\boldsymbol{e}$.

(4 marks)

QUESTION CONTINUES ON FOLLOWING PAGE

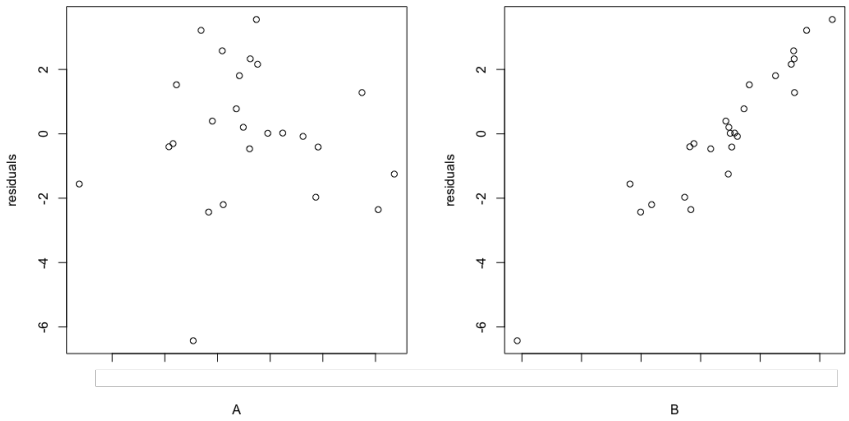

Figure 1: One plot above shows residuals from a normal linear model plotted against the fitted values; the other shows residuals plotted against the response variable Y .

(Total: 20 marks)

2. This question concerns the Poisson generalized linear model, in which the random variables $Y_i \sim \text{POISSON}(\mu_i)$ are independent for $i = 1, \ldots, n$, and $\boldsymbol{\mu}$ is related to the linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$ by the canonical link function.

(a) Write the Poisson mass function $f_Y(y)$ in exponential family form, indicating clearly the form of the canonical parameter.

(3 marks)

(b) Define the *score function* $\boldsymbol{U}(\boldsymbol{\beta})$ and show that it has the form

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

(3 marks)

(c) Define the *Fisher information* $I(\boldsymbol{\beta})$ and find a diagonal matrix W such that $I(\boldsymbol{\beta}) = X^T W X$. State the approximate asymptotic relationship between $I(\boldsymbol{\beta})$ and the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$.

(3 marks)

(d) Define what is meant by the *deviance* of a generalized linear model and show that in this case,

$$D = 2 \sum_{i=1}^{n} y_i \log \left(\frac{y_i}{\widehat{\mu_i}} \right) - (y_i - \widehat{\mu_i}).$$

State the approximate asymptotic sampling distribution of the deviance.

(3 marks)

(e) Suppose the covariate X_j has the property that $X_{ij} = 0$ for all observations i such that $y_i > 0$, and $X_{ij} > 0$ for all observations such that $y_i = 0$. Show that the maximum likelihood estimator for β_j does not exist.

(3 marks)

(f) A simulation study is conducted into the effects of model misspecification on Poisson GLMs. In the code at the end of the question, data are simulated from a binomial distribution with a large, fixed number of trials and a constant success probability. Then a Poisson intercept-only GLM is fit to the data, and an estimate of the dispersion parameter is extracted. The plot in Figure 2 shows the estimated dispersion parameter when the Poisson GLM is fitted to independently generated datasets, each with a different value of the success probability p. Comment on the features of this plot.

(5 marks)

QUESTION CONTINUES ON FOLLOWING PAGE

```
y <- rbinom(n_data, size = n_trials, prob = p)
fit0 <- glm(y ~ 1, family ="poisson")
phi <- deviance(fit0)/fit0$df.residual
```

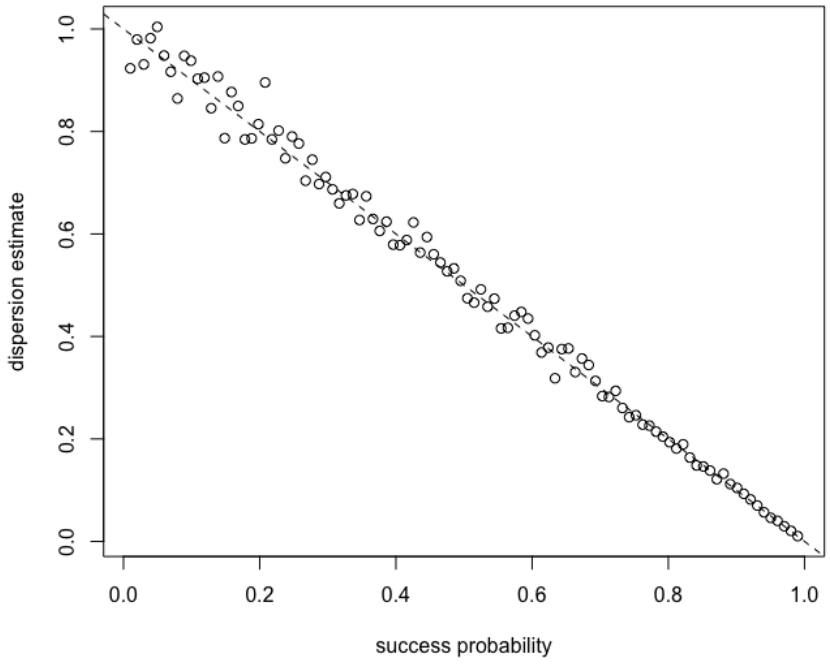

Figure 2: Dispersion estimate based on the deviance. Each point in the plot represents an estimate from an independent dataset, with fixed number of trials, and success probability as given on the x-axis.

(Total: 20 marks)

3. A study is conducted into the incidence of a form of respiratory disease. The covariates are body mass index (bmi), which is a numerical measure of weight, and smoking status, a binary variable indicating whether or not the subject smokes. For each subject, the response variable is a binary indicator, taking the value 1 if the subject has the disease, and 0 otherwise.

The output below is a 2×2 contingency table, showing the incidence of the disease conditional on smoker status, and a corresponding model. Further relevant code is included at the end of the question.

```
> xtabs(~ dat$y + dat$smoker)
     dat$smoker
dat$y   0   1
    0 618 151
    1 164 389
> fit0 <- glm(y ~ smoker, family = binomial, data = dat)
> summary(fit0)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.32662    0.08784  -15.10   <2e-16 ***
smoker       2.27292    0.13003   17.48   <2e-16 ***
```

(a) Identify the link function that was used in `fit0` and state an advantage of using this link function.

(2 marks)

(b) Write down an expression, in terms of the values in the 2×2 table, for the estimated effect of smoking in the model `fit0`. Assuming the model is reasonable, give a plain language summary of the effect of smoking.

(3 marks)

(c) Explain the difference between the models `fit1` and `fit2`. Assuming that asymptotic results are reliable, use the summary information to evaluate the evidence for the additional effect estimated in `fit2`.

(5 marks)

(d) With reference to the summary information and the data context, comment on the reliability of the asymptotic results.

(3 marks)

(e) The code at the end of the question gives another approach to inference that does not rely on asymptotic results. Explain what the code achieves, and interpret the output. Your response should address the purpose of the code; a line-by-line description is not needed.

(5 marks)

QUESTION CONTINUES ON FOLLOWING PAGE

(f) Suggest a practical difficulty that could arise when using the numerical routine above for a binomial GLM with a much smaller sample.

(2 marks)

```
## Improved models
> fit1 <- glm(y ~ smoker + bmi, family = binomial, data = dat)
> fit2 <- glm(y ~ smoker * bmi, family = binomial, data = dat)

> summary(fit2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.77841    0.57162 -13.608  < 2e-16 ***
smoker      -0.77551    1.04363  -0.743  0.45743
bmi          0.29211    0.02455  11.900  < 2e-16 ***
smoker:bmi   0.15560    0.05003   3.110  0.00187 **

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1797.23  on 1321  degrees of freedom
Residual deviance:  960.84  on 1318  degrees of freedom
AIC: 968.84

## Test statistic
test_stat <- as.numeric(logLik(fit2)) - as.numeric(logLik(fit1))

## Bootstrap likelihood ratio test
n_boot <- 10000
lrt_boot <- rep(0, times = n_boot)
dat_boot <- dat

## Pr(Y_i = 1) estimated using fit1
pstar <- predict(fit1, type = "response")

for(i in 1:n_boot){
  dat_boot$y <- rbinom(n, size = 1, prob = pstar)
  fit1boot <- glm(y ~ smoker + bmi, family = binomial, data = dat_boot)
  fit2boot <- glm(y ~ smoker * bmi, family = binomial, data = dat_boot)
  lrt_boot[i] <- as.numeric(logLik(fit2boot)) - as.numeric(logLik(fit1boot))
}
> mean(lrt_boot > test_stat)
[1] 0.0012
```

(Total: 20 marks)

4. The one-way random effects model can be written as

$$y_{ij} = \mu + \nu_j + \epsilon_{ij}, \qquad i = 1, \ldots, K_j, \quad j = 1, \ldots, m,$$

where $\nu_j \sim N(0, \sigma_\nu^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ are independent random variables.

(a) This part concerns the *balanced* case in which all groups have an equal number of observations, so that $K_j = K$ for all $j = 1, \ldots, m$.

(i) Determine the distribution of $\bar{Y}_{\cdot j} = \frac{1}{K}\sum_{i=1}^{K} Y_{ij}$ and $\bar{Y} = \frac{1}{m}\sum_{i=1}^{K} \bar{Y}_{\cdot j}$.

(4 marks)

(ii) Determine the distribution of $SSE = \sum_{i=1}^{K}\sum_{j=1}^{m}\left(Y_{ij} - \bar{Y}_{\cdot j}\right)^2$ and $SSA = \sum_{i=1}^{K}\sum_{j=1}^{m}\left(\bar{Y}_{\cdot j} - \bar{Y}\right)^2$. Determine whether or not these random variables are independent.

(4 marks)

(iii) Show that a confidence interval for the mean $\alpha_j = \mu + \nu_j$ of the jth group can be written in terms of the quantity

$$\frac{\bar{y}_{\cdot j} - \alpha_j}{[SSE/mK(K-1)]^{\frac{1}{2}}}.$$

What is the distribution of this quantity?

(5 marks)

(b) The R output at the end of the question concerns an *unbalanced* one-way random effects model, in which groups are either small, with $K_j = 2$, or large, with $K_j = 50$.

(i) State the method that has been used to estimate the variance components, and explain briefly how it works.

(2 marks)

(ii) Use the output provided to write down a numerical expression for the intraclass correlation coefficient. No simplification is required.

(2 marks)

(iii) Figure 3 shows the relationship between the best linear predictor of the random effects $\widehat{\nu}_j = \mathrm{E}(\nu_j|\boldsymbol{y})$ and $\bar{y}_{\cdot j} - \bar{y}$. Comment on the appearance of the plot, indicating in particular which size of group corresponds to which style of points.

(3 marks)

QUESTION CONTINUES ON FOLLOWING PAGE

```
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ 1 + (1 | id)

REML criterion at convergence: 7679.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2392 -0.6726  0.0090  0.6506  3.2152

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.8237   0.9076
 Residual             1.0254   1.0126

Fixed effects:
            Estimate Std. Error t value
(Intercept)   2.9389     0.1017   28.89
```

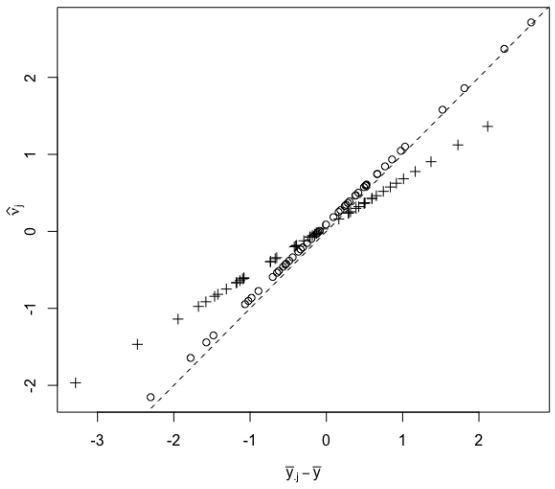

Figure 3: Relationship between the best linear unbiased estimator $\widehat{\nu}_j$ and $\bar{y}_{.j} - \bar{y}$. Point style represents group size. The identity line is shown dashed.

(Total: 20 marks)

5. (a) Give a brief summary of the aim of the experiment analysed in the article by Efron, and the main statistical challenge associated with data on the scale considered. In your response, you should explain the problem that is caused by multiple testing when evaluating many candidate hypotheses.

(4 marks)

(b) Equation 3.2 of the article defines the effect size z_i for gene i by

$$z_i = \Phi^{-1}\left(F_{100}(t_i)\right),$$

where Φ and F_{100} are the cumulative distribution functions of the standard normal distribution and t- distribution with 100 degrees of freedom, respectively, and t_i is the t-statistic for gene i. Explain from first principles why z_i should follow a standard normal distribution for null genes.

(3 marks)

(c) Consider a test with power β and significance level α. Write down an expression for the false discovery rate if the proportion of null genes is p_0, justifying your answer briefly.

(2 marks)

(d) In the article, the marginal probability density function of the effect size is given by an expression of the form

$$f_Z(z) = \exp\left(\sum_{j=0}^{d} \beta_j z^j\right)$$

in which $d = 7$. Table 1 at the end of the question gives the log likelihood of models for different values of the polynomial degree d. Use the information to suggest the value of d that would be preferred, explaining your conclusions.

(3 marks)

QUESTION CONTINUES ON FOLLOWING PAGE

degree	llk	change
2	-186.51	
3	-186.25	0.26
4	-148.96	37.29
5	-148.42	0.53
6	-148.31	0.12
7	-148.16	0.15
8	-146.93	1.23

Table 1: Log likelihood and change in log likelihood for different polynomial degree. Any discrepancies are due to rounding.

(e) The aim of this part of the question is to justify the use of the Poisson GLM applied to histogram counts. Writing π_k for the probability that a z score lands in bin k, state the joint distribution of $Y_1, \ldots, Y_m$, where Y_k is the number of scores that fall in bin k and m is the number of bins. Write down the joint probability distribution of m independent but not necessarily identically distributed Poisson random variables, conditional on the event that their sum is N. Show that this has the same form as the joint distribution of $Y_1, \ldots, Y_m$.

(3 marks)

(f) Explain the *zero* assumption made in the article, and how it is used to estimate the parameters of an empirical null distribution $z_i \sim N(\delta_0, \sigma_0^2)$ as well as the proportion p_0 of null genes. Write down equations that can be used to estimate these parameters in terms of the parameters β_j above.

(5 marks)

(Total: 20 marks)

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2023

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

MATH60044/70044

Statistical Modelling 2 (Solutions)

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

1. (a) seen ⇓

$$X^T X\widehat{\beta} = X^T \boldsymbol{y},$$

2, A

or, since X has full rank,

$$\widehat{\beta} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Now, $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, so $\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$. As a linear transformation of $\boldsymbol{Y}$, $\hat{\boldsymbol{\beta}}$ is multivariate normal, and so its distribution is determined by its mean and covariance matrix.

First, by linearity,

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}) = (X^T X)^{-1} X^T \mathrm{E}(\boldsymbol{Y}) = (X^T X)^{-1} X^T X\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Then

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\beta}) &= \mathrm{Var}((X^T X)^{-1} X^T \boldsymbol{Y}) = (X^T X)^{-1} X^T \mathrm{Var}(\boldsymbol{Y}) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.\end{aligned}$$

3, A

(b) The fitted values are given by $\widehat{\boldsymbol{y}} = X\widehat{\beta} = X(X^T X)^{-1} X^T \boldsymbol{y}$, so the matrix $P = X(X^T X)^{-1} X^T$.

We check that P is a projection matrix by checking it is both symmetric and idempotent.

It is symmetric, since $P^T = (X(X^T X)^{-1} X^T)^T = (X^T)^T ((X^T X)^{-1})^T X^T = X(X^T X)^{-1} X^T = P$, using the facts that $(AB)^T = B^T A^T$ and $(A^{-1})^T = (A^T)^{-1}$.

It is idempotent since

$$P^2 = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X(X^T X)^{-1} I X^T = X(X^T X)^{-1} X^T = P.$$

seen ⇓

(c) The vector of residuals $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - P\boldsymbol{Y} = (I_n - P)\boldsymbol{Y}$, so that $Q = I_n - P$.

1, B

[Not needed for the mark but included for completeness as used later] We show that Q is symmetric and idempotent. $(Q^T)_{ii} = 1 - P_{ii} = Q_{ii}$ and $(Q^T)_{ij} = Q_{ji} = -P_{ji} = -P_{ij} = Q_{ij}$ as $P = P^T$. Thus $Q = Q^T$.
For idempotence, $Q^2 = QQ = Q^T Q$ as Q is symmetric by above. Then $Q^T Q = (I_n - P)^T (I_n - P) = I_n - 2P + \underbrace{P^T P}_{:=P} = I_n - P = Q$.

(d) $\widehat{\boldsymbol{Y}} = P\boldsymbol{Y}$. As a linear transformation of $\boldsymbol{Y}$, $\hat{\boldsymbol{Y}}$ is multivariate normal, and so its distribution is determined by its mean and covariance matrix. The same is true for $\boldsymbol{e} = (I_n - P)\boldsymbol{Y}$, so that $(\widehat{\boldsymbol{Y}}, \boldsymbol{e})$ is a multivariate normal vector. This is a $2n$-dimensional vector whose distribution is supported on an n-dimensional subspace, i.e. the variance-covariance matrix has rank n.

2, B
~~3, D~~

unseen ⇓

We compute the expectations, variances and covariance.

$$\mathrm{E}(PY) = P\mathrm{E}(Y) = X(X^TX)^{-1}X^TX\beta = X\beta.$$

For the covariance matrix

$$\mathrm{Var}(PY) = P\mathrm{Var}(Y)P^T = P\sigma^2 I_n P^T = \sigma^2 P^2 = \sigma^2 P.$$

Similarly,

$$\mathrm{E}((I_n - P)Y) = (I_n - P)\mathrm{E}(Y) = (I - X(X^TX)^{-1}X^T)X\beta = X\beta - X\beta = \mathbf{0}.$$

$$\mathrm{Var}((I_n - P)Y) = (I - P)\mathrm{Var}(Y)(I_n - P)^T = (I_n - P)\sigma^2 I_n (I_n - P)^T = \sigma^2(I - P)^2 = \sigma^2(I_n - P).$$

To compute the covariance,

$$\mathrm{Cov}(\widehat{\boldsymbol{Y}}, \boldsymbol{e}) = \mathrm{E}\left(PY\left[(I_n - P)Y\right]^T\right) - E(PY)\left[\mathrm{E}((I_n - P)Y)\right]^T = \mathrm{E}\left(PY\left[(I_n - P)Y\right]^T\right).$$

Then

$$\mathrm{E}\left(PY\left[(I_n - P)Y\right]^T\right) = \mathrm{E}\left(PYY^T(I - P)\right) = P\mathrm{E}(YY^T)(I_n - P) = P\left(\sigma^2 I_n + X\boldsymbol{\beta}\boldsymbol{\beta}^T X^T\right)(I_n - P) = 0,$$

since $P(I_n - P) = 0$ and $X^T(I_n - P) = 0$.

Hence the joint distribution of $(\widehat{\boldsymbol{Y}}, \boldsymbol{e})$ is (degenerate) multivariate normal with mean vector and variance-covariance matrix

$$\begin{pmatrix} X\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \qquad \sigma^2 \left(\begin{array}{c|c} P & \mathbf{0} \\ \hline \mathbf{0} & I_n - P \end{array}\right)$$

(e)

3, C
~~2, D~~

* Note that the (sample) covariance between the residuals and fitted values is zero, since these two vectors are necessarily orthogonal to each other: $\widehat{\boldsymbol{y}}$ is in the column space of X and $\boldsymbol{e}$ is in its orthogonal complement.
* On the other hand, $\boldsymbol{e} = (I_n - P)\boldsymbol{y}$, so necessarily $\boldsymbol{e}$ and $\boldsymbol{y}$ are linearly related, and will typically show strong correlation.
* Hence, plot A shows residuals against fitted values and plot B shows residuals against the response variable.
* Plot A is what R does, and is the better choice of residual plot, because it will show no pattern when the model is fitting well, and so structure in the plot can be attributed to model misspecification.
* Plot B is less useful precisely because it will contain structure even for a model that is fitting well.
* Plot A is useful (e.g.) for identifying non-linear relationships between Y and the covariates. Such a relationship will be visible in the residuals after a linear model is fit.

(f)

2, C
2, D

* The normal QQ plot is a check on the distributional assumptions of the model, i.e. the normal, constant variance distribution of errors.
* If the distribution of standardized residuals suggests the errors are not plausibly normally distributed, this might invalidate inference that relies on the normality assumption.
* A normal QQ-plot of the response variable is not generally useful for model criticism - in general, $\boldsymbol{Y}$ is a mixture of normal distributions with different means, and so will not typically look normal.
* A normal QQ-plot of the raw residuals is not generally useful because different residuals may have very different variances. This is precisely why we work with standardized residuals.

2. (a)

seen ⇓

3, A

$$\Pr(Y = y; \mu) = \exp(-\mu)\frac{\mu^y}{y!} = \exp\left(-\mu + y\log\mu - \log y!\right).$$

We identify $\log\mu$ as the canonical parameter.

seen ⇓

(b) The score function is the gradient of the log likelihood $U(\boldsymbol{\beta}) = \nabla l(\beta)$.

3, A

For the Poisson generalized linear model, the canonical link function is $\theta = \eta = \log\mu$.

Then the likelihood has the form

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

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 $j = 1, \ldots, p$,

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

This is indeed the jth entry of the vector $X^T(\boldsymbol{y} - \boldsymbol{\mu})$, as required.

seen ⇓

(c) The Fisher information is given by

3, 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^TWX, 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 $(X^TWX)^{-1}$.

(d) The deviance of a generalized linear model is the expression

2, A
1, B

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

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

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

Asymptotically, the 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.)

1, B
2, D

(e) Consider the jth component of the score equation defined in Part (b). By the condition in the question, this is

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} -X_{ij} \exp(\eta_i) + y_i X_{ij} = \sum_{i: y_i > 0} -X_{ij} \exp(\eta_i).$$

But this expression is clearly negative, and so this component of the score equation has no solution.

2, C
3, D

(f)
- The plot shows evidence of misspecification of the mean-variance relationship, leading to underdispersion as a linear function of p.
- The Poisson model assumes a mean-variance relationship $\mathrm{Var}(Y) = \mu$ but in fact the variance function is $\mathrm{Var}(Y) = np(1-p) = \mu(1-p)$, where n is the number of trials.
- As $p \to 0$, the binomial distribution goes over into a Poisson distribution, and so the extent of misspecification decreases: in this limit we see a dispersion parameter estimated close to 1.
- For p close to zero, see the variability expected from the sampling variance of the deviance (which is approximately χ^2).
- As $p \to 1$, get increasingly severe underdispersion.

3. (a) Since no link function is specified when the model is fit, R uses the canonical link, which for the binomial family is the **logit link**. $\eta = \log\left(\frac{p}{1-p}\right)$.

seen ⇓

2, A

With this choice of link function (any single plausible reason for the second mark)

* The fitted values lie in an appropriate range for probabilities.
* The effects of predictors are interpretable on the odds scale
* The link is canonical, so the observed information is the same as the expected information.
* Effects can be estimated from either a prospective or retrospective study design.

sim. seen ⇓

3, A

(b)

$$\widehat{\beta}_{\mathsf{smoker}} = \log(389/151) - \log(164/618) \qquad (\sim 2.27)\,.$$

For smokers, the odds of developing the disease are larger by a multiplicative factor of $\exp(2.27) \sim 9.7$. This is statistically significant, with a vanishingly small p-value.

sim. seen ⇓

2, A

3, B

(c) In `fit1`, the model is

$$\eta_i = \beta_0 + \beta_{\mathsf{smoker}}\mathsf{smoker}_i + \beta_{\mathsf{bmi}}\mathsf{bmi}_i.$$

In `fit2`, the model is

$$\eta_i = \gamma_0 + \gamma_{\mathsf{smoker}}\mathsf{smoker}_i + \gamma_{\mathsf{bmi}}\mathsf{bmi}_i + \gamma_{\mathsf{interaction}}\mathsf{bmi}_i \times \mathsf{smoker}_i.$$

This means that in `fit1`, the linear relationship between the log odds of developing the disease and bmi is common between the groups. In contrast, `fit2` allow for the relationship between the log odds of developing the disease and BMI to be different for smokers and non-smokers.

In this part, we assume that the asymptotic results hold, so that the sampling distribution of the parameter estimate is normally distributed around its true value, with variance-covariance matrix given by the inverse of the Fisher information. On this assumption, the p-value for the interaction parameter `smoker:bmi` is 0.00187, which is significant at the 5% level. Hence there is evidence that the relationship between BMI and the log odds of having the disease depends on whether or not the subject smokes.

(d)
meth seen ⇓
2, B
1, C

* The residual deviance is ~ 960, on 1318 residual degrees of freedom.
* For a model that fits well, in the asymptotic regime we would expect that the distribution of the residual (scaled) deviance is close to $\chi^2(n-p)$. This means that the quantity

$$\widehat{\phi} = \frac{D}{n-p} \sim 1.$$

* However, for these data this quantity is around $960.84/1318 \sim 0.72$, substantially smaller than 1.
* This could be sampling uncertainty (could be checked e.g. with bootstrap).
* Alternatively, it suggests either that we are not in the asymptotic regime, or that some model misspecification is present, e.g. underdispersion due to unmodelled correlation or missing variables.
* A priori, the data context suggests that the asymptotic results may be unreliable, since we are doing binary logistic regression: likely to have only a small number of observations for a given bmi/smoker combination.

(e)
meth seen ⇓
2, A
3, C

* The code is a **parametric bootstrap** routine.
* It uses simulation to evaluate the distribution of the likelihood ratio test statistic under the null hypothesis that the smaller model (`fit1`) generated the data.
* It takes the estimated parameters of the model `fit1` as fixed, and uses them to produce estimates of π_i, the probability that the ith subject has the disease, given their covariates, for each i.
* These probabilities are then repeatedly used to generate independent samples from the response distribution.
* For each simulated dataset, both models are fitted and the log likelihood ratio is extracted.
* An empirical p-value is then obtained, as the proportion of simulated instances for which the log likelihood ratio exceeds the value for the data.
* The conclusion in this case is that the increase in the log likelihood when fitting the larger model is greater than we would expect if the data were generated from the smaller model. Hence we reject the null hypothesis - it appears that the interaction term is needed.

sim. seen ⇓

2, D

(f)
- ∗ Particularly with small datasets, it is possible for the observations to be *linearly separable* in covariate space.
- ∗ This means that there exists a linear function of the covariates $f(\boldsymbol{x}) = \sum_{j=1}^{p} \beta_j x_j$ such that all observations with $y_i = 1$ have $f(\boldsymbol{x}) > 0$ and all observations with $y_i < 0$ have $f(\boldsymbol{x}) < 0$.
- ∗ In this case, the maximum likelihood estimator of $\boldsymbol{\beta}$ does not exist, and estimates of the log likelihood from R will be unreliable.
- ∗ Even for a (small) dataset that is not linearly separable, it is likely that some of the parametric bootstrap samples will be linearly separable; the resulting log likelihoods would contaminate the bootstrap distribution.

4. (a) (i) Since linear combinations of normal random variables are normal, and ν_j and ϵ_{ij} are independent, we have that

seen ⇓

2, A

2, B

$$\bar{Y}_{\cdot j} = \mu + b_j + \bar{\epsilon}_{\cdot j} \sim N\left(\mu, \sigma_\nu^2 + \frac{\sigma_\epsilon^2}{K}\right),$$

$$\bar{Y} = \mu + \bar{b} + \bar{\epsilon} \sim N\left(\mu, \frac{\sigma_\nu^2}{m} + \frac{\sigma_\epsilon^2}{mK}\right).$$

.

seen ⇓

2, A

2, B

(ii) We note the standard distributional result that if $X_1, \ldots X_n \sim N(\mu, \sigma^2)$ are independent then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \qquad \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n-1)\sigma^2} \sim \chi^2(n-1),$$

and these random variables are independent.
We now apply this result repeatedly. For each $j = 1, \ldots, m$, we have $Y_{ij} - \bar{Y}_{\cdot j} = \epsilon_{ij} - \bar{\epsilon}_{\cdot j}$ and so

$$\sum_{i=1}^{K} (Y_{ij} - \bar{Y}_{\cdot j})^2 \sim \sigma_\epsilon^2 \chi^2(K-1).$$

These variables for each group are mutually independent, so it follows that (with slight abuse of notation)

$$SSE = \sum_{j=1}^{m} \sum_{i=1}^{K} \left(Y_{ij} - \bar{Y}_{\cdot j}\right)^2 = \sum_{j=1}^{m} \sigma_\epsilon^2 \chi^2(K-1) = \sigma_\epsilon^2 \chi^2\left(m(K-1)\right).$$

Now, since the variables $\bar{Y}_{\cdot j}$ are independent, with sample mean $\bar{Y}$, the random variable

$$SSA = \sum_{j=1}^{m} \sum_{i=1}^{K} \left(\bar{Y}_{\cdot j} - \bar{Y}\right)^2 \sim K\left(\sigma_b^2 + \frac{\sigma_\epsilon^2}{K}\right) \chi^2(m-1).$$

It is straightforward to check that $Y_{ij} - \bar{Y}_{\cdot j}$ is independent of $\bar{Y}_{\cdot j} - \bar{Y}$ for each j, and so it is clear that (as functions of these variables alone), SSA and SSE are independent.

(iii) From the statement above, $\bar{Y}_{\cdot j} - \alpha_j \sim N(0, \frac{\sigma_\epsilon}{K}^2)$. It follows (since $\bar{Y}_{\cdot j}$ is independent of SSE) that

seen ⇓

4, B

1, C

$$\frac{\sqrt{K}(\bar{Y}_{\cdot j} - \alpha_j)/\sigma_\epsilon}{\sqrt{SSE/\left(\sigma_\epsilon^2 m(K-1)\right)}} = \frac{\bar{Y}_{\cdot j} - \alpha_j}{\sqrt{SSE/\left(mK(K-1)\right)}} \sim t_{m(K-1)}.$$

This distribution statement can now be manipulated to give a confidence interval

$$\left(\bar{y}_{\cdot j} - t_\star \sqrt{\frac{SSE}{mK(K-1)}}, \bar{y}_{\cdot j} + t_\star \sqrt{\frac{SSE}{mK(K-1)}}\right),$$

in which $t_\star$ is an appropriate quantile of the t-distribution with $m(K-1)$ degrees of freedom.

seen ⇓

2, A

(b) (i) Restricted Maximum Likelihood (REML) was used to fit the model. To estimate the variance components using this method, 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 $(\sigma_\nu^2, \sigma_\epsilon^2)$, to obtain unbiased estimates of these variances.

meth seen ⇓

2, B

(ii)

$$ICC = \frac{0.8237}{0.8237 + 1.0254} \sim 0.45.$$

unseen ⇓

3, D

(iii)

- We see in the plot that the best linear predictor is a linear function of the difference between the sample mean $\bar{y}_{\cdot j}$ and the grand mean $\bar{y}$.
- In fact, the best linear predictor satisfies

$$\tilde{\nu}_j = \mathrm{E}(\nu_j | \boldsymbol{y}) = \frac{\bar{y}_{\cdot j} - \bar{y}}{1 + \frac{\sigma_\epsilon^2}{n_j \sigma_b^2}}.$$

- In this expression, we see a *shrinkage* effect that is dependent on group size. The prediction for small groups is shrunk back to zero.
- This shows the random effects model as an interpolation between a *complete pooling* model which is unaware of the group structure, and a *no pooling* model in which the group structure is modelled using fixed effects.
- The circular points lie very close to the identity line; they have hardly been shrunk at all, and so must correspond to the larger group. The cross points experience rather more shrinkage, as we would expect given that the smaller groups contain only two observations.

5. (a)
* The aim of the experiment is to identify a small list of candidate genes that might be implicated in prostate cancer.
* In a hypothesis testing framework, a threshold would be chosen such that all genes with a sufficiently large standardized mean difference in expression between groups would be followed up.
* The difficulty is that there is a large number of genes to test, and a priori only a relatively small number of them will be involved.
* Following up a gene is costly in time and resources for the experimental team, and so it is important to choose a threshold with an acceptable false discovery rate.

unseen ⇓

4, M

(b) Under the null hypothesis, the t-statistic is t_{100} distributed, because there are 100 observations and 2 estimated parameters. This means that $F_1 00(t_i) \sim \textsc{Uniform}(0,1)$, by the probability integral transform. Applying Φ^{-1} then gives a $N(0,1)$ quantity, again by the probability integral transform.

meth seen ⇓

3, M

(c) If the proportion of null genes is p_0, then a proportion $p_0\alpha$ of all genes will be false positives and a proportion $(1-p_0)\beta$ will be true positives. This means that the probability that a gene is a false positive, given that it is positive, is

$$\frac{p_0\alpha}{p_0\alpha + (1-p_0)\beta}.$$

meth seen ⇓

2, M

(d) We can use the AIC to compare between models. Need to identify the lowest AIC from the information given. We can see from the table that the log likelihood increases substantially going from degree 3 to degree 4, which more than compensates for the additional parameter when calculating the AIC. Increasing the degree further increases the log likelihood by less than 1, and so does not compensate in the AIC for the additional parameter. Hence the AIC will increase. The AIC will continue to increase for each additional parameter, until a final decrease from degree 7 to degree 8. However, the AIC for degree 8 will still be higher than that for degree 4. This is because the cumulative increase in likelihood is $0.5 + 0.12 + 0.15 + 1.23 = 2$, but the number of parameters has increased by 4.

meth seen ⇓

3, M

(e) Let Y_k be the number of z scores in bin k, for $k = 1, \ldots m$. Then the vector $\boldsymbol{Y} = (Y_1, \ldots Y_m)$ is $\text{MULTINOMIAL}\,(N, (\pi_1, \ldots \pi_k))$.

sim. seen ⇓

3, M

We start from the joint distribution of independent Poisson random variables, conditional on their sum. Note first that $\sum_{j=1}^m Y_j \sim \text{POISSON}(\sum_{j=1}^m \lambda_j)$.

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

This is indeed the probability mass function of a multinomial, setting $\pi_j = \frac{\lambda_j}{\sum_{k=1}^m \lambda_k}$.

meth seen ⇓

5, M

(f) The *zero* assumption asserts that most of the z-values near zero come from null genes. Then, for $z \sim 0$, we have

$$f(z) = p_0 f_0(z) + (1 - p_0) f_1(z) \approx p_0 f_0(z).$$

Making the assumption that f_0 is the density of a $N(\delta_0, \sigma_0^2)$ random variable, we get that

$$\begin{aligned}
f(z) = \exp\left(\sum_{j=0}^d \beta_j z^j\right) &\approx p_0 \exp\left(-\frac{1}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}(z - \delta_0)^2\right) \\
&= \exp\left(\log p_0 - \frac{1}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\left(z^2 - 2\delta_0 z + \delta_0^2\right)\right).
\end{aligned}$$

Equating coefficients of the Taylor expansion of $\log f(z)$ then gives

$$\beta_2 = -\frac{1}{2\sigma_0^2}$$

$$\beta_1 = \frac{\delta_0}{\sigma_0^2}$$

$$\beta_0 = \log(p_0) - \frac{1}{2}\log(2\pi\sigma_0^2) - \frac{\delta_0^2}{\sigma_0^2}.$$

The first equation gives σ_0; the second gives δ_0 and the third gives p_0.

Review of mark distribution:
Total A marks: 31 of 32 marks
Total B marks: 20 of 20 marks
Total C marks: 12 of 12 marks
Total D marks: 17 of 16 marks
Total marks: 100 of 80 marks
Total Mastery marks: 20 of 20 marks

If your module is taught across multiple year levels, you might have received this form for each level of the module. You are only required to fill this out once for each question.		
ExamModuleCode	**QuestionNumber**	**Comments for Students**
MATH60044/70044	1	No Comments Received
MATH60044/70044	2	No Comments Received
MATH60044/70044	3	No Comments Received
MATH60044/70044	4	No Comments Received
MATH70044	5	No Comments Received