IMPERIAL

MATH70044

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
Summer 2025

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

Applied Statistical Inference

Date: Friday, May 16, 2025

Time: Start time 14:00 – End time 16:30 (BST)

Time Allowed: 2.5 hours

This paper has 5 Questions.

Please Answer Each Question in a Separate Answer Booklet

This is a closed book examination.

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.

Credit will be given for all questions attempted.

Allow margins for marking.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO DO SO

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

1. (a) For a normal linear model of the form $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$ and X has full rank, show that the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ of the model parameters satisfies the linear system

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

Hence write down the expectation and variance of $\widehat{\boldsymbol{\beta}}$.

(6 marks)

The remainder of this question concerns the data used in Tutorial 1. The data come from a designed experiment in which fruit flies were randomized in equal numbers to one of four different diet conditions: either a control diet or one of three test diets labelled `FD` in the output at the end of the question. The response variable is the ratio of fat to protein.

(b) Give the form of the row of the design matrix for (i) an individual in the control diet group and (ii) an individual in the S24Y3 group. Hence write down expressions for the mean protein level in these two groups in terms of values given in the output.

(4 marks)

(c) Write down the form of a 95% confidence interval for the mean difference in the response between the control diet and diet S24Y3 in terms of information available in the output.

(2 marks)

(d) Carry out a test of the null hypothesis that all diets have equal mean response. State the value of the test statistic, and its null distribution and degrees of freedom. Explain the conclusion of the test in plain language.

(4 marks)

(e) Explain why the standard errors for the three experimental diets are equal.

(1 mark)

(f) Justifying your answer, calculate the value of the largest standardized residual.

(3 marks)

(Total: 20 marks)

Information for Question 1 can be found on the following page

```
Call:
lm(formula = tag_protein ~ diet, data = df)

Residuals:
Min        1Q         Median     3Q        Max
-0.14341   -0.08586   -0.00863   0.04319   0.31732

Coefficients:
              Estimate   Std. Err  t val   Pr(>|t|)
(Intercept)   0.13156    0.03027   4.347   8.05e-05 ***
dietFD S24Y3  0.16284    0.04281   3.804   0.000435 ***
dietFD S30Y3  0.16893    0.04281   3.946   0.000282 ***
dietFD SY10   0.12997    0.04281   3.036   0.004015 **

Residual standard error: 0.1049 on 44 degrees of freedom
Multiple R-squared:  0.3163, Adjusted R-squared:  0.2696
F-statistic: 6.784 on 3 and 44 DF,  p-value: 0.0007376
```

2. This question concerns a gamma distributed response variable Y, with probability density function parameterized in terms of its mean $\mu > 0$ and known shape parameter $\nu > 0$,

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} e^{-\frac{\nu y}{\mu}}, \quad y > 0.$$

(a) Write the probability density of Y in exponential family form.

(4 marks)

(b) Stating any standard results you use, determine the form of the mean-variance relationship, i.e. find the function $V(\mu)$ such that

$$\mathrm{Var}[Y] = \phi V(\mu),$$

where ϕ depends only on ν.

(3 marks)

(c) In a common biological model for biochemical reactions, the reaction rate y can be modelled as

$$y = \frac{\alpha_0 x}{1 + \alpha_1 x},$$

where α_1 and α_2 are constants and x is a covariate (called substrate concentration) that can be measured without error. If this model describes the relationship between $\mathrm{E}[Y]$ and the non-random covariate x, show that we can form a gamma GLM with the canonical link for the linear predictor

$$\eta = \beta_0 + \frac{\beta_1}{x}.$$

In your answer, state the link function clearly and specify how the regression coefficients β_0 and β_1 are related to the parameters α_0 and α_1. (3 marks)

(d) Given a random sample (x_i, y_i), $i = 1, \ldots, n$, explain briefly how a numerical method can be used to estimate parameters and produce approximate standard errors.

(5 marks)

(e) Derive the form of the deviance for the Gamma GLM and state the asymptotic sampling distribution of the scaled deviance.

(2 marks)

(f) Explain how the shape parameter could be estimated if it were not known, assuming any necessary asymptotic results. Comment on how estimating the shape parameter would impact inference. (3 marks)

(Total: 20 marks)

3. This question concerns a binomial generalized linear model (GLM) with the canonical link function. The independent response variables $Y_1, \ldots, Y_n$ are distributed as $Y_i \sim \text{Binomial}(n_i, \pi_i)$, where $0 < \pi_i < 1$. For a binomial GLM, the canonical link is the logit function, defined in terms of π_i as

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right).$$

(a) For the two-parameter model with linear predictors

$$\eta_i = \beta_1 + \beta_2 x_i \quad (i = 1, \ldots, n),$$

where x_i is a real-valued covariate, show that the log-likelihood can be written in terms of the parameters β_1 and β_2, as

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i(\beta_1 + \beta_2 x_i) - n_i \log(1 + \exp\{\beta_1 + \beta_2 x_i\}) + \log \binom{n_i}{y_i} \right].$$

(3 marks)

(b) Show that the score vector $\boldsymbol{U}(\boldsymbol{\beta})$ is $X^T(\boldsymbol{y} - \boldsymbol{\mu})$, where X is the design matrix with columns $\mathbf{1}$ and $\boldsymbol{x}$ and $\boldsymbol{\mu} = E[\boldsymbol{Y}]$. Find the Fisher information matrix.

(7 marks)

(c) For binary logistic regression, where $n_i = 1$ for all i, explain how the maximum likelihood estimator for β_2 can fail to exist.

(2 marks)

The remainder of the question concerns the following variables from the Framingham Heart Study, which was introduced in Tutorial 7.

- `PREVHYP`, binary response variable, taking values 0 (no hypertension) and 1 (hypertension).
- `SEX`, binary covariate taking values 0 (male) and 1 (female).
- `BMI`, numerical covariate giving body mass index, mass / height ^{2}, in kg/m^2.

(d) Write down the proportion of subjects in the sample with hypertension in terms of values given in the output.

(2 marks)

(e) Give a plain language interpretation of the effect of BMI on the prevalence of hypertension in `fit_BMI`.

(2 marks)

(f) Stating any necessary asymptotic results, determine whether there is evidence that the effect of BMI differs by sex. Comment on the validity of the asymptotic results, and suggest an alternative way of obtaining standard errors, identifying any limitations.

(4 marks)

Information for Question 3 can be found on the following page

(Total: 20 marks)

```
fit0 <- glm(PREVHYP ~ 1, family = binomial, data = dat)
summary(fit0)

             Estimate   Std. Err  z val   Pr(>|z|)
(Intercept) -0.74227    0.03213   -23.1   <2e-16 ***

fit_BMI <- glm(PREVHYP ~ BMI, family = binomial, data = dat)
summary(fit_BMI)

             Estimate   Std. Err  z val    Pr(>|z|)
(Intercept) -5.056265   0.238769  -21.18   <2e-16 ***
BMI          0.164654   0.008948   18.40   <2e-16 ***

fit_SEX <- glm(PREVHYP ~ BMI*SEX, family = binomial, data = dat)
  summary(fit_SEX)

             Estimate   Std. Err  z val  Pr(>|z|)
(Intercept) -3.26872    0.86571  -3.776  0.00016 ***
BMI          0.09365    0.03231   2.899  0.00375 **
SEX         -1.08202    0.50476  -2.144  0.03206 *
BMI:SEX      0.04325    0.01890   2.289  0.02209 *
```

4. This question uses the `Oxboys` dataset introduced in Tutorial 8. It is a longitudinal dataset consisting of observations of the heights of 26 boys. Relevant code and abridged output are provided at the end of the question.

(a) With reference to the data context, explain what is meant by the complete pooling and no pooling models. Explain how they can be considered as limiting cases of the linear mixed model.

(5 marks)

(b) State the method used to fit the model in `fit_mixed1`. Explain in detail how it works, and give an advantage and a disadvantage of this method.

(4 marks)

(c) State the difference between `fit_mixed1` and `fit_mixed2` and explain how the two mixed models can be compared .

(4 marks)

(d) Give an interpretation in plain language of the fixed and random effects for `age` in `fit_mixed2`.

(2 marks)

(e) Write down the estimated variance-covariance matrix for the random effects in `fit_mixed2` in terms of values given in the output.

(2 marks)

(f) Comment on the goodness of fit of `fit_mixed2`.

(3 marks)

Information for Question 4 can be found on the following page

```
fit_pool <- lm(height ~ age, data = Oxboys)
fit_nopool <- lmList(height ~ age|Subject, data = Oxboys)
fit_mixed1 <- lmer(height ~ age + (1|Subject), data = Oxboys)
fit_mixed2 <- lmer(height ~ age + (age|Subject), data = Oxboys)
```

```
  Formula: height ~ age + (age | Subject)
   Data: Oxboys

REML criterion at convergence: 724.1

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.65092 -0.57493 -0.02843  0.59605  2.60497

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 65.3041  8.0811
          age          2.8248  1.6807   0.64
 Residual              0.4355  0.6599
Number of obs: 234, groups:  Subject, 26

Fixed effects:
            Estimate Std. Error t value
(Intercept) 149.3718     1.5854   94.22
age           6.5255     0.3363   19.40
```

(Total: 20 marks)

5. This question concerns Fisher's dilution assay, which can be used to estimate the concentration of a micro-organism in a solution. At each stage, the solution is progressively diluted and a number of plates are inoculated with the diluted solution. On some plates, the organism grows, while others remain uninfected. For each plate, the response variable Y takes the value 1 if the plate is infected, and 0 otherwise.

A model for the experimental set up is as follows. Suppose there are ρ_0 micro-organisms per unit volume. Using a dilution factor of 2, at stage x the concentration of micro-organisms is

$$\rho_x = \frac{\rho_0}{2^x}.$$

At stage x, when a unit volume of solution is applied to a plate, the number of micro-organisms present will be Poisson distributed with mean ρ_x. Plates may be assumed to be independent. Let π_x be the probability under this model that a plate is infected.

(a) State the key assumptions underpinning the model above.

(3 marks)

(b) Write down the relationship between π_x and ρ_0, and hence justify the choice of link function in `fit1`.

(2 marks)

(c) The model in `fit0` is for a preliminary experiment that only uses the original undiluted solution to inoculate plates, i.e. the data consists of n observations with $x = 0$. Write down the maximum likelihood estimator of π_0 in this case and state its variance.

(3 marks)

(d) Show that the intercept in `fit0` has an approximate variance

$$\frac{\pi_0}{(1-\pi_0)\log(1-\pi_0)^2 n}.$$

(4 marks)

(e) Suppose the initial concentration is such that $\pi_0 \approx 1$. By considering the approximate Fisher information of the intercept estimator in Part (d), explain how serial dilution enables more precise estimation of the concentration.

(4 marks)

(f) The alternative approach `fit2` estimates π_x for each dilution as in Part (c), and then uses a weighted linear model to produce a combined estimate of ρ_0. Suggest how the weights should be chosen, and comment on the relative efficiency of this approach versus the generalised linear model. (4 marks)

Information for Question 5 can be found on the following page

(Total: 20 marks)

```
# data from a single dilution
fit0 <- glm(cbind(y, n_plates - y) ~ 1, family = binomial(link = "cloglog"),
         data = single_dilution)

# data from a dilution series
fit1 <- glm(cbind(y, n_plates - y) ~ x, family = binomial(link = "cloglog"),
        data = series_dilution)

fit2 <- lm(log(-log(1 - y/n_plates)) ~ x, weights = w, data = series_dilution)
```

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2025

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

MATH60044/70044

Applied Statistical Inference(Solutions)

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

1. (a) 6, 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

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

.

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

Now, for the expectation, using linearity and noting that $X^T X$ is invertible since X has full rank,

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

For the variance, we use bilinearity,

$$\text{Var}\left[\widehat{\boldsymbol{\beta}}\right] = \text{Var}\left[(X^TX)^{-1}X^T\boldsymbol{Y}\right] = (X^TX)^{-1}X^T\text{Var}\left[\boldsymbol{Y}\right]X(X^TX)^{-1} = (X^TX)^{-1}X^T\sigma^2 I_n X(X^TX)^{-1}$$
$$= \sigma^2(X^TX)^{-1}.$$

seen ⇓

(b) (i) For an observation in the control group, $\boldsymbol{x}_c = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}$. Hence

2, A

$$\widehat{y}_c = \boldsymbol{x}_c\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 = 0.13156.$$

2, B

seen ⇓

(ii) For an observation in the S24Y3 group, $\boldsymbol{x}_1 = \begin{pmatrix}1 & 1 & 0 & 0\end{pmatrix}$. Hence

$$\widehat{y}_1 = \boldsymbol{x}_1\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 = 0.13156 + 0.16284.$$

sim. seen ⇓

(c)

2, C

$$\frac{\widehat{\beta}_1 - \beta_1}{SE(\widehat{\beta}_1)} \sim t_{n-p}.$$

Manipulating this expression, in terms of information available in the output, a 95% confidence interval for β_1 is

$$\widehat{\beta}_1 \pm SE(\widehat{\beta}_1)t_{44}(0.025) = 0.16284 \pm 0.04281t_{44}(0.025).$$

sim. seen ⇓

(d)

4, C

- The null hypothesis is that $\beta_1 = \beta_2 = \beta_3 = 0$. This is a comparison of the model given in the question and the intercept-only model.
- The relevant part of the output is the F-statistic. Under the null hypothesis, this has the $F(3, 44)$ distribution.
- The observed value of 6.784 is far into the right tail of this distribution - the corresponding p-value is ~ 0.0007.
- There is substantial evidence against the null hypothesis. The observed data provide strong evidence that the group means are not all equal.

1, D

(e) The equal standard errors reflect a balanced design: an equal number of observations in each group, and a symmetry in the columns of the design matrices corresponding to the three non-control groups.

unseen ⇓

(f)

3, D

- Since the design is balanced, all observations have equal leverage.
- The leverage of observation i is p_{ii}, the ith diagonal entry of the projection matrix $P = X(X^TX)^{-1}X^T$. Since P is a projection matrix, $trace(P) = 4$, its rank, given by the number of model parameters.
- Since there are $12 \times 4 = 48$ observations overall, the leverage of an individual observation is $4/48 = 1/12$.
- Hence the value of largest standardized residual is

$$\frac{0.31732}{\widehat{\sigma}\sqrt{1 - p_{ii}}} = \sqrt{\frac{12}{11}} \times 0.1049 \times 0.31732.$$

2. (a) Want to put the Gamma pdf into exponential (dispersion) family form:

meth seen ⇓

4, A

$$\exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}$$

For the gamma distribution as parameterized:

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

We then identify $\theta = -1/\mu$ and $\phi = 1/\nu$. Then

$$\begin{aligned} b(\theta) &= \log(\mu) = -\log(-\theta) \\ c(y,\phi) &= (\phi^{-1}-1)\log(y) - \phi^{-1}\log(\phi) - \log\Gamma(\phi^{-1}). \end{aligned}$$

meth seen ⇓

(b) For an exponential family distribution, $b''(\theta)$ gives the variance function. Hence we differentiate $b(\theta) = -log(-\theta)$ twice to obtain

3, B

$$b'(\theta) = -\frac{1}{\theta}; \qquad b''(\theta) = \frac{1}{\theta^2} = \mu^2 = V(\mu).$$

Hence $\mathrm{Var}[Y] = \phi\mu^2 = \frac{\mu^2}{\nu}$.

unseen ⇓

(c) By inspecting the exponential family representation, the canonical link function is: $\eta = \ g(\mu) = -1/\mu$.

3, B

Given in the question that $y = \frac{\alpha_0 x}{1+\alpha_1 x}$, hence

$$\frac{-1}{y} = \frac{1+\alpha_1 x}{\alpha_0 x} = \frac{\alpha_1}{\alpha_0} + \frac{1}{\alpha_0 x},$$

hence $\beta_0 = \frac{\alpha_1}{\alpha_0}$ and $\beta_1 = \frac{-1}{\alpha_0}$ (minus sign not important so long as consistent).

sim. seen ⇓

(d)

5, A

- Parameters are estimated by maximum likelihood. In this case, we set the gradient of the log likelihood (score function) to zero and solve.
- The resulting nonlinear equations need to be solved numerically. The log likelihood function is convex and (so long as the MLE exists), it can be found by a version of Newon's method (iterated weighted least squares.
- The method proceeds by successively optimising quadratic approximations to the log likelihood. This can be shown to be equivalent to fitting a succession of weighted linear models, where the weights account for the mean-variance relationship established in part (b).
- The method can be initialized e.g. with a moment estimator, and (in R at least) terminates when the deviance is no longer changing.
- At convergence, the hessian of the log likelihood $X^T W X$ is approximately the Fisher information, and its inverse approximates the variance of the parameter estimates. Taking the square root of diagonal entries gives approximate standard errors.

sim. seen ⇓

(e) We first obtain the log likelihood of the saturated model. Hence we set the expected values μ_i to be the observations y_i.

2, C

$$l(\boldsymbol{y},\phi,\boldsymbol{y}) = \sum_{i=1}^{n}\left[\frac{y_i(-1/y_i) - \log(y_i)}{\phi} + c(y_i,\phi)\right] = \sum_{i=1}^{n}\left[\frac{-1-\log(y_i)}{\phi} + c(y_i,\phi)\right]$$

For a model with expected values $\widehat{\mu}_i$, the log-likelihood is

$$l(\widehat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) = \sum_{i=1}^{n} \left[\frac{y(-1/\widehat{\mu}_i) - \log(\widehat{\mu}_i)}{\phi} + c(y_i, \phi) \right]$$

Now substitute in to the definition of deviance to get

$$\begin{aligned} D &= 2\phi \left\{ \ell(\boldsymbol{y}, \phi, \boldsymbol{y}) - l(\widehat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) \right\} \\ &= 2\phi \left\{ \sum_{i=1}^{n} \left[\frac{-1 - \log(y_i)}{\phi} + c(y_i, \phi) - \frac{y_i(-1/\widehat{\mu}_i) - \log(\widehat{\mu}_i)}{\phi} - c(y_i, \phi) \right] \right\} \\ &= 2\phi \left\{ \sum_{i=1}^{n} \left[\frac{-\log(y_i/\widehat{\mu}_i) + (y - \widehat{\mu}_i)/\widehat{\mu}_i}{\phi} \right] \right\} \\ &= 2 \sum_{i=1}^{n} \left[-\log\left(\frac{y_i}{\widehat{\mu}_i} \right) + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} \right] \end{aligned}$$

Very roughly, the scaled Deviance $D/\phi \sim \chi^2(n-p)$ when p parameters are estimated; here $p = 2$.

unseen ⇓

3, D

(f) Using the result in the previous part, we can form a moment estimator. $E[D/\phi] = n - p$, so that we can set $\widehat{\phi} = \frac{D}{n-2}$, giving $\widehat{\nu} = \frac{n-2}{D}$.

Where this parameter is estimated from data, the appropriate approximate confidence interval for β is based on quantiles of the t distribution, rather than the normal distribution. For moderate sample sizes, the difference between the two is small.

3. (a) First, the binomial mass function is

seen ⇓

3, A

$$\Pr(Y_i = y_i|\pi_i) = \binom{n_i}{y_i}\pi_i^{y_i}\,(1-\pi_i)^{n_i-y_i}\,.$$

Take the logarithm of the given probability mass function and using independence, sum the log-likelihoods for the individual observations

$$\ell(\boldsymbol{\pi};\boldsymbol{y}) = \sum_{i=1}^{n}\left[y_i\log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\log(1-\pi_i) + \log\binom{n_i}{y_i}\right].$$

Using the canonical link we have

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

Rearranging to make π the subject, we obtain

$$\pi = \frac{1}{1+e^{-\eta}}.$$

For the given linear predictor we have

$$\pi_i = \frac{1}{1+e^{-\eta_i}} = \frac{1}{1+\exp\{-(\beta_1+\beta_2 x_i)\}}.$$

Start by rewriting the log-likelihood:

$$\begin{aligned}&\sum_{i=1}^{n}\left[y_i\eta_i + n_i\log\left(1-\frac{1}{1+e^{-\eta_i}}\right) + \log\binom{n_i}{y_i}\right]\\ =&\sum_{i=1}^{n}\left[y_i\eta_i - n_i\log\left(1+e^{\eta_i}\right) + \log\binom{n_i}{y_i}\right].\end{aligned}$$

Next, plug in the linear predictors, η_i, given in the question:

$$\ell(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{i=1}^{n}\left[y_i(\beta_1+\beta_2 x_i) - n_i\log\left(1+\exp\{\beta_1+\beta_2 x_i\}\right) + \log\binom{n_i}{y_i}\right]$$

which is the stated result.

meth seen ⇓

(b) The score vector is given by:

3, B

4, C

$$\boldsymbol{U} = \begin{pmatrix}\dfrac{\partial\ell}{\partial\beta_1} & \dfrac{\partial\ell}{\partial\beta_2}\end{pmatrix}^T.$$

To derive $\boldsymbol{U}$, take the partial derivatives in turn:

$$\begin{aligned}\frac{\partial\ell}{\partial\beta_1} &= \sum_{i=1}^{n}\left[y_i - n_i\left(\frac{e^{\beta_1+\beta_2 x_i}}{1+e^{\beta_1+\beta_2 x_i}}\right)\right]\\ &= \sum_{i=1}^{n}\left[y_i - n_i\left(\frac{1}{1+e^{-(\beta_1+\beta_2 x_i)}}\right)\right]\\ &= \sum_{i=1}^{n}\left[y_i - n_i\pi_i\right]\end{aligned}$$

since

$$\frac{1}{1+e^{-(\beta_1+\beta_2 x_i)}} = \frac{1}{1+e^{-\eta_i}} = \pi_i.$$

Similarly,

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_2} &= \sum_{i=1}^{n}\left[y_i x_i - n_i x_i\left(\frac{e^{\beta_1+\beta_2 x_i}}{1+e^{\beta_1+\beta_2 x_i}}\right)\right] \\ &= \sum_{i=1}^{n} x_i(y_i - n_i\pi_i),\end{aligned}$$

using the same substitution used for the first partial derivative.

This is exactly $\boldsymbol{U} = X^T(\boldsymbol{y}-\boldsymbol{\mu})$, noting the mean of the binomial is $\mu_i = n_i\pi_i$.

For the Fisher Information, recall that

$$\mathcal{J} = -\mathrm{E}\begin{pmatrix} \frac{\partial^2\ell}{\partial\beta_1^2} & \frac{\partial^2\ell}{\partial\beta_1\partial\beta_2} \\ \frac{\partial^2\ell}{\partial\beta_2\partial\beta_1} & \frac{\partial^2\ell}{\partial\beta_2^2}\end{pmatrix} = -\mathrm{E}\begin{pmatrix} \frac{\partial}{\partial\beta_1}U_1 & \frac{\partial}{\partial\beta_1}U_2 \\ \frac{\partial}{\partial\beta_2}U_1 & \frac{\partial}{\partial\beta_2}U_2\end{pmatrix}.$$

Since U_1 and U_2 have already been computed we only need

$$\begin{aligned}\frac{\partial\pi_i}{\partial\beta_1} &= \frac{e^{-(\beta_1+\beta_2 x_i)}}{(1+e^{-(\beta_1+\beta_2 x_i)})^2} = \pi_i(1-\pi_i) \\ \frac{\partial\pi_i}{\partial\beta_2} &= \frac{x_i e^{-(\beta_1+\beta_2 x_i)}}{(1+e^{-(\beta_1+\beta_2 x_i)})^2} = x_i\pi_i(1-\pi_i)\end{aligned}$$

to obtain

$$\begin{aligned}\mathcal{J} &= \mathrm{E}\begin{pmatrix} \sum_{i=1}^{n} n_i\frac{\partial}{\partial\beta_1}\pi_i & \sum_{i=1}^{n} x_i n_i\frac{\partial}{\partial\beta_1}\pi_i \\ \sum_{i=1}^{n} n_i\frac{\partial}{\partial\beta_2}\pi_i & \sum_{i=1}^{n} x_i n_i\frac{\partial}{\partial\beta_2}\pi_i\end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n} n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} n_i x_i\pi_i(1-\pi_i) \\ \sum_{i=1}^{n} n_i x_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} n_i x_i^2\pi_i(1-\pi_i)\end{pmatrix}.\end{aligned}$$

This can be written more cleanly as X^TWX where W is a diagonal matrix with $w_{ii} = n_i\pi_i(1-\pi_i) = \mathrm{Var}[Y_i]$.

2, D

(c) When the data are *linearly separable*, the MLE for β_2 does not exist. For the single covariate model considered here, this means that there is a value x_0 such that $y_i = 1$ whenever $x_i > x_0$ and $y_0 = 0$ whenever $x_i < x_0$ (or vice versa). In this case, the fitted probabilities π_i are either 0 or 1. Clearly we cannot estimate the effect of a unit change in x in this case.

sim. seen ⇓

2, A

(d) The intercept of `fit0` gives the log odds of having hypertension in the sample. Hence, applying the inverse of the link function,

$$p = \frac{1}{1 + \exp(0.74227)}.$$

sim. seen ⇓

2, A

2, C

2, D

(e) Assuming the model fit is adequate, for each unit increase in BMI, the odds of having hypertension increase by factor of $\exp(0.164654) \sim 18\%$.

(f) In large samples, the asymptotic sampling distribution of the maximum likelihood estimator is normally distributed around the true parameter value. Assuming this holds here, we can use the standard errors given in the output to form a confidence interval. We are interested in determining whether the effect of BMI on hypertension differs by sex. The relevant parameter is the interaction term, `BMI:SEX`, with coefficient 0.04, which is significantly different from zero, giving evidence that the effect differs by sex.

The asymptotic results can be unreliable for binary linear regression. An alternative would be to use bootstrap to approximate the sampling distribution of the parameter estimate. One difficulty that could be encoutered is that in some bootstrap samples, the maximum likelihood estimator may fail to exist, as in Part(c), where the data become separable.

4. (a) The *complete pooling* model neglects the correlation between repeated measurements on the same individual. The model assumes a common (fixed) intercept and slope,

seen ⇓

5, A

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2).$$

The complete pooling model assumes observations are independent, so that even if the resulting parameter estimates are reasonable, the uncertainty estimates will be too small. Equivalently, complete pooling ignores individual-to-individual variability.

A *no pooling* analysis treats individuals as entirely separate entities, assuming unrelated (fixed) intercepts β_{0j} and slopes β_{1j} for each individual j in the sample,

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma_j^2).$$

This makes it difficult to generalize conclusions to unseen individuals. By fitting independent models, we ignore potentially relevant information - the resulting coefficients will have large standard errors, particularly for individuals for which there are fewer observations.

The *linear mixed model* assumes that the intercept and slope are random variables $\beta_{0j} \sim N(0, \sigma_0^2)$ and $\beta_{1j} \sim N(0, \sigma_1^2)$, and estimates the variances of these distributions. The complete pooling and no pooling models correspond to the limiting cases where the variances are zero or infinite, respectively. The linear mixed model should shrink extreme regression coefficients from individuals with very few observations back to the mean of the coefficient distribution.

sim. seen ⇓

(b) The default method for `lmer` is *Restricted Maximum Likelihood* . Suppose we have a linear mixed model,

2, A

2, B

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\epsilon}, \qquad \boldsymbol{b} \sim N(0, \sigma_b^2 I_m), \quad \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 I_n).$$

To perform 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. $L\boldsymbol{Y} = LZ\boldsymbol{b} + L\boldsymbol{\epsilon}$ is then independent of the fixed effects β, and we can maximize the resulting restricted likelihood for the parameters $(\sigma_b^2, \sigma_\epsilon^2)$, obtaining unbiased estimates of these variances. These estimates can then be plugged in to the likelihood, which can then be maximized over the fixed effects $\boldsymbol{\beta}$.

An advantage of REML is that it produces unbiased estimates of the variance components. A disadvantage is that (unlike maximum likelihood), it does not use all of the information in the data. Another disadvantage is that models fit with REML with different fixed effects structures cannot straightforwardly be compared, because the restricted likelihoods are defined on spaces of different dimensions.

2, B

(c) `fit_mixed1` allows for a random intercept, but fixed slope. The more general model `fit_mixed2` allows random intercept and random slope. Comparing these

2, D

nested models amounts to testing the null hypothesis $\sigma_1^2 = 0$ that the random slope variance is zero. One way to do this is parametric bootstrap: simulate many samples from `fit_mixed1`, fit both models to each simulated data set using maximum likelihood, and obtain the empirical distribution of the improvement in the likelihood ratio. Reject H_0 if the observed likelihood ratio statistic is in the upper tail of the empirical distribution.

2, B

(d) On the average, a boy's height increases by $6.5cm/year$, but there is an individual-to-individual difference in rate of increase of $2.83cm/year$. The standard error of the mean age effect is $0.34cm/year$. The t-value is sufficiently large that the effect is statistically significant. (Note that the relevant test statistic is not exactly t-distributed in this case, but the very large t value in this case means we can comfortably reject the null.)

unseen ⇓

(e) Reading off directly from the output,

2, C

$$\Sigma = \begin{pmatrix} 8.0811^2, & 8.0811 \times 1.6807 \times 0.64 \\ 8.0811 \times 1.6807 \times 0.64 & 1.6807^2 \end{pmatrix}$$

unseen ⇓

(f) The distribution of the scaled residuals appears symmetrical: the median is close to zero and the first and third quartiles are symmetrical. The maximum and minimum are consistent with an approximate normal distribution of residuals (with somewhat heavier tails). No reason to be concerned about goodness of fit from the output presented.

3, D

5. (a)
   - ∗ Assumes solution is well-stirred, i.e. homogenous distribution of micro-organisms.
   - ∗ Poisson distribution reasonable if sufficiently small volumes contain at most one micro-organism and nearby small volumes contain organisms independently.

seen ⇓

3, M

meth seen ⇓

(b) The status of an individual plate $Z_x \sim \text{BERNOULLI}(\pi_x)$, so $\mathrm{E}[Z_x] = \pi_x$.
Since the number of organisms is Poisson distributed,

2, M

$$\mathrm{E}[Z_x] = \pi_x = 1 - \exp(-\rho_x),$$

so

$$\log(1 - \pi_x) = -\rho_x = -\frac{\rho_0}{2^x}.$$

Taking logs again gives

$$\log(-\log(1 - \pi_x)) = \log \rho_0 - x \log 2 = \beta_0 + \beta_1 x.$$

Hence can apply the complementary log-log link function to get a GLM in which the intercept $\beta_0 = \log \rho_0$ can be used to estimate the unknown concentration.

3, M

(c) For a single dilution with n plates, the number of infected plates is $Y \sim \text{BINOMIAL}(n, \pi_0)$, hence (standard result),

$$\widehat{\pi}_0 = \frac{y}{n}.$$

A standard result also gives the variance of this estimator,

$$\mathrm{Var}[\widehat{\pi}_0] = \frac{\pi_0(1 - \pi_0)}{n}.$$

unseen ⇓

4, M

(d) We know $\mathrm{Var}[\widehat{\pi}_0] = \frac{\pi_0(1-\pi_0)}{n}$.
In the intercept-only model `fit0`, we have

$$\log(-\log(1 - \widehat{\pi}_0)) = \widehat{\beta}_0.$$

The delta method gives the approximate variance:

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\beta}_0\right] &\approx \frac{\pi_0(1 - \pi_0)}{n} \left| \frac{d}{d\pi_0} \{\log(-\log(1 - \pi_0))\} \right|^2 \\
&= \frac{\pi_0(1 - \pi_0)}{n} \frac{1}{\log(1 - \pi_0)^2} \left| \frac{d}{d\pi_0} \{\log(1 - \pi_0)\} \right|^2 \\
&= \frac{\pi_0}{(1 - \pi_0)\log(1 - \pi_0)^2 n},
\end{aligned}$$

as required.

4, M

(e) Suppose the initial concentration is such that most plates are infected, so that $\pi_0 \approx 1$. We claim that the Fisher information using just the initial plates is small, and subsequent dilutions, as π_x decreases, contribute more information up to some critical value, which we determine.

The Fisher Information can be approximated by $1/\mathrm{Var}\left[\widehat{\beta}_0\right]$, which is

$$I(\pi) = \frac{(1-\pi)\log(1-\pi)^2}{\pi} \qquad \text{per observation.}$$

Clearly $I(\pi) \to 0$ as $\pi \to 0$ since $\log(1-\pi) \sim \pi$ in this limit.

As $\pi \to 1$, we have $(1-\pi)\log(1-\pi)^2 \to 0$ (e.g. by l'Hopital's rule), so that again $I(\pi) \to 0$ here.

The derivative of the information is (by the product rule)

$$\begin{aligned} I'(\pi) &= -2\log(1-\pi)\frac{1}{\pi}\left(\frac{1}{\pi}-1\right) - \log(1-\pi)^2\frac{1}{\pi^2} \\ &= -\frac{\log(1-\pi)}{\pi^2}\left(2(1-\pi)+\log(1-\pi)\right). \end{aligned}$$

Hence $I(\pi)$ has a single critical point in $(0,1)$ where $\log(1-\pi) = 2(\pi-1)$. In view of the limiting behaviour at 0 and 1, this must be a maximum.

It therefore follows that the information per observation increases from close to zero up to the critical point. The purpose of the serial dilution is to shift the concentration into the (a priori unknown) range where reasonably precise estimates of π_x, and therefore of ρ_0, can be obtained.

4, M

(f) For a linear model, observations should be weighted according to their relative variance. This could be estimated directly from the result in Part(d).

However, this approach is still less efficient than using a generalized linear model. Since the GLM parameters are estimated by maximum likelihood, this model uses all of the information in the data, and produces estimates that have smaller variance than those using the linear model. In the concentration range where the binomial is well-approximated by the normal, the difference in the two methods should be small. However, where the binomial distribution is more skewed, the linear model will lose some information (contained in higher moments of the distribution).

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

MATH70044 Applied Statistical Inference Markers Comments

Question 1	No Comment
Question 2	No Comment
Question 3	Well done overall. Some candidates struggled in finding the Fisher information matrix and part (c) was not always complete.
Question 4	Many candidates skipped several parts. Most mistakes were in part (e) and in particular in the off-diagonal elements of Sigma.
Question 5	Many candidates showed good understanding of the principle of the experiment discussed, and were able to deploy mastery of undergraduate statistics to solve unfamiliar problems. There were a few excellent attempts, showing insight into the principle of maximum likelihood at work in this non-standard example.