Linear models and their mathematical foundations: Bayesian inference for the linear regression model

Steffen Unkel

Department of Medical Statistics
University Medical Center Göttingen
The name ‘Bayesian’ refers to Reverend Thomas Bayes, an English 18th century Presbyterian minister and amateur mathematician and scientist.

Contrary to frequentist inference, in Bayesian statistics any relevant information external to the data is also used.

The modern Bayesian movement began in the second half of the 20th century.

Bayesian methods are being increasingly used in the field of statistics.
Bayes’ theorem

- **Bayes’ theorem** underlies the Bayesian approach to statistics and is formally stated as follows.

- For two events $A$ and $B$, provided that $P(B) > 0$,

  $$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} ,$$

  where $P(B) = P(B|A) \times P(A) + P(B|\overline{A}) \times P(\overline{A})$.

- Bayes’ theorem gives a method of revising probability estimated as additional information becomes available.
Bayes’ theorem in practice

\[ P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \]

- Prior \( P(A) \)
- Information \( B \)
- Posterior \( P(A|B) \)
Sequential updating

\[ P(A) \xrightarrow{B} P(A|B) \xrightarrow{C} P(A|C, B) \]
Bayesian approach to statistical inference

- **Statistical inference** is concerned with learning about the parameters of interest using data.

- In the Bayesian approach to inference, parameters are viewed as random variables with probability distributions.

- It requires an **initial distribution** to represent belief about the parameter **before** collecting data.

- The data obtained from a sample can be combined with the prior to form a new belief about the parameter.
Prior, likelihood and posterior distribution

Scientific enquiry is about our prior beliefs regarding a situation in the light of new evidence to provide an assessment of our current state of knowledge.

Ingredients of a Bayesian approach for data $y$ and unknown parameter $\theta$:

1. Prior distribution: $f(\theta)$
2. Likelihood function: $f(y|\theta)$
3. Posterior distribution: $f(\theta|y)$

Bayes’ theorem for distributions:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}.$$ 

In other words: posterior $\propto$ likelihood $\times$ prior.
The process of a Bayesian analysis

Prior distribution \rightarrow \text{Likelihood} \rightarrow \text{Bayes’ theorem} \rightarrow \text{Posterior distribution} \rightarrow \text{Estimates and inferences}
Bayesian updating

**Figure:** The prior, likelihood and posterior for a proportion $\theta$: three analyses.
The normal prior $\mathcal{N}(a, b)$ with mean $a$ and variance $b$ may be used to represent beliefs about $\theta$ that are symmetric about a single most likely value.

- Mode = median = mean = $a$.
- Variance = $b$.
- All values of $\theta$ in the range $-\infty < \theta < \infty$ are possible, but only those in the range $a \pm 3\sqrt{b}$ are likely.
Using standard distributions as priors (2)

The uniform prior

The uniform prior $\mathcal{U}(a, b)$ with parameters $a$ and $b$ may be used to represent beliefs that the value of $\theta$ lies between $a$ and $b$ when it is not known which values within the interval $[a, b]$ are more likely than others.

- $\mathcal{U}(a, b)$ is noninformative if the interval $[a, b]$ necessarily includes all values in the range of $\theta$.
- Improper uniform priors on $(-\infty, \infty)$, $(-\infty, b]$ or on $[a, \infty)$ may be used to represent lack of prior information about $\theta$ and its range.
The beta prior

The beta prior $\text{Be}(a, b)$ with parameters $a > 0$ and $b > 0$ may be used to represent beliefs about a proportion $\theta$, $0 \leq \theta \leq 1$.

- When $a > 1$ and $b > 1$, the beta density has the single mode $\frac{a-1}{a+b-2}$.
- When $a < 1$ ($b < 1$), the mode is at 0 (1). When $a < 1$ and $b < 1$, the density has two modes – at 0 and 1.
- The $\text{Be}(1, 1)$ distribution is the same as the $\mathcal{U}(0, 1)$ distribution.
- The mean and variance of $\text{Be}(a, b)$ are given by

  \[
  \text{mean} = \frac{a}{a + b}, \quad \text{variance} = \frac{ab}{(a + b)^2(a + b + 1)}.
  \]

- The larger the value of $a + b$ is, the stronger are the beliefs represented by the beta prior.
The gamma prior and inverse gamma prior

The gamma prior $\mathcal{G}(a, b)$ with parameters $a > 0$ and $b > 0$ may be used to represent prior beliefs about a parameter $\theta$ which takes only non-negative values. The parameter $a$ is the shape parameter.

- When $a > 1$, the density has the single mode $\frac{a-1}{b}$.
- When $0 < a \leq 1$, there is single mode at 0.
- The mean and variance of $\mathcal{G}(a, b)$ are given by

$$\text{mean} = \frac{a}{b}, \quad \text{variance} = \frac{a}{b^2}.$$

If $\theta \sim \mathcal{G}(a, b)$, then $1/\theta \sim \text{Inv}\mathcal{G}(a, b)$, where $\text{Inv}\mathcal{G}(\cdot, \cdot)$ denotes an inverse gamma distribution.
Specifying the prior

How to specify the prior distribution $f(\theta)$?

1. Assess the location of $f(\theta)$.
2. Assess the spread of $f(\theta)$.
3. Calculate the values of $a$ and $b$ that give the assessed location and spread.
Specifying the prior: assessing the location

**Figure:** Two prior densities: (a) symmetric (b) right-skew.
To assess the spread of the prior for a parameter $\theta$, an equal-tailed $100(1 - \alpha)\%$ interval $[L, U]$ is obtained, where $L$ is the $\alpha/2$-quantile and $U$ is the $(1 - \alpha/2)$-quantile of the prior.

Commonly used values of $1 - \alpha$ are 0.5, in which case $L$ and $U$ are the lower and upper quartiles of the prior for $\theta$, and 0.66.
Specifying the prior: determining the parameters

**Specifying a normal prior $\mathcal{N}(a, b)$**

The mean $a$ and variance $b$ of a normal prior may be assessed as follows:

$$a = \text{assessed mode or median},$$

$$b = \left(\frac{U-L}{2z}\right)^2,$$

where $L$ is the assessed $\alpha/2$-quantile and $U$ is the assessed $(1 - \alpha/2)$-quantile of the prior for $\theta$, and $z$ is the $(1 - \alpha/2)$-quantile of $\mathcal{N}(0, 1)$. 

Conjugate analyses

- Recall that posterior $= k \times$ likelihood $\times$ prior, where $k$ is a constant, not involving $\theta$, which ensures that the area under the posterior is equal to 1.

- For some likelihoods, a conjugate prior can be used which produces a posterior of the same form as the prior distribution.

- In this case, the constant $k$ does not need to be calculated and the prior to posterior analysis is called a conjugate analysis.

- Conjugate models are for example the gamma/Poisson model for estimating a Poisson mean and the normal/normal model for estimating a normal mean.
Example: Conjugate analyses for a proportion

The beta/binomial model

- Suppose that $\theta$ is a proportion, with beta prior distribution with parameters $a$ and $b$:

  $$\theta \sim \mathcal{B}e(a, b)$$

- Suppose that $Y$ may be modelled by the binomial distribution $\mathcal{B}(n, \theta)$:

  $$Y \sim \mathcal{B}(n, \theta)$$

- If data $Y = y$ are collected, then the posterior distribution for $\theta$, given $y$, is

  $$\theta|\text{data} \sim \mathcal{B}e(a + y, b + n - y)$$
Using uniform priors

- Although uniform priors are not conjugate priors, their use is widespread.

- When combined with binomial, Poisson and normal data, suitably chosen uniform priors produce beta, gamma and normal posteriors, respectively.

- The posteriors produced when using these uniform priors are the same as the distributions produced with conjugate analyses.

- However, the posterior parameters are different.
Summarising the posterior distribution

- The **marginal posterior distribution** can be displayed using kernel density estimation.

- **Measures of location** that are used with the posterior are the mean, median and the mode.

- The **posterior mode** is the most likely value for $\theta$ after observing the data.

- The **spread** of the posterior represents the posterior uncertainty about $\theta$ and is commonly summarized using the variance, standard deviation and quantiles.
Credible intervals

- An interval \([l, u]\) is a 100(1 − \(\alpha\))% credible interval for a parameter \(\theta\) if the posterior probability that \(l \leq \theta \leq u\), given the data, is equal to 1 − \(\alpha\):

\[
P(l \leq \theta \leq u | \text{data}) = 1 - \alpha.
\]

- The probability 1 − \(\alpha\) is the credibility level of the interval.

- For any specified credibility level 1 − \(\alpha\), there are usually many 100(1 − \(\alpha\))% credible intervals for \(\theta\).
Credible intervals (2)

- **Highest posterior density interval**: Whenever $\theta_1$ is within the interval $[l, u]$ and $\theta_2$ is outside the interval, then $f(\theta_1|\text{data}) \geq f(\theta_2|\text{data})$.

![Probability density curve](image)

**Figure**: A posterior density for $\theta$ along with a $100(1 - \alpha)$% highest posterior density interval.

- If a $100(1 - \alpha)$% credible interval $[l, u]$ is equal-tailed, then

$$P(\theta < l|\text{data}) = P(\theta > u|\text{data}) = \frac{1}{2} \alpha.$$
How to obtain the posterior distribution?

Typically it will not be possible to write the posterior distribution in closed form or sample independent values from it directly.

The posterior distribution is usually being computed using Markov chain Monte Carlo (MCMC) simulation:

1. Set up a Markov chain that has the posterior distribution as its equilibrium distribution.
2. Sample from the Markov chain.
3. Use the (dependent) samples values (from the equilibrium distribution and not from the transient phase) to make inferences about the unknown quantities of interest.

Several sampling methods are available, e.g. Gibbs sampling.
A simulated Markov chain

Figure: First 500 values simulated from a Markov chain.
Sampling variability and MC error

- The random variation of samples from stochastic simulations is known as **sampling variability**.

- The **Monte Carlo standard error** (MC error) measures the amount of uncertainty in the sampling-based estimate of the posterior mean.

- Rule of thumb: The sample size $N$ used in a stochastic simulation should be chosen so that the estimate of the MC error is less than 5% of the estimated posterior standard deviation of the parameter.

- The estimated MC errors reported by software packages usually allow for the dependence between the sampled values.
Convergence assessment

Figure: (a) No evidence against convergence, (b) Clear evidence against convergence.
Convergence assessment

**Figure:** Trace plots of simulated values from three simulations.
Dealing with samples from MCMC

**Standard procedure for MCMC**

- Run a small number of separate simulations, say three to five, each started from a different initial value.

- For each parameter in the model, inspect trace plots of the simulated values of the parameter from all the simulations.

- Assess when it is safe to assume that convergence to the equilibrium distribution has been reached.

- Discard the sampled values that are deemed to come from the transient phase (burn-in period) of the Markov chain simulation.

- Use the remaining sampled values for producing graphical and numerical summaries of the parameters of interest.
Model for the data

- We consider the **multiple linear regression model** \( y = X\beta + \epsilon \) with \( \text{E}(\epsilon) = 0 \) and \( \text{Cov}(\epsilon) = \sigma^2 I \).

- In contrast to frequentist inference, the Bayesian approach considers the unknown parameters \( \beta \) and \( \sigma^2 \) as random variables.

- Thus, the distribution of the response \( y \) can be understood as conditional on the parameters \( \beta \) and \( \sigma^2 \), and we obtain the following **model for the observations**:

\[
y | \beta, \sigma^2 \sim N_n(X\beta, \sigma^2 I)
\]
## Normal-inverse gamma prior

- We assume that $\beta | \sigma^2 \sim \mathcal{N}_{k+1}(m, \sigma^2 V)$ with known expectation $m$ and covariance matrix $V$.
- For $\sigma^2$ we specify an inverse gamma distribution with hyperparameters $a$ and $b$, i.e., $\sigma^2 \sim \text{InvG}(a, b)$.
- Alternatively, we might specify a $\mathcal{G}(a, b)$ prior distribution for the precision $\tau = 1/\sigma^2$.
- The joint prior for $\beta$ and $\sigma^2$ is a **normal-inverse gamma distribution** with density
  $$p(\beta, \sigma^2) = p(\beta | \sigma^2)p(\sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{k+1}{2} + a + 1}} \exp \left( -\frac{1}{2\sigma^2} (\beta - m)^\top V^{-1} (\beta - m) - \frac{b}{\sigma^2} \right).$$
- We write $\beta, \sigma^2 \sim \text{NIG}(m, V, a, b)$. 
Unconditional prior distribution of $\beta$

- We integrate the joint prior $p(\beta, \sigma^2)$ with respect to $\sigma^2$ and obtain

$$p(\beta) \propto \left( 1 + \frac{1}{2a} (\beta - m)\top \left( \frac{b}{a} V \right)^{-1} (\beta - m) \right)^{-\left( a + \frac{k+1}{2} \right)}.$$

- This is the density of a multivariate t-distribution with $2a$ degrees of freedom, having location parameter $m$ and dispersion matrix $b/aV$.

- We write $\beta \sim t_{k+1}(2a, m, b/aV)$. 
Introduction
Bayesian inference
Conjugate analysis for the linear regression model

Posterior distribution

Results:

- The joint posterior, \( g(\beta, \sigma^2 | y) \), can be written as
  \( \beta, \sigma^2 | y \sim NIG(\tilde{m}, \tilde{V}, \tilde{a}, \tilde{b}) \) with parameters
  \[
  \tilde{V} = (X^\top X + V^{-1})^{-1}, \quad \tilde{m} = \tilde{V}(V^{-1}m + X^\top y), \\
  \tilde{a} = a + \frac{n}{2}, \quad \tilde{b} = b + \frac{1}{2} \left( y^\top y + m^\top V^{-1}m - \tilde{m}^\top \tilde{V}^{-1} \tilde{m} \right).
  \]

- The conditional posterior distribution of \( \beta \) given \( \sigma^2 \) is
  \( \beta | \sigma^2, y \sim N_{k+1}(\tilde{m}, \sigma^2 \tilde{V}) \).

- The marginal posterior of \( \beta \) is
  \( \beta | y \sim t_{k+1}(2\tilde{a}, \tilde{m}, \tilde{b}/\tilde{a} \tilde{V}) \).
One noninformative prior in the linear model is the improper prior $p(\beta, \sigma^2) \propto 1/\sigma^2$, which maximizes the influence of the data on the posterior.

The prior can be expressed as the product of a uniform prior $p(\beta)$ over the $(k+1)$-dimensional space and the prior $p(\sigma^2) \propto 1/\sigma^2$ for $\sigma^2$.

Note that the prior for $\sigma^2$ is equivalent to a uniform prior for $\ln(\sigma^2)$ on $(-\infty, \infty)$.

Technically, we can identify the prior $p(\beta, \sigma^2) \propto 1/\sigma^2$ by taking $m \to 0$, $V^{-1} \to O$, $a \to -(k+1)/2$ and $b \to 0$. 

Winter term 2018/19
Posterior with noninformative prior

**Results:**

- In case of a noninformative prior with $\mathbf{m} \to \mathbf{0}$, $\mathbf{V}^{-1} \to \mathbf{0}$, $a \to -(k+1)/2$ and $b \to 0$, we obtain
  
  \[
  \tilde{\mathbf{V}} = (\mathbf{X}^\top \mathbf{X})^{-1}, \quad \tilde{\mathbf{m}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \hat{\beta},
  \]

  \[
  \tilde{a} = (n - k - 1)/2, \quad \tilde{b} = \frac{1}{2} \left( \mathbf{y}^\top \mathbf{y} + \hat{\beta}^\top \mathbf{X}^\top \mathbf{X} \hat{\beta} \right).
  \]

- The conditional posterior distribution of $\beta$ given $\sigma^2$ is
  \[
  \beta | \sigma^2, \mathbf{y} \sim \mathcal{N}_{k+1}(\hat{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}).
  \]

- The marginal posterior of $\beta$ is
  \[
  \beta | \mathbf{y} \sim t_{k+1}(n - k - 1, \hat{\beta}, s^2 (\mathbf{X}^\top \mathbf{X})^{-1}),
  \]

  where $s^2 = \mathbf{y}^\top (\mathbf{I} - \mathbf{H}) \mathbf{y} / (n - k - 1)$ and $\mathbf{H}$ the hat matrix.
Suppose we have observed a new $n \times (k + 1)$ matrix $X_0$, and we wish to predict the corresponding outcome $y_0$.

If $\beta$ and $\sigma^2$ were known, then the probability law for the predicted outcomes would be described as $y_0 \sim \mathcal{N}_n(X_0\beta, \sigma^2 I)$ and would be independent of $y$.

However, these parameters are not known. Instead they are summarized through their posterior samples.

Therefore, all predictions for the data must follow from the posterior predictive distribution.
Conjugate analysis for the linear regression model

Posterior predictive distribution

Results:

- For the **normal-inverse gamma prior**, the posterior predictive density can be written as

\[ p(y_0 | y) = \int p(y_0 | \beta, \sigma^2) p(\beta, \sigma^2 | y) \, d\beta \, d\sigma^2 . \]

- Since \( y_0 | \beta, \sigma^2 \sim \mathcal{N}(X_0 \beta, \sigma^2 I) \) and \( \beta, \sigma^2 | y \sim \text{NIG}(\tilde{m}, \tilde{V}, \tilde{a}, \tilde{b}) \), it turns out that

\[ y_0 | y \sim t_n \left( 2\tilde{a}, X_0 \tilde{m}, \tilde{b} / \tilde{a}(I + X_0 \tilde{V} X_0^T) \right) . \]

- For the **improper prior** discussed above, we obtain the posterior predictive density as a

\[ t_n \left( n - k - 1, X_0 \hat{\beta}, s^2(I + X_0 (X^T X)^{-1} X_0^T) \right) . \]
Gibbs sampling for the linear regression model

1. Specify a starting value $\sigma^2_{(0)}$ [for example, use $s^2$].

2. Sample $\beta_{(t)}$ by drawing from the full conditional distribution $\mathcal{N}_{k+1}(\tilde{m}, \sigma^2_{(t-1)} \tilde{V})$.

3. Sample $\sigma^2_{(t)}$ by drawing from the full conditional distribution $\text{InvG}(a', b')$, where $a' = a + n/2 + (k + 1)/2$ and $b' = b + \frac{1}{2}(y - X\beta_{(t)})^\top (y - X\beta_{(t)}) + \frac{1}{2}(\beta_{(t)} - m)^\top V^{-1}(\beta_{(t)} - m)$.

4. Stop if $t = T$, otherwise set $t = t + 1$ and go to step 2.

5. Discard the first $B$ draws (as burn-in), and consider the last $T - B$ draws $(\beta_{(t)}, \sigma^2_{(t)})$ to be draws from the joint posterior distribution.