MAS8303 Modern Bayesian Inference
Part 2

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

Chapter 1

The Normal Linear Model

1.1 Regression and the normal linear model

1.1.1 Introduction

A model which describes how the conditional distribution of one variable, often called the *dependent* variable, given some other variables, depends on the values taken by these other variables, is called a *regression*. Typically we are interested in how the conditional mean of the dependent variable depends on the values of the other variables but other features of the distribution may also change. Various names are used to describe these other variables, including *regressors*, *explanatory variables* and *covariates*.

There are many different kinds of regression models. One of the simplest is described by the equation

$$Y = \alpha + \beta x + \varepsilon. \tag{1.1}$$

This might be used in situations where an observation consists of a pair of values $(x_i,\ y_i)$ where

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

- y_i is observation number i on the dependent variable,
- x_i is observation number i on a single explanatory variable,
- ε_i is a random "error" and
- α and β are parameters the values of which are usually unknown.

Our data might consist of n such pairs.

We also need to specify a sampling distribution for ε_i. In this chapter we assume the following *(conditional on model parameters)*:

Normality : $\varepsilon \sim N(0,\ \sigma^2)$.

Independence : $\varepsilon_1, \ldots, \varepsilon_n$ are independent , given the parameters of their distribution (typically the variance σ^2).

Equality of variance each of $\varepsilon_1, \ldots, \varepsilon_n$ has the same variance σ^2 (equivalently, the same precision τ).

Another way to express this model is to say that the conditional distribution of Y given x (and the model parameters) is normal with mean $\alpha + \beta x$ and variance σ^2 and that, given x_i and x_j and the model parameters, Y_i and Y_j are independent for $i \neq j$.

This model, with the relationship given in (1.1) and these assumptions about the errors is called an ordinary linear regression on a single covariate with normal errors.

1.1.2 Example

Here is a simple example. We wish to be able to predict the height of a student if we know the student's shoe size.

Suppose that we are prepared to accept (1.1) as a reasonable description of the relationship. That is, the conditional mean height, given shoe size, is a linear function of shoe size and the actual heights, given a particular shoe size, have a distribution centered on this mean. The "errors" ε are the differences between the actual height values and the conditional mean given by our linear function of shoe size. Suppose that we are also prepared to accept the usual assumptions of normality, independence and equal variance, that is that the conditional variance of height given shoe size does not depend on shoe size. These are issues of *model choice*. One way to think of a regression model like this is as a device which allows us to use information from many different values of the regressor X to help us to make predictions about Y for other values of X, in a way which seems to be appropriate to us according to our prior beliefs.

We need to specify our prior distribution for the parameters. There are three parameters in this model, α, the intercept, β, the slope of the *regression line*, and $\tau = \sigma^{-2}$, the error precision.

There are many possibilities, including the following.

- The value of τ is known and we give α and β a bivariate normal prior.
- The value of τ is unknown but we use a conjugate prior. We give τ a gamma prior and we give α and β a bivariate normal conditional prior, given τ, where the precision matrix is proportional to τ.
- We use a semi-conjugate prior in which τ has a gamma prior and α and β have a bivariate normal prior independently of τ.
- A non-conjugate prior.

For illustration, consider a semi-conjugate prior.

As it stands, our model says that the conditional mean height for a student with shoe size x is $\alpha + \beta x$. This makes α a rather unnatural paramater because it represents the mean height for students with shoe size zero, a shoe size which is well outside the usual range for students. This makes it difficult for us to think about our prior beliefs about α and also creates a rather awkward relationship between α and β in our beliefs since a change in α would require a change in β to make the regression line continue to pass through the region where we think (X, Y) points will typically be found. It is better to change the origin of X to a more usable reference value x_{ref}. I know my own shoe size and height so let us use my shoe size, 11, as a reference value. Let $z = x - x_{\text{ref1}} = x - 11$ then our regression equation becomes

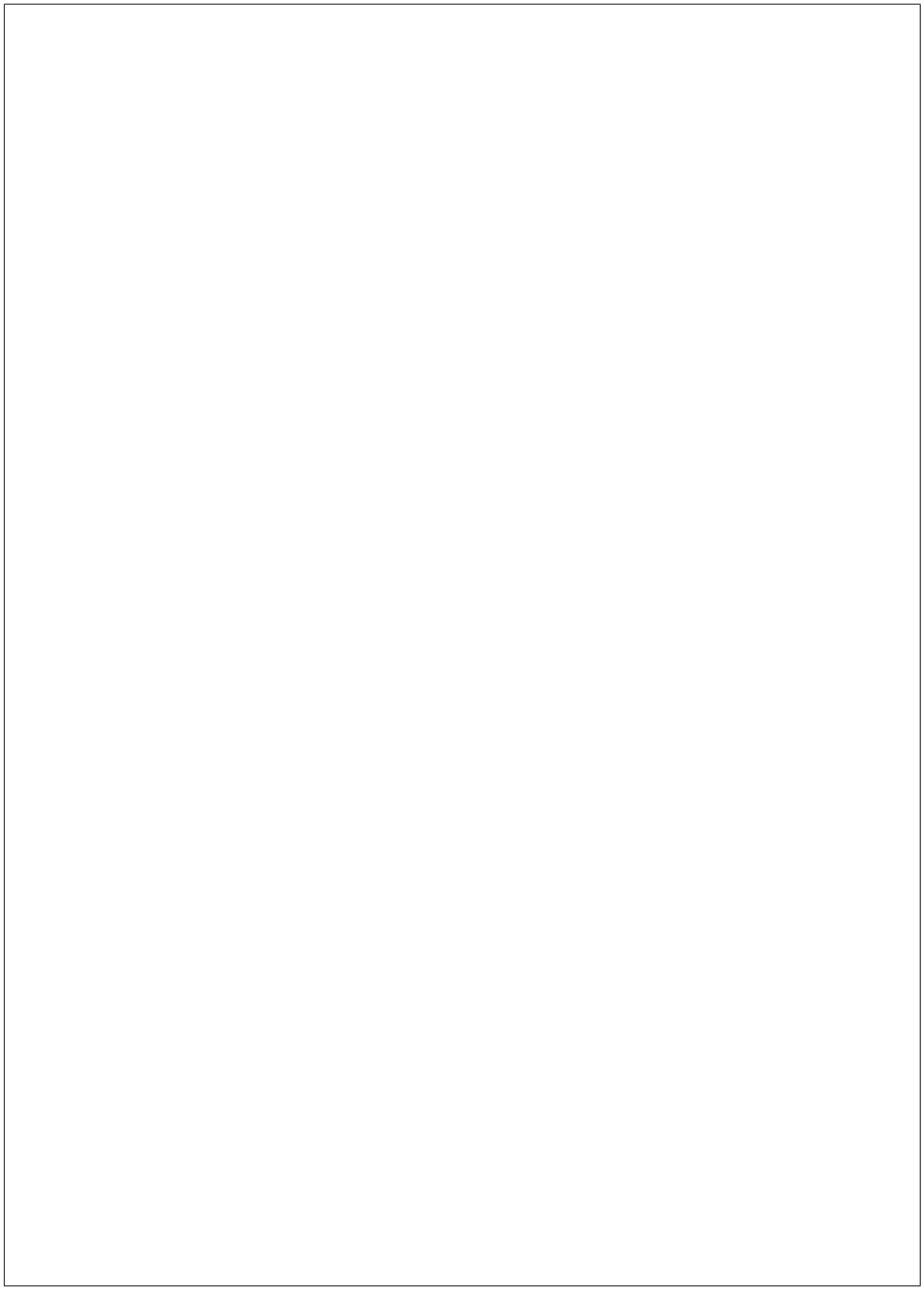
$$Y = \tilde{\alpha}_1 + \beta z + \varepsilon,$$

where $\tilde{\alpha}_1 = \alpha + \beta x_{\text{ref1}} = \alpha + 11\beta$ now represents the mean height for students who take size 11 shoes. I can now use my own height, 74 inches, as a guide to the likely value of $\tilde{\alpha}$. Let us give $\tilde{\alpha}$ a prior distribution which is normal with mean 74. Of course I can not assume that I am exactly the average height for size 11 shoe-wearers so we need a suitable prior standard deviation for $\tilde{\alpha}_1$. I think that, even bearing in mind that it is a long time since I was a first year student, I am unlikely to be more than six inches from the conditional mean so let us make the standard deviation 3, giving a variance of 9 and a precision of 0.111. We could choose to make the standard deviation larger, of course, if we felt less confident about the value of our prior information.

Now we need a prior for β. How much does the mean height change when we change the shoe size by one unit? As a guide, my wife is 64 inches tall and takes size 5 shoes. This suggests a change of 10 inches in 6 shoes sizes or $\beta = 10/6 \approx 1.7$. Let us use $x_{\text{ref2}} = 5$ as a second reference value. Let $\tilde{\alpha}_2 = \alpha + \beta x_{\text{ref2}} = \alpha + 5\beta$ now represent the mean height for students who take size 5 shoes. Let us give $\tilde{\alpha}_2$ a $N(64,\ 9)$ prior distribution.

In this example it seems reasonable to make $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ independent in our prior distribution and this is what we will do. In other examples we might, for example, feel that we are likely to have misjudged both conditional means in the same direction and so give them a positive covariance. So, let us write

$$\underset{\sim}{\tilde{\beta}} = \begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{pmatrix}.$$



1.1.3 The normal linear model

The normal linear model is a more general class of models which includes (1.1) and many more kinds of model, as special cases.

First of all let us rewrite (1.1), using slightly different notation, as

$$Y = \beta_0 + \beta_1 x + \varepsilon. \tag{1.2}$$

Now suppose that we want to relate the dependent variable Y to the values, $x_1, \ldots, x_k$, of two or more regressors $X_1, \ldots, X_k$. One way to do this is to write

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$$

So, our model for observation i is

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \varepsilon_i, \tag{1.3}$$

where $x_{i,j}$ is the value of regressor X_j in observation i.

It is convenient to make a further change to the notation. We relabel the regressors and coefficients $1, \ldots, p$ instead of $0, \ldots, k$. So $k = p - 1$. Then

$$Y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i = \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i. \tag{1.4}$$

We seem to have lost the intercept term β_0 in (1.2) and (1.3). However this is easily overcome by defining X_1 so that $x_{i,1} = 1$ for all i. Then we can rewrite (1.2) as

$$Y = \beta_1 1 + \beta_2 x + \varepsilon$$

and define $X_1 \equiv 1$ and $X_2 = X$.

Example: one-way layout We observe several samples from normal distributions (as in the "one-way ANOVA"). Model: $Y_{i,j} \sim N(\mu_j,\ \tau^{-1})$ for the i^{th} observation in sample j. Let us rename μ_j as β_j. Then we can write the model as

$$Y_{i,j} = \beta_j + \varepsilon_{i,j} \tag{1.5}$$

where $\varepsilon_{i,j} \sim N(0,\ \tau^{-1})$. Now, suppose that, instead of numbering the observations within each sample, we number them all in one long sequence $Y_1, \ldots, Y_n$, where $n = \sum_{j=1}^{J} n_j$. We need a way to indicate to which sample an observation belongs so we define regressors $X_1, \ldots, X_J$ where $x_{i,j} = 1$ if observation i is in sample j and $x_{i,j} = 0$ otherwise. Then our model is exactly of the form (1.4) if we set $p = J$.

Notice that, for fixed values of the regressors $x_{i,1}, \ldots, x_{i,p}$, (1.4) is linear in the coefficients $\beta_1, \ldots, \beta_p$. This is therefore called a *linear model* or a *linear regression*. It is called a *normal linear model* because of our assumption that the "errors" ε are normally distributed. The normal linear model includes a great variety of models which are commonly used in statistics. Generalisations and extensions allow an even greater variety but we will leave these for later. Just to illustrate that the linearity refers to the coefficients and not to the shape of a graph which we might draw to represent how Y changes, consider a model in which we want to describe the way that Y changes over time t using a cubic function of t. We simply write $x_{i,1} = 1,\ x_{i,2} = t_i,\ x_{i,3} = t_i^2$ and $x_{i,4} = t_i^3$. Then

$$Y_i = \beta_1 + \beta_2 t_i + \beta_3 t_i^2 + \beta_4 t_i^3 + \varepsilon_i.$$

Matrix notation

It is convenient to use matrix notation. We rewrite (1.4) as

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{1.6}$$

where $\underline{Y} = (Y_1, \ldots, Y_n)^T$ is a $n \times 1$ vector of observations on Y,

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

is a $n \times p$ matrix whose elements are known x-values. (In some cases all of the elements are 0 or 1). We call X the *design matrix*. This name reflects the fact that sometimes, that is is designed experiments, the elements of X are deliberately chosen and X then represents the *design* of the experiment. The $p \times 1$ vector of unknown parameters is $\underline{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\underline{\varepsilon} = (\varepsilon_1 \ldots, \varepsilon_n)^T$ is a $n \times 1$ error vector. The vector of random errors has a multivariate normal distribution (given τ):

$$\underline{\varepsilon} \sim N_n(\underline{0},\ \tau^{-1} I)$$

where $\underline{0}$ is a vector of zeroes and I is a $n \times n$ identity matrix.

Given τ and $\underline{\beta}$, the vector of observations $\underline{y}$ is an observation from a multivariate normal distribution:

$$\underline{Y} \sim_n N(X\underline{\beta},\ \tau^{-1} I).$$

Example: Regression on a single covariate Here $Y_i = \alpha + \beta x_i + \varepsilon_i$. We have

$$X = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_1 & x_2 & x_3 & \ldots & x_n \end{pmatrix}^T$$

and

$$\underline{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

Example: one-way layout (as above). Suppose, for illustration, that we have four samples, each with three observations. Then we have

$$\underline{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \qquad \text{and} \qquad X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

(There is, in fact, more than one way to *parameterise*, that is express in terms of parameters, this model and it is sometimes convenient to do it in a different way).

Notice that the design matrix contains one column corresponding to each of the coefficients $\beta_1, \ldots, \beta_p$.

1.2 Inference for the normal linear model

1.2.1 Likelihood and sufficient statistics

Given the model in (1.6) and a data vector $\underline{y}$ containing n observations, the likelihood is

$$L = (2\pi)^{-n/2}\tau^{n/2}\exp\left\{-\frac{\tau}{2}(\underline{y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta})\right\}.$$

We will assume in what follows that the design matrix X is of full rank and that therefore $(X^TX)^{-1}$ exists. If X is not of full rank then this does not mean that the likelihood does not exist nor that no Bayesian inference is possible. However it does mean that there is at least one linear function of $\underline{\beta}$ about which the data will tell us nothing. In such a case it may be best to reconsider the model. For example, suppose that, instead of the model in (1.5), we had $Y_{i,j} = \mu + \beta_j + \varepsilon_{i,j}$ where μ is meant to represent a sort of overall mean. Then, when we put this in the form (1.4), μ becomes, in effect, β_{J+1} and we have an extra regressor X_{J+1} where $x_{i,J+1} = 1$. However, for all i, $x_{i,J+1} = x_{i,1} + \cdots + x_{i,J}$ so the rank of X is still J, not $J+1$ even though it now has $J+1$ columns. It is easy to see that, in this case, we have too many parameters and they can not all be *identified*. If we replaced $\beta_1, \ldots, \beta_J$ with $\tilde{\beta}_1, \ldots, \tilde{\beta}_J$, where $\tilde{\beta}_j = \beta_j + \delta$, and β_{J+1} with $\tilde{\beta}_{J+1} = \beta_{J+1} - \delta$, then we would get exactly the same model and exactly the same likelihood so the data can not tell us about δ and therefore not about the complete set of values of $\beta_1, \ldots, \beta_J$. We could, however, learn about the differences $\beta_j - \beta_{J+1}$ for $j = 1, \ldots, J$.

So, assuming that $(X^TX)^{-1}$ exists, let us write

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

We call $\hat{\underline{\beta}}$ the *least squares estimates* of $\underline{\beta}$.

Then

$$\begin{aligned}(\underline{y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta}) &= (\underline{y}-X\hat{\underline{\beta}}-X[\underline{\beta}-\hat{\underline{\beta}}])^T(\underline{y}-X\hat{\underline{\beta}}-X[\underline{\beta}-\hat{\underline{\beta}}]) \\ &= (\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}}) + (\underline{\beta}-\hat{\underline{\beta}})^TX^TX(\underline{\beta}-\hat{\underline{\beta}}) - 2(\underline{\beta}-\hat{\underline{\beta}})^TX^T(\underline{y}-X\hat{\underline{\beta}})\end{aligned}$$

but

$$(\underline{\beta}-\hat{\underline{\beta}})^TX^T(\underline{y}-X\hat{\underline{\beta}}) = (\underline{\beta}-\hat{\underline{\beta}})^T\{X^T\underline{y} - X^TX(X^TX)^{-1}X^T\underline{y}\} = 0.$$

Thus

$$L = (2\pi)^{-n/2}\tau^{n/2}\exp\left\{-\frac{\tau}{2}[S_d + (\underline{\beta}-\hat{\underline{\beta}})^TX^TX(\underline{\beta}-\hat{\underline{\beta}})]\right\} \tag{1.7}$$

and S_d and $\hat{\underline{\beta}}$ are sufficient for τ and $\underline{\beta}$, where

$$S_d = (\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}}).$$

Moreover, if τ is known, then $\hat{\underline{\beta}}$ is sufficient for $\underline{\beta}$.

The sampling distribution of $\underline{Y}$ is

$$\underline{Y} \mid \tau, \underline{\beta} \sim N_n(X\underline{\beta},\ \tau^{-1}I)$$

so the sampling distribution of $\hat{\underline{\beta}}$ is

$$\hat{\underline{\beta}} \mid \tau, \underline{\beta} \sim_p N(\underline{\beta},\ \tau^{-1}[X^TX]^{-1})$$

since $(X^TX)^{-1}X^TX\underline{\beta} = \underline{\beta}$ and $(X^TX)^{-1}X^T[\tau^{-1}I]X(X^TX)^{-1} = \tau^{-1}[X^TX]^{-1}$. Thus the "data precision" is τX^TX.

1.2.2 Inference with known error precision

Suppose that the error precision is known and that our prior distribution for $\underline{\beta}$ is a multivariate normal distribution with mean $\underline{b}_0$ and variance $V_0 = P_0^{-1}$. Then the posterior distribution is a multivariate normal distribution with mean $\underline{b}_1$ and variance $V_1 = P_1^{-1}$ where

$$\begin{aligned} \underline{b}_1 &= P_1^{-1}(P_0\underline{b}_0 + P_d\underline{\hat{\beta}}), \\ P_1 &= P_0 + P_d \\ \text{and} \qquad P_d &= \tau X^T X. \end{aligned}$$

The matrices P_0 and P_1 are the prior and posterior *precision matrices* respectively.

Proof: The prior density is proportional to

$$\exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{b}_0)^T P_0(\underline{\beta} - \underline{b}_0)\right\}.$$

The posterior density is therefore proportional to

$$\begin{aligned} h(\underline{\beta}) &= \exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{b}_0)^T P_0(\underline{\beta} - \underline{b}_0)\right\} \exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{\hat{\beta}})^T P_d(\underline{\beta} - \underline{\hat{\beta}})\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\underline{\beta}^T(P_0 + P_d)\underline{\beta} - 2(\underline{b}_0^T P_0 + \underline{\hat{\beta}}^T P_d)\underline{\beta} + \underline{b}_0^T P_0\underline{b}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}}\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\underline{\beta}^T(P_0 + P_d)\underline{\beta} - 2(\underline{b}_0^T P_0 + \underline{\hat{\beta}}^T P_d)(P_0 + P_d)^{-1}(P_0 + P_d)\underline{\beta} + \underline{b}_0^T P_0\underline{\beta}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}}\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\underline{\beta}^T(P_0 + P_d)\underline{\beta} - 2\underline{b}_1^T(P_0 + P_d)\underline{\beta} + \underline{b}_0^T P_0\underline{b}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}}\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\underline{\beta}^T(P_0 + P_d)\underline{\beta} - 2\underline{\beta}_1^T(P_0 + P_d)\underline{\beta} + \underline{b}_1^T(P_0 + P_d)\underline{b}_1\right]\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}\left[\underline{b}_0^T P_0\underline{b}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}} - \underline{b}_1^T(P_0 + P_d)\underline{b}_1\right]\right\} \\ &= \exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{b}_1)^T(P_0 + P_d)(\underline{\beta} - \underline{b}_1)\right\} \times \exp\left\{-\frac{1}{2}\left[\underline{b}_0^T P_0\underline{b}_0 + \underline{\hat{\beta}}^T P_d\underline{\hat{\beta}} - \underline{b}_1^T(P_0 + P_d)\underline{b}_1\right]\right\} \end{aligned}$$

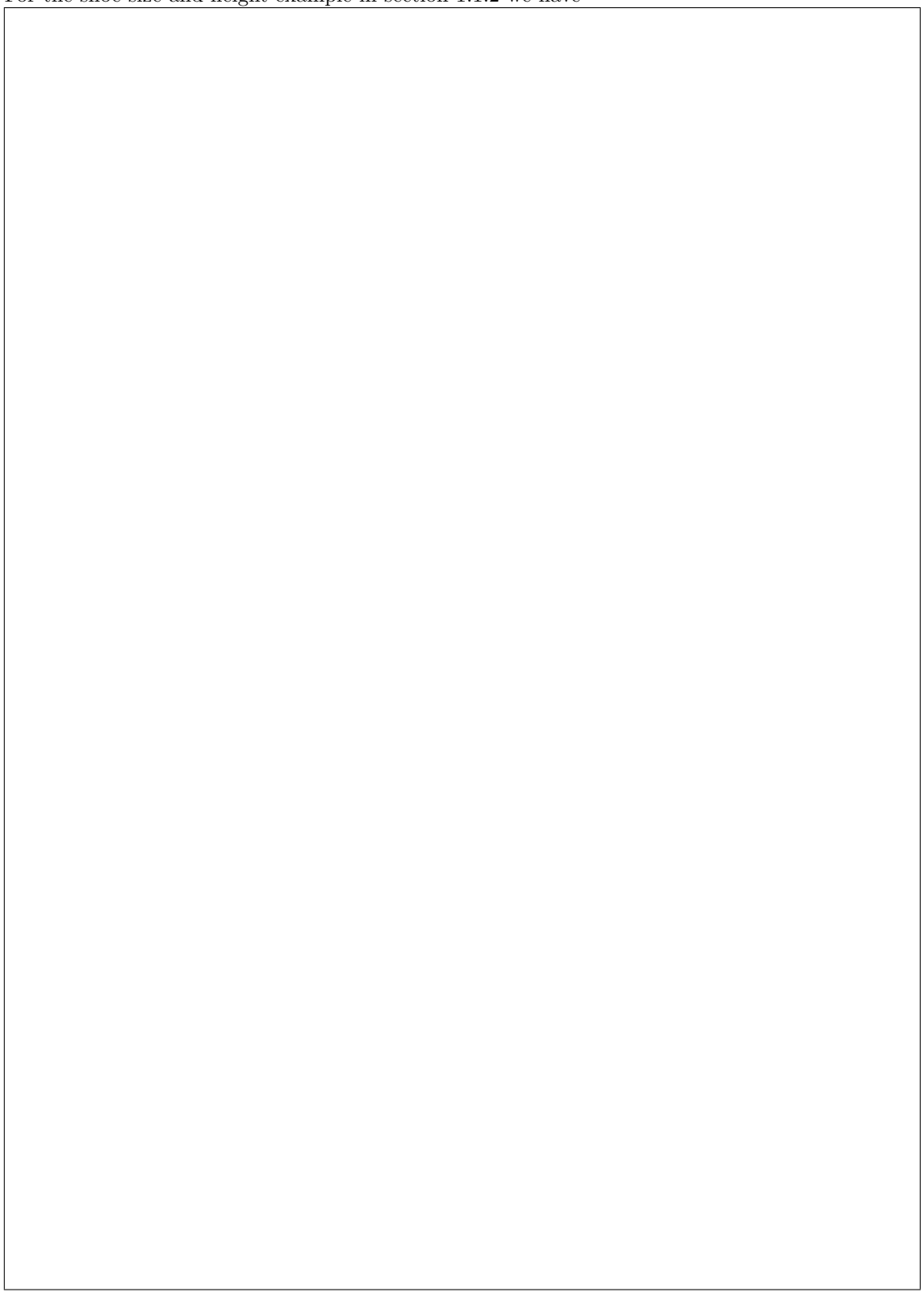
which is proportional to

$$\exp\left\{-\frac{1}{2}(\underline{\beta} - \underline{b}_1)^T(P_0 + P_d)(\underline{\beta} - \underline{b}_1)\right\}$$

which, in turn, is proportional to the pdf of a normal distribution with mean $\underline{b}_1$ and precision matrix $P_0 + P_d$.

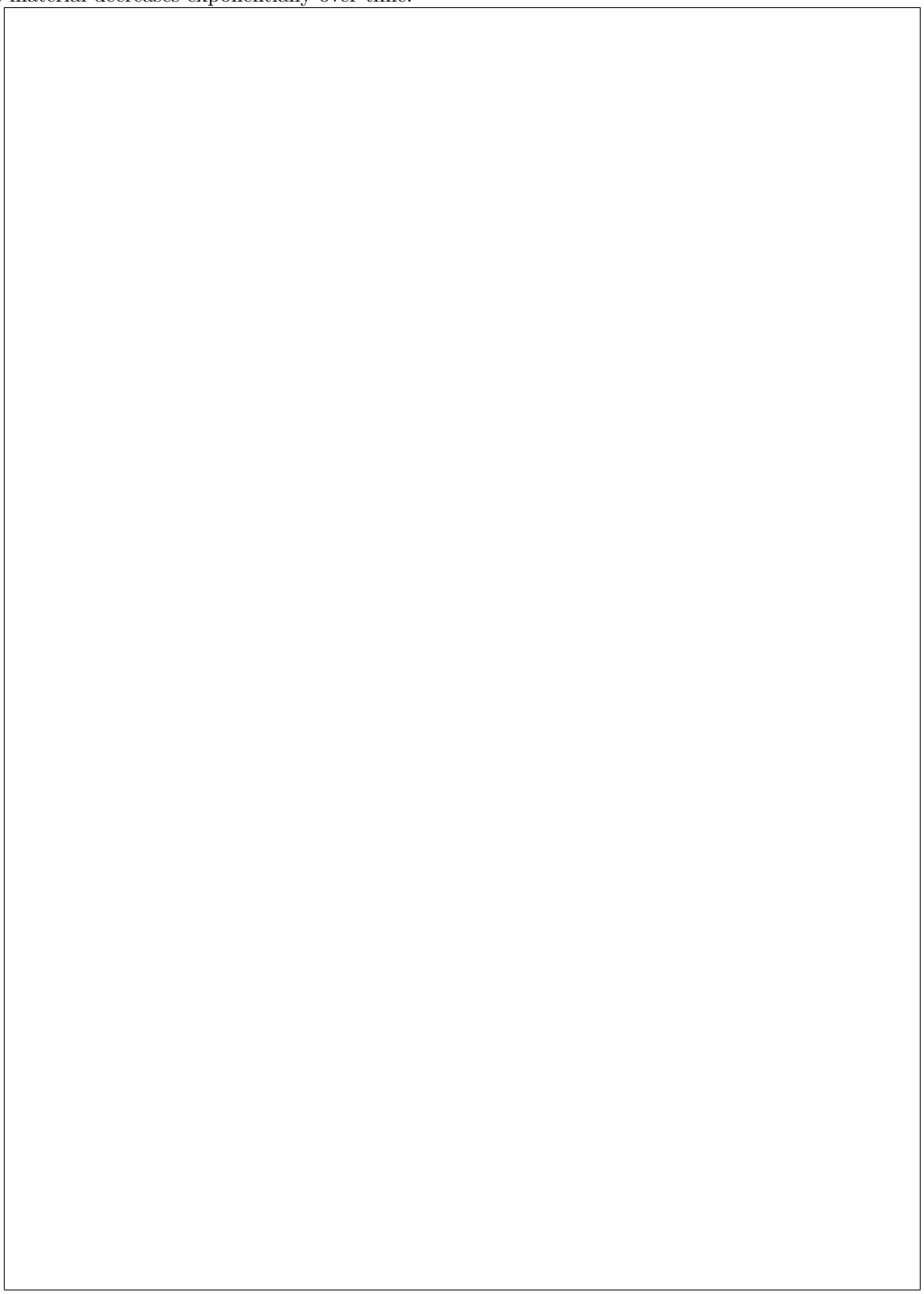
Example 1

For the shoe-size and height example in section 1.1.2 we have



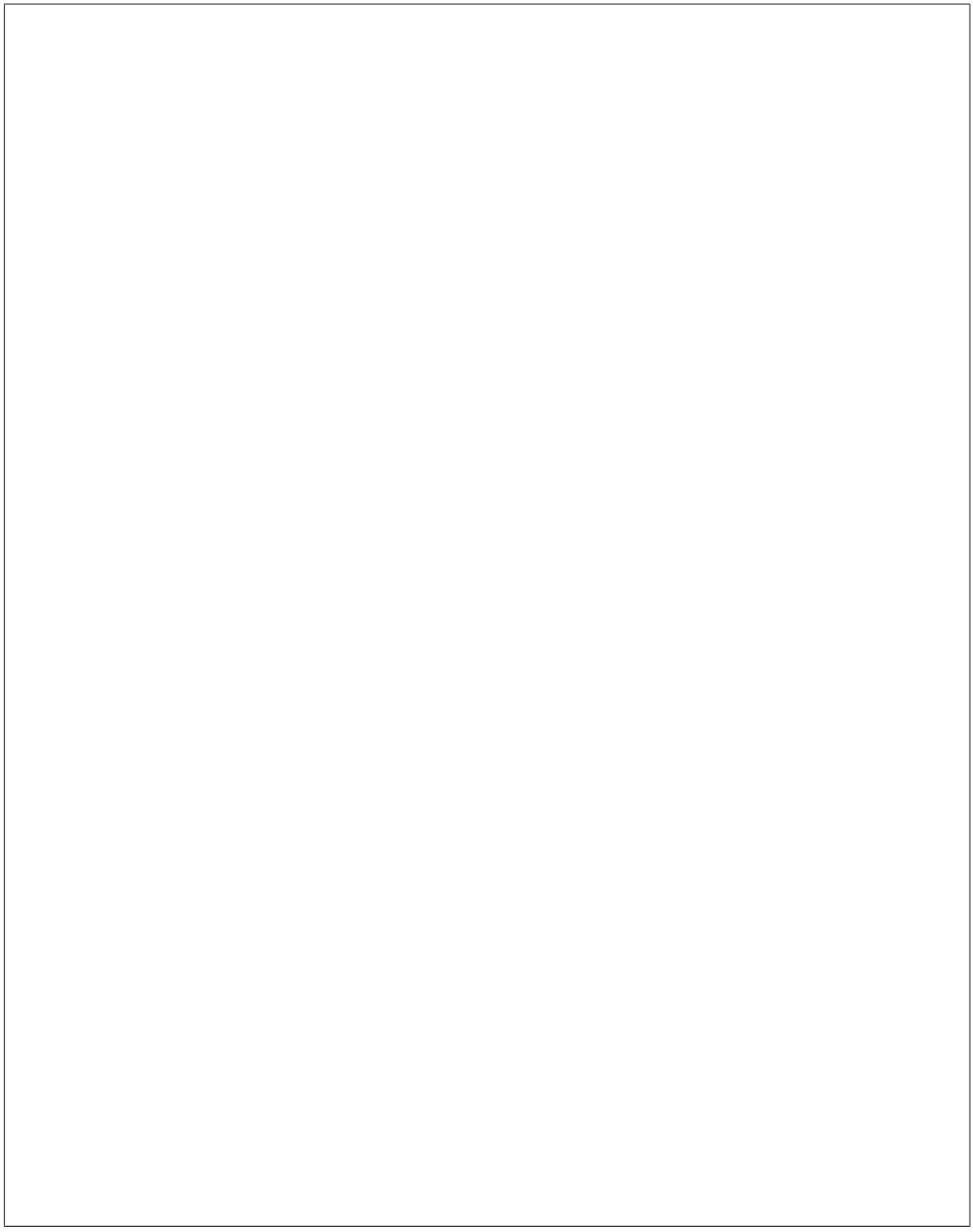
Example 2

After certain material is extracted from an organism, the concentration of a certain compound in the material decreases exponentially over time.



The data are as follows.

i	1	2	3	4	5	6
Time t_i	25	50	75	100	125	150
Measured Concentration $\tilde{Z}$	113	81	74	52	43	36
Log Concentration Y	4.73	4.39	4.30	3.95	3.76	3.58



1.3 Inference with a conjugate prior

1.3.1 Prior and posterior

Suppose now that τ is unknown. There is a conjugate prior.

We give τ a gamma Ga$(d_0/2,\ d_0v_0/2)$ prior. Then $d_0v_0\tau$ has a $\chi^2_{d_0}$ distribution.

We then define the *conditional* prior distribution of $\underline{\beta}$ given τ as a multivariate normal distribution with mean $\underline{b}_0$ and precision $P_0 = C_0\tau$, where the value of C_0 is specified. Thus the prior precision of $\underline{\beta}$ is proportional to the error precision τ. It is easily shown that the marginal prior distribution of β_j is such that

$$\frac{\beta_j - b_{0,j}}{\sqrt{v_0/c_{0,j,j}}} \sim t_{d_0}$$

where $b_{0,j}$ is the j^{th} element of $\underline{b}_0$ and $c^{-1}_{0,j,j}$ is the j^{th} diagonal element of C_0^{-1}.

The prior density is then proportional to

$$\tau^{d_0/2-1}e^{-\tau(d_0v_0/2)}\tau^{p/2}\exp\left\{-\frac{\tau}{2}(\underline{\beta}-\underline{b}_0)^TC_0(\underline{\beta}-\underline{b}_0)\right\}.$$

From (1.7) the likelihood is proportional to

$$\tau^{n/2}\exp\left\{-\frac{\tau}{2}S_d\right\}\exp\left\{-\frac{\tau}{2}(\hat{\underline{\beta}}-\underline{\beta})^TC_d(\hat{\underline{\beta}}-\underline{\beta})\right\}$$

where $C_d = X^TX$.

The posterior density is therefore proportional to

$$\tau^{(d_0+n)/2-1}e^{-\tau(d_0v_0+S_d)/2}\tau^{p/2}\exp\left\{-\frac{\tau}{2}[(\underline{\beta}-\underline{b}_0)^TC_0(\underline{\beta}-\underline{b}_0)+(\hat{\underline{\beta}}-\underline{\beta})^TC_d(\hat{\underline{\beta}}-\underline{\beta})]\right\}.$$

Some further algebra shows that the posterior density is proportional to

$$\tau^{d_1/2-1}e^{-\tau d_1v_1/2}\tau^{p/2}|C_1|^{1/2}\exp\left\{-\frac{\tau}{2}(\underline{\beta}-\underline{b}_1)^TC_1(\underline{\beta}-\underline{b}_1)\right\}$$

where

$$\begin{aligned}
d_1 &= d_0+n\\
v_1 &= \frac{d_0v_0+nv_d}{d_0+n}\\
v_d &= \frac{S_d+R}{n}\\
R &= \underline{b}_0^TC_0\underline{b}_0+\hat{\underline{\beta}}^TC_d\hat{\underline{\beta}}-\underline{b}_1^TC_1\underline{b}_1\\
C_1 &= C_0+C_d\\
\underline{b}_1 &= (C_0+C_d)^{-1}(C_0\underline{b}_0+C_d\hat{\underline{\beta}})
\end{aligned}$$

Thus

- The marginal posterior distribution of τ is gamma Ga$(d_1/2,\ d_1v_1/2)$.
  So $d_1v_1\tau \sim \chi^2_{d_1}$.
- The conditional posterior distribution of $\underline{\beta}$ given τ is multivariate normal with mean $\underline{b}_1$ and variance $P_1^{-1} = \tau^{-1}C_1^{-1}$.

It is convenient to use a R function to do the calculations. A suitable function is shown in figure 1.1. The prior specification is supplied as a list containing $d_0,\ v_0,\ b_0$ and V_0, where $(V_0/v_0)^{-1} = C_0$. The function returns a list containing $d_1,\ v_1,\ b_1$ and V_1, where $(V_1/v_1)^{-1} = C_1$. The data are supplied as a matrix X and a vector $\underline{y}$. The function can be called with a command such as the following.

```
posterior<-linmod(prior,X,y)
```

```
linmod<-function(prior,X,y,unknown=TRUE)
{Xt<-t(X)
 Cd<-Xt%*%X
 Xty<-Xt%*%y
 b0<-prior$b
 betahat<-solve(Cd,Xty)
 n<-length(y)
 C0<-solve(prior$V/prior$v)
 C1<-C0+Cd
 b1<-solve(C1,(C0%*%b0+Cd%*%betahat))
 res<-y-X%*%betahat
 Sd<-sum(res^2)
 if (unknown)
    {d1<-prior$d+n
     R<-t(b0)%*%C0%*%b0 + t(betahat)%*%Cd%*%betahat - t(b1)%*%C1%*%b1
     nvd<-Sd+R
     v1<-(prior$d*prior$v + nvd)/d1
     v1<-v1[1,1]
     }
 else
    {v1<-prior$v
     d1<-0
     }
 V1<-v1*solve(C1)
 result<-list(d=d1,v=v1,b=b1,V=V1)
 result
 }
```

Figure 1.1: R function for the normal linear model

Optionally, we can use a command such as the following.

```
posterior<-linmod(prior,X,y,unknown=FALSE)
```

In this latter case the calculations for the known-τ case are used and the `prior` argument is a list containing $v_0 = \tau^{-1}$, b_0 and $V_0 = P_0^{-1}$. Similarly the result is a list containing $v_1 = v_0 = \tau^{-1}$, b_0 and $V_1 = P_1^{-1}$. The result in this case also contains the value $d_1 = 0$.

The use of the function is illustrated in the following examples.

Example 1

This is the example involving shoe sizes and heights of students, as in section 1.2.2. The only difference here is that we make τ unknown with $d_0 = 2$ and $v_0 = 2$. We can use the R function as follows.

```
> Xshoe<-matrix(c(rep(1,152),shoesize),ncol=2)
> d0shoe<-2
> v0shoe<-2
> b0shoe<-matrix(c(55.7,1.7),ncol=1)
> V0shoe<-matrix(c(36.5,-4,-4,0.5),ncol=2)
> priorshoe<-list(d=d0shoe,v=v0shoe,b=b0shoe,V=V0shoe)
> postshoe<-linmod(priorshoe,Xshoe,height)
> postshoe
$d
[1] 154

$v
```

Treatment	Diet	Weight gain									
1	Beef Low	90	76	90	64	86	51	72	90	95	78
2	Beef High	73	102	118	104	81	107	100	87	117	111
3	Cereal Low	107	95	97	80	98	74	74	67	89	58
4	Cereal High	98	74	56	111	95	88	82	77	86	92

Table 1.1: Weight gains in rats given different diets

```
[1] 3.515502

$b
          [,1]
[1,] 56.189352
[2,]  1.561022

$V
           [,1]          [,2]
[1,]  0.3947624 -0.046750199
[2,] -0.0467502  0.005879935
```

The posterior means are exactly the same as in the known-τ case. This is a property of the conjugate prior when it is specified in this way, with everything unchanged and v_0 equal to the previous "known" value. It seems though that the error variance may be a little greater than our "known" value.

Example 2

The data in table 1.1 are from Snedecor and Cochran (1967) and are also given by Hand *et al.* (1994). They give the gains in weight of rats fed on four different diets. The diets differ in terms of the amount of protein ("low" or "high") and the source of the protein ("beef" or "cereal").

Suppose that our prior beliefs are as follows. Given parameters $\underline{\mu} = (\mu_1, \ldots, \mu_4)^T$, τ, the weight gains $Y_{1,1}, \ldots, Y_{10,4}$ are independent with $Y_{i,j} \sim N(\mu_j, \tau^{-1})$. Our prior distribution for τ is gamma $\mathrm{Ga}(d_0/2,\ d_0 v_0/2)$ with $d_0 = 2$ and $v_0 = 60$. Our conditional prior distribution for $\underline{\mu}$ is $N_4(\underline{M}_0, (\tau C_0)^{-1})$ with $\underline{M}_0 = (80, 80, 80, 80)^T$ and

$$C_0 = \frac{1}{8}\begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}^{-1} = \frac{1}{40}\begin{pmatrix} 4 & -1 & -1 & -1 \\ -1 & 4 & -1 & -1 \\ -1 & -1 & 4 & -1 \\ -1 & -1 & -1 & 4 \end{pmatrix}.$$

Consider an alternative way of formulating this example. Instead of working directly in terms of the four means $\mu_1, \ldots, \mu_4$, we can use different parameters. We can write

$$\begin{aligned} \mu_1 &= \mu - \beta_a - \beta_s + \gamma, \\ \mu_2 &= \mu + \beta_a - \beta_s - \gamma, \\ \mu_3 &= \mu - \beta_a + \beta_s - \gamma, \\ \mu_4 &= \mu + \beta_a + \beta_s + \gamma. \end{aligned}$$

Here μ is an *overall mean*, β_a is an *effect* due to the amount of protein, β_s is an effect due to the source of protein. The interaction effect allows the treatment means to be unrestricted. It allows for the mean for, eg., "cereal high" not to be obtained simply by adding the source effect and the amount effect to the overall mean. It is easily seen that

$$\underline{\beta} = (\mu,\ \beta_a,\ \beta_s,\ \gamma)^T = H\underline{\mu}$$

where

$$H = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

So, if our prior mean and conditional prior variance for $\underline{\mu}$ were $\underline{M}_0$ and $(\tau C_{0,\mu})^{-1}$ respectively, then our prior mean and prior variance for $\underline{\beta}$ are

$$\underline{b}_0 = H\underline{M}_0 = \begin{pmatrix} 80 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad (\tau C_0)^{-1} = H(\tau C_{0,\mu})^{-1}H^T = \tau^{-1}\begin{pmatrix} 10 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

Hence

$$C_0 = \begin{pmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{pmatrix}.$$

In practice we might assess the prior distribution for $\underline{\beta}$ directly rather than through a prior distribution for $\underline{\mu}$. Also we might well wish to give γ a smaller prior variance than β_a or β_s since we might judge that such an interaction effect is likely to be less important than the *main* effects of amount and source of protein.

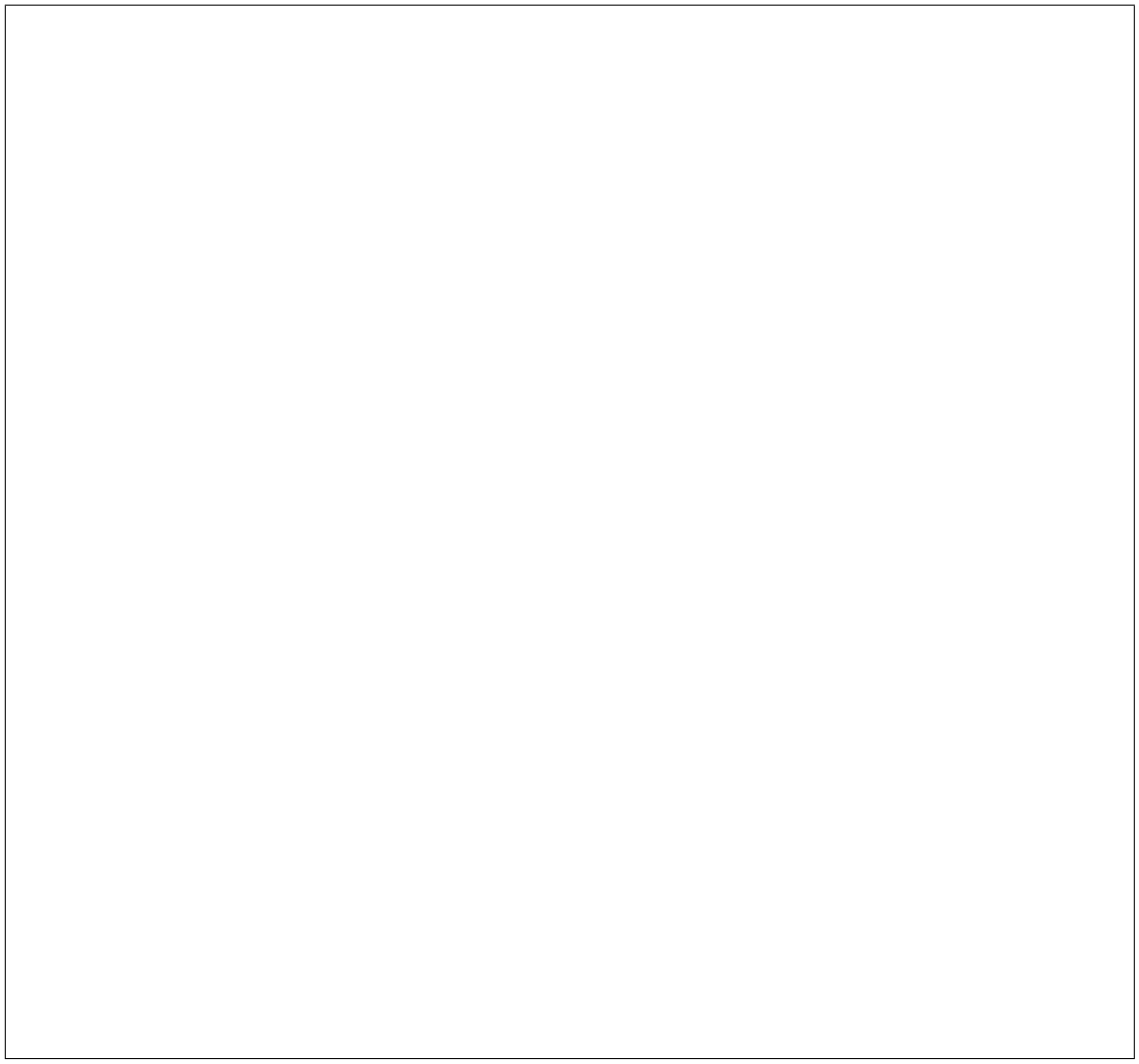
Here is the calculation of the posterior distribution using the R function `linmod`. The vector `ratgain` contains the weight gains $(90, 76, 90, \ldots, 86, 92)$.

```
> x1<-rep(1,40)
> x2<-rep(c(-1,1,-1,1),c(10,10,10,10))
> x3<-rep(c(-1,-1,1,1),c(10,10,10,10))
> x4<-rep(c(1,-1,-1,1),c(10,10,10,10))
> Xrat<-cbind(x1,x2,x3,x4)
> d0rat<-2
> v0rat<-60
> b0rat<-matrix(c(80,0,0,0),ncol=1)
> V0rat<-60*diag(c(10,2,2,2))
> priorrat<-list(d=d0rat,v=v0rat,b=b0rat,V=V0rat)
> postrat<-linmod(priorrat,Xrat,ratgain)
> postrat
$d
[1] 42

$v
[1] 195.3410

$b
        [,1]
x1 87.231920
x2  5.629630
x3 -2.320988
x4 -4.641975

$V
         x1       x2       x3       x4
x1 4.871347 0.000000 0.000000 0.000000
x2 0.000000 4.823235 0.000000 0.000000
x3 0.000000 0.000000 4.823235 0.000000
x4 0.000000 0.000000 0.000000 4.823235
```



1.3.2 Linear functions of coefficients

In our posterior distribution $\tau \sim \text{Ga}(d_1/2,\ d_1 v_1/2)$ and $\beta \mid \tau \sim N_p(\underline{b}_1,\ (\tau C_1)^{-1})$.

Suppose that we are interested in some linear function of $\underline{\beta}$. For example, with $\underline{\beta} = (\beta_1, \beta_2, \beta_3)^T$, we might be interested in $\delta = \underline{x}\underline{\beta} = 4\beta_1 + 3\beta_2 - 5\beta_3$. This is, of course, the mean of Y when $\underline{x} = (4, 3, -5)$.

Then

$$\delta \mid \tau \sim N(\underline{x}\underline{b}_1,\ \underline{x}(\tau C_1)^{-1}\underline{x}^T).$$

That is

$$\delta \mid \tau \sim N(\underline{x}\underline{b}_1,\ (\tau c_{\delta,1})^{-1})$$

where $c_{\delta,1}^{-1} = \underline{x} C_1^{-1} \underline{x}^T$.

So the marginal posterior for δ is such that

$$\frac{\delta - \underline{x}\underline{b}_1}{\sqrt{v_1/c_{\delta,1}}} = \frac{\delta - \underline{x}\underline{b}_1}{\sqrt{\underline{x} V_1 \underline{x}^T}} \sim t_{d_1},$$

where $V_1 = v_1 C_1^{-1}$.

1.3.3 Prediction

Very often our purpose in using a regression is to be able to make predictions. That is, we want to find the distribution of a future observation on the dependent variables, or perhaps a collection of future observations. In the case of the normal linear model this is usually straightforward.

Suppose that we are going to make a new observation on Y and the covariate values will be $x_{0,1}, \ldots, x_{0,p}$. We arrange these covariate values into a vector $\underline{x}_0$. For convenience, we regard this as a *row* vector rather than the more usual column vector. That is, its dimension is $(1 \times p)$ rather than $(p \times 1)$. Then we can write

$$Y = \underline{x}_0 \underline{\beta} + \varepsilon,$$

where ε is a *new* error which is conditionally independent of any data which we have observed, given τ, and therefore also conditionally independent of the unknown value of $\underline{\beta}$, given τ. Given τ, the distribution of ε is $N(0,\ \tau^{-1})$. Let us assume that our distribution for $\underline{\beta}$ is normal. Then, given τ, the distribution of Y is normal with mean given by the mean of $\underline{x}_0 \underline{\beta}$ and variance given by the sum of the variance of $\underline{x}_0 \underline{\beta}$ and the variance of ε.

Suppose that we are making a *posterior* prediction. That is, we are making our prediction after we have observed some data and our conditional posterior distribution for $\underline{\beta} \mid \tau$ is $N(\underline{b}_1,\ \tau^{-1} C_1^{-1})$. Then we can write

$$\begin{aligned} Y \mid \tau &\sim N(\underline{x}_0 \underline{b}_1,\ \underline{x}_0 [C_1 \tau]^{-1} \underline{x}_0^T + \tau^{-1}) \\ &\sim N(\underline{x}_0 \underline{b}_1,\ [c_p \tau]^{-1}), \end{aligned}$$

where

$$c_p = \{1 + \underline{x}_0 C_1^{-1} \underline{x}_0^T\}^{-1}.$$

In the conjugate case, where our posterior distribution for τ is $\tau \sim \mathrm{Ga}(d_1/2,\ d_1 v_1/2)$, it follows that the marginal distribution for Y is given by

$$\frac{Y - \underline{x}_0 \underline{b}_1}{\sqrt{v_1 / c_p}} = \frac{Y - \underline{x}_0 \underline{b}_1}{\sqrt{v_1 + \underline{x}_0 V_1 \underline{x}_0^T}} \sim t_{d_1}. \tag{1.8}$$

This is our predictive distribution for the new observation Y. It includes both the uncertainty due to our lack of knowledge of the model parameters and our uncertainty associated with the new error.

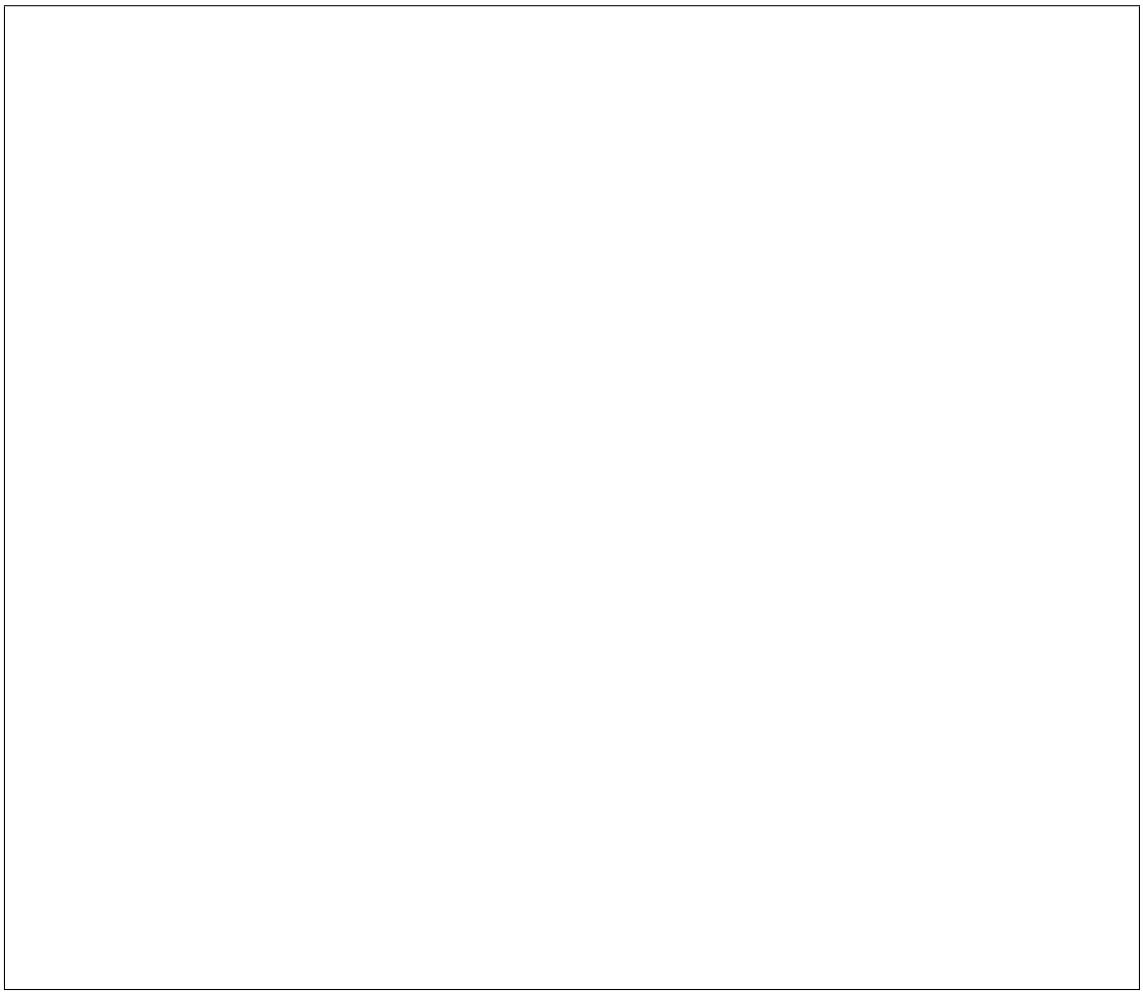
More generally we might want a joint predictive distribution for a vector $\underline{Y}$ of new observations with sampling distribution $N_m(X_0 \beta,\ \tau^{-1} I)$, where I is an identity matrix and the covariate values for the i^{th} element of $\underline{Y}$ give the i^{th} row of X_0. Then

$$\begin{aligned} \underline{Y} \mid \tau &\sim N_m(X_0 \underline{b}_1,\ X_0 [C_1 \tau]^{-1} X_0^T + \tau^{-1} I) \\ &\sim N_m(X_0 \underline{b}_1,\ [C_p \tau]^{-1}), \end{aligned}$$

where

$$C_p = \{I + X_0 C_1^{-1} X_0^T\}^{-1}.$$

Example



1.3.4 Other cases

We have looked in detail at the conjugate case. We can also analyse linear models with a semi-conjugate prior or with a non-conjugate prior. In the semi-conjugate case we need numerical integration in one dimension, that of τ. In the non-conjugate case we usually need more difficult numerical integration and it is usually easier to use MCMC.

1.4 Practical 1

1.4.1 Abrasion Loss

The data in Table 1.2 are taken from Davies and Goldsmith (1972). They come from an experiment to investigate how the resistance of rubber to abrasion is affected by other properties. These are X_1, its hardness, in degrees Shore, and X_2, its tensile strength (in kg per square cm). The dependent variable Y is abrasion loss in g per hour. This is the weight loss due to abrasion which was measured over a fixed time.

You are to fit a linear regression of Y on X_1 and X_2. The model is

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

where, given τ, the errors ε_i are independent with $\varepsilon_i \sim N(0, \tau^{-1})$.

1. Install the function `linmod`. It is available from the Web page at

Abrasion loss Y	Hardness X_1	Tensile strength X_2
372	45	162
206	55	233
175	61	232
154	66	231
136	71	231
112	71	237
55	81	224
45	86	219
221	53	203
166	60	189
164	64	210
113	68	210
82	79	196
32	81	180
228	56	200
196	68	173
128	75	188
97	83	161
64	88	119
249	59	161
219	71	151
186	80	165
155	82	151
114	89	128
341	51	161
340	59	146
283	65	148
267	74	144
215	81	134
148	86	127

Table 1.2: Abrasion loss data

```
http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/
```

You can install it by copying and pasting or by

```
source("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/linmod.txt")
```

2. The data are available in the file `abrasion.txt` which is available from the Web page. You can read the data using commands such as the following.

```
abrasion<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/abrasion.txt")
loss<-abrasion[,1]
hard<-abrasion[,2]
tens<-abrasion[,3]
```

3. Construct a design matrix X as follows.

```
X<-matrix(c(rep(1,30),hard,tens),ncol=3)
```

4. Our prior distribution is as follows. We give τ a $\mathrm{Ga}(d_0/2,\ d_0v_0/2)$ distribution with $d_0 = 4$ and $v_0 = 1600$. Conditional on τ we give $\underline{\beta} = (\beta_0, \beta_1, \beta_2)^T$ a multivariate normal prior distribution with mean vector $\underline{b}_0 = (150, 0, 0)^T$ and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and we construct V_0 as follows. Consider first a reference value with $x_1 = 60$ and $x_2 = 200$. If we consider the model $\mathrm{E}(Y) = \tilde{\beta}_0 + \beta_1(x_1 - 60) + \beta_2(x_2 - 200)$ we obtain for the parameters $\underline{\tilde{\beta}} = (\tilde{\beta}_0,\ \beta_1,\ \beta_2)^T$ the matrix

$$\tilde{V}_0 = 1600 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 0.25 \end{pmatrix}.$$

Now, since $\underline{\beta} = H\underline{\tilde{\beta}}$ where

$$H = \begin{pmatrix} 1 & -60 & -200 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

we can construct $V_0 = H\tilde{V}_0H^T$ as follows.

```
V0tilde<-matrix(c(1600,0,0,0,400,0,0,0,400),ncol=3)
H<-matrix(c(1,0,0,-60,1,0,-200,0,1),ncol=3)
V0<-H%*%V0tilde%*%t(H)
```

Put all of the elements of the prior together.

```
d0<-4
v0<-1600
b0<-matrix(c(150,0,0),ncol=1)
priorabloss<-list(d=d0,v=v0,b=b0,V=V0)
```

5. Find the posterior.

```
postabloss<-linmod(priorabloss,X,loss)
```

6. Find a 95% posterior predictive interval for the abrasion loss in a new observation with $x_1 = 80$ and $x_2 = 150$.

Ayrshire		Canadian	
Mature	2-yr-old	Mature	2-yr-old
3.74	4.44	3.92	4.29
4.01	4.37	4.95	5.24
3.77	4.25	4.47	4.43
3.78	3.71	4.28	4.00
4.10	4.08	4.07	4.62
4.06	3.90	4.10	4.29
4.27	4.41	4.38	4.85
3.94	4.11	3.98	4.66
4.11	4.37	4.46	4.40
4.25	3.53	5.05	4.33

Table 1.3: Butterfat percentages in milk

```
v1<-postabloss$v
V1<-postabloss$V
x0<-matrix(c(1,80,150),nrow=1)
mean<-x0%*%postabloss$b
var<-v1+x0%*%V1%*%t(x0)
tval<-qt(0.975,postabloss$d)
mean-tval*sqrt(var)
mean+tval*sqrt(var)
```

1.4.2 Butterfat

Table 1.3 shows part of a set of data taken from Sokal and Rohlf (1981). The table shows average butterfat percentages in the milk of forty cows. Twenty of the cows belong to each of two breeds, Ayrshire and Canadian. Within each breed, ten of the cows are mature (i.e. at least five years old) and ten are two-year-olds.

We adopt the following model.

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i$$

where Y_i is the butterfat percentage for cow i. We make the usual assumptions about $\varepsilon_1, \ldots, x_{40}$. That is, given τ, they are independent and $\varepsilon_i \sim N(0, \tau^{-1})$. The explanatory variables are as follows.

- Breed, where $x_{i,1} = -1$ if the breed of cow i is Ayrshire and $x_{i,1} = 1$ if the breed of cow i is Canadian.
- Age, where $x_{i,2} = -1$ if cow i is mature and $x_{i,2} = 1$ if cow i is a 2-year-old.
- Breed by age interaction, $x_{i,3} = x_{i,1}x_{i,2}$.

1. If you have not already done so, install the function `linmod` . (See above).

   ```
   source("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/linmod.txt")
   ```

2. The data are available in the file `butter.txt` which is available from the Web page.

   You can read the data using commands such as the following.

   ```
   butter<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/butter.txt")
   butter<-c(butter[,1],butter[,2],butter[,3],butter[,4])
   ```

3. Construct a design matrix X as follows.

```
z<-rep(10,4)
x0<-rep(1,40)
x1<-rep(c(-1,-1,1,1),z)
x2<-rep(c(-1,1,-1,1),z)
x3<-x1*x2
X<-cbind(x0,x1,x2,x3)
```

4. Our prior distribution is as follows. We give τ a $\mathrm{Ga}(d_0/2,\ d_0v_0/2)$ distribution with $d_0 = 6$ and $v_0 = 0.1$. Conditional on τ we give $\underline{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ a multivariate normal prior distribution with mean vector $\underline{b}_0 = (4, 0, 0, 0)^T$ and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and

$$V_0 = \begin{pmatrix} 40 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 2.5 \end{pmatrix}.$$

   We can construct V_0 as follows.

```
V0<-diag(c(40,10,10,2.5))
```

   Put all of the elements of the prior together.

```
d0<-6
v0<-0.1
b0<-matrix(c(4,0,0,0),ncol=1)
priorbutter<-list(d=d0,v=v0,b=b0,V=V0)
```

5. Find the posterior.

```
postbutter<-linmod(priorbutter,X,butter)
```

6. Find a 90% posterior interval for the mean butterfat percentage for 2-yr-old Ayrshire cows.

```
v1<-postbutter$v
V1<-postbutter$V
x0<-matrix(c(1,-1,1,-1),nrow=1)
mean<-x0%*%postbutter$b
var<-x0%*%V1%*%t(x0)
tval<-qt(0.95,postbutter$d)
mean-tval*sqrt(var)
mean+tval*sqrt(var)
```

1.5 Exercises

1. Table 1.4 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 are available in the file `height.txt` on the Web page.

   You can read the data using commands such as the following.

```
eleven<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/height.txt")
height<-eleven[,1]
weight<-eleven[,2]
```

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 1.4: Heights and weights of eleven-year-old girls

(a) You should work in terms of the logarithms of both height and weight. So, let Y be the natural logarithm of the weight and X be the natural logarithm of the height. Calculate these and plot a graph to show the data.

Our model is

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

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

(b) Our prior distribution is as follows. We give τ a $\text{Ga}(d_0/2,\ d_0 v_0/2)$ distribution with $d_0 = 6$ and $v_0 = 0.02$. Conditional on τ we give $\underline{\beta} = (\alpha, \beta)^T$ a bivariate normal prior distribution with mean vector $\underline{b}_0 = (-10, 3)^T$ and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and

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

Find the posterior distribution. (I.e. explain it as I have explained the prior distribution but with the appropriate parameter values).

(c) 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.

2. Table 1.5 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 are available in the file `salinity.txt` on the Web page.

You can read the data using commands such as the following.

```
bimini<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/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)
```

(a) Our model is

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 1.5: Salinity measurements (parts per thousand)

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

where, given the value of τ, the errors ε_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 τ a Ga$(d_0/2,\ d_0 v_0/2)$ distribution with $d_0 = 4$ and $v_0 = 0.3$. Conditional on τ 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 τC_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

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

Find the posterior distribution. (I.e. explain it as I have explained the prior distribution but with the appropriate parameter values).

(c) Find a 95% posterior interval for the difference in mean salinity between water mass I and water mass II.

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.

1.6 Problems 3

Solutions to all questions are to be submitted in the Homework Letterbox no later than 4.00pm on Wednesday November 28th. Please note that you should give some attention to the presentation of your work. Describe the data, model, prior etc. and explain what you have done. Comment on your conclusions. A listing of the output from a R session with one or two things written on it will not get a very good mark on its own.

In questions 2 and 3, each student is given different data. For this purpose each student is given a reference number according to the table below. Please use the correct data and write your reference number on your work. In these questions you may, of course, use R functions such as `linmod` *for calculations.*

Reference numbers

Browning,	Bethany Megan	11
Bulmer,	Rebecca Louise	12
Chaffey,	Adam John	13
Cherlin,	Svetlana	14
Clawson,	Rebecca	15
Consul,	Juliana Iworikumo	16
Dickens,	Jordan Mark	17
Goodall,	Elizabeth Adeline	18
Halliwell,	James William	19
Jones,	Dean Robert Matthew	20
Moffatt,	Joseph Michael	21
Mossop,	Helen	22
Sofro,	A'Yunin	23
Sutherland,	Fiona	24
Varey,	Emma Catherine	25
Wong,	Goldie Sin Man	26

Problems

1. Prior Elicitation

   Some household contents insurance policies require an estimate to be made of what it would cost to replace the existing contents. Suppose that a person has a large collection of books. We might attempt to predict the replacement cost of all of the books by looking at a sample. We might improve this prediction by taking into account an auxiliary variable such as the width of the spine of the book. (We might also distinguish between hardback and paperback books so suppose that we are only considering hardback books). Let C_i be the replacement cost, in £, of book i, and let w_i be its spine width in mm.

   Let

   $$Y_i = \log_e(C_i) \qquad \text{and} \qquad x_i = \log_e(w_i).$$

   It is believed that Y is related to X by

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

   where Y_i and x_i refer to book i for $i = 1, \ldots, n$, $\varepsilon_i \sim N(0, \tau^{-1})$ and $\varepsilon_1, \ldots, \varepsilon_n$ are conditionally independent given τ.

   We give α and β a bivariate normal prior distribution. Find the parameters of this distribution based on the following prior judgments.

   Suppose that we could observe a very large number of books, each of which has a spine $w = 20$mm wide, and a very large number of books, each of which has a spine $w = 30$mm wide. Let the median replacement costs at these two spine widths be M_{20} and M_{30} respectively.

Our prior median for M_{20} is 25 and our prior median for M_{30} is 35. Our prior upper quartile for M_{20} is 40 and our prior upper quartile for M_{30} is 55.

Let

$$m_{20} = \log_e(M_{20}) \quad \text{and} \quad m_{30} = \log_e(M_{30}).$$

Our prior correlation for m_{20} and m_{30} is 0.75.

Find the prior means, prior variances and prior covariance of $\alpha,\ \beta$.

(10 marks)

2. Lowering blood pressure during surgery

It is sometimes necessary to lower a patient's blood pressure during surgery, using a hypotensive drug. The length of time over which the drug is administered varies and therefore so does the total dose. This, in turn, might affect the time it takes for the patient's blood pressure to return to normal.

The data provided are as follows, for $n = 53$ patients.

- The natural logarithm of the recovery time, T, in minutes.
- The natural logarithm of the dose, d, in milligrams.
- The average systolic blood pressure, b, in millimetres of mercury, during administration.

Let $Y = \ln(T)$, $x_1 = \ln(d) - 5$ and $x_2 = b - 60$. We will use a regression model with

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where $\beta_0, \beta_1, \beta_2$ are unknown parameters and, conditional on the values of the parameters, $\varepsilon_1, \ldots, \varepsilon_{53}$ are independent with $\varepsilon_i \sim N(0, \tau^{-1})$.

Our prior distribution is as follows.

We give τ a gamma prior, $\tau \sim \text{Ga}(1.5,\ 0.6)$. Conditional on τ we give $\underline{\beta} = (\beta_0, \beta_1, \beta_2)^T$ a multivariate normal prior distribution with mean vector $\underline{b}_0 = (3.0, -0.03, 0.5)^T$ and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and

$$V_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 10^{-4} & 0 \\ 0 & 0 & 0.04 \end{pmatrix}.$$

You can read the data using a command such as the following.

```
surgery<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/surgerydata.txt")
```

There are thirty columns.

- The log doses $\ln(d)$ are in column 1.
- The blood pressures b are in column 2.
- Your log recovery times t are in the column corresponding to your reference number. For example, if your reference number is 20 then your data are in column 20.

(a) Find the posterior distribution of $\beta_0, \beta_1, \beta_2, \tau$ (in the same form as the prior distribution).

(4 marks)

(b) Find and plot the posterior predictive probability density of the logarithm of the recovery time for a patient with log dose 4.0 and blood pressure 70 during administration.

(4 marks)

(c) Find and plot the posterior predictive probability density of the recovery time for a patient with log dose 4.0 and blood pressure 70 during administration.

(4 marks)

(d) Present, explain and comment on your findings clearly.

(8 marks)

Hint: You can use the following R commands to build the design matrix.

```
x1<-surgery[,1]-5
x2<-surgery[,2]-60
x0<-rep(1,53)
X<-cbind(x0,x1,x2)
```

3. Yields of barley

An experiment was conducted to investigate the effect of manure on the yield of barley. Four different levels of manure were compared: 1: no manure, 2: 0.01 tons per acre, 3: 0.02 tons per acre, 4: 0.04 tons per acre. Three different varieties of barley were used. The experimental plots were arranged in six blocks. (A "block" is an area of land).

You can read the data using a command such as the following.

```
barley<-read.table("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas8303/splitdata.txt")
```

There are thirty columns. Your barley yields y are in the column corresponding to your reference number. For example, if your reference number is 20 then your data are in column 20. Columns 1, 2 and 3 contain the block, variety and manure level respectively.

You can construct a suitable design matrix using the following R commands.

```
block<-barley[,1]
variety<-barley[,2]
manure<-barley[,3]
X<-matrix(nrow=72,ncol=11)
for (col in 1:4)
    {X[,col]<-ifelse(manure==col,1,0)
     }
b<-rep(12,6)
X[,5]<-rep(c(1, 1, 1,-1,-1,-1),b)
X[,6]<-rep(c(2,-1,-1, 0, 0, 0),b)
X[,7]<-rep(c(0, 1,-1, 0, 0, 0),b)
X[,8]<-rep(c(0, 0, 0, 2,-1,-1),b)
X[,9]<-rep(c(0, 0, 0, 0, 1,-1),b)
v<-rep(4,3)
x<-rep(c(2,-1,-1),v)
X[,10]<-rep(x,6)
x<-rep(c(0, 1,-1),v)
X[,11]<-rep(x,6)
```

The first four columns of X correspond to the four levels of manure. Columns 5-9 are for the block effects. (There are five degrees of freedom between the six blocks). Columns 10-11 are for the variety effects. (There are two degrees of freedom between the three varieties). (We could also fit interaction effects but we will leave that for now).

Let the parameters corresponding to the eleven columns of X be $\beta_1, \ldots, \beta_{11}$. Then the mean yield, $\mu_{m,b,v}$, for manure level m in block b with variety v is defined as follows.

$$\mu_{m,b,v} = \beta_m + \sum_{j=5}^{9} \beta_j w_{b,j} + \sum_{j=10}^{11} \beta_j z_{v,j}$$

Here $w_{b,j}$ and $z_{v,j}$ are defined as follows.

$w_{b,j}$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$
$b=1$	1	2	0	0	0
$b=2$	1	−1	1	0	0
$b=3$	1	−1	−1	0	0
$b=4$	−1	0	0	2	0
$b=5$	−1	0	0	−1	1
$b=6$	−1	0	0	−1	−1

$z_{v,j}$	$j=10$	$j=11$
$v=1$	2	0
$v=2$	−1	1
$v=3$	−1	−1

The actual yield for for manure level m in block b with variety v is

$$y_{m,b,v} = \mu_{m,b,v} + \varepsilon_{m,b,v}$$

where $\varepsilon_{m,b,v} \sim N(0, \tau^{-1})$ and $\varepsilon_{m,b,v}$ is independent of $\varepsilon_{m',b',v'}$ unless $(m,b,v) = (m',b',v')$.

Our prior distribution is as follows.

We give τ a gamma prior, $\tau \sim \text{Ga}(d_0/2,\ d_0 v_0/2)$ with $d_0 = 2.1$ and $v_0 = 250$. Conditional on τ we give $\underline{\beta} = (\beta_0, \ldots, \beta_{11})^T$ a multivariate normal prior distribution with mean vector

$$\underline{b}_0 = 100(1,1,1,1,0,0,0,0,0,0,0,0)^T$$

and precision matrix τC_0 where $C_0 = (V_0/v_0)^{-1}$ and

$$V_0 = \frac{1}{6}\begin{pmatrix} 24 & 12 & 12 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & 24 & 12 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & 12 & 24 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & 12 & 12 & 24 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}.$$

You can construct V_0 in R, for example using the following commands.

```
V0<-matrix(0,nrow=11,ncol=11)
V0[1:4,1:4]<-matrix(2,nrow=4,ncol=4)+diag(2,4)
V0[5:11,5:11]<-diag(c(2,2,6,2,6,1,3))/6
```

(a) Find the posterior distribution of $\beta_0, \ldots, \beta_{11}, \tau$ (in the same form as the prior distribution).

(6 marks)

(b) Find a symmetric 95% posterior interval for the mean yield for Manure level 1 in Block 1 with Variety 1.

(6 marks)

(c) Present, explain and comment on your findings clearly.

(8 marks)