Imperial College
London

M4/5S2

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2018

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

Statistical Modelling II

Date: Friday, 25 May 2018

Time: 2:00 PM - 4:30 PM

Time Allowed: 2.5 hours

This paper has 5 questions.

Candidates should use ONE main answer book.

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

All required additional material will be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Each question carries equal weight.
- Calculators may not be used.

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

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

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 $\widehat{\beta}$.

(b) Give 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.

(c) Show that $I - P$ is also a projection matrix.

(d) Show that the vector of residuals, e, can be written as $(I_n - P)\boldsymbol{y}$, and find the expectation and variance-covariance matrix of e.

(e) Stating clearly any properties of projection matrices that you use, explain why

$$\frac{e^T e}{n - p}$$

is an unbiased estimator of σ^2.

In an experimental study, the model above is fitted to some data. Figure 1 shows leverage against standardized residuals for this model.

(f) State, with brief justification, which of the two labelled points A and B in Figure 1:

(i) has a residual with a higher variance.

(ii) would result in a greater change in the estimates of the model coefficients if removed.

(g) State one property of the unlabelled residuals in Figure 1 that indicates a poor model fit. Suggest another plot that could be used to confirm this.

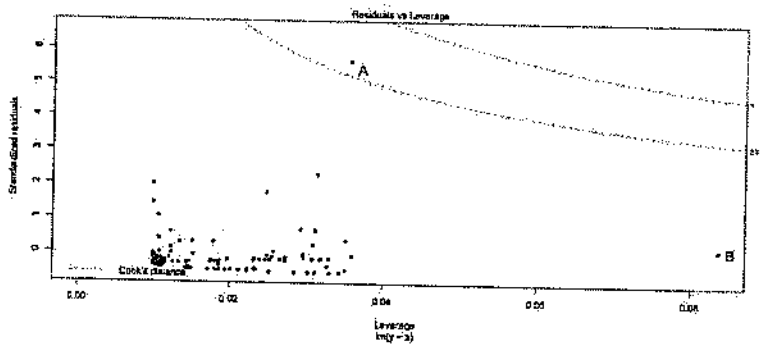

Figure 1: Standardized residuals plotted against leverage for the Normal linear model.

2. (a) Explain what is meant by

   (i) A generalized linear model.

   (ii) The canonical link function for a generalized linear model.

   (iii) A saturated model.

   (b) Explain why a saturated model would not usually be fit to data.

The random variable Y has a density function written in exponential family form as follows

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right).$$

(c) Show that $\mathrm{E}(Y) = b'(\theta)$.

(d) Determine an expression for $\mathrm{Var}(Y)$ in terms of θ and ϕ.

(e) The inverse Gaussian distribution has density function

$$f(y;\lambda,\mu) = \left[\frac{\lambda}{2\pi y^3}\right]^{\frac{1}{2}} \exp\left\{\frac{-\lambda(y-\mu)^2}{2\mu^2 y}\right\}, \quad \lambda,\mu,y > 0.$$

Show that the inverse Gaussian density can be written in exponential family form, stating clearly the forms of θ, $b(\theta)$ and ϕ in terms of λ and μ.

(f) State the canonical link for the inverse Gaussian distribution.

(g) Derive the form of the deviance for this model, in terms of $\widehat{\mu}$.

(h) Suppose that an inverse Gaussian GLM is fitted to a large random sample of size n, with linear predictor $\eta = X\beta$, where β is of length p. Suggest a test statistic that can be used to carry out a test of the hypothesis that

$$\beta_{q+1} = \ldots = \beta_p = 0,$$

and state the approximate distribution it follows under the null hypothesis. (Relevant asymptotic results may be assumed to hold.)

3. A Normal linear mixed model is written as

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

where $\boldsymbol{\nu} \sim N(0, \sigma_\nu^2 I_m)$ is a random vector of length m, $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 I_n)$ is a vector of length n, independent of $\boldsymbol{\nu}$, X is the design matrix for the fixed effects, and Z is the model matrix for the random effects. Assume that the design matrix has full rank, p.

(a) Determine the distribution of the random vector $\boldsymbol{Y}$.

(b) Calculate the intra-class correlation coefficient for this model.

(c) Show that left multiplying $\boldsymbol{Y}$ by the matrix $L^T = I_n - X(X^TX)^{-1}X^T$ gives a quantity that does not depend on the fixed effect coefficients β, and explain the relevance of this construction to the estimation of σ_ν^2.

(d) If a different matrix P, such that $\text{rank}(P) = n - p$ and $P^TX = 0$, were instead applied to $\boldsymbol{Y}$, what would be the effect on the estimation?

The code below is used to test the null hypothesis that the random effects variance $\sigma_\nu^2 = 0$.

```
> mylm <- lm(response~1,data=dat)
> mylme2 <- lmer(response~1+(1|group),data=dat,REML=FALSE)
> d<-as.numeric(2*(logLik(mylme2)-logLik(mylm)))
> d
[1] 2.54379

> ds <- numeric(1000)
> for(i in 1:1000){
+     y <- unlist(simulate(mylm))
+     nullmod <- lm(y~1)
+     altmod <- lmer(y~1+(1|group),data=dat,REM=FALSE)
+     ds[i] <- as.numeric(2*(logLik(altmod)-logLik(nullmod)))
}
> mean(ds < 0.00001)
[1] 0.81
mean(ds>d)
[1] 0.01
```

(e) Explain briefly what the code does. (Your response should address the statistical purpose of the code, rather than simply describing it line-by-line.)

(f) Use the output given to explain how the usual asymptotic result for the generalized likelihood ratio test can be seen to be invalid here. State the assumption that is violated.

(g) State with reasons which of the models is to be preferred.

(h) Describe how the sampling uncertainty in the p-value above could be estimated.

4. The R output below concerns a binomial GLM, $Y_i \sim \text{BINOMIAL}(m_i, \pi_i)$ for $i = 1, \ldots n$, where π is related to covariates X and Z. Y_i is the number of people who suffer from a disease. Throughout, m_i may be assumed to be large.

```
Call:
glm(formula = y ~ x + z, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2999  -1.6093   0.0299   1.8933   4.8947

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.23923    0.06404   3.736 0.000187 ***
x            0.38472    0.08242   4.668 3.05e-06 ***
z            0.44893    0.05360   8.376  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 643.64  on 99  degrees of freedom
Residual deviance: 541.87  on 97  degrees of freedom
AIC: 1016.4

Number of Fisher Scoring iterations: 4
```

(a) State the link function used for this model, and one desirable property that results from this choice of link.

(b) Explain the effect of a one unit change in x on the odds of having the disease.

(c) Stating any distributional assumptions necessary, and taking the fitted model to be adequate, give an approximate 95% confidence interval for the coefficient of x in the model. You need not simplify your answer.

(d) Use appropriate values from the output to comment on the goodness of fit. Justify briefly any approximations you use.

(e) Explain what is meant by the AIC, and how it could be used to distinguish between the model here and a model that excluded the variable Z, with AIC 1015.2.

(f) State whether (and if so, how) the estimates of the parameters, and their standard errors, would change if a plug-in estimator of the dispersion were used. State the distribution that would be used to compute the probabilities in the fourth column in this case.

[CONTINUED]

(g) Clustering is a common source of overdispersion. Suppose the sample is modelled as consisting of k clusters, each of size $\frac{m}{k}$, with

$$Y_i = \sum_{j=1}^{\frac{m}{k}} Z_{ij},$$

where the variables $Z_{ij}|p_{ij} \sim \text{BINOMIAL}(k, p_{ij})$, independently, and the probabilities p_{ij} are such that $\mathrm{E}(p_{ij}) = \pi_i$ and $\mathrm{Var}(p_{ij}) = \tau^2\pi_i(1-\pi_i)$.

Show that in this case, $\mathrm{E}(Y_i) = m\pi_i$ and $\mathrm{Var}(Y_i) = \phi V(\pi_i)$, where V is the variance function for the generalized linear model and ϕ depends on τ and k in a way that you should determine.

5. (a) Briefly explain why feature selection is desirable when fitting a linear model with a large number of predictors.

(b) Explain why R^2 is not a suitable goodness of fit measure for selecting amongst models with different numbers of predictors.

(c) Explain why ridge regression is only applied after standardizing predictor variables.

(d) For the model

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

where $\epsilon \sim N(0, \sigma^2 I_n)$, write down the log likelihood function and show that if independent Normal priors with zero mean and variance τ^2 are assumed for the parameters, the estimates that maximize the resulting posterior distribution are identical to the parameter estimates from ridge regression.

(e) Show that the ridge regression estimators $\hat{\beta}_r$ satisfy $\hat{\beta}_r = (X^T X + \lambda I_n)^{-1} X^T y$, where λ is the tuning parameter.

(f) Consider a lab that is trying to design a clinical screen for diabetes. A large number of blood markers are included as predictors in a Normal linear model in which the response variable is blood glucose level. The clinicians' understanding is that the level of glucose is strongly determined by a small number of markers, but they have no knowledge of which markers are likely to be involved. Explain whether ridge regression or the lasso is likely to perform better in this context.

(g) Consider the problem of selecting from amongst two candidate models. Model A has 3 parameters and model B has 4 parameters. Two statisticians propose different approaches. The first statistician chooses the simpler model unless the test statistic

$$2(l(B) - l(A))$$

is greater than the 95th percentage point of the distribution of a $\chi^2(1)$ variable. The second statistician always chooses the model with the smaller AIC. Determine which statistician has the lower type I error rate.

Note that $\Pr(\chi^2(1) > 3.84) \approx 0.95$.

Imperial College
London

Course: M3S2/M4S2
Setter:
Checker:
Editor:
External:
Date: April 24, 2018

MSc EXAMINATIONS (MATHEMATICS)

May 2018

M3S2/M4S2

Statistical Modelling II [SOLUTIONS]

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

1. (a) [1 mark] $\widehat{\beta}$ satisfies

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

or, since X has full rank,

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

[Seen]

(b) [4 marks] The fitted values are given by $\widehat{y} = X\widehat{\beta} = X(X^T X)^{-1} X^T y$, so the matrix $P = X(X^T X)^{-1} X^T$.

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

It is symmetric, since $P^T = (X(X^T X)^{-1} X^T)^T = (X^T)^T((X^T X)^{-1})^T X^T = X(X^T X)^{-1} X^T$, 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) [2 marks] Let $Q = I - P$. We show it 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 $Q^T Q = (I_n - P)^T (I_n - P) = I_n - 2P + \underbrace{P^T P}_{:=P} = I_n - P = Q$. *[Seen]*

(d) [3 marks] The residuals are given by $e = y - \widehat{y} = y - Py = (I_n - P)y$.
To compute the mean of the residuals,

$$\mathrm{E}(e) = \mathrm{E}((I_n - P)y) = (I_n - P)\mathrm{E}(y) = X\beta - PX\beta = X\beta - X(X^T X)^{-1} X^T X\beta = X\beta - X I_p \beta = 0.$$

To compute the variance of the residuals,

$$\begin{aligned}\mathrm{Cov}(e) &= \mathrm{Cov}((I_n - P)y) = (I_n - P)\mathrm{Cov}(y)(I_n - P)^T = (I_n - P)\sigma^2 I_n (I_n - P)^T \\ &= \sigma^2 (I_n - P)(I_n - P)^T = \sigma^2 (I_n - P),\end{aligned}$$

using that $I_n - P$ is both symmetric and idempotent. *[Seen]*

(e) [4 marks]

$$\begin{aligned}
\mathrm{E}(e^Te) &= \mathrm{E}(Y^T(I_n-P)^T(I_n-P)Y) = \mathrm{E}(Y^T\underbrace{(I_n-P)}_{=Q}Y) \\
&= \mathrm{E}\left(\sum_i\sum_j Y_iY_jQ_{ij}\right) = \mathrm{E}(\mathrm{trace}(QYY^T)) \\
&= \mathrm{trace}(Q\mathrm{E}(YY^T)) = \mathrm{trace}(Q\left\{\mathrm{Cov}(Y)+\mathrm{E}(Y)\mathrm{E}(Y)^T\right\}) \\
&= \mathrm{trace}(Q\mathrm{Cov}(Y)) + \mathrm{trace}(QE(Y)\mathrm{E}(Y)^T) \\
&= \mathrm{trace}(Q\sigma^2) + \mathrm{trace}(\underbrace{QX\beta}_{=0}(X\beta)^T) \\
&= \sigma^2\mathrm{trace}(I_n-P) \\
&= \sigma^2(n-p). \quad \text{as } \mathrm{trace}(P)=\mathrm{rank}(P)=\mathrm{rank}(X)=p.
\end{aligned}$$

[Seen]

(f) [2 marks] Standardized residuals are asymmetrically distributed about zero, but they should be approximately standard normal, and so symmetric. A normal Q-Q plot would confirm this. *[Unseen]*

(g) (i) [2 marks] $\mathrm{Var}(e_i)=(1-P_{ii})\sigma^2$, where P_{ii} is the leverage. High leverage corresponds to small variance, so point B has smallest variance. *[Unseen]*

(ii) [2 marks] Cook's distance gives change in the fit as a result of removing a particular observation. Since observation A has a higher Cook's distance, removing it will cause the larger difference in fit. *[Seen Similar]*

2. (a) (i) [3 marks] A generalized linear model is defined by

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

(ii) [1 mark] If the density of Y is written in exponential family form,

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

then the canonical link is defined by the relation

$$\theta = \eta. [Seen]$$

(iii) [1 mark] A model is said to be saturated if there as many independent parameters as observations. *[Seen]*

(b) [1 mark] A saturated model involves estimating n parameters from n data points. Typically, this means that the model can fit the data perfectly. The act of fitting a statistical model presupposes that the process under consideration is subject to random fluctuations, which we cannot expect to model. A saturated model is therefore fit to noise as well as systematic variation of interest, and inferences and predictions made from it will, on average, be subject to larger errors than a model with fewer parameters. *[Seen]*

(c) [3 marks]

Note first, that since f is a density function,

$$\int f(y;\theta,\phi)dy = 1.$$

Hence,

$$\frac{\partial}{\partial\theta}\int f(y;\theta,\phi)dy = \frac{\partial}{\partial\theta}1 = 0.$$

Differentiating the density function, we see that

$$\frac{\partial}{\partial\theta}f(y;\theta,\phi) = \frac{y-b'(\theta)}{a(\phi)}f(y;\theta,\phi).$$

Assuming sufficient regularity that integration and partial diferentiation with respect to θ can be interchanged, this gives

$$\int\left(\frac{y-b'(\theta)}{a(\phi)}\right)f(y;\theta,\phi)dy = 0,$$

which simplifies to

$$\frac{1}{a(\phi)}\left(\mathrm{E}(Y)-b'(\theta)\right) = 0,$$

by definition of the expected value. Thus

$$\mu \equiv \mathrm{E}(Y) = b'(\theta). \tag{1}$$

[Seen]

(d) [2 marks] Differentiating again with respect to θ,

$$\frac{\partial^2}{\partial\theta^2}f(y;\theta,\phi) = -\frac{b''(\theta)}{a(\phi)}f(y;\theta,\phi) + \left(\frac{y-b'(\theta)}{a(\phi)}\right)^2 f(y;\theta,\phi).$$

Integrating over y then gives

$$\begin{aligned}0 = \int\frac{\partial^2}{\partial\theta^2}f(y;\theta,\phi)dy &= -\frac{b''(\theta)}{a(\phi)} + \int\left(\frac{y-b'(\theta)}{a(\phi)}\right)^2 f(y;\theta,\phi)dy \\ &= -\frac{b''(\theta)}{a(\phi)} + \frac{\mathrm{Var}(Y)}{a(\phi)^2}.\end{aligned}$$

Rearranging yields

$$\mathrm{Var}(Y) = b''(\theta)a(\phi).$$

[Seen]

(e) [4 marks] The density function is

$$f(y;\lambda,\mu) = \left[\frac{\lambda}{2\pi y^3}\right]^{\frac{1}{2}} \exp\left\{\frac{-\lambda(y-\mu)^2}{2\mu^2 y}\right\}, \quad \lambda,\mu,y>0.$$

Putting this all inside an exponential gives

$$\begin{aligned} f(y;\lambda,\mu) &= \exp\left\{\frac{1}{2}(\log\lambda - \log 2\pi) - \frac{3}{2}\log y - \frac{\lambda(y-\mu)^2}{2\mu^2 y}\right\} \\ &= \exp\left\{\frac{1}{2}(\log\lambda - \log 2\pi) - \frac{3}{2}\log y - \frac{\lambda(y^2-2\mu y+\mu^2)}{2\mu^2 y}\right\} \\ &= \exp\left\{\frac{1}{2}(\log\lambda - \log 2\pi) - \frac{3}{2}\log y - \frac{\lambda y}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2y}\right\} \end{aligned}$$

By considering the term in $\frac{1}{y}$, this is clearly an exponential family with $\phi \propto \frac{2}{\lambda}$. Taking the constant of proportionality as 1 (we only have uniqueness up to a constant), using the form of the exponential family gives:

$$\begin{aligned} \theta &= -\frac{1}{\mu^2} \\ b(\theta) &= -\frac{2}{\mu} = -2\sqrt{-\theta}. \end{aligned}$$

[Unseen]

(f) [1 mark] For the canonical link, we need $\eta = \theta = -\frac{1}{\mu^2}$. *[Seen Similar]*

(g) [2 marks]

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

The log likelihood for a single observation is

$$\frac{1}{2}(\log\lambda - \log 2\pi) - \frac{3}{2}\log y - \frac{\lambda(y-\mu)^2}{2\mu^2 y}$$

Note first that terms in the log likelihood involving y and λ alone cancel. Moreover, the remaining term for the saturated model is clearly zero, since it is a function of $(y_i - \mu_i)$, and for the saturated model $\widehat{\mu}_i = y_i$. Hence, since $\phi = \frac{\lambda}{2}$ here, the deviance is

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

[Seen Similar]

(h) [2 marks] Under the null hypothesis we have (approximately, for large samples) that the scaled deviances $D^* = D/\phi$ are χ^2:

$$D_1^* \sim \chi^2_{n-p} \quad \text{and} \quad D_0^* - D_1^* \sim \chi^2_{p-q}.$$

If we consider D_1^* and $D_0^* - D_1^*$ as asymptotically independent, then

$$\frac{(D_0^* - D_1^*)/(p-q)}{D_1^*/(n-p)} \sim F_{(p-q),(n-p)}.$$

We can multiply top and bottom by ϕ to get a test statistic based on the deviance:

$$\frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} \sim F_{(p-q),(n-p)}.$$

[Seen Similar]

3. (a) [3 marks] As a linear combination of multivariate normal variables, Y must be multivariate normal. It suffices to compute its mean vector and variance-covariance matrix.
For the mean

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

And for the variance-covariance matrix,

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{Y}) &= \mathrm{Cov}(X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon}) \\ &= Z\mathrm{Cov}(\boldsymbol{\nu})Z^T + \mathrm{Cov}(\boldsymbol{\epsilon}) \quad \text{(since } \boldsymbol{\epsilon}, \boldsymbol{\nu} \text{ indep.)} \\ &= ZI_m\sigma_\nu^2 Z^T + I_n\sigma_\epsilon^2 \\ &= \sigma_\epsilon^2\left(ZI_m\frac{\sigma_\nu^2}{\sigma_\epsilon^2}Z^T + I_n\right) \\ &= \sigma_\epsilon^2\left(I_n + Z\Psi Z^T + I_m\right),\end{aligned}$$

where $\Psi = \frac{\sigma_\nu^2}{\sigma_\epsilon^2}I_n$. *[Seen]*

(b) [3 marks]
The correlation for observations in the same group is:

$$\mathrm{corr}(Y_{1,j}, Y_{2,j}) = \frac{\mathrm{E}\left[\left(Y_{1,j} - \mathrm{E}(Y_{1,j})\right)\left(Y_{2,j} - \mathrm{E}(Y_{2,j})\right)\right]}{\sqrt{\mathrm{Var}(Y_{1,j})\mathrm{Var}(Y_{2,j})}}$$

We have that $\mathrm{Var}(Y_{1,j}) = \mathrm{Var}(Y_{2,j}) = \sigma_\epsilon^2 + \sigma_\nu^2$, and thus the denominator is $\sigma_\epsilon^2 + \sigma_\nu^2$. Next, we have $Y_{i,j} - \mathrm{E}(Y_{i,j}) = \nu_j + \epsilon_{ij}$, by noting that $\mathrm{E}(Y_{i,j}) = \mu$ and rearranging the model equation. Therefore the numerator is

$$\begin{aligned}\mathrm{E}\left[\left(Y_{1,j} - \mathrm{E}(Y_{1,j})\right)\left(Y_{2,j} - \mathrm{E}(Y_{2,j})\right)\right] &= \mathrm{E}\left[\left(\nu_j + \epsilon_{1,j})\right)\left(\nu_j + \epsilon_{2,j})\right)\right] \\ &= \mathrm{E}(\nu_j^2) + \mathrm{E}(\nu_j(\epsilon_{1,j} + \epsilon_{2,j})) + \mathrm{E}(\epsilon_{1,j}\epsilon_{2,j}) \\ &= \mathrm{E}(\nu_j^2) = \sigma_\nu^2,\end{aligned}$$

where we used the independence between the ν_j and ϵ_{ij}. Plugging in these results gives the intraclass correlation coefficient as

$$\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\epsilon^2},$$

[Seen]

(c) [3 marks] Let $L^T = I_n - X(X^TX)^{-1}X^T$, then

$$\begin{aligned} L^TY &= L^TX\beta + L^TZ\nu + L^T\epsilon \\ \left(I_n - X(X^TX)^{-1}X^T\right)Y &= X\beta - X(X^TX)^{-1}(X^TX)\beta + L^TZ\nu + L^T\epsilon \\ Y - X\widehat{\beta} &= L^TZ\nu + L^T\epsilon, \end{aligned}$$

so we see that the term in $X\beta$ cancels. Now let $B = L^TY$, then

$$B \sim N(0, \sigma_\epsilon^2 L^TVL),$$

where $V = I_n + \frac{\sigma_\nu^2}{\sigma_\epsilon^2}ZZ^T$.

Then the joint pdf of B (note that strictly this is an improper density on the $n-p$ dimensional subspace orthogonal to the columns of X) is

$$\frac{1}{(2\pi)^{(n-p)/2}\left|\sigma_\epsilon^2 L^TVL\right|^{1/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}b^T(L^TVL)^{-1}b\right\}$$

so that the log-likelihood of the variance components vector $\tau = (\sigma_\epsilon^2, \sigma_\nu^2)$ is

$$\ell_R(\tau; b) = -\frac{(n-p)}{2}\log(2\pi) - \frac{1}{2}\log\left|\sigma_\epsilon^2 L^TVL\right| - \frac{1}{2\sigma_\epsilon^2}b^T(L^TVL)^{-1}b.$$

The REML (restricted maximum likelihood) estimator of τ is defined as the maximiser of ℓ_R. This maximiser can now be found by solving the REML equation

$$\frac{\partial \ell_R}{\partial \tau} = 0.$$

[Seen]

(d) [2 marks] To show that the REML estimators are invariant to the choice of L, we show that $\ell(\tau; y)$ is invariant to the choice of L. To show this, suppose that we have another L_2 such that $\text{rank}(L_2) = n-p$ and $L_2^TX = 0$. Then, there exists a non-singular matrix M such that $L^T = ML_2^T$. Further

$$\left|L^TVL\right| = \left|ML_2^TVL_2M^T\right| = |M|^2\left|L_2^TVL_2\right|$$

and

$$\begin{aligned} b^T(L^TVL)^{-1}b &= y^TL(L^TVL)^{-1}L^Ty \\ &= y^TL_2M^T(ML_2^TVL_2M^T)^{-1}ML_2^Ty \\ &= y^TL_2M^T(M^T)^{-1}(L_2^TVL_2)^{-1}M^{-1}ML_2^Ty \\ &= y^TL_2(L_2^TVL_2)^{-1}L_2^Ty. \end{aligned}$$

Substituting these into the restricted log-likelihood, we get

$$\ell(\tau; y) = -\frac{(n-p)}{2}\log(2\pi) - \log|M| - \frac{1}{2}\log\left|\sigma_\epsilon^2 L_2^TVL_2\right| - \frac{1}{2\sigma_\epsilon^2}y^TL_2(L_2^TVL_2)^{-1}L_2^Ty.$$

[Seen Similar]

(e) [2 marks] The code is a parametric bootstrap routine to estimate the probability, under the null hypothesis that $\sigma_\nu^2 = 0$, that the log likelihood ratio test statistic is as large as the one observed. It achieves this by generating 1000 independent samples from the null model, and computing the observed value of the test statistic. *[Seen Similar]*

(f) [3 marks] The output shows the distribution of the test statistic is concentrated near zero. This is not consistent with the usual asymptotic result that the GLRT statistic has an approximate χ^2 distribution. To derive the asymptotic χ^2 distribution, we assume the parameter lies in the interior of the parameter space. But since a variance cannot be negative, $\sigma_\nu^2 = 0$ is on the boundary, and so the assumption does not hold. *[Seen Similar]*

(g) [2 marks] The improvement in likelihood is larger than would be expected due to chance, so prefer the model with $\sigma_\nu^2 > 0$. *[Seen Similar]*

(h) [2 marks] Standard error for the binomial proportion given by

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{0.01 \times 0.99/1000}$$

(negligible here). *[Unseen]*

4. (a) [3 marks] Since the link function is not explicitly specified in the output, it must be the canonical link. For the binomial family this is the logit link, $\eta = \log \frac{\pi}{1-\pi}$.

With this choice of link function, the fitted values are normalized to lie in an appropriate range for probabilities. Also, the effects of predictors are interpretable on the odds scale. Moreover, since the link is canonical, the observed information is the same as the expected information. *[Seen]*

(b) [2 marks] With the logit link,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=1}^{p} x_{ij}\beta_j.$$

so on the odds scale, the predictors act multiplicatively. Hence a one unit increase in x leads to the odds of having the disease being multiplied by $\exp(\widehat{\beta_x}) = \exp(0.38)$. *[Seen Similar]*

(c) [2 marks] Assuming that the estimator for the parameter is normally distributed, a 95% confidence interval is given by $0.38 \pm 1.96 \times 0.08$. *[Seen Similar]*

(d) [3 marks] When m (number of binomial samples per observation) is large, the scaled deviance should be be roughly $\chi^2(n-p)$, where n is the number of observations and p is the number of parameters. But this means that $\frac{D}{n-p}$ should estimate ϕ, the dispersion parameter. But this is just 1 for the binomial model, which does not match up well with the value $542/97$ from the table. *[Seen Similar]*

(e) [2 marks] AIC is given by $-2l(\widehat{\beta}) + 2p$. It is an attempt to penalize models that contain too many parameters - we select the model with the lower AIC. Therefore, we prefer the model without Z. *[Seen Similar]*

(f) [3 marks] The estimates themselves would not change - these are independent of ϕ. However, the covariance matrix of $\widehat{\beta}$ is $\phi(X^TWX)^{-1}$, where W is the matrix of weights from iterated weighted least squares. The standard errors are the square roots of the diagonal entries of the covariance matrix, so the standard errors would be multiplied by a factor of $\sqrt{\phi}$. The t distibution would be used, since the dispersion parameter is estimated from data. *[Seen Similar]*

(g) [5 marks] First note that

$$\mathrm{E}(Y_i) = \mathrm{E}_{\pi_i}\mathrm{E}(Y_i|\pi_i) = \mathrm{E}\left(\sum_{j=1}^{\frac{m}{k}} kp_{ij}\right) = \frac{m}{k} \times k\pi_i = m\pi_i.$$

Use law of total variance

$$\begin{aligned}\mathrm{Var}(Y_i) &= \mathrm{E}\left(\sum_{j=1}^{\frac{m}{k}} \mathrm{Var}(Z_{ij})|p_{ij}\right) + \mathrm{Var}\left(\sum_{j=1}^{\frac{m}{k}} \mathrm{E}(Z_{ij})|p_{ij}\right) \\ &= \mathrm{E}\left(\sum_{j=1}^{\frac{m}{k}} kp_{ij}(1-p_{ij})|p_{ij}\right) + \mathrm{Var}\left(\sum_{j=1}^{\frac{m}{k}} \mathrm{E}(Z_{ij})|p_{ij}\right)\end{aligned}$$

The first term is

$$\begin{aligned}E\left(\sum_{j=1}^{\frac{m}{k}} \mathrm{Var}(Z_{ij})|p_{ij}\right) &= \mathrm{E}\left(\sum_{j=1}^{\frac{m}{k}} kp_{ij}(1-p_{ij})\right) = \mathrm{E}\left(\sum_{j=1}^{\frac{m}{k}} kp_{ij}\right) - \mathrm{E}\left(\sum_{j=1}^{\frac{m}{k}} kp_{ij}^2\right). \\ &= m\pi_i - \sum_{j=1}^{\frac{m}{k}} k\mathrm{Var}(p_{ij}) - k\mathrm{E}(p_{ij})^2 \\ &= m\pi_i - m\tau\pi_i(1-\pi_i) - m\pi_i^2 = m\pi_i(1-\pi_i) - m\tau\pi_i(1-\pi_i).\end{aligned}$$

The second term is

$$\mathrm{Var}\left(\sum_{j=1}^{\frac{m}{k}} \mathrm{E}(Z_{ij})|p_{ij}\right) = \sum_{j=1}^{\frac{m}{k}} \mathrm{Var}(kp_{ij}) = \frac{m}{k} \times k^2\tau\pi_i(1-\pi_i) = mk\tau\pi_i(1-\pi_i).$$

Combining these two gives

$$\mathrm{Var}(Y_i) = m\pi_i(1-\pi_i)\left(1 + (k-1)\tau^2\right).$$

so define $\phi = 1 + (k-1)\tau^2$. *[Unseen]*

5. (a) [2 marks] By choosing a subset of predictors, we reduce the tendency of least squares to overfitting. For a model with $p \approx n$, i.e. number of predictors comparable to number of observations, the parameter estimates will have high variance. *in text*

(b) [3 marks] When a predictor is added, the residual sum of squares must decrease - even an irrelevant predictor explains some variability. Since $R^2 = 1 - \frac{RSS}{TSS}$, R^2 necessarily increases when a new predictor is added. *[Seen]*

(c) [3 marks] As stated in the text, ridge regression is not invariant under scale transformation - changing the scale of the predictors would lead to very different parameter estimates, as it is the total l^2 norm of the parameter estimates that it penalized. *in text*

(d) [4 marks] The log likelihood function is given by

$$-\frac{1}{2\sigma^2}(\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta}).$$

Independent normal priors on $\beta_1, \ldots \beta_p$ give rise to a prior on the log scale

$$-\frac{1}{2\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

Bayes' theorem then gives the log posterior as

$$-\frac{1}{2\sigma^2}(\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta}) - \frac{1}{2\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta} + K,$$

where K is the log of a normalizing constant that does not depend on $\boldsymbol{\beta}$.
Maximizing this function is clearly equivalent to minimizing the ridge regression criterion

$$(\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta},$$

where $\lambda = \frac{\sigma^2}{\tau^2}$.

(e) [3 marks] Let $S(\boldsymbol{\beta}) = (\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$. Differentiating with respect to $\boldsymbol{\beta}$ gives

$$-2X^T\boldsymbol{Y} + 2X^TX\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta}.$$

Setting this expression equal to zero for a critical point gives

$$(X^TX + \lambda I_n)\boldsymbol{\beta} = X^T\boldsymbol{Y}.$$

[Unseen] an exercise later in the text, not in the extract given

(f) [2 marks] The lasso is equivalent to a prior under which most predictors have essentially zero effect, but a small number may have large effects. Ridge regression on the other hand supports a large number of variables contributing in moderate amounts. Lasso therefore better respects the clinicians' intuitions about this data. *[Unseen]*

(g) [3 marks] $AIC = -2l(\widehat{\boldsymbol{\beta}}) + 2p$, so selecting the model with the smaller AIC in this case (where B has one more parameter than A) corresponds to rejecting the null hypothesis that model A is correct whenever $2(l(B) - l(A)) > 2$. This means that the first statistician will have the lower type I error rate, since they are using a more extreme point of the $\chi^2(1)$ distribution. *[Unseen]*