Bayesian Inference

Chapter 9. Linear models and regression

M. Concepcion Ausin
Universidad Carlos III de Madrid

Master in Business Administration and Quantitative Methods
Master in Mathematical Engineering

Chapter 9. Linear models and regression

Objective

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.

Chapter 9. Linear models and regression

AFM Smith

AFM Smith developed some of the central ideas in the theory and practice of modern Bayesian statistics.

Contents

The multivariate normal distribution

Firstly, we review the definition and properties of the multivariate normal distribution.

Definition

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:

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right),$$

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 multivariate normal distribution

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

- Any subset of $\mathbf{X}$ has a (multivariate) normal distribution.
- Any linear combination $\sum_{i=1}^{k} \alpha_i X_i$ is normally distributed.
- If $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$ is a linear transformation of $\mathbf{X}$, then:

$$\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T\right)$$

- If

$$\mathbf{X} = \left(\begin{array}{c} \mathbf{X}_1 \\ \mathbf{X}_2 \end{array} \middle| \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) \sim \mathcal{N}\left(\left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{array} \right), \left(\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right) \right),$$

then, the conditional density of $\mathbf{X}_1$ given $\mathbf{X}_2 = x_2$ 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 distribution

The likelihood function given a sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of data from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is:

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

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 $tr(\mathbf{M})$ represents the trace of the matrix $\mathbf{M}$.

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

Conjugate Bayesian inference when $\boldsymbol{\Sigma} = \frac{1}{\phi}\mathbf{C}$

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

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

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

Conjugate Bayesian inference when $\boldsymbol{\Sigma} = \frac{1}{\phi}\mathbf{C}$

Definition

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

$$\boldsymbol{\mu}|\phi \sim \mathcal{N}\left(\mathbf{m}, \frac{1}{\phi}\mathbf{V}\right)$$

$$\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right).$$

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

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

Conjugate Bayesian inference when $\mathbf{\Sigma} = \frac{1}{\phi}\mathbf{C}$

Definition

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

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

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

Theorem

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

$$\boldsymbol{\mu} \sim \mathcal{T}(\mathbf{m}, \frac{b}{a}\mathbf{V}, a)$$

Conjugate Bayesian inference when $\boldsymbol{\Sigma} = \frac{1}{\phi}\mathbf{C}$

Theorem

Let $\mathbf{X}|\boldsymbol{\mu}, \phi \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{\phi}\mathbf{C})$ and assume a priori $(\boldsymbol{\mu}, \phi) \sim \mathcal{NG}(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2})$. Then, given a sample data, $\mathbf{x}$, we have that,

$$\boldsymbol{\mu}|\mathbf{x}, \phi \sim \mathcal{N}\left(\mathbf{m}^*, \frac{1}{\phi}\mathbf{V}^*\right)$$

$$\phi|\mathbf{x} \sim \mathcal{G}\left(\frac{a^*}{2}, \frac{b^*}{2}\right).$$

where,

$$\begin{aligned}
\mathbf{V}^* &= \left(\mathbf{V}^{-1} + n\mathbf{C}^{-1}\right)^{-1} \\
\mathbf{m}^* &= \mathbf{V}^*\left(\mathbf{V}^{-1}\mathbf{m} + n\mathbf{C}^{-1}\bar{x}\right) \\
a^* &= a + nk \\
b^* &= b + tr\left(\mathbf{S}\mathbf{C}^{-1}\right) + \mathbf{m}^T\mathbf{V}^{-1} + n\bar{\mathbf{x}}^T\mathbf{C}^{-1}\bar{\mathbf{x}} + \mathbf{m}^*\mathbf{V}^{*-1}m^*
\end{aligned}$$

Conjugate Bayesian inference when $\boldsymbol{\Sigma} = \frac{1}{\phi}\mathbf{C}$

Theorem

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

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

Then, $\boldsymbol{\mu}, \phi \mid \mathbf{x} \sim \mathcal{NG}(\bar{\mathbf{x}}, n\mathbf{C}^{-1}, \frac{(n-1)k}{2}, \frac{tr(\mathbf{S}\mathbf{C}^{-1})}{2})$ which implies that,

$$\boldsymbol{\mu} \mid \mathbf{x}, \phi \sim \mathcal{N}\left(\bar{\mathbf{x}}, \frac{1}{n\phi}\mathbf{C}\right)$$

$$\phi \mid \mathbf{x} \sim \mathcal{G}\left(\frac{(n-1)\,k}{2}, \frac{tr\left(\mathbf{S}\mathbf{C}^{-1}\right)}{2}\right)$$

$$\boldsymbol{\mu} \mid \mathbf{x} \sim \mathcal{T}\left(\bar{\mathbf{x}}, \frac{tr\left(\mathbf{S}\mathbf{C}^{-1}\right)}{n\,(n-1)\,k}\mathbf{C}, (n-1)\,k\right)$$

Conjugate Bayesian inference when $\boldsymbol{\Sigma}$ is unknown

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

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

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

Conjugate Bayesian inference when $\boldsymbol{\Sigma}$ is unknown

Definition

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,

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

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

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 $E\left[\mathbf{W}^{-1}\right] = \frac{1}{d-k-1}\mathbf{V}^{-1}$

Conjugate Bayesian inference when $\boldsymbol{\Sigma}$ is unknown

Theorem

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

$$\boldsymbol{\mu} \mid \boldsymbol{\Phi}, \mathbf{x} \sim \mathcal{N}\left(\frac{\alpha\mathbf{m} + \bar{\mathbf{x}}}{\alpha + n}, \frac{1}{\alpha + n}\boldsymbol{\Phi}^{-1}\right)$$

$$\boldsymbol{\Phi} \mid \mathbf{x} \sim \mathcal{W}\left(d + n, \mathbf{W}^{-1} + \mathbf{S} + \frac{\alpha n}{\alpha + n}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})^T\right)$$

Theorem

Given the limiting prior, $p(\boldsymbol{\Phi}) \propto |\boldsymbol{\Phi}|^{\frac{k+1}{2}}$, the posterior distribution is,

$$\boldsymbol{\mu} \mid \boldsymbol{\Phi}, \mathbf{x} \sim \mathcal{N}\left(\bar{\mathbf{x}}, \frac{1}{n}\boldsymbol{\Phi}^{-1}\right)$$

$$\boldsymbol{\Phi} \mid \mathbf{x} \sim \mathcal{W}(n - 1, \mathbf{S})$$

Conjugate Bayesian inference when $\boldsymbol{\Sigma}$ is unknown

The conjugacy assumption that the prior precision of $\boldsymbol{\mu}$ is proportional to the model precision ϕ 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 $\boldsymbol{\Phi}$, say $\boldsymbol{\Phi} \sim \mathcal{W}(\mathbf{d}, \mathbf{W})$ as earlier.

In this case, the conditional posterior distributions are:

$$\boldsymbol{\mu} \mid \boldsymbol{\Phi}, \mathbf{x} \sim \mathcal{N}\left(\left(\mathbf{V}^{-1} + n\boldsymbol{\Phi}\right)^{-1}\left(\mathbf{V}^{-1}\mathbf{m} + n\boldsymbol{\Phi}\bar{\mathbf{x}}\right), \left(\mathbf{V}^{-1} + n\boldsymbol{\Phi}\right)^{-1}\right)$$

$$\boldsymbol{\Phi} \mid \boldsymbol{\mu}, \mathbf{x} \sim \mathcal{W}\left(d + n, \mathbf{W}^{-1} + \mathbf{S} + n\left(\boldsymbol{\mu} - \bar{\mathbf{x}}\right)\left(\boldsymbol{\mu} - \bar{\mathbf{x}}\right)^{T}\right)$$

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

Normal linear models

Definition

A normal linear model is of form:

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

where $\mathbf{y} = (y_1, \ldots, y_n)'$ are the observed data, $\mathbf{X}$ is a given design matrix of size $n \times k$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ are the set of model parameters. Finally, it is usually assumed that:

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}_n, \frac{1}{\phi}\mathbf{I}_n\right).$$

Normal linear models

A simple example of normal linear model is the simple linear regression model where $\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}^T$ and $\boldsymbol{\theta} = (\alpha, \beta)^T$.

It is easy to see that there is a conjugate, multivariate normal-gamma prior distribution for any normal linear model.

Conjugate Bayesian inference

Theorem

Consider a normal linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, and assume a normal-gamma priori distribution,

$$\boldsymbol{\theta}, \phi \sim \mathcal{NG}(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}).$$

Then, the posterior distribution given $\mathbf{y}$ is also a normal-gamma distribution:

$$\boldsymbol{\theta}, \phi \mid \mathbf{y} \sim \mathcal{NG}(\mathbf{m}^*, \mathbf{V}^{*-1}, \frac{a^*}{2}, \frac{b^*}{2})$$

where:

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

Conjugate Bayesian inference

Given a normal linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, and assuming a normal-gamma priori distribution, $\boldsymbol{\theta}, \phi \sim \mathcal{NG}(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2})$, it is easy to see that the predictive distribution of $\mathbf{y}$ is:

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

To see this, note that the distribution of $\mathbf{y} \mid \phi$ is:

$$\mathbf{y} \mid \phi \sim \mathcal{N}\left(\mathbf{Xm}, \frac{1}{\phi}\left(\mathbf{XVX}^T + \mathbf{I}\right)\right),$$

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

$$\mathbf{y}, \phi \sim \mathcal{NG}\left(\mathbf{Xm}, (\mathbf{XVX}^T + \mathbf{I})^{-1}, \frac{a}{2}, \frac{b}{2}\right)$$

and therefore, the marginal of $\mathbf{y}$ is the multivariate t-distribution obtained before.

Interpretation of the posterior mean

We have that:

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

where $\hat{\theta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ is the maximum likelihood estimator.

Thus, this expression may be interpreted as a weighted average of the prior estimator, $\mathbf{m}$, and the MLE, $\hat{\theta}$, 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 $\hat{\theta} \mid \phi \sim \mathcal{N}\left(\theta, \frac{1}{\phi}(\mathbf{X}^T\mathbf{X})^{-1}\right)$

Conjugate Bayesian inference

Theorem

Consider a normal linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, and assume the limiting prior distribution, $p(\boldsymbol{\theta}, \phi) \propto \frac{1}{\phi}$. Then, we have that,

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

$$\phi \mid \mathbf{y} \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T\left(\mathbf{X}^T\mathbf{X}\right)\hat{\boldsymbol{\theta}}}{2}\right).$$

And then,

$$\boldsymbol{\theta} \mid \mathbf{y} \sim \mathcal{T}\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}, n-k\right),$$

where,

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T\left(\mathbf{X}^T\mathbf{X}\right)\hat{\boldsymbol{\theta}}}{n-k}$$

Conjugate Bayesian inference

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

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 of full rank.

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

ANOVA model

The ANOVA model is an example of normal lineal model where:

$$y_{ij} = \theta_i + \epsilon_{ij},$$

where $\epsilon_{ij} \sim \mathcal{N}(0, \frac{1}{\phi})$, for $i = 1, \ldots, k$, and $j = 1, \ldots, n_i$.

Thus, the parameters are $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, the observed data are $\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}, \ldots, y_{k1}, \ldots, y_{kn_k})^T$, the design matrix is:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \\ 1_{n_1} & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1_{n_2} & & 0 \\ \vdots & & \vdots & \vdots \\ 0 & 0 & & 1 \end{pmatrix}$$

ANOVA model

If we use conditionally independent normal priors, $\theta_i \sim \mathcal{N}\left(m_i, \frac{1}{\alpha_i \phi}\right)$, for $i = 1, \ldots, k$, and a gamma prior $\phi \sim \mathcal{G}(\frac{a}{2}, \frac{b}{2})$. Then, we have that:

$$(\boldsymbol{\theta}, \phi) \sim \mathcal{NG}\left(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$$

where $\mathbf{m} = (m_1, \ldots, m_k)$ and $\mathbf{V} = \begin{pmatrix} \frac{1}{\alpha_1} & & \\ & \ddots & \\ & & \frac{1}{\alpha_k} \end{pmatrix}$

Noting that, $\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1} = \begin{pmatrix} n_1 + \alpha_1 & & \\ & \ddots & \\ & & n_k + \alpha_k \end{pmatrix}$, we can obtain the following posterior distributions.

ANOVA model

It is obtained that,

$$\boldsymbol{\theta} \mid \mathbf{y}, \phi \sim \mathcal{N}\left(\begin{pmatrix} \frac{n_1 \bar{y}_{1\cdot} + \alpha_1 m_1}{n_1 + \alpha_1} \\ \vdots \\ \frac{n_1 \bar{y}_{1\cdot} + \alpha_1 m_1}{n_1 + \alpha_1} \end{pmatrix}, \frac{1}{\phi} \begin{pmatrix} \frac{1}{\alpha_1 + n_1} & & \\ & \ddots & \\ & & \frac{1}{\alpha_k + n_k} \end{pmatrix}\right)$$

and

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

ANOVA model

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

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

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

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

ANOVA model

If we assume alternatively the reference prior, $p(\boldsymbol{\theta}, \phi) \propto \frac{1}{\phi}$, we have:

$$\boldsymbol{\theta} \mid \mathbf{y}, \phi \sim \mathcal{N}\left(\begin{pmatrix} \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{k\cdot} \end{pmatrix}, \frac{1}{\phi}\begin{pmatrix} \frac{1}{n_1} & & \\ & \ddots & \\ & & \frac{1}{n_k} \end{pmatrix}\right),$$

$$\phi \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{(n-k)\,\hat{\sigma}^2}{2}\right),$$

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

A 95% posterior interval for $\theta_1 - \theta_2$ is given by:

$$\bar{y}_{1\cdot} - \bar{y}_{2\cdot} \pm \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{n-k}\left(0.975\right),$$

which is equal to the usual, classical interval.

Simple linear regression model

Consider the simple 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:

$$p(\alpha, \beta, \phi) \propto \frac{1}{\phi}.$$

Simple linear regression model

Then, we have that:

$$\left.\begin{matrix}\alpha \\ \beta\end{matrix}\right| \mathbf{y}, \phi \sim \mathcal{N}\left(\begin{pmatrix}\hat{\alpha} \\ \hat{\beta}\end{pmatrix}, \frac{1}{\phi n s_x}\begin{pmatrix}\sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n\end{pmatrix}\right)$$

$$\phi \mid \mathbf{y} \sim \mathcal{G}\left(\frac{n-2}{2}, \frac{s_x\left(1-r^2\right)}{2}\right)$$

$$\left.\begin{matrix}\alpha \\ \beta\end{matrix}\right| \mathbf{y} \sim \mathcal{T}\left(\begin{pmatrix}\hat{\alpha} \\ \hat{\beta}\end{pmatrix}, \frac{\hat{\sigma}^2}{n}\frac{1}{s_x}\begin{pmatrix}\sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n\end{pmatrix}, n-2\right)$$

where

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{s_{xy}}{s_x},$$

$$s_x = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2, \qquad s_y = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2,$$

$$s_{xy} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right), \qquad r = \frac{s_{xy}}{\sqrt{s_x s_y}}, \qquad \hat{\sigma}^2 = \frac{s_y\left(1-r^2\right)}{n-2}.$$

Simple linear regression model

Thus, the marginal distributions of α and β are:

$$\alpha \mid \mathbf{y} \sim \mathcal{T}\left(\hat{\alpha}, \frac{\hat{\sigma}^2}{n}\frac{\sum_{i=1}^{n} x_i^2}{s_x}, n-2\right)$$

$$\beta \mid \mathbf{y} \sim \mathcal{T}\left(\hat{\beta}, \frac{\hat{\sigma}^2}{s_x}, n-2\right)$$

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

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

equal to the usual classical interval.

Simple linear regression model

Suppose now that we wish to predict a future observation:

$$y_{new} = \alpha + \beta x_{new} + \epsilon_{new}.$$

Note that,

$$\begin{aligned} E\left[y_{new} \mid \phi, \mathbf{y}\right] &= \hat{\alpha} + \hat{\beta} x_{new} \\ V\left[y_{new} \mid \phi, \mathbf{y}\right] &= \frac{1}{\phi}\left(\frac{\sum_{i=1}^{n} x_i^2 + n x_{new}^2 - 2n\bar{x}x_{new}}{n s_x} + 1\right) \\ &= \frac{1}{\phi}\left(\frac{s_x + n\bar{x}^2 + n x_{new}^2 - 2n\bar{x}x_{new}}{n s_x} + 1\right) \end{aligned}$$

Therefore,

$$y_{new} \mid \phi, \mathbf{y} \sim \mathcal{N}\left(\hat{\alpha} + \hat{\beta} x_{new}, \frac{1}{\phi}\left(\frac{\left(\bar{x} - x_{new}\right)^2}{s_x} + \frac{1}{n} + 1\right)\right)$$

Simple linear regression model

And then,

$$y_{new} \mid \mathbf{y} \sim \mathcal{T}\left(\hat{\alpha} + \hat{\beta} x_{new}, \hat{\sigma}^2 \left(\frac{(\bar{x} - x_{new})^2}{s_x} + \frac{1}{n} + 1\right), n-2\right)$$

leading to the following 95% credible interval for y_{new} :

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

which coincides with the usual, classical interval.

Generalized linear models

The generalized linear model generalizes the normal linear model by allowing the possibility of non-normal error distributions and by allowing for a non-linear relationship between **y** and **x**.

A generalized linear model is specified by two functions:

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

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

Generalized linear models

The following are generalized linear models with the canonical link function which is the natural parameterization to leave the exponential family distribution in canonical form.

Logistic regression

It is often used for predicting the occurrence of an event given covariates:

$$Y_i \mid p_i \sim \mathcal{B}in(n_i, p_i)$$
$$\log \frac{p_i}{1 - p_i} = \mathbf{x}_i \boldsymbol{\theta}$$

Poisson regression

It is used for predicting the number of events in a time period given covariates:

$$Y_i \mid p_i \sim \mathcal{P}(\lambda_i)$$
$$\log \lambda_i = \mathbf{x}_i \boldsymbol{\theta}$$

Generalized linear models

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.