Chapter 5

Bayesian Approaches

5.1 Introduction

In statistics there are two fundamentally different philosophies. The main difference is in their understanding of the term **probability**.

- **Frequentists** understand probability as a measure of how often a certain event happens.
- **Bayesians** use probability as a tool to quantify uncertainty about a certain event. Uncertainty and its perception can vary between different individuals. This fact has been a big point for criticizing Bayesian approaches.

All methods that were presented in this course so far are Frequentist concepts. The relevant differences between Bayesians and Frequentists can be found in the following points.

- understanding of probability
- differentiation between components of a model and the data
- techniques to estimate parameters.

The following table gives an overview over the differences.

Table 5.1: Differences between Frequentists and Bayesians

Topic	Frequentists	Bayesians
Probability	Ratio between cardinalities of sets	Measure of uncertainty
Model and Data	Parameter are unknown, data are known	Differentiation between knowns and unknowns
Parameter Estimation	ML or REML are used for parameter estimation	MCMC techniques to approximate posterior distributions

5.2 Linear Model

The Bayesian way to estimate parameter is shown with the following simple linear model[1]. Let us assume the following linear model for a single observation y_i

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \tag{5.1}$$

where β_0 is the intercept and x_{i1} is a predictor variable. The error term is denoted by ϵ_i with a variance σ^2.

5.2.1 Knowns and Unknowns

In Bayesian statistics, the separation into knowns and unknowns replaces the differentiation between data and parameter from frequentist statistics. Whenever there are no missing data the separation into knowns and unknowns correspond to the differentiation into data and parameter. For our model (5.1) the separation into knowns and unknowns is given in the following table.

Table 5.2: Separation Into Knowns And Unknowns

Term	Known	Unknown
y_i	X	
x_1	X	
β_0		X
β_1		X
σ^2	X	

5.2.2 Bayesian Parameter Estimation

Bayesians base their estimation of unknowns on the **posterior distribution** of the unknowns given the knowns. The posterior distribution is computed using **Bayes Theorem** based on the prior distribution of all unknowns and based on the likelihood. The terms "prior" and "posterior" are to be understood relative to the point in time where the data to be analysed was collected. This concept is shown in Figure 5.1.

For the linear model (5.1), we define the vector β as

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

[1] In Bayesian statistics there is no separation into fixed and random effects. Hence, we call this model just linear model.

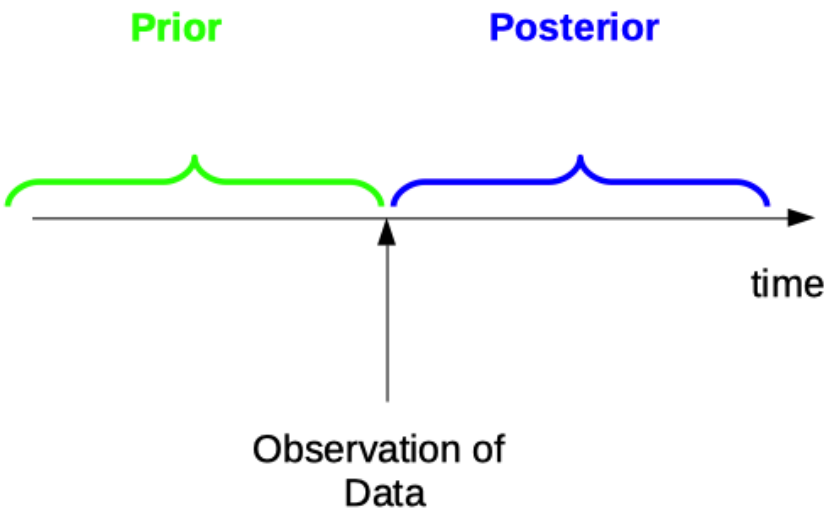

Figure 5.1: Distinctions between Prior and Posterior in Bayesian Statistics

The observations are collected in the vector y. For this simple example, we assume that the variance σ^2 is known[2]. In Bayesian statistics the estimate of the unknown β is based on the posterior distribution $f(\beta|y)$. Using Bayes Theorem, the posterior distribution can be written as

$$\begin{aligned} f(\beta|y) &= \frac{f(\beta, y)}{f(y)} \\ &= \frac{f(y|\beta)f(\beta)}{f(y)} \\ &\propto f(y|\beta)f(\beta) \end{aligned} \tag{5.2}$$

In equation (5.2) the posterior distribution $f(\beta|y)$ is expressed as a product of the prior distribution $f(\beta)$ and the likelihood $f(y|\beta)$. The factor $f(y)^{-1}$ corresponds to the normalizing constant and is not of further interest to us. Hence in the final result, the posterior distribution is given as a proportionality relation.

The posterior distribution $f(\beta|y)$ can in many cases not be expressed explicitly. For a long time this has been a problem restricting the use of Bayesian methods. Two important developments have contributed important solutions to this problem.

1. In [Besag, 1974], it was shown that every posterior distribution can be expressed in terms of their full conditional distribution. For our example of the linear model (5.1), the full conditional distributions are: (i) the conditional distribution of β_0 given all other parameters, hence $f(\beta_0|\beta_1, y)$ and (ii) the conditional distribution of β_1 given all other parameters which corresponds to $f(\beta_1|\beta_0, y)$.
2. The second important development consists of the development of efficient pseudo-random number generators that are easy to use on computers.

5.3 Gibbs Sampler

The implementation of the above two mentioned developments has lead to a procedure that is referred to as the **Gibbs Sampler**. Most of the times [Geman and Geman, 1984] is given credit for a first application of the described parameter estimation technique. When the Gibbs Sampling technique is applied to a simple linear model, the following procedure can be derived. In general, a data analysis with the Gibbs Sampler can always be done by going through the following steps.

[2] As a consequence, σ^2 is omitted from all subsequent derivations.

1. Determine the prior distributions for the unknowns
2. Determine the likelihood
3. Determine the full conditional distributions.

5.3.1 Prior Distributions

In our example of the simple linear model, the prior distribution corresponds to $f(\beta)$. In most cases when a certain type of dataset is analysed for the first time, there is no prior information about the unknowns available. In such a case an uninformative prior is chosen. That means $f(\beta)$ is chosen as a constant. For our example, we would use an uninformative prior for β and hence, we set $f(\beta) = c$ where c is a constant.

A well established alternative to uninformative priors are prior distributions of unknowns that have been used in many data analyses. As a result such prior distributions can be considered as de-facto standard due to their wide-spread usage.

5.3.2 Likelihood

Similarly to frequentist statistics, the likelihood is defined as the conditional distribution $f(y|\beta)$ of the data y given the parameter β. In the case where not data are missing, the Bayesian likelihood is the same.

5.3.3 Full Conditional Distribution

With full conditional distributions, we mean that for every unknown, the conditional distribution of that unknown given everything else has to be determined. In our example of the simple linear model, there are two unknowns β_0 and β_1. Hence, we have two full conditional distributions. Assuming that the data y follow a normal distribution, the full conditional distribution can be written as shown in the following table.

Unknown	full conditional	resulting distribution
β_0	$f(\beta_0\|\beta_1, y)$	$\mathcal{N}\left(\widehat{\beta}_0, var(\widehat{\beta}_0)\right)$
β_1	$f(\beta_1\|\beta_0, y)$	$\mathcal{N}\left(\widehat{\beta}_1, var(\widehat{\beta}_1)\right)$

Based on a series of computations not shown here the full conditional distributions can be converted into the resulting distributions. The symbol $\mathcal{N}$ stands for normal distribution where the first argument is the mean and the second argument is the variance. To compute the mean and the variance that are included in the full conditional distributions, the model (5.1) has to be re-formulated as follows.

$$y = 1\beta_0 + x\beta_1 + \epsilon \tag{5.3}$$

The model (5.3) is now written such that we have a new linear model with β_0 as its only parameter. This means

$$w_0 = 1\beta_0 + \epsilon \tag{5.4}$$

with $w_0 = y - x\beta_1$. The least squares estimate $\hat{\beta}_0$ can be written as

$$\hat{\beta}_0 = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{w}_0 \tag{5.5}$$

The variance of $\hat{\beta}_0$ is

$$var(\hat{\beta}_0) = (\mathbf{1}^T\mathbf{1})^{-1}\sigma^2 \tag{5.6}$$

What was shown for β_0 can also be done for β_1.

$$\hat{\beta}_1 = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{w}_1 \tag{5.7}$$

where $\mathbf{w}_1 = \mathbf{y} - \mathbf{1}\beta_0$

$$var(\hat{\beta}_1) = (\mathbf{x}^T\mathbf{x})^{-1}\sigma^2 \tag{5.8}$$

5.3.4 Implementation Of The Gibbs Sampler

The Gibbs Sampler is implemented by repeated drawing of random samples from the full conditional distributions. That means, we use starting values for all unknowns. For β_0 and β_1 we use 0 as a starting value. In the second step, we compute the expected value and the variance for the full conditional distributions and we draw a random sample from this distribution. The random sample are then used for the computation of the moments of the full conditional distributions in the next round. This procedure of computing expectations and variance of the full conditional distributions and drawing random samples from these distributions is repeated about 10000 times. All drawn samples for β_0 and β_1 are stored. From the drawn sample, we compute the mean and the standard deviation. These are used as representatives of Bayesian parameter estimates and standard deviation of these estimates.

The following R code chunk gives an implementation of the Gibbs Sampler for the unknowns β_0 and β_1. For reasons of simplicity σ^2 was assumed to be constant with a value of $\sigma^2 = 1$.

```
# ### starting values for beta0 and beta1
beta <- c(0, 0)
# ### set the number of iterations
niter <- 10000
# ### initialize the vector of results
meanBeta <- c(0, 0)
### # loop over iterations
for (iter in 1:niter) {
  # get expected value and variance for
  # full conditional of beta_0
  w <- y - X[, 2] * beta[2]
  x <- X[, 1]
  xpxi <- 1/(t(x) %*% x)
  betaHat <- t(x) %*% w * xpxi
  # ### draw random value for beta0
  beta[1] <- rnorm(1, betaHat, sqrt(xpxi))
  # expected valuea nd variance for beta1
  w <- y - X[, 1] * beta[1]
  x <- X[, 2]
  xpxi <- 1/(t(x) %*% x)
  betaHat <- t(x) %*% w * xpxi
  # ### new random number for beta1
  beta[2] <- rnorm(1, betaHat, sqrt(xpxi))
  meanBeta <- meanBeta + beta
}
# ### Output of results
cat(sprintf("Achsenabschnitt = %6.3f \n", meanBeta[1]/iter))
cat(sprintf("Steigung = %6.3f \n", meanBeta[2]/iter))
```

The application of this procedure to a real data set will be the topic of an exercise.