

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May-June 2020

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

Statistical Modelling 2

Date: 22nd May 2020

Time: 13.00pm - 15.30pm (BST)

Time Allowed: 2 Hours 30 Minutes

Upload Time Allowed: 30 Minutes

This paper has 5 Questions.

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

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

Only use 1 side of the paper.

Allow margins for marking.

Any required additional material(s) will be provided.

Credit will be given for all questions attempted.

Each question carries equal weight.

SUBMIT YOUR ANSWERS AS SEPARATE PDFs TO THE RELEVANT DROPBOXES ON BLACKBOARD (ONE FOR EACH QUESTION) WITH COMPLETED COVERSHEETS WITH YOUR CID NUMBER, QUESTION NUMBERS ANSWERED AND PAGE NUMBERS PER QUESTION.

Throughout this paper, numerical answers need not be simplified.

1. Consider the linear model with n observations,

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

- (a) Show that the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$X^T(\mathbf{y} - X\boldsymbol{\beta}) = 0,$$

and give a geometric interpretation of this result. (4 marks)

- (b) Derive an expression in terms of X for a matrix P such that the fitted values $\hat{\mathbf{y}} = P\mathbf{y}$ and the residuals $\mathbf{e} = (I_n - P)\mathbf{y}$. (4 marks)

- (c) Explain what is meant by the *leverage* of an observation, and state its relationship to the variance of the corresponding residual. (2 marks)

- (d) In the case of simple linear regression with an intercept, where

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

give the condition for X to have full rank, and interpret this condition in practical terms.

(2 marks)

- (e) In the setting of part (d), show that the leverage of the i th observation can be written as

$$\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2},$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

(5 marks)

- (f) Explain how leverages can be used in model criticism.

(3 marks)

(Total: 20 marks)

2. Consider a Poisson regression model, in which the random variables $Y_i \sim \text{POISSON}(\mu_i)$ are independent, and μ is related to the linear predictor $\eta = X\beta$ by the canonical link function.

(a) Write the Poisson mass function $f_Y(y)$ in exponential family form, identifying the canonical parameter. (2 marks)

(b) Show that the score and Fisher information can be written as

$$U = X^T (\mathbf{y} - \boldsymbol{\mu}), \quad I = X^T W X,$$

respectively, where W is a matrix to be determined.

(6 marks)

(c) Define what is meant by the *deviance* of a generalized linear model and show that in this case, (3 marks)

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

Consider a Poisson GLM for the number of new cases of a disease as a function of time.

$$\mu_i = \exp(\beta_0 + \beta_1 t_i + \beta_2 t_i^2).$$

The R output at the end of the question represents the result of fitting this model to observed data.

(d) Assuming that the model is adequate, carry out a hypothesis test for $\beta_2 = 0$, stating your conclusion in plain language. Describe the conclusions that can be drawn from your preferred model, in the original data context. (4 marks)

(e) State a feature of the output that suggests that the model fit is not adequate. Propose an approximate solution and comment on how this would affect your conclusions in part (d). (3 marks)

(f) Suggest features of the data context that may violate the modelling assumptions employed here. (2 marks)

```

Call:
glm(formula = cases ~ year + I(year^2), family = poisson)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.408883   0.400405   1.021   0.3072
year          0.134293   0.058848   2.282   0.0225 *
I(year^2)    -0.002327   0.001971  -1.181   0.2378
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 127.624  on 24  degrees of freedom
Residual deviance:  94.204  on 22  degrees of freedom
AIC: 176.44

```

(Total: 20 marks)

3. The output below shows the result of fitting a model to the the pulp dataset that was considered in a tutorial class. It consists of 20 observations, balanced between four operators labelled *a* to *d*. Some output has been obscured with ####. For questions that refer to obscured output, you should briefly justify your answers.

Call:

```
lm(formula = bright ~ operator, data = pulp)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	60.2400	0.1458	413.243	<2e-16	***
operatorb	-0.1800	0.2062	-0.873	0.3955	
operatorc	0.3800	0.2062	1.843	0.0839	.
operatord	0.4400	0.2062	2.134	0.0486	*

Residual standard error: 0.326 on ## degrees of freedom

Multiple R-squared: 0.4408, Adjusted R-squared: 0.3359

F-statistic: 4.204 on ## and ## DF, p-value: 0.02261

- (a) Explain what is meant by the *Adjusted R-squared*, and how it can be used. (2 marks)
- (b) Give the forms of the row of the design matrix for an observation from operator *a* and an observation from operator *b*. (2 marks)
- (c) State the null hypothesis and the number of degrees of freedom for the F test given in the output. (3 marks)
- (d) Explain why the three model parameters for operators b, c and d have equal standard errors. (2 marks)

Question continues on the next page

Suppose now we fit the linear mixed model as given in the code below.

```
Linear mixed model fit by REML ['lmerMod']  
Formula: bright ~ 1 + (1 | operator)  
Data: pulp
```

REML criterion at convergence: 18.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.4666	-0.7595	-0.1244	0.6281	1.6012

Random effects:

Groups	Name	Variance	Std.Dev.
operator	(Intercept)	0.06808	0.2609
Residual		#####	#####

Number of obs: 20, groups: operator, 4

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	60.4000	0.1494	404.2

- (e) State, with justification, the value of the residual standard deviation. (2 marks)
- (f) Explaining your reasoning, determine an estimate of the intra-class correlation (you need not simplify your answer). (3 marks)
- (g) Explain the difficulty that arises when using standard asymptotic results for the null distribution of the likelihood ratio test statistic to compare the two models that have been fit, and suggest an alternative approach. (3 marks)
- (h) Suppose that the mixed model here is to be compared with one in which an additional fixed effect is included. Explain how this could be done, noting any changes to the fitting procedure that would be required.

(3 marks)

(Total: 20 marks)

4. (a) Define the three components of a generalized linear model (GLM). (2 marks)
- (b) Suppose that it is desired to estimate the concentration ρ_0 of bacteria per unit volume of a solution, by means of a *dilution assay*. At dilution stage x , a sample of the solution is diluted so that it contains

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

bacteria per unit volume on average.

A unit volume of solution is applied to n plates at each dilution stage. The number of bacteria on a plate at dilution stage x is then Poisson distributed with mean ρ_x . A plate is said to be infected if any bacteria are present.

- (i) Write down in terms of ρ_x the probability π_x that a plate at stage x is infected. (1 mark)
- (ii) Write down the distribution of the number Y_x out of n plates that are infected, stating relevant modelling assumptions. (2 marks)
- (iii) Show that ρ_0 can be estimated using a GLM with the complementary log-log link $g(\pi_x) = \log(-\log(1 - \pi_x))$, where the linear predictor $\eta = \beta_0 + \beta_1 x$ depends on ρ_0 in a way that should be determined. (2 marks)
- (iv) The R output below is from a GLM as described in part (iii). Select values from the output to give a point estimate of ρ_0 , together with an approximate 95% confidence interval. State any assumptions needed. (3 marks)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.7443	0.5207	7.191	6.42e-13 ***
x	-0.8185	0.1044	-7.841	4.47e-15 ***

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 379.3628 on 24 degrees of freedom
 Residual deviance: 5.9539 on 23 degrees of freedom
 AIC: 28.014

Question continues on the next page

- (c) A study is conducted to determine the association between smoking and circulatory disease. The table below shows the number of disease sufferers (D) and non-sufferers (\bar{D}), by smoking status.

	D	\bar{D}	Total
Smoker	18	19	37
Non-smoker	38	175	213
Total	56	194	250

Table 1: Smoking and circulatory disease data.

To analyze this data, a GLM was fit in R as follows. Note that the output has been abridged.

```
fit0<-glm(disease~smoker,family=binomial,data=circ_dat)
summary(fit0)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.5272	0.1790	-8.533	< 2e-16 ***
smoker	####1	####2	3.934	8.35e-05 ***

- State the link function that was used to fit the model. (1 mark)
- Explain how to use the values in the table to determine the intercept reported in the output. (2 marks)
- Write down the log odds ratio Δ , marked ####1 for the effect of smoking, leaving your answer in terms of fractions. (3 marks)
- Give the standard error for the log ratio marked ####2, leaving your answer in terms of fractions. (2 marks)
- State which (if any) of the parameter estimates would change if these data had come from a retrospective rather than a prospective study. State an important assumption needed when interpreting results from a retrospective study. (2 marks)

(Total: 20 marks)

5. Example 2.1.1 of the extract discusses an experiment in which tree seedlings are grown under two different concentration regimes for carbon dioxide. Three trees are assigned to each of the two conditions, and the stomatal area is measured at four random locations on each plant.

Models are fit in R as follows

```
m0 <- lm(area ~ C02, stomata)
m1 <- lm(area ~ C02 + tree, stomata)
anova(m0,m1)
Analysis of Variance Table
Model 1: area ~ C02
Model 2: area ~ C02 + tree
Res.Df  RSS   Df Sum of Sq
1   22  2.1348
2   18  0.8604  4   1.2744
```

- (a) Explain the difficulties in Example 2.1.1 with using a fixed effects model

$$y_i = \alpha_j + \beta_k + \epsilon_i,$$

where observation i is of tree k , exposed to CO_2 level j .

(4 marks).

- (b) Explain why m0 and m1 have 22 and 18 residual degrees of freedom, respectively. (4 marks)
- (c) Write down a numerical expression, in terms of the values in the table, for the F statistic. State the distribution followed by the F statistic under the null hypothesis. (3 marks)
- (d) Describe how least squares is used in the example above to obtain an unbiased estimate of the random effects variance σ_b^2 . (3 marks)

Section 2.2.1 of the extract discusses numerical methods for parameter estimation.

- (e) Explain briefly how Newton's method is used to obtain maximum likelihood estimators. (2 marks)
- (f) Show by means of a sketch in the case of a one-dimensional optimization that it is possible for a Newton step to *decrease* the log likelihood. Explain why a sufficiently small step in the Newton direction must increase the likelihood, so long as the Hessian matrix is negative definite. (3 marks)
- (g) Give an example of a model for which Newton's method would converge after precisely one iteration. Justify your answer briefly. (1 mark)

(Total: 20 marks)

Course: MATH96051/MATH97082
Setter:
Checker:
Editor:
External:
Date: April 18, 2020

MSc EXAMINATIONS (MATHEMATICS)

May 2020

MATH96051/MATH97082

Statistical Modelling II [SOLUTIONS]

Setter's signature

.....

Checker's signature

.....

Editor's signature

.....

1. (a) [Seen] The likelihood is given by

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

Hence the log likelihood is

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

For fixed σ^2 , we can maximize this log likelihood by expanding the expression

$$(\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T X^T \mathbf{y} + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta}$$

(or by the chain rule) and taking the gradient with respect to $\boldsymbol{\beta}$ to see that

$$-2X^T \mathbf{y} + 2X^T X \boldsymbol{\beta} = 0$$

at a stationary point. Since the gram matrix $X^T X$ is positive (semi-) definite, this stationary point is a minimum, yielding a maximum for the original log likelihood.

Geometrically, this says that the vector $\mathbf{e} = \mathbf{y} - X\boldsymbol{\beta}$ of residuals is orthogonal to all columns of the design matrix X . $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$ is the orthogonal projection of \mathbf{y} onto the columns of the design matrix.

- (b) [Seen] Taking

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

so that if X is full rank, we have

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

Then if $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$ we see that $P = X(X^T X)^{-1} X^T$ is the desired matrix.

Now $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (I - P)\mathbf{y}$, so the residuals are given by the projection $I - P$ onto the orthogonal complement of the column space of X .

- (c) [Seen] The leverage h_i of the i th observation is defined to be the i th diagonal entry of the matrix P given in part (b).

Note that $E(\mathbf{e}) = E(I - P)\mathbf{y} = (I - P)E(\mathbf{y}) = (I - P)X\boldsymbol{\beta} = 0$, since $I - P$ maps all columns of X to $\mathbf{0}$.

Hence the variance-covariance matrix of \mathbf{e} is given by $E(\mathbf{e}\mathbf{e}^T) = (I - P)^T E(\mathbf{y}\mathbf{y}^T) (I - P) = (I - P)\sigma^2$, using the fact that $(I - P)^T = (I - P) = (I - P)^2$. Hence the variance of the i th residual is given by $(1 - h_i)\sigma^2$.

- (d) [Seen Similar] X has full rank when the vectors $(1, \dots, 1)^T$ and $(x_1, \dots, x_n)^T$ are not scalar multiples. This is the case when the x_i values are not all equal.

Practically, we cannot learn how y changes when x changes if we have only a single x value.

(e) [Unseen] In this case,

$$X = \begin{pmatrix} \uparrow & \uparrow \\ 1 & x_i \\ \downarrow & \downarrow \end{pmatrix},$$

so that $X^T X = \begin{pmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{pmatrix},$

and

$$(X^T X)^{-1} = \frac{1}{n \sum x_j^2 - (\sum x_j)^2} \begin{pmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{pmatrix} = \frac{1}{n \sum (x_j - \bar{x})^2} \begin{pmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{pmatrix}.$$

Now the i th diagonal entry of P is given by

$$\begin{aligned} \begin{pmatrix} 1 & 0 \end{pmatrix} P \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= \begin{pmatrix} 1 & 0 \end{pmatrix} X (X^T X)^{-1} X^T \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & x_i \end{pmatrix} (X^T X)^{-1} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \\ &= \frac{1}{n \sum (x_j - \bar{x})^2} \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \sum x_j^2 - x_i \sum x_j \\ -\sum x_j + n x_i \end{pmatrix} \\ &= \frac{\sum x_j^2 - 2x_i \sum x_j + n x_i^2}{n \sum (x_j - \bar{x})^2} \\ &= \frac{\sum (x_j - \bar{x})^2 + n \bar{x}^2 - 2n x_i \bar{x} + n x_i^2}{n \sum (x_j - \bar{x})^2} = \frac{\sum (x_j - \bar{x})^2 + n (x_i - \bar{x})^2}{n \sum (x_j - \bar{x})^2} \\ &= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}. \end{aligned}$$

(f) [Seen Similar] Leverage can be used to identify observations that have the potential to have a substantial effect on the fit, in virtue of their position in covariate space. Specifically, it identifies points with a large (Mahalanobis) distance from the centroid of covariate space.

High leverage points with a large residual correspond to a large Cook's distance. This is a measure of how much the predictions from the model would change if a particular observation were omitted.

So a plot of leverage against residual, marked with contours of Cook's distance (which is a function of these), allows points with a substantial effect on the fit to be identified.

2. (a) [Seen]

$$f_Y(y; \lambda) = \exp(-\lambda + y \log \lambda - \log(y!))$$

is in exponential family form with canonical parameter $\theta = \log \lambda$.

(b) [Seen Similar]

with the canonical link $\theta_i = \eta_i = X_i \beta$, the log likelihood takes the form of a sum over contributions from the individual observations

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

so that the j th entry of the gradient is given by

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^n y_i x_{ij} - x_{ij} \exp(\eta_i).$$

Hence the gradient vector is of the form $X^T(\mathbf{y} - \boldsymbol{\mu})$.

We get the (j, k) th entry of the observed information by differentiating the expression above with respect to β_k :

$$\frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_j} = \sum_{i=1}^n -x_{ij} x_{ik} \exp(\eta_i).$$

We see that this is independent of \mathbf{y} , so taking expectation with respect to \mathbf{y} gives

$$I_{jk} = E \left(-\frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_j} \right) = \sum_{i=1}^n x_{ij} x_{ik} w_{ii},$$

where W is a diagonal matrix with entries $w_{ii} = \exp(\eta_i)$. Hence $I = X^T W X$.

(c) [Seen] The deviance of a model is twice the difference between the log likelihood for the saturated model, and the log likelihood evaluated at the MLE of the model parameters.

$$D = 2(l(\mathbf{y}, \mathbf{y}) - l(\mathbf{y}, \hat{\boldsymbol{\mu}})).$$

In the case of the Poisson, this gives

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

(d) [Seen Similar] We have no reason to reject the null hypothesis $\beta_2 = 0$, because the p-value for this test is 0.2378. This suggests that the data are consistent with a model in which $\mu_i = \exp(\beta_0 + \beta_1 t_i)$. In this model, since $\beta_1 > 0$, the spread of the disease is essentially unchecked as a function of time - exponential increase in new cases.

- (e) *[Seen Similar]* Assuming that the sample is large enough that the deviance is roughly $\chi^2(n-p)$, the residual deviance is rather large. This suggests overdispersion - the variance is larger than can be accounted for by the functional dependency on the mean.

Could use a quasi-Poisson model to address this, estimating the dispersion parameter from the data e.g. using the residual deviance $\hat{\phi} = \frac{D}{n-p}$.

This would not change the point estimate $\hat{\beta}$, but the standard errors for the entries of $\hat{\beta}$ would change by a factor of $\sqrt{\hat{\phi}} \approx 2$. Hence the coefficient for year would no longer be significantly different from zero. The quadratic term would remain non-significant.

- (f) *[Unseen]*
- * The data are structured in time and so there may be autocorrelation between observations in successive years.
 - * Poisson-distributed response suggests new cases as independent, rare events. Instead, more likely to see clusters of cases - hence overdispersion.

3. (a) [Seen] The adjusted R^2 is a measure of goodness of fit defined by

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1},$$

where n is the number of observations and p the number of identifiable parameters. It is an attempt to adjust R^2 , which is not comparable between models with different numbers of predictors.

It can be used to compare the goodness of fit of two linear models fit to the same data, with different numbers of variables.

- (b) [Seen Similar] For an observation with operator a, the design matrix row would be (1, 0, 0, 0). For an observation with operator b, the design matrix row would be (1, 1, 0, 0).
- (c) [Seen Similar] $H_0 : \beta_b = \beta_c = \beta_d = 0$. The test statistic has an $F(3, 16)$ distribution under the null hypothesis.
- (d) [Seen Similar] Balanced design, so the matrix $X^T X$ is invariant under permutation of the b, c, d class labels. Hence the standard errors arising from the diagonal entries of the inverse of this matrix must be identical.
- (e) [Seen Similar] Standard error is the same as for the linear model, $\hat{\sigma} = 0.3260$.
- (f) [Seen Similar] The intra-class correlation is given by

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} = \frac{0.06808}{0.06808 + 0.10625}.$$

- (g) [Seen Similar] The asymptotic null distribution of the generalized likelihood ratio test statistic is valid only when entries of the parameter vector are restricted to the interior of the parameter space - here the value under consideration is on the boundary and the asymptotic result no longer applies.

Instead, could use a parametric bootstrap routine to estimate the probability, under the null hypothesis that $\sigma_b^2 = 0$, that the log likelihood ratio test statistic is as large as the one observed.

To do this, generate a large number of independent samples from the null model, and compute the observed value of the test statistic.

- (h) [Seen Similar] Can use a likelihood ratio test, with asymptotic chi square distribution, to compare models on the same data but different fixed effect structures. However, the model given was estimated using REML, and REML likelihoods cannot be compared between models with different fixed effect structures. (REML involves the likelihood of transformed data, and the transformation is a function of the fixed effect design matrix - in general may not even be probability densities on the same dimension.) Hence need to fit with maximum likelihood rather than REML.

4. (a) [Seen]

- * The **random component** specifies the probability distribution of the response variables. Specifically, the components of \mathbf{y} have pdf or pmf from an exponential family of distributions, with $E(\mathbf{Y}) = \boldsymbol{\mu}$.
- * The **systematic component** specifies a linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$ as a function of the covariates and the unknown parameters.

- * The **link function** g may be any monotonic differentiable function. The link function provides a functional relationship between the systematic component and the expectation of the response in the random component; namely $\eta = g(\mu)$. [Seen]

- (b) (i) [Seen Similar] $\pi_x = 1 - \exp(-\rho_x)$.
(ii) [Seen Similar] Assuming that the solution is well-mixed, and growth on different plates proceeds independently,

$$Y_x \sim \text{BINOMIAL}(n, \pi_x).$$

- (iii) [Unseen] By part (i), $\log(1 - \pi_x) = -\rho_x = -\frac{\rho_0}{2^x}$.
Taking logs gives

$$\log(-\log(1 - \pi_x)) = \log \rho_0 - x \log 2$$

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.

- (iv) [Unseen] Assuming that the the number n of plates at each dilution is sufficiently large that maximum likelihood estimators of the β parameters can be taken to be normally distributed, an approximate 95% confidence interval for $\log \rho_0$ is given by $3.7443 \pm 1.96 \times 0.5207$. Then a 95% confidence interval for ρ_0 is given by $(\exp(3.7443 - 1.96 \times 0.5207), \exp(3.7443 + 1.96 \times 0.5207))$.
(c) (i) [Seen] Since the no link function is specified, R uses the canonical link. For the binomial family, this is the logit link.
(ii) [Seen Similar] Intercept is the log odds for non-smokers

$$\log\left(\frac{38}{175}\right) = (-1.5272).$$

- (iii) [Seen Similar]

$$\Delta = \log\left(\frac{38}{175}\right) - \log\left(\frac{18}{19}\right) (= 1.4731)$$

- (iv) [Seen Similar]

$$SE(\Delta) = \sqrt{\frac{1}{18} + \frac{1}{19} + \frac{1}{38} + \frac{1}{175}}$$

- (v) [Seen Similar] The intercept will change by an additive factor relating to the relative probability of being sampled within the two disease conditions. The coefficient for smoking will be unchanged. An important assumption is that the sampling probability depends only on disease status, and not on the covariate (in this case, smoking status).

5. (a) * Since trees are nested within treatment, the α and β parameters are not identifiable: the individual tree effects and the treatment effects are confounded.
* The model assumes that trees are wholly unrelated - we cannot use the result obtained to generalize to the population of trees beyond the six that have been studied.

- (b) There are $4 \times 3 \times 2 = 24$ observations.

The design matrix for \mathbf{m}_0 contains an intercept column and a column indicating which observations are in the second CO_2 condition. These two columns are linearly independent, giving a design matrix of rank 2, so there are $24 - 2 = 22$ residual degrees of freedom.

The design matrix for \mathbf{m}_1 contains an intercept column and a column indicating which observations are in the second CO_2 condition. It contains five further columns specifying (by means of a binary indicator) which observations come from tree $k = 2, 3, 4, 5, 6$, however, the sum of the $k = 4, 5, 6$ columns is equal to the column indicating the second CO_2 condition. Hence the rank of the design matrix is $1 + 1 + 5 - 1 = 6$ and so there are $24 - 6 = 18$ residual degrees of freedom.

- (c)

$$F = \frac{(2.1348 - 0.8604)/4}{0.8604/18} \quad (= 6.665).$$

Under H_0 , this statistic has the $F(4, 18)$ distribution.

- (d) Suppose that we use a random effect to model differences between trees,

$$y_i = \alpha_j + b_k + \epsilon_i,$$

where $b_k \sim N(0, \sigma_b^2)$ and $\epsilon_i \sim N(0, \sigma^2)$. We can obtain an unbiased estimate of σ^2 from least squares in the usual way.

For balanced data, averaging over levels of a random effect yields a simplified model. Consider sample average for each tree, this is just

$$\bar{y}_k = \alpha_j + e_k,$$

where tree k is in condition j , and the errors $e_k \sim N(0, \sigma_b^2 + \sigma^2/4)$ are independent.

Least squares can now be used to obtain an unbiased estimate of $\sigma_b^2 + \sigma^2/4$.

We can now obtain an unbiased estimate of σ_b^2 by combining the two variance estimates.

- (e) Suppose we have an initial value θ_0 , close to the optimum. The method approximates the log likelihood by a quadratic function of the parameters θ (Taylor expansion to second order), and seeks the value of θ that maximizes the quadratic approximant. This can be determined analytically to be

$$\theta_1 = \theta_0 - (\nabla^2 l)^{-1} \nabla l.$$

This procedure is iterated using θ_1 as an initial value, and in regular circumstances, will converge to a local maximizer of the log likelihood.

- (f) Should see sketch similar to fig 2.5 in the reading material. If the inverse of the Hessian is a negative definite matrix Q , then by Taylor expansion of l to first order, for sufficiently small $\alpha > 0$,

$$l(\theta_0 - \alpha Q \nabla l) \approx l(\theta_0) - \alpha (\nabla l)^T Q (\nabla l) > l(\theta_0),$$

since $(\nabla l)^T Q (\nabla l) < 0$.

- (g) For a Normal linear model, the log likelihood is a quadratic function of its parameters, so the Newton step performs exact maximization in a single step.