9. Linear models and regression

AFM Smith

Objective

To illustrate the Bayesian approach to fitting normal and generalized linear models.

Recommended reading

- Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society B*, **34**, 1–41.

- Broemeling, L.D. (1985). *Bayesian Analysis of Linear Models*, Marcel-Dekker.

- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, Chapter 8.

- Wiper, M.P., Pettit, L.I. and Young, K.D.S. (2000). Bayesian inference for a Lanchester type combat model. *Naval Research Logistics*, **42**, 609–633.

Introduction: the multivariate normal distribution

Definition 22

A random variable $\mathbf{X} = (X_1, \ldots, X_k)^T$ is said to have a *multivariate normal* distribution with mean $\boldsymbol{\mu}$ and variance / covariance matrix $\boldsymbol{\Sigma}$ if

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \quad \text{for } \mathbf{x} \in \mathbb{R}^k.$$

In this case, we write $\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The following properties of the multivariate normal distribution are well known.

i. Any subset of $\mathbf{X}$ has a (multivariate) normal distribution.

ii. Any linear combination $\sum_{i=1}^{k} \alpha_i X_i$ is normally distributed

iii. If $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$ is a linear transformation of $\mathbf{X}$, then $\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\Sigma\mathbf{B}^T\right)$.

iv. If $\mathbf{X} = \left. \begin{array}{c} \mathbf{X}_1 \\ \mathbf{X}_2 \end{array} \right| \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{array}, \left(\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right) \right)$ then the conditional density of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is

$$\mathbf{X}_1|\mathbf{x}_2, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

The multivariate normal likelihood function

Suppose that we observe a sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of data from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the likelihood function is given by

$$
\begin{aligned}
l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}) &= \frac{1}{(2\pi)^{nk/2}|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right) \\
&\propto \frac{1}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\left[\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}})\right]\right) \\
&\propto \frac{1}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\left[\mathrm{tr}\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}\right) + n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}})\right]\right)
\end{aligned}
$$

where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\mathbf{S} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ and $\mathrm{tr}(\mathbf{M})$ represents the trace of the matrix $\mathbf{M}$.

It is possible to carry out Bayesian inference with conjugate priors for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. We shall consider two cases which reflect different levels of knowledge about the variance-covariance matrix $\boldsymbol{\Sigma}$.

Conjugate Bayesian inference for the multivariate normal distribution I: $\mathbf{\Sigma} = \frac{1}{\phi}\mathbf{C}$

Firstly, consider the case where the variance-covariance matrix is known up to a constant, i.e. $\mathbf{\Sigma} = \frac{1}{\phi}\mathbf{C}$ where $\mathbf{C}$ is a known matrix. Then, we have $\mathbf{X}|\boldsymbol{\mu}, \phi \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{\phi}\mathbf{C}\right)$ and the likelihood function is

$$l(\boldsymbol{\mu}, \phi|\mathbf{x}) \propto \phi^{\frac{nk}{2}} \exp\left(-\frac{\phi}{2}\left[\operatorname{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right) + n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T \mathbf{C}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}})\right]\right).$$

Analogous to the univariate case, it can be seen that a multivariate normal-gamma prior distribution is conjugate.

The multivariate normal gamma distribution

We say that $\boldsymbol{\mu}, \phi$ have a *multivariate normal gamma prior* with parameters $\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}$ if

$$\begin{aligned} \boldsymbol{\mu}|\phi &\sim \mathcal{N}\left(\mathbf{m}, \frac{1}{\phi}\mathbf{V}\right) \\ \phi &\sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right). \end{aligned}$$

In this case, we write $\boldsymbol{\mu}, \phi \sim \mathcal{N}\mathcal{G}\left(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$.

The marginal distribution of $\boldsymbol{\mu}$

In this case, the marginal distribution of $\boldsymbol{\mu}$ is a multivariate, non-central t distribution.

Definition 23

A (k-dimensional) random variable, $\mathbf{T} = (T_1, \ldots, T_k)$, has a *multivariate t distribution* with parameters $d, \boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T$ if

$$f(\mathbf{t}) = \frac{\Gamma\left(\frac{d+k}{2}\right)}{(\pi d)^{\frac{k}{2}} |\boldsymbol{\Sigma}_T|^{1/2} \Gamma\left(\frac{d}{2}\right)} \left(1 + \frac{1}{d}(\mathbf{t} - \boldsymbol{\mu}_T)^T \boldsymbol{\Sigma}_T^{-1} (\mathbf{t} - \boldsymbol{\mu}_T)\right)^{-\frac{d+k}{2}}.$$

In this case, we write $\mathbf{T} \sim \mathcal{T}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T, d)$.

The following theorem gives the density of $\boldsymbol{\mu}$.

Theorem 34

Let $\boldsymbol{\mu}, \phi \sim \mathcal{NG}\left(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$. Then the marginal density of $\boldsymbol{\mu}$ is $\boldsymbol{\mu} \sim \mathcal{T}\left(\mathbf{m}, \frac{b}{a}\mathbf{V}, a\right)$.

Proof

$$\begin{aligned}
p(\boldsymbol{\mu}) &= \int_0^\infty p(\boldsymbol{\mu}, \phi)\, d\phi \\
&= \int_0^\infty p(\boldsymbol{\mu}|\phi) p(\phi)\, d\phi \\
&= \int_0^\infty \frac{1}{(2\pi)^{\frac{k}{2}} |\mathbf{V}|^{\frac{1}{2}}} \exp\left(-\frac{\phi}{2} (\boldsymbol{\mu} - \mathbf{m})^T \mathbf{V}^{-1} (\boldsymbol{\mu} - \mathbf{m}) \right) \frac{\left(\frac{b}{2}\right)^{\frac{a}{2}}}{\Gamma\left(\frac{a}{2}\right)} \phi^{\frac{a}{2}-1} \exp\left(-\frac{b}{2}\phi \right)\, d\phi \\
&= \frac{1}{(2\pi)^{\frac{k}{2}} |\mathbf{V}|^{\frac{1}{2}}} \frac{\left(\frac{b}{2}\right)^{\frac{a}{2}}}{\Gamma\left(\frac{a}{2}\right)} \int_0^\infty \phi^{\frac{a+k}{2}-1} \exp\left(-\frac{\phi}{2} \left[b + (\boldsymbol{\mu} - \mathbf{m})^T \mathbf{V}^{-1} (\boldsymbol{\mu} - \mathbf{m}) \right] \right)\, d\phi \\
&= \frac{1}{(2\pi)^{\frac{k}{2}} |\mathbf{V}|^{\frac{1}{2}}} \frac{\left(\frac{b}{2}\right)^{\frac{a}{2}}}{\Gamma\left(\frac{a}{2}\right)} \Gamma\left(\frac{a+k}{2}\right) \left(\frac{b + (\boldsymbol{\mu} - \mathbf{m})^T \mathbf{V}^{-1} (\boldsymbol{\mu} - \mathbf{m})}{2} \right)^{-\frac{a+k}{2}}
\end{aligned}$$

and, reordering terms proves the result. ■

The posterior distribution of $\boldsymbol{\mu}, \phi$

Theorem 35

Let $\mathbf{X}|\boldsymbol{\mu}, \phi \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{\phi}\mathbf{C}\right)$ and assume the prior distributions $\boldsymbol{\mu}|\phi \sim \mathcal{N}\left(\mathbf{m}, \frac{1}{\phi}\mathbf{V}\right)$ and $\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$. Then, given sample data $\mathbf{x}$, we have

$$
\begin{aligned}
\boldsymbol{\mu}|\mathbf{x}, \phi &\sim \mathcal{N}\left(\mathbf{m}^\star, \frac{1}{\phi}\mathbf{V}^\star\right) \\
\phi|\mathbf{x} &\sim \mathcal{G}\left(\frac{a^\star}{2}, \frac{b^\star}{2}\right) \quad \text{where} \\
\mathbf{V}^\star &= \left(\mathbf{V}^{-1} + n\mathbf{C}^{-1}\right)^{-1} \\
\mathbf{m}^\star &= \mathbf{V}^\star\left(\mathbf{V}^{-1}\mathbf{m} + n\mathbf{C}^{-1}\bar{\mathbf{x}}\right) \\
a^\star &= a + nk \\
b^\star &= b + \mathrm{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right) + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} + n\bar{\mathbf{x}}^T\mathbf{C}^{-1}\bar{\mathbf{x}} - \mathbf{m}^\star\mathbf{V}^{\star-1}\mathbf{m}^\star.
\end{aligned}
$$

Proof Exercise. The proof is analogous to the univariate case. ■

A simplification

In the case where $\mathbf{V} \propto \mathbf{C}$, we have a simplified result, similar to that for the univariate case.

Theorem 36

Let $\boldsymbol{\mu}, \phi \sim \mathcal{NG}\left(\mathbf{m}, \alpha\mathbf{C}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$. Then,

$$\begin{aligned}
\boldsymbol{\mu}|\phi, \mathbf{x} &\sim \mathcal{N}\left(\frac{\alpha\mathbf{m} + n\bar{\mathbf{x}}}{\alpha + n}, \frac{1}{(\alpha + n)\phi}\mathbf{C}\right) \\
\phi|\mathbf{x} &\sim \mathcal{G}\left(\frac{a+n}{2}, \frac{b + \mathrm{tr}\left(\left(\mathbf{S} + \frac{\alpha n}{\alpha+n}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})^T\right)\mathbf{C}^{-1}\right)}{2}\right).
\end{aligned}$$

Proof This follows from the previous theorem substituting $\mathbf{V} = \frac{1}{\alpha}\mathbf{C}$. ∎

Results with the reference prior

Theorem 37

Given the prior $p(\boldsymbol{\mu}, \phi) \propto \frac{1}{\phi}$, then the posterior distribution is

$$\begin{aligned}
p(\boldsymbol{\mu}, \phi | \mathbf{x}) &\propto \phi^{\frac{nk}{2}-1} \exp\left(-\frac{\phi}{2}\left[\mathrm{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right) + n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T \mathbf{C}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}})\right]\right) \\
\boldsymbol{\mu} | \mathbf{x}, \phi &\sim \mathcal{N}\left(\bar{\mathbf{x}}, \frac{1}{n\phi}\mathbf{C}\right) \\
\phi | \mathbf{x} &\sim \mathcal{G}\left(\frac{(n-1)k}{2}, \frac{\mathrm{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right)}{2}\right) \\
\boldsymbol{\mu} | \mathbf{x} &\sim \mathcal{T}\left(\bar{\mathbf{x}}, \frac{\mathrm{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right)}{n(n-1)k}\mathbf{C}, (n-1)k\right).
\end{aligned}$$

Proof $\boldsymbol{\mu}, \phi | \mathbf{x} \sim \mathcal{N}\mathcal{G}\left(\bar{\mathbf{x}}, \frac{1}{n}\mathbf{C}, \frac{(n-1)k}{2}, \frac{\mathrm{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right)}{2}\right)$ and the rest follows. ▮

Conjugate inference for the multivariate normal distribution II: $\mathbf{\Sigma}$ unknown

In this case, it is useful to reparameterize the normal distribution in terms of the precision matrix $\mathbf{\Phi} = \mathbf{\Sigma}^{-1}$ when the normal likelihood function becomes

$$l(\boldsymbol{\mu}, \mathbf{\Phi}|\mathbf{x}) \propto |\mathbf{\Phi}|^{\frac{n}{2}} \exp\left(-\frac{1}{2}\left[\operatorname{tr}\left(\mathbf{S}\mathbf{\Phi}\right) + n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T \mathbf{\Phi}(\boldsymbol{\mu} - \bar{\mathbf{x}})\right]\right)$$

It is clear that a conjugate prior for $\boldsymbol{\mu}$ and $\mathbf{\Phi}$ must take a similar form to the likelihood. This is a normal-Wishart distribution.

The normal Wishart distribution

Definition 24

A $k \times k$ dimensional symmetric, positive definite random variable $\mathbf{W}$ is said to have a *Wishart distribution* with parameters d and $\mathbf{V}$ if

$$f(\mathbf{W}) = \frac{|\mathbf{W}|^{\frac{d-k-1}{2}}}{2^{\frac{dk}{2}}|V|^{\frac{d}{2}}\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}\Gamma\left(\frac{d+1-i}{2}\right)}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{W}\right)\right)$$

where $d > k-1$. In this case, $E[\mathbf{W}] = d\mathbf{V}$ and we write $\mathbf{W} \sim \mathcal{W}(d, \mathbf{V})$.

If $\mathbf{W} \sim \mathcal{W}(d, \mathbf{V})$, then the distribution of $\mathbf{W}^{-1}$ is said to be an *inverse Wishart distribution*, $\mathbf{W}^{-1} \sim \mathcal{IW}\left(d, \mathbf{V}^{-1}\right)$ with mean $E\left[\mathbf{W}^{-1}\right] = \frac{1}{d-k-1}\mathbf{V}^{-1}$.

Theorem 38

Suppose that $\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Phi} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Phi}^{-1}\right)$ and let $\boldsymbol{\mu}|\boldsymbol{\Phi} \sim \mathcal{N}\left(\mathbf{m}, \frac{1}{\alpha}\boldsymbol{\Phi}^{-1}\right)$ and $\boldsymbol{\Phi} \sim \mathcal{W}(d, \mathbf{W})$. Then:

$$\begin{aligned}
\boldsymbol{\mu}|\boldsymbol{\Phi}, \mathbf{x} &\sim \mathcal{N}\left(\frac{\alpha\mathbf{m} + n\bar{\mathbf{x}}}{\alpha + n}, \frac{1}{\alpha + n}\boldsymbol{\Phi}^{-1}\right) \\
\boldsymbol{\Phi}|\mathbf{x} &\sim \mathcal{W}\left(d + nk, \mathbf{W}^{-1} + \mathbf{S} + \frac{\alpha n}{\alpha + n}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})^T\right)
\end{aligned}$$

Proof Exercise. ■

We can also derive a limiting prior distribution by letting $d \rightarrow 0$ when $p(\boldsymbol{\Phi}) \propto |\boldsymbol{\Phi}|^{\frac{k+1}{2}}$ when the posterior distribution is

$$\boldsymbol{\mu}|\boldsymbol{\Phi}, \mathbf{x} \sim \mathcal{N}\left(\bar{\mathbf{x}}, \frac{1}{n}\boldsymbol{\Phi}^{-1}\right) \quad \boldsymbol{\Phi}|\mathbf{x} \sim \mathcal{W}\left(n(k-1), \mathbf{S}\right).$$

Semi-conjugate inference via Gibbs sampling

The conjugacy assumption that the prior precision of $\boldsymbol{\mu}$ is proportional to the model precision $\mathbf{\Sigma}$ is very strong in many cases. Often, we may simply wish to use a prior distribution of form $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$ where $\mathbf{m}$ and $\mathbf{V}$ are known and a Wishart prior for $\mathbf{\Phi}$, say $\boldsymbol{\phi} \sim \mathcal{W}(d, \mathbf{W})$ as earlier.

In this case, the conditional posterior distributions are

$$\begin{aligned} \boldsymbol{\mu}|\mathbf{\Phi}, \mathbf{x} &\sim \mathcal{N}\left(\left(\mathbf{V}^{-1} + n\mathbf{\Phi}\right)^{-1}\left(\mathbf{V}^{-1}\mathbf{m} + n\mathbf{\Phi}\bar{\mathbf{x}}\right), \left(\mathbf{V}^{-1} + n\mathbf{\Phi}\right)^{-1}\right) \\ \mathbf{\Phi}|\boldsymbol{\mu}, \mathbf{x} &\sim \mathcal{W}\left(d + n, \mathbf{W}^{-1} + \mathbf{S} + n(\boldsymbol{\mu} - \bar{x})(\boldsymbol{\mu} - \bar{\mathbf{x}})^T\right) \end{aligned}$$

and therefore, it is straightforward to set up a Gibbs sampling algorithm to sample the joint posterior, as in the univariate case.

Aside: sampling the multivariate normal, multivariate t and Wishart distributions

Samplers for the multivariate normal distribution (usually based on the *Cholesky decomposition*) are available in most statistical packages such as `R` or `Matlab`. Sampling the multivariate t distribution is only slightly more complicated. Assume that we wish to sample from $\mathbf{T} \sim \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d})$. Then from Theorem 34, the distribution of $\mathbf{T}$ is the same as the marginal distribution of $\mathbf{T}$ in the two stage model

$$\mathbf{T}|\phi \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{\phi}\boldsymbol{\Sigma}\right) \quad \phi \sim \mathcal{G}\left(\frac{d}{2}, \frac{d}{2}\right).$$

Thus, sampling can be undertaken by first generating values of ϕ and then generating values of $\mathbf{T}$ from the associated normal distribution.

Sampling from a Wishart distribution can be done in a straightforward way if the degrees of freedom is a natural number. Thus, assume $\mathbf{W} \sim \mathcal{W}(d, \mathbf{V})$ where $d \in \mathbb{N}$. Then the following algorithm generates a Wishart variate.

```
1. Simulate $\mathbf{z}_1, \dots, \mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$

2. Define $\mathbf{W} = \sum_{i=1}^{k} \mathbf{z}_i \mathbf{z}_i^T$.
```

Normal linear models

Definition 25

A *normal linear model* is of form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$, and we will assume initially that $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\phi}\mathbf{I}\right)$.

This framework includes regression models, defining, e.g. $\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}^T$ and $\boldsymbol{\theta} = (\alpha, \beta)^T$.

It is easy to see that for such a model, a conjugate, multivariate normal-gamma prior distribution is available.

Conjugate Bayesian inference

Theorem 39

Let $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ and assume a multivariate normal-gamma prior distribution $\boldsymbol{\theta}, \phi \sim \mathcal{NG}\left(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$. Then, the predictive distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{T}\left(\mathbf{X}\mathbf{m}, \frac{b}{a}(\mathbf{X}\mathbf{V}\mathbf{X}^T + \mathbf{I}), a\right)$$

and the posterior distribution of $\boldsymbol{\theta}, \phi$ given $\mathbf{y}$ is

$$\begin{aligned}
\boldsymbol{\theta}, \phi|\mathbf{y} &\sim \mathcal{NG}\left(\mathbf{m}^\star, \mathbf{V}^{\star-1}, \frac{a^\star}{2}, \frac{b^\star}{2}\right) \quad \text{where} \\
\mathbf{m}^\star &= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right) \\
\mathbf{V}^\star &= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1} \\
a^\star &= a + n \\
b^\star &= b + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}^{\star T}\mathbf{V}^{\star-1}\mathbf{m}^\star
\end{aligned}$$

Proof First we shall prove the predictive distribution formula. We have $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ and therefore, the distribution of $\mathbf{y}|\phi$ is

$$\mathbf{y}|\phi \sim \mathcal{N}\left(\mathbf{X}\mathbf{m}, \frac{1}{\phi}\left(\mathbf{X}\mathbf{V}\mathbf{X}^T + \mathbf{I}\right)\right)$$

and the joint distribution of $\mathbf{y}$ and ϕ is multivariate normal-gamma

$$\mathbf{y}, \phi \sim \mathcal{N}\mathcal{G}\left(\mathbf{X}\mathbf{m}, \left(\mathbf{X}\mathbf{V}\mathbf{X}^T + \mathbf{I}\right)^{-1}, \frac{a}{2}, \frac{b}{2}\right)$$

and therefore, the marginal distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{T}\left(\mathbf{X}\mathbf{m}, \frac{b}{a}(\mathbf{X}\mathbf{V}\mathbf{X}^T + \mathbf{I}), a\right).$$

Now we shall evaluate the posterior distribution

$$
\begin{aligned}
p(\boldsymbol{\theta}, \phi | \mathbf{y}) &\propto \phi^{\frac{a+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + (\boldsymbol{\theta} - \mathbf{m})^T \mathbf{V}^{-1}(\boldsymbol{\theta} - \mathbf{m})\right]\right) \phi^{\frac{n}{2}} \exp\left(-\frac{\phi}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\right) \\
&\propto \phi^{\frac{a+n+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + \boldsymbol{\theta}^T\left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\left(\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right) + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + \boldsymbol{\theta}^T\mathbf{V}^{\star-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\left(\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right) + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + \boldsymbol{\theta}^T\mathbf{V}^{\star-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{V}^{\star-1}\mathbf{V}^\star\left(\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right) + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + \boldsymbol{\theta}^T\mathbf{V}^{\star-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{V}^{\star-1}\mathbf{m}^\star + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + (\boldsymbol{\theta} - \mathbf{m}^\star)^T\mathbf{V}^{\star-1}(\boldsymbol{\theta} - \mathbf{m}^\star) + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}^\star\mathbf{V}^{\star-1}\mathbf{m}^\star\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b^\star + (\boldsymbol{\theta} - \mathbf{m}^\star)^T\mathbf{V}^{\star-1}(\boldsymbol{\theta} - \mathbf{m}^\star)\right]\right)
\end{aligned}
$$

which is the kernel of the required normal-gamma distribution. ■

Interpretation of the posterior mean

We have

$$
\begin{aligned}
E[\boldsymbol{\theta}|\mathbf{y}] &= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right) \\
&= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right) \\
&= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\theta}} + \mathbf{V}^{-1}\mathbf{m}\right)
\end{aligned}
$$

where $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the maximum likelihood estimator. Thus, this expression may be interpreted as a weighted average of the prior estimator and the MLE, with weights proportional to precisions, as we can recall that, conditional on ϕ, the prior variance was $\frac{1}{\phi}\mathbf{V}$ and that the distribution of the MLE from the classical viewpoint is $\hat{\boldsymbol{\theta}}|\phi \sim \mathcal{N}\left(\boldsymbol{\theta}, \frac{1}{\phi}(\mathbf{X}^T\mathbf{X})^{-1}\right)$.

Relation to ridge regression

The classical least squares regression solution $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ does not exist if $\mathbf{X}^T\mathbf{X}$ is not of full rank. In this case, an often employed technique is to use *ridge regression*, see Hoerl and Kennard (1970).

The ridge regression estimator is the value of which minimizes

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^2 + \alpha^2||\boldsymbol{\theta}||^2$$

and the solution can be shown to be

$$\hat{\theta}_\alpha = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

If we use a Bayesian approach with prior, $\boldsymbol{\mu}|\phi \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\alpha\phi}\mathbf{I}\right)$, then the posterior mean is

$$E[\boldsymbol{\mu}|\mathbf{y}] = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

which is equal to the ridge regression estimate.

Example: an ANOVA type model

Example 65

Consider the ANOVA model: $y_{ij} = \theta_i + \epsilon_{ij}$ where $\epsilon_{ij}|\phi \sim \mathcal{N}\left(0, \frac{1}{\phi}\right)$, for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$.

Thus $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$, $\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}, \ldots, y_{kn_k})^T$,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1_{n_1} & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}$$

and $n = \sum_{i=1}^{k} n_i$ is the model dimension.

If we use conditionally independent normal priors, $\theta_i|\phi \sim \mathcal{N}\left(m_i, \frac{1}{\alpha_i\phi}\right)$ for $i = 1, \ldots, k$ and a gamma prior $\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$, then $\mathbf{m} = (m_1, \ldots, m_k)^T$ and $\mathbf{V} = \begin{pmatrix} \frac{1}{\alpha_1} & 0 & 0 & \ldots & 0 \\ 0 & \frac{1}{\alpha_2} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \frac{1}{\alpha_k} \end{pmatrix}$. Also,

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= \begin{pmatrix} n_1 & 0 & 0 & \ldots & 0 \\ 0 & n_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & n_k \end{pmatrix} \\ \mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1} &= \begin{pmatrix} n_1+\alpha_1 & 0 & 0 & \ldots & 0 \\ 0 & n_2+\alpha_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & n_k+\alpha_k \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
\left(\mathbf{X}^T\mathbf{X}+\mathbf{V}^{-1}\right)^{-1} &= \begin{pmatrix} \frac{1}{n_1+\alpha_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{n_2+\alpha_2} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{n_k+\alpha_k} \end{pmatrix} \\
\mathbf{X}^T\mathbf{y} &= (n_1\bar{y}_{1\cdot},\dots,n_k\bar{y}_{k\cdot})^T \\
\mathbf{V}^{-1}\mathbf{m} &= (\alpha_1 m_1,\dots,\alpha_k m_k)^T \\
\mathbf{X}^T\mathbf{y}+\mathbf{V}^{-1}\mathbf{m} &= (n_1\bar{y}_{1\cdot}+\alpha_1 m_1,\dots,n_k\bar{y}_{k\cdot}+\alpha_k m_k)^T \\
\left(\mathbf{X}^T\mathbf{X}+\mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{y}+\mathbf{V}^{-1}\mathbf{m}\right) &= \left(\frac{n_1\bar{y}_{1\cdot}+\alpha_1 m_1}{n_1+\alpha_1},\dots,\frac{n_k\bar{y}_{k\cdot}+\alpha_k m_k}{n_k+\alpha_k}\right)^T \quad \text{so}
\end{aligned}$$

$$\boldsymbol{\theta}|\mathbf{y},\phi \sim \mathcal{N}\left(\begin{pmatrix} \frac{n_1\bar{y}_{1\cdot}+\alpha_1 m_1}{n_1+\alpha_1} \\ \frac{n_2\bar{y}_{2\cdot}+\alpha_2 m_2}{n_2+\alpha_2} \\ \vdots \\ \vdots \\ \frac{n_k\bar{y}_{k\cdot}+\alpha_k m_k}{n_k+\alpha_k} \end{pmatrix}, \frac{1}{\phi}\begin{pmatrix} \frac{1}{n_1+\alpha_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{n_2+\alpha_2} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{n_k+\alpha_k} \end{pmatrix}\right).$$

Now we can calculate the posterior distribution of ϕ.

$$\begin{aligned}
\mathbf{y}^T\mathbf{y} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 \\
\mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} &= \sum_{i=1}^{k}\alpha_i m_i^2 \\
\mathbf{m}^{\star T}\mathbf{V}^{\star -1}\mathbf{m}^\star &= \sum_{i=1}^{k}\frac{(n_i\bar{y}_{i\cdot}+\alpha_i m_i)^2}{n_i+\alpha_i} \\
\mathbf{y}^t\mathbf{y}+\mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}-\mathbf{m}^{\star T}\mathbf{V}^{\star -1}\mathbf{m}^\star &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2+\sum_{i=1}^{k}\frac{n_i\alpha_i}{n_i+\alpha_i}(\bar{y}_{i\cdot}-m_i)^2 \text{ so}
\end{aligned}$$

$$\phi|\mathbf{y} \sim \mathcal{G}\left(\frac{a+n}{2}, \frac{b+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2+\sum_{i=1}^{k}\frac{n_i\alpha_i}{n_i+\alpha_i}(\bar{y}_{i\cdot}-m_i)^2}{2}\right)$$

Often we are interested in the differences in group means, e.g. $\theta_1 - \theta_2$. Here we have

$$\theta_1 - \theta_2 | \mathbf{y}, \phi \sim \mathcal{N}\left(\frac{n_1\bar{y}_{1\cdot} + \alpha_1 m_1}{n_1 + \alpha_1} - \frac{n_2\bar{y}_{2\cdot} + \alpha_2 m_2}{n_2 + \alpha_2}, \frac{1}{\phi}\left(\frac{1}{\alpha_1 + n_1} + \frac{1}{\alpha_2 + n_2}\right)\right)$$

and therefore, a posterior, 95% interval for $\theta_1 - \theta_2$ is given by

$$\frac{n_1\bar{y}_{1\cdot} + \alpha_1 m_1}{n_1 + \alpha_1} - \frac{n_2\bar{y}_{2\cdot} + \alpha_2 m_2}{n_2 + \alpha_2} \pm \sqrt{\left(\frac{1}{\alpha_1 + n_1} + \frac{1}{\alpha_2 + n_2}\right)\frac{b + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k}\frac{n_i\alpha_i}{n_i+\alpha_i}(\bar{y}_{i\cdot} - m_i)^2}{a+n}} t_{a+n}(0.975).$$

Limiting results for the linear model

Assume that we use the limiting prior $p(\boldsymbol{\theta}, \phi) \propto \frac{1}{\phi}$. Then, we have

$$
\begin{aligned}
p(\boldsymbol{\theta}, \phi|\mathbf{y}) &\propto \phi^{\frac{n}{2}-1} \exp\left(-\frac{\phi}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\right) \\
&\propto \phi^{\frac{n}{2}-1} \exp\left(-\frac{\phi}{2}\left[\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\right]\right) \\
&\propto \phi^{\frac{n}{2}-1} \exp\left(-\frac{\phi}{2}\left[\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\right]\right) \\
&\propto \phi^{\frac{n}{2}-1} \exp\left(-\frac{\phi}{2}\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\theta}}\right]\right) \\
\boldsymbol{\theta}|\mathbf{y}, \phi &\sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \frac{1}{\phi}(\mathbf{X}^T\mathbf{X})^{-1}\right) \quad \text{and} \quad \phi|\mathbf{y} \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\theta}}}{2}\right) \\
\boldsymbol{\theta}|\mathbf{y} &\sim \mathcal{T}\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}, n-k\right) \quad \text{where } \hat{\sigma}^2 = \frac{\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\theta}}}{n-k}.
\end{aligned}
$$

Note that $\hat{\sigma}^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})$ is the usual classical estimator of σ^2.

In this case, Bayesian credible intervals, estimators etc. will coincide with their classical counterparts.

One should note however that the propriety of the posterior distribution in this case relies on two conditions:

1. $n > k$,

2. $\mathbf{X}^T\mathbf{X}$ is off full rank.

If either of these two conditions is not satisfied, then the posterior distribution will be improper.

Example 66

In Example 65, suppose that we use the reference prior $p(\boldsymbol{\theta}, \phi) \propto \frac{1}{\phi}$. Then, we have

$$\begin{aligned} \boldsymbol{\theta}|\mathbf{y} &\sim \mathcal{N}\left(\begin{pmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \vdots \\ \bar{y}_{k\cdot} \end{pmatrix}, \frac{1}{\phi} \begin{pmatrix} \frac{1}{n_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{n_2} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{n_k} \end{pmatrix} \right) \\ \phi|\mathbf{y} &\sim \mathcal{G}\left(\frac{n-k}{2}, \frac{(n-k)\hat{\sigma}^2}{2} \right) \end{aligned}$$

where $\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$ is the classical variance estimate for this problem.

A 95% posterior interval for $\theta_1 - \theta_2$ is given by $\bar{y}_{1\cdot} - \bar{y}_{2\cdot} \pm \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{n-k}(0.975)$ which is equal to the usual, classical interval.

Simple linear regression

Example 67

Consider the simple linear regression model, $y_i = \alpha + \beta x_i + \epsilon_i$, for $i = 1, \ldots, n$, where $\epsilon_i \sim \mathcal{N}\left(0, \frac{1}{\phi}\right)$ and suppose that we use the limiting prior $p(\alpha, \beta, \phi) \propto \frac{1}{\phi}$. Then, we have

$$
\begin{aligned}
\begin{array}{c} \alpha \\ \beta \end{array} \left| \mathbf{y}, \phi \right. &\sim \mathcal{N}\left(\begin{array}{c} \hat{\alpha} \\ \hat{\beta} \end{array}, \frac{1}{\phi n S_{xx}} \left(\begin{array}{cc} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{array} \right) \right) \\
\phi | \mathbf{y} &\sim \mathcal{G}\left(\frac{n-2}{2}, \frac{S_{yy}(1-r^2)}{2} \right) \\
\begin{array}{c} \alpha \\ \beta \end{array} \left| \mathbf{y} \right. &\sim \mathcal{T}\left(\begin{array}{c} \hat{\alpha} \\ \hat{\beta} \end{array}, \frac{\hat{\sigma}^2}{n} \frac{1}{S_{xx}} \left(\begin{array}{cc} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{array} \right), n-2 \right)
\end{aligned}
$$

where $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ and $\hat{\sigma}^2 = \frac{S_{yy}(1-r^2)}{n-2}$.

Thus, the marginal distributions of α and β are

$$\begin{aligned} \alpha|\mathbf{y} &\sim \mathcal{T}\left(\hat{\alpha}, \hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}, n-2\right) \\ \beta|\mathbf{y} &\sim \mathcal{T}\left(\hat{\beta}, \frac{\hat{\sigma}^2}{S_{xx}}, n-2\right) \end{aligned}$$

and therefore, for example, a 95% credible interval for β is given by

$$\hat{\beta} \pm \frac{\hat{\sigma}}{\sqrt{S_{xx}}} t_{n-2}(0.975)$$

equal to the usual classical interval.

Suppose now that we wish to predict a future observation $y_{new} = \alpha + \beta x_{new} + \epsilon_{new}$. Then

$$\begin{array}{rcl}
E[y_{new}|\phi,\mathbf{y}] & = & \hat{\alpha} + \hat{\beta} x_{new} \\
V[y_{new}|\phi,\mathbf{y}] & = & \frac{1}{\phi}\left(\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} + \frac{x_{new}^2}{nS_{xx}} - 2\frac{x_{new}\bar{x}}{nS_{xx}} + 1\right) \\
 & = & \frac{1}{\phi}\left(\frac{(x_{new}-\bar{x})^2}{nS_{xx}} + \frac{1}{n} + 1\right) \\
y_{new}|\mathbf{y} & \sim & \mathcal{T}\left(\hat{\alpha} + \hat{\beta} x_{new}, \hat{\sigma}^2\left(\frac{(x_{new}-\bar{x})^2}{nS_{xx}} + \frac{1}{n} + 1\right), n-2\right)
\end{array}$$

and thus, a 95% credible interval for y_{new} is

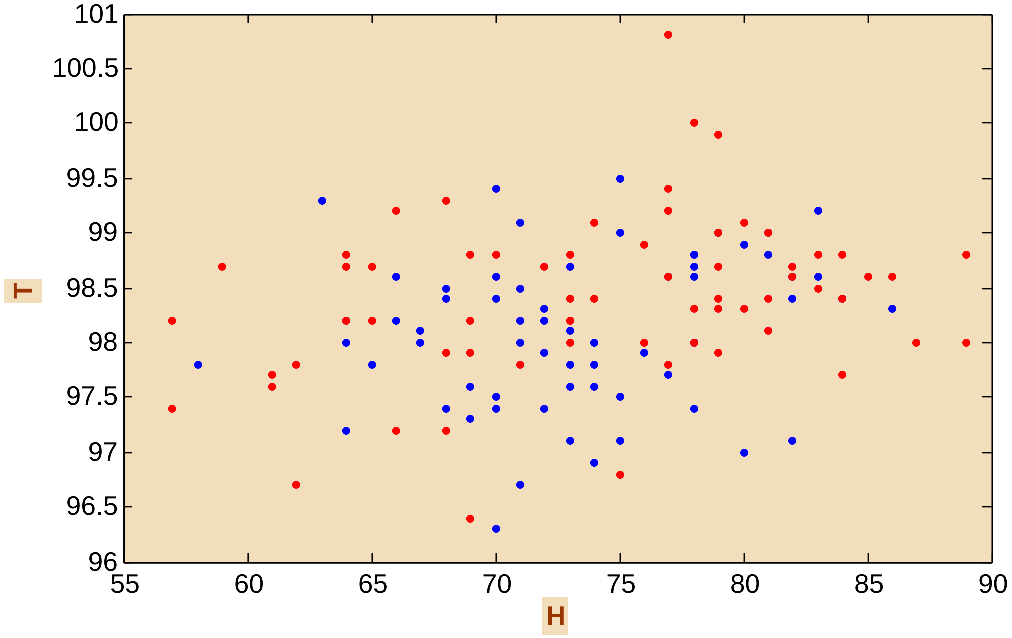
$$\hat{\alpha} + \hat{\beta} x_{new} \pm \hat{\sigma}\left(\frac{(x_{new}-\bar{x})^2}{nS_{xx}} + \frac{1}{n} + 1\right) t_{n-2}(0.975)$$

which coincides with the usual, classical interval.

The normal body temperature example again

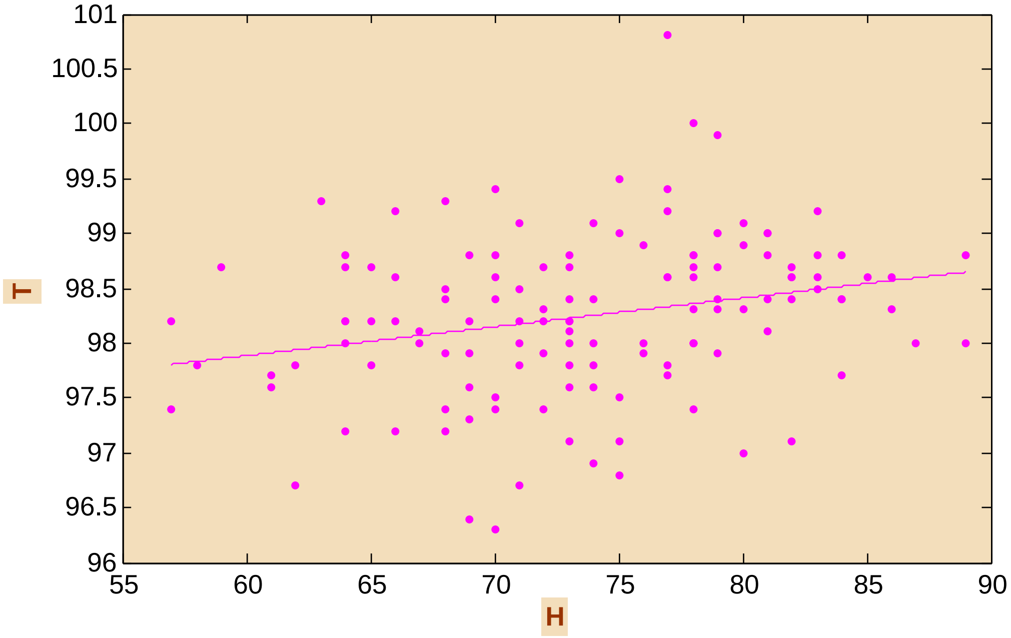
Example 68

In Example 20, we studied the average body temperatures of humans and we saw in Example 21 that these were different for men and women. Now we shall consider the effects of introducing a covariate. In the following diagram, temperature is plotted against heart rate (beats per minute) for male (blue) and female (red) subjects



It is reasonable to assume that body temperature would be related to heart rate. Thus, we might assume the global linear model, $T_i = \alpha + \beta H \epsilon_i$, independent of temperature. Fitting this model with a non-informative prior leads to the estimated regression equation

$$E[T|H, \mathrm{data}] = 96.31 + 0.021H.$$



However, earlier we supposed that gender also influenced body temperature and therefore, a model taking gender into account might be considered. Thus, we assume

$$T_{ij} = \alpha_i + \beta H_{ij} + \epsilon_{ij}$$

where T_{ij} represents the temperature of subject j in group $i = 1$ (men) or $i = 2$ (women) and H_{ij} is the subject's heartrate and $\epsilon_{ij} \sim \mathcal{N}\left(0, \frac{1}{\phi}\right)$ is a random error.

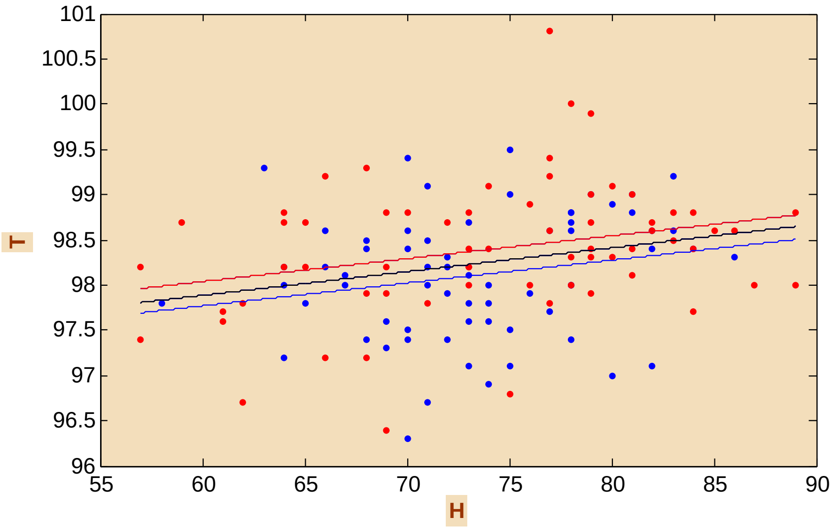
Representing this model in linear form, we have $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta)^T$, $\mathbf{y} = (T_{11}, \ldots, T_{1n_1}, T_{21}, \ldots, T_{2n_2})^T$ where n_1 and n_2 are the numbers of men and women sampled respectively and

$$\mathbf{X}^T = \left(\begin{array}{cccccc} 1 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 \\ H_{11} & \ldots & H_{1n_1} & H_{21} & \ldots & H_{2n_2} \end{array} \right)$$

Assume that we use the prior $p(\alpha_1, \alpha_2, \beta, \phi)$ posterior mean parameter values for $\boldsymbol{\alpha}, \beta$ are

$$\begin{aligned} E[\alpha_1|\mathbf{T}] &= 96.2508 \\ E[\alpha_2|\mathbf{T}] &= 96.5202 \\ E[\beta|\mathbf{T}] &= 0.0253 \end{aligned}$$

The posterior distribution of ϕ is $\phi|\mathbf{y} \sim \mathcal{G}\left(\frac{127}{2}, \frac{62.5}{2}\right)$. The following diagram shows the results of fitting the simple and combined models.



It is interesting to assess whether or not the difference between the sexes is still important. Thus, we can calculate the posterior distribution of $\alpha_1 - \alpha_2$. We have $\alpha_1 - \alpha_2|\mathbf{y}, \phi \sim \mathcal{N}\left(-0.2694, \frac{0.031}{\phi}\right)$ and therefore a 95% posterior credible interval is

$$-0.2694 \pm \sqrt{0.031 * 62.5/127} * t_{127}(0 - 975) = (-0.5110, -0.0278).$$

Thus, it seems likely that the combined model is superior to the simple regression model.

Note that in order to undertake a formal analysis of this, we could use fractional or intrinsic Bayes factors as the prior distributions in this case were improper.

Including covariance in the linear model

It is possible to fit a more general linear model where we do not assume that the model variance is proportional to the identity. One possibility is to assume that the model variance is proportional to a known matrix $(\mathbf{C})$. Lindley and Smith (1972) then demonstrate the following theorem.

Theorem 40

Let $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}|\phi \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\phi}\mathbf{C}\right)$ with prior distribution $\boldsymbol{\theta}, \phi \sim \mathcal{NG}\left(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$. Then, the predictive distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{T}\left(\mathbf{Xm}, \frac{b}{a}\left(\mathbf{XVX}^T + \mathbf{C}\right), a\right)$$

and the posterior distribution of $\boldsymbol{\mu}, \phi|\mathbf{y}$ is

$$\boldsymbol{\theta}|\mathbf{y}, \phi \sim \mathcal{N}\left(\mathbf{m}^\star, \frac{1}{\phi}\mathbf{V}^\star\right) \quad \phi|\mathbf{y} \sim \mathcal{G}\left(\frac{a^\star}{2}, \frac{b^\star}{2}\right) \quad \text{where}$$

$$
\begin{aligned}
\mathbf{m}^\star &= \left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X}+\mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{y}+\mathbf{V}^{-1}\mathbf{m}\right), \\
\mathbf{V}^\star &= \left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X}+\mathbf{V}^{-1}\right)^{-1} \\
a^\star &= a+n \\
b^\star &= b+\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}+\mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}-\mathbf{m}^{\star T}\mathbf{V}^{\star -1}\mathbf{m}^\star
\end{aligned}
$$

Proof Exercise. ■

In the limiting case, given the prior $p(\boldsymbol{\theta},\phi)\propto\frac{1}{\phi}$, it is easy to show that these posterior distributions converge to produce posterior distributions which lead to the same numerical results as in standard classical inference, e.g.

$$
\boldsymbol{\theta}|\mathbf{y},\phi\sim\mathcal{N}\left(\left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{C}^{-1}\mathbf{y},\frac{1}{\phi}\left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\right)
$$

and the posterior mean is the classical MLE of $\boldsymbol{\theta}$.

The SUR model

A more general approach is to assume an unknown model variance-covariance matrix $\mathbf{\Sigma}$ and set an inverse Wishart prior distribution, e.g. $\mathbf{\Phi} = \mathbf{\Sigma}^{-1} \sim \mathcal{W}(d, \mathbf{W})$.

Example 69

Seemingly unrelated regression (SUR) is a well known econometric model. In the traditional SUR model, we have M equations of form

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$$

for $i = 1, \ldots, M$ where $\mathbf{y}_i$ is an N dimensional vector of observations on a dependent variable, $\mathbf{X}_i$ is a $(N \times K_i)$ matrix of observations on K_i independent variables, $\boldsymbol{\beta}_i$ is a K_i dimensional vector of unknown regression coefficients and $\boldsymbol{\epsilon}_i$ is an N dimensional, unobserved error vector.

The M equations can be written as

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & & & \\ & \mathbf{X}_2 & & \\ & & \ddots & \\ & & & \mathbf{X}_M \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_M \end{pmatrix}$$

and written compactly as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ has dimension $(NM \times 1)$, $\mathbf{X}$ has dimension $(NM \times K)$ where $K = \sum_{i=1}^{M} K_i$, $\boldsymbol{\beta}$ is $(K \times 1)$ and $\boldsymbol{\epsilon}$ has dimension $(NM \times 1)$. Assume the distribution of $\boldsymbol{\epsilon}$ is

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_N\right).$$

The likelihood function is now

$$
\begin{aligned}
l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}) &\propto \boldsymbol{\Sigma}^{-\frac{N}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_N (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \\
&\propto \boldsymbol{\Sigma}^{-\frac{N}{2}} \exp\left(-\frac{1}{2} tr(\mathbf{A}\boldsymbol{\Sigma}^{-1})\right)
\end{aligned}
$$

where $\mathbf{A}$ is a $(M \times M)$ matrix with (i,j)'th element

$$
(\mathbf{A})_{ij} = (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_i)^T (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta}_j)
$$

A natural, uninformative prior is $p(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\frac{M+1}{2}}$ and then, given this prior, we have immediately that

$$\begin{aligned} \boldsymbol{\beta}|\mathbf{y}, \mathbf{\Sigma} &\sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \left[\mathbf{X}^T(\mathbf{\Sigma}^{-1} \otimes \mathbf{I}_N)\mathbf{X}\right]^{-1}\right) \\ \mathbf{\Sigma} &\sim \mathcal{IW}\left(N, \mathbf{A}\right) \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = \left[\mathbf{X}^T(\mathbf{\Sigma}^{-1} \otimes \mathbf{I}_N)\mathbf{X}\right]^{-1} \mathbf{X}^T(\mathbf{\Sigma}^{-1} \otimes \mathbf{I}_N)\mathbf{y}$ is the standard, least squares estimator.

The three stage linear model and ideas of hierarchical models

Up to now, we have used direct priors on the regression parameters $\boldsymbol{\theta}$, ϕ. In some cases, it may be more appropriate to use *hierarchical priors*. One example is the three stage linear model of Lindley and Smith (1972) who propose the structure

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \\ \boldsymbol{\theta} &\sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\beta}, \mathbf{V}\right) \\ \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{m}, \mathbf{W}) \end{aligned}$$

so that the prior distribution of $\boldsymbol{\theta}$ is defined hierarchically in two stages. Lindley and Smith (1972) demonstrate how to calculate the posterior distribution in this case when all the variances are known.

Example 70

In Example 65, we assumed direct, independent priors on the group mean parameters θ_i. Often however, we may have little information about these paramaters except that we have no prior evidence that they are different. This would suggest the use of a hierarchical prior, for example,

$$\begin{aligned} \theta_i|\theta_0,\phi &\sim \mathcal{N}\left(\theta_0, \frac{1}{\alpha\phi}\right) \\ p(\theta_0,\phi) &\propto \frac{1}{\phi}. \end{aligned}$$

We shall illustrate how Bayesian inference using hierarchical priors can be carried out in the following chapter.

Generalized linear models

The generalized linear model (Nelder and Wedderburn 1972) generalizes the normal linear model by allowing the possibility of non-normal error distributions and by allowing for a non-linear relationship between $\mathbf{y}$ and $\mathbf{x}$.

Definition 26

A generalized linear model is specified by two functions:

i a conditional, exponential family density function of y given $\mathbf{x}$, parameterized by a mean parameter, $\mu = \mu(\mathbf{x}) = E[Y|\mathbf{x}]$ and (possibly) a dispersion parameter, $\phi > 0$ that is independent of $\mathbf{x}$,

ii a (one-to-one) *link function*, g, which relates the mean, $\mu = \mu(\mathbf{x})$ to the covariate vector, $\mathbf{x}$, as $g(\mu) = \mathbf{x}\boldsymbol{\theta}$.

Example 71

The logistic regression model is often used for predicting the occurrence of an event given covariates. It is assumed that

$$\begin{aligned} Y_i|p_i &\sim \mathcal{BI}(n_i, p_i) \quad \text{for } i = 1, \ldots, m, \text{ and} \\ \log \frac{p_i}{1-p_i} &= \mathbf{x}_i \boldsymbol{\theta} \end{aligned}$$

Example 72

In Poisson regression, it is supposed that

$$\begin{aligned} Y_i|\lambda_i &\sim \mathcal{P}(\lambda_i) \\ \log \lambda_i &= \mathbf{x}_i \boldsymbol{\theta} \end{aligned}$$

In both examples, we have assumed the *canonical link function* which is the natural parameterization to leave the exponential family distribution in canonical form.

The Bayesian specification of a GLM is completed by defining (typically normal or normal gamma) prior distributions $p(\boldsymbol{\theta}, \phi)$ over the unknown model parameters. As with standard linear models, when improper priors are used, it is then important to check that these lead to valid posterior distributions.

Clearly, these models will not have conjugate posterior distributions, but, usually, they are easily handled by Gibbs sampling.

In particular, the posterior distributions from these models are usually log concave and are thus easily sampled via adaptive rejection sampling, see e.g. Gilks and Wild (1992).

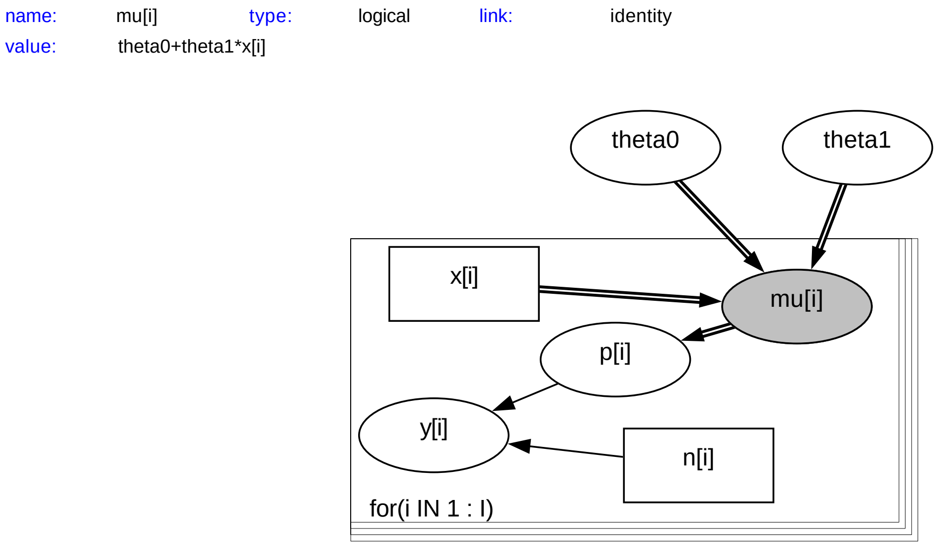
Example 73

The table shows the relationship, for 64 infants, between gestational age of the infant (in weeks) at the time of birth (x) and whether the infant was breast feeding at the time of release from hospital (y).

x	28	29	30	31	32	33
$\#y = 0$	4	3	2	2	4	1
$\#y = 1$	2	2	7	7	16	14

Let x_i represent the gestational age and n_i the number of infants with this age. Then we can model the probability that y_i infants were breast feeding at time of release from hospital via a standard binomial regression model.

It is easy to set this model up in Winbugs.



name: mu[i] type: logical link: identity

value: theta0+theta1*x[i]

The doodle gives the following model:

```
model; {
   theta0 ~ dnorm( 0.0,1.0E-6)
   theta1 ~ dnorm( 0.0,1.0E-6)
   for( i in 1 : I ) {
      mu[i] <- theta0+theta1 * x[i]
   }
   for( i in 1 : I ) {
      logit(p[i]) <- mu[i]
   }
   for( i in 1 : I ) {
      y[i] ~ dbin(p[i],n[i])
   }
}
```

and we can read in the data using

```
list(I=6,x=c(28,29,30,31,32,33),n=c(6,5,9,9,20,15),
y=c(2,2,7,7,16,14))
```

In this case, we set the initial values to `theta0=theta1=1`. Given 30000 iterations to burn in and 30000 in equilibrium we have the following posterior estimates for `p` and `theta`.

```
 node    mean    sd      MC error   2.5%    median  97.5%
p[1]    0.3783  0.1376  0.008533   0.1332  0.3714  0.6544
p[2]    0.5089  0.1118  0.00595    0.2842  0.5117  0.7173
p[3]    0.646   0.07742 0.002298   0.484   0.6501  0.7858
p[4]    0.7636  0.0577  9.063E-4   0.6435  0.7667  0.8673
p[5]    0.8483  0.05263 0.002359   0.7332  0.8528  0.9374
p[6]    0.9032  0.04844 0.002742   0.789   0.9108  0.9747
theta0  -16.85  6.479   0.4591     -30.71  -16.42  -5.739
theta1  0.5823  0.2112  0.01497    0.2222  0.567   1.036
```

The posterior mean values are quite close to the sample proportions.

Application V: Inference for Lanchester's combat models

Lanchester

Lanchester (1916) developed a system of equations for modeling the losses of combating forces.

Lanchester's equations

The Lanchester equations for modern warfare (aimed combat, without reinforcements) between armies of size $x(t)$ and $y(t)$ are

$$\begin{aligned} \frac{\partial x}{\partial t} &= -\alpha y \\ \frac{\partial y}{\partial t} &= -\beta x \end{aligned}$$

for $x(t), y(t) > 0$.

These equations lead to the well-known Lanchester square law

$$\alpha \left(x(0)^2 - x(t)^2 \right) = \beta \left(y(0)^2 - y(t)^2 \right) .$$

This law has been fitted to some real combat situations such as the battle of Iwo Jima (Engel 1954).

Different types of warfare

A more general system of differential equations which includes the square law can be used to represent different forms of warfare as follows

$$\begin{aligned}
\frac{\partial x}{\partial t} &= -\beta x^{\phi_1} y^{\phi_2} \\
\frac{\partial y}{\partial t} &= -\alpha y^{\phi_1} x^{\phi_2}
\end{aligned}$$

Here, $\boldsymbol{\phi} = (0,1)$ gives the square law, $\boldsymbol{\phi} = (1,1)$ gives a linear law representing unaimed fire combats, $\boldsymbol{\phi} = (0,0)$ leads to a different linear law representing hand-to hand combats and $\boldsymbol{\phi} = (1,0)$ leads to a logistic law which has been used to represent large scale combats such as the American Civil War. See e.g. Weiss (1966).

Introducing uncertainty

In modeling combat situations, interest lies in:

- classifying battle types (historical analysis),
- assessing relative fighting strengths of the two armies (i.e. the ratio α/β),
- predicting casualties,
- predicting who will win the battle.

However, the basic Lanchester models are deterministic and thus, for instance, the battle winner is predetermined given the model parameters. Thus, we need to introduce a random element into the Lanchester systems.

One possibility is to consider stochastic Lanchester models based on Poisson process type assumptions, see e.g. Clark (1969), Goldie (1977) or Pettit et al (2003). Following Wiper et al (2000), an alternative is to discretize time, linearize the Lanchester systems and fit a regression model.

Discretization and Linearization

Given daily casualty data, we can discretize the Lanchester equations to give:

$$\Delta x_t \approx \beta x_{t-1}^{\phi_1} y_{t-1}^{\phi_2} \quad \Delta y_t \approx \alpha y_{t-1}^{\phi_1} x_{t-1}^{\phi_2}$$

where Δx_t and Δy_t are the daily casualties recorded in the two armies.

Bracken (1995) attempted to fit this model directly to combat data from the Ardennes campaign. However, it seems more natural to linearize the model. Taking logs and introducing an error term, we have:

$$\mathbf{z}_t = \boldsymbol{\theta} + \mathbf{P}_{t-1}\boldsymbol{\phi} + \boldsymbol{\epsilon}_t$$

where $\mathbf{z}_t$ are the logged casualties of the two armies on day t, $\boldsymbol{\phi} = (\phi_1, \phi_2)^T$ and $\mathbf{P}_{t-1} = \begin{pmatrix} \log y_{t-1} & \log x_{t-1} \\ \log x_{t-1} & \log y_{t-1} \end{pmatrix}$.

Assuming that the error distribution is normal, $\boldsymbol{\epsilon}_t | \boldsymbol{\Phi} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Phi}^{-1}\right)$, then we have a (multivariate) linear model.

Analysis of the Ardennes campaign data

The Battle of the Bulge

The *Battle of the Bulge* in 1944, was one of the largest battles in the Second World War involving, for example, over 250000 US troops with nearly 20000 US casualties and even more German casualties. The Germans launched an initial surprise offensive under cloud cover and attacked during 5 days when the allied troops mounted a counterattack which lead to the eventual defeat of the German army 23 days later.

For a full description, see:

http://en.wikipedia.org/wiki/Battle_of_the_Bulge

The forces involved in the battle were troops, cannons and tanks and for the purposes of this analysis, we consider a composite force for each army which is a weighted measure of these components.

As it seems likely that casualties may be affected by whether an army is attacking or not, the general regression model was modified to include a factor $\boldsymbol{\delta}$ which indicated which army was attacking so that the full model is

$$\mathbf{z}_t = \boldsymbol{\theta} + \mathbf{P}_{t-1}\boldsymbol{\phi} + \mathbf{I}_t\boldsymbol{\delta} + \boldsymbol{\epsilon}_t.$$

Here $\mathbf{I}_t = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ if the Germans were attacking on day t and $\mathbf{I}_t = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ if the Americans were attacking.

Relatively informative prior distributions were used, with a Wishart prior structure for $\mathbf{\Phi}$. The model was then fitted using Gibbs sampling. The posterior mean parameter estimates are given below.

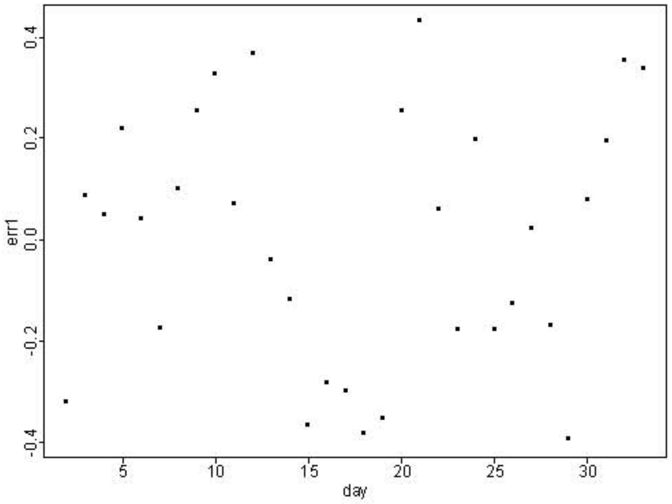
Parameter	Mean (standard deviation)	
$\boldsymbol{\theta}$	7.88 $(.09)$	8.04 $(.08)$
$\boldsymbol{\phi}$	0.84 $(.41)$	0.27 $(.40)$
$\boldsymbol{\delta}$	-0.38 $(.09)$	-0.01 $(.09)$

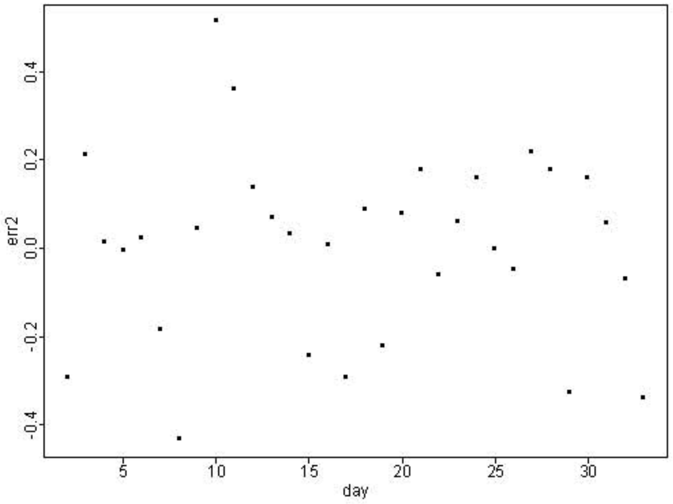
There is no clear evidence in favour of any one of the standard Lanchester laws as against any other although, using Bayes factors to compare the specific models $\boldsymbol{\phi} = (0,1)$, $(0,0)$, $(1,0)$ and $(1,1)$ suggests that the logistic law, $\boldsymbol{\phi} = (1,0)$ is the most likely.

There is strong evidence that the American casualties were typically higher when the Germans were attacking although the German casualties did not seem to be influenced by this.

Goodness of fit

In order to assess the fit of the model, we calculated the predicted errors $\mathbf{z}_t - E[\mathbf{z}_t|\mathrm{data}]$. A plot of these errors for the first American (left) and German (right) armies is given below.





There is slight evidence of lack of fit for the German forces. More complex models might be preferable. See e.g. Lucas and Turkes (2003).

Conclusions and Extensions

- Bayesian inference for Lanchester models is straightforward to implement.
- There is high colinearity in these models and proper prior information is necessary.
- Many models fit the Ardennes data almost equally well.
- The Lanchester model has been proposed as a model for competition between ants, see e.g. McGlynn (2000) and for business competition, see e.g. Campbell and Roberts (2006).
- The Lanchester equations are similar to the Lotka-Volterra equations for predator prey systems. It is possible to extend the approach to these systems.

References

Bracken, J. (1995). Lanchester models of the Ardennes campaign. *Naval Research Logistics*, **42**, 559-577.

Campbell, N.C.G. and Roberts, K.J. (2006). Lanchester market structures: A Japanese approach to the analysis of business competition. *Strategic Management Journal*, **7**, 189–200.

Clark, G.M. (1969). *The combat analysis model*. Ph.D. thesis, The Ohio State University.

Engel, J.H. (1954). A verification of Lanchester's law. *Operations Research*, **2**, 53–71.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Goldie, C.M. (1977). Lanchester square-law battles: Transient and terminal distributions. *Journal of Applied Probability*, **14**, 604-610.

Hoerl, A.E. and Kennard R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Lanchester, F.W. (1916). *Aircraft in warfare the dawn of the fourth arm*. London: Constable.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society Series B*, **34**, 1-41.

Lucas, T.W. and Turkes, T. (2003). Fitting Lanchester equations to the battles of Kursk and Ardennes. *Naval Research Logistics*, **51**, 95–116.

McGlynn, T.P. (2000). Do Lanchester's laws of combat describe competition in ants? *Behavioral Ecology*, **11**, 686–690.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, **132**, 107–120.

Pettit, L.I., Wiper, M.P. and Young, K.D.S. (2003). Bayesian inference for some Lanchester combat laws. *European Journal of Operational Research*, **148**, 152–165.

Weiss, H.K. (1966). Combat models and historical data: The US Civil War. *Operations Research*, **14**, 759-790.

Wiper, M.P., Pettit, L.I. and Young, K.D.S. (2000). Bayesian inference for a Lanchester type combat model. *Naval Research Logistics*, **42**, 609-633.