

# MAS4303

## Solutions to Problems 3

School of Mathematics & Statistics

November 2008

1. Table 1 shows the heights and weights of thirty eleven-year-old girls attending Heaton Middle School, Bradford. The data are taken from Open University (1983).

The data were read from the file `height.txt` on the Web page using the following R commands.

```
eleven<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas4303/height.txt")
height<-eleven[,1]
weight<-eleven[,2]
```

- (a) Let  $Y$  be the natural logarithm of the weight and  $X$  be the natural logarithm of the height. These are shown in figure 1. The commands used were as follows.

```
logh<-log(height)
logw<-log(weight)
plot(logh,logw,xlab="Log height",ylab="Log weight")
```

Our model is

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

where, given the value of  $\tau$ , the errors  $\varepsilon_i$  are independent and  $\varepsilon_i \sim N(0, \tau^{-1})$ .

- (b) Our prior distribution is as follows. We give  $\tau$  a gamma( $d_0/2$ ,  $d_0 v_0/2$ ) distribution with  $d_0 = 6$  and  $v_0 = 0.02$ . Conditional on  $\tau$  we give  $\underline{\beta} = (\alpha, \beta)^T$  a bivariate normal prior distribution with mean vector  $\underline{b}_0 = (-10, 3)^T$  and precision matrix  $\tau C_0$  where  $C_0 = (V_0/v_0)^{-1}$  and

$$V_0 = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}.$$

We construct the design matrix in R as follows.

```
X<-cbind(rep(1,30),logh)
```

We construct the prior specification as follows.

```
d0<-6
v0<-0.02
b0<-matrix(c(-10,3),ncol=1)
V0<-matrix(c(25,0,0,1),ncol=2)
prioreleven<-list(d0=d0,v0=v0,b0=b0,V0=V0)
```

We compute the posterior distribution as follows.

```
posteleven<-linmod(prioreleven,X,logw)
posteleven
$d1
```

Height (cm)	Weight (kg)	Height (cm)	Weight (kg)
135	26	133	31
146	33	149	34
153	55	141	32
154	50	164	47
139	32	146	37
131	25	149	46
149	44	147	36
137	31	152	47
143	36	140	33
146	35	143	42
141	28	148	32
136	28	149	32
154	36	141	29
151	48	137	34
155	36	135	30

Table 1: Heights and weights of eleven-year-old girls

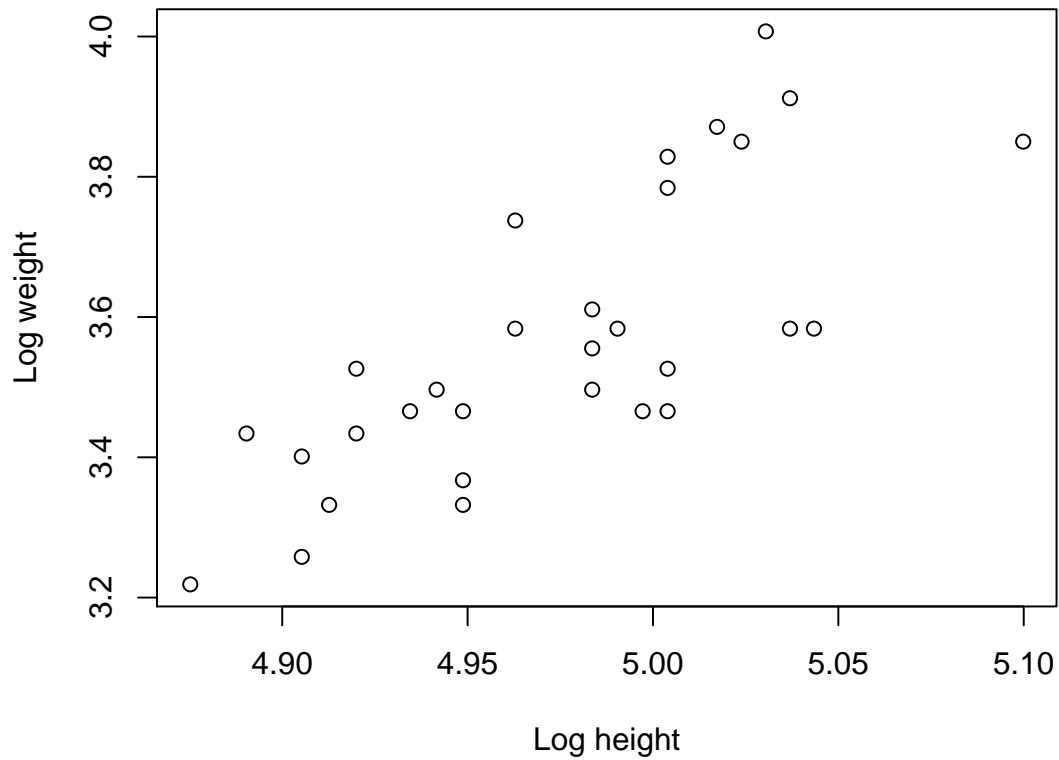


Figure 1: Log weight against log height for eleven-year-old girls.

```

[1] 36

$v1
[1] 0.01713107

$b1
      [,1]
      -10.92102
logh   2.91288

$V1
      logh
      3.4981088 -0.7031835
logh -0.7031835  0.1413758

```

Thus our posterior distribution is as follows. For  $\tau$  we have a  $\text{gamma}(d_1/2, d_1 v_1/2)$  distribution with  $d_1 = 36$  and  $v_1 = 0.01713107$ . Conditional on  $\tau$ , for  $\underline{\beta} = (\alpha, \beta)^T$  we have a bivariate normal posterior distribution with mean vector  $\underline{b}_1 = (-10.92102, 2.91288)^T$  and precision matrix  $\tau C_1$  where  $C_1 = (V_1/v_1)^{-1}$  and

$$V_0 = \begin{pmatrix} 3.4981088 & -0.7031835 \\ -0.7031835 & 0.1413758 \end{pmatrix}.$$

- (c) We find a 95% posterior predictive interval for the natural logarithm of the weight of an eleven-year-old girl who is 145 cm tall and, convert this into a 95% posterior predictive interval for the actual weight of such a girl as follows.

```

> x0<-matrix(c(1,log(145)),ncol=2)
> k1<-posteleven$v1+x0%*%posteleven$V1%*%t(x0)
> t<-qt(0.975,36)
> mean<-x0%*%posteleven$b1
> lower<-mean-t*sqrt(k1)
> upper<-mean+t*sqrt(k1)
> interval<-c(lower,upper)
> interval
[1] 3.305765 3.845455
> exp(interval)
[1] 27.26940 46.77999

```

Thus the interval for the log weight is

$$3.306 < y < 3.845$$

and the interval for the actual weight in kg is

$$27.27 < e^y < 46.78.$$

2. Table 2 gives some data from Till (1974). They give measured salinity values (parts per thousand) for three separate water masses in the Bimini Lagoon in the Bahamas.

The data were read from the file `salinity.txt` on the Web page using the following commands.

```

bimini<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas4303/salinity.txt")
salinity<-bimini[,1]
location<-bimini[,2]
mass1<-ifelse((location==1),1,0)
mass2<-ifelse((location==2),1,0)
mass3<-ifelse((location==3),1,0)

```

I	II	III
37.54	40.17	39.04
37.01	40.80	39.21
36.71	39.76	39.05
37.03	39.70	38.24
37.32	40.79	38.53
37.01	40.44	38.71
37.03	39.79	38.89
37.70	39.38	38.66
37.36		38.51
36.75		40.08
37.45		
38.85		

Table 2: Salinity measurements (parts per thousand)

(a) Our model is

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i$$

where, given the value of  $\tau$ , the errors  $\varepsilon_i$  are independent and  $\varepsilon_i \sim N(0, \tau^{-1})$  and  $x_{i,j} = 1$  if observation  $i$  is from location  $j$  with  $x_{i,j} = 0$  otherwise.

(b) Our prior distribution is as follows. We give  $\tau$  a gamma( $d_0/2$ ,  $d_0 v_0/2$ ) distribution with  $d_0 = 4$  and  $v_0 = 0.3$ . Conditional on  $\tau$  we give  $\underline{\beta} = (\beta_1, \beta_2, \beta_3)^T$  a multivariate normal prior distribution with mean vector  $\underline{b}_0 = (40, 40, 40)^T$  and precision matrix  $\tau C_0$  where  $C_0 = (V_0/v_0)^{-1}$  and

$$V_0 = H \tilde{V}_0 H^T$$

where

$$\tilde{V}_0 = \begin{pmatrix} 40 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 25 \end{pmatrix}$$

and

$$H = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

We construct the design matrix as follows.

```
X<-cbind(mass1,mass2,mass3)
```

We construct the matrix  $V_0$  as follows.

```
V0tilde<-diag(c(40,25,25,25))
H<-matrix(c(1,1,1,1,0,0,0,1,0,0,0,1),ncol=4)
V0<-H%*%V0tilde%*%t(H)
```

We construct the prior specification as follows.

```
d0<-4
v0<-0.3
b0<-matrix(c(40,40,40),ncol=1)
priorbimini<-list(d0=d0,v0=v0,b0=b0,V0=V0)
```

We compute the posterior distribution as follows.

```

postbimini<-linmod(priorbimini,X,salinity)
postbimini
$d1
[1] 34

$v1
[1] 0.2703041

$b1
      [,1]
mass1 37.31500
mass2 40.10207
mass3 38.89211

$V1
      mass1      mass2      mass3
mass1 2.250905e-02 9.307062e-06 7.447881e-06
mass2 9.307062e-06 3.375137e-02 1.116624e-05
mass3 7.447881e-06 1.116624e-05 2.700695e-02

```

Thus our prior distribution is as follows. For  $\tau$  we have a  $\text{gamma}(d_1/2, d_1 v_1/2)$  distribution with  $d_1 = 34$  and  $v_1 = 0.27030$ . Conditional on  $\tau$ , for  $\underline{\beta} = (\beta_1, \beta_2, \beta_3)^T$  we have a multivariate normal posterior distribution with mean vector  $\underline{b}_1 = (37.31500, 40.10207, 38.89211)^T$  and precision matrix  $\tau C_1$  where  $C_1 = (V_1/v_1)^{-1}$  and

$$V_1 = \begin{pmatrix} 2.250905 \times 10^{-2} & 9.307062 \times 10^{-6} & 7.447881 \times 10^{-6} \\ 9.307062 \times 10^{-6} & 3.375173 \times 10^{-2} & 1.116624 \times 10^{-5} \\ 7.447881 \times 10^{-6} & 1.116624 \times 10^{-5} & 2.700695 \times 10^{-2} \end{pmatrix}.$$

- (c) To find a 95% posterior interval for the difference in mean salinity between water mass I and water mass II we calculate as follows.

```

> x0<-matrix(c(1,-1,0),ncol=3)
> k<-x0*%postbimini$V1*%t(x0)
> mean<-x0*%postbimini$b1
> t<-qt(0.975,34)
> lower<-mean-t*sqrt(k)
> upper<-mean+t*sqrt(k)
> c(lower,upper)
[1] -3.269023 -2.305116

```

Hence the interval is

$$-3.269 < \beta_1 - \beta_2 < -2.305.$$

Note that, as an alternative to using the function `linmod` in this question, you could use the function `oneway` which is also available from the Web page.

## References

Open University, 1983. *MDST242 Statistics in Society, Unit C3: Is my child normal?*. Milton Keynes: The Open University.

Till, R., 1974. *Statistical Methods for the Earth Scientist*. London: MacMillan.