

UNIVERSITY OF WOLLONGONG
School of Mathematics and Applied Statistics
STAT902. Advanced Data Analysis

ASSIGNMENT 7

Due: 5:00pm Monday 3rd May, 2010.

Local students: please put under lecturer's door.

Remote students: please e-mail or post by this date.

1. Consider the *intercept-only* linear regression model

$$y_i = \beta_0 + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2)$$

for $1 \leq i \leq n$. Note that this model may be written in general linear model notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \beta_0.$$

- (a) The profile log-likelihood for estimation of σ_ε^2 in general linear models (1) is

$$\ell_P(\sigma_\varepsilon^2) = -\text{RSS}/(2\sigma_\varepsilon^2) - \frac{n}{2} \ln(2\pi\sigma_\varepsilon^2)$$

where $\text{RSS} = \mathbf{y}^T \{ \mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \} \mathbf{y}$ is the residual sum of squares. Show that, in the special case of the intercept-only model, maximisation of the profile log-likelihood leads to

$$\hat{\sigma}_{\varepsilon, \text{ML}}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

- (b) The *restricted* profile log-likelihood for estimation of σ_ε^2 in general linear models (1) is

$$\ell_R(\sigma_\varepsilon^2) = -\text{RSS}/(2\sigma_\varepsilon^2) - \frac{n}{2} \ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2} \ln |(1/\sigma_\varepsilon^2) \mathbf{X}^T \mathbf{X}|$$

Show that, in the special case of the intercept-only model, maximisation of the residual profile log-likelihood leads to

$$\hat{\sigma}_{\varepsilon, \text{REML}}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

2. A small longitudinal study has $m = 3$ subjects and $n_1 = 2$, $n_2 = 3$, $n_3 = 2$ measurements on each subject, respectively. The measurements are bone density (y) and age (x).

The response vector and error vectors are:

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}.$$

The fixed effects and random effects vectors are

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{and} \quad \mathbf{u} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix}.$$

The design matrices are

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{21} \\ 1 & x_{22} \\ 1 & x_{23} \\ 1 & x_{31} \\ 1 & x_{32} \end{bmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Starting with the general linear mixed model set-up:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

obtain expression for each of the y_{ij} ($1 \leq i \leq 3, 1 \leq j \leq n_i$). Be sure to show all steps of the working.

Hint: the answer for the first one is

$$y_{11} = \beta_0 + \beta_1 x_{11} + U_1 + \varepsilon_{11}.$$

3. Consider the general linear model set-up:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

where

$$E \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad \text{Cov} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}.$$

Show that

$$\text{Cov}(\mathbf{y}) = \mathbf{ZGZ}^T + \mathbf{R}.$$

4. This question illustrates standard error calculations in mixed model analysis. The random intercept model

$$\text{weight}_{ij} = \beta_0 + U_i + \beta_1 \text{weeks}_{ij} + \varepsilon_{ij}$$

$1 \leq i \leq 48, 1 \leq j \leq 9$, for the pig weights data will be used.

- (a) Make sure that the data file `pigweights.txt` from Assignment 6 is available.
- (b) Issue the following commands to fit the above model and print a summary table:

```
library(nlme)
pigweights <- read.table("pigweights.txt", header=T)
pigweights <- groupedData(weight~weeks|idnum, data=pigweights)
fit <- lme(weight~weeks, random=~1, data=pigweights)
print(summary(fit))
```

Include the output in your submission.

- (c) Issue the following commands to extract the estimated variance components (I don't know why the developers of `lme()` made this part so complicated):

```
sigmaSqHat.eps <- fit$sigma^2
sigmaSqHat.U <- sigmaSqHat.eps*exp(2*unlist(fit$modelStruct)[1])
```

- (d) Issue the following commands to set up the matrices, corresponding to the general linear mixed model set-up:

```
X <- cbind(rep(1, 432), pigweights$weeks)
Z <- kronecker(diag(48), rep(1, 9))
G <- sigmaSqHat.U*diag(48)
R <- sigmaSqHat.eps*diag(432)
```

- (e) Obtain the matrix $V = ZGZ^T + R$:

```
V <- Z%*%G%*%t(Z) + R
```

- (f) V is a 432×432 matrix; the estimated covariance matrix of the response vector. It is too large to visualise numerically. Issue the following command to visualise it graphically:

```
image(V[432:1, ], col=grey(seq(0, 1, length=1001)), xaxt="n", yaxt="n")
```

The shades of grey are from black for the lowest numbers in V (zeroes) to white for the highest numbers.

Include the graph in your submission.

- (g) Issue the commands:

```
FishInfo <- t(X)%*%solve(V, X)
invFishInfo <- solve(FishInfo)
print(invFishInfo)
print(sqrt(diag(invFishInfo)))
```

Comment on these numbers in relation to the regression output from part (b).

5. Consider again the situation from Assignment 5 where random variables X and Y have joint density function:

$$[x, y] = \frac{2}{3}(x + 2y), \quad 0 < x < 1, \quad 0 < y < 1.$$

There it was shown that the best predictor for X given Y is:

$$\text{BP}(X) = E(X|Y) = \frac{2(1 + 3Y)}{3(1 + 4Y)}.$$

A drawback of best prediction is that it requires knowledge of the joint distribution of X and Y . A commonly-used alternative is *best linear prediction* where the predictor is restricted to a linear function of the observed data. For the current example the best linear predictor for X takes the form

$$\text{BLP}(X) = a + bY.$$

where a and b are chosen to minimise the mean squared error

$$\text{MSE}(a, b) = E[\{X - (a + bY)\}^2].$$

The solution is:

$$b = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}, \quad a = E(X) - bE(Y)$$

which depends only on the second-order moments of (X, Y) .

Find $\text{BLP}(X)$ for the current example and use it to predict X when Y is observed to be $1/2$.