WIKIPEDIA

Bayesian linear regression

In statistics, **Bayesian linear regression** is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. When the regression model has errors that have a normal distribution, and if a particular form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

Contents

Model setup

Consider a standard linear regression problem, in which for $i = 1, \dots, n$ we specify the mean of the conditional distribution of y_i given a $k \times 1$ predictor vector $\mathbf{x}_i$:

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector, and the ε_i are independent and identically normally distributed random variables:

$$\varepsilon_i \sim N(0, \sigma^2).$$

This corresponds to the following likelihood function:

$$\rho(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right).$$

The ordinary least squares solution is used to estimate the coefficient vector using the Moore-Penrose pseudoinverse:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

where $\mathbf{X}$ is the $n \times k$ design matrix, each row of which is a predictor vector $\mathbf{x}_i^{\mathrm{T}}$; and $\mathbf{y}$ is the column n-vector $[y_1 \ \cdots \ y_n]^{\mathrm{T}}$.

This is a frequentist approach, and it assumes that there are enough measurements to say something meaningful about $\boldsymbol{\beta}$. In the Bayesian approach, the data are supplemented with additional information in the form of a prior probability distribution. The prior belief about the parameters is combined with the data's likelihood function according to Bayes theorem to yield the posterior belief about the parameters $\boldsymbol{\beta}$ and σ. The prior can take different functional forms depending on the domain and the information that is available *a priori*.

With conjugate priors

Conjugate prior distribution

For an arbitrary prior distribution, there may be no analytical solution for the posterior distribution. In this section, we will consider a so-called conjugate prior for which the posterior distribution can be derived analytically.

A prior $\rho(\boldsymbol{\beta}, \sigma^2)$ is conjugate to this likelihood function if it has the same functional form with respect to $\boldsymbol{\beta}$ and σ. Since the log-likelihood is quadratic in $\boldsymbol{\beta}$, the log-likelihood is re-written such that the likelihood becomes normal in $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$. Write

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

The likelihood is now re-written as

$$\rho(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{v}{2}} \exp\left(-\frac{vs^2}{2\sigma^2}\right)(\sigma^2)^{-\frac{n-v}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right),$$

where

$$vs^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad \text{and} \quad v = n - k,$$

where k is the number of regression coefficients.

This suggests a form for the prior:

$$\rho(\boldsymbol{\beta}, \sigma^2) = \rho(\sigma^2)\rho(\boldsymbol{\beta}|\sigma^2),$$

where $\rho(\sigma^2)$ is an inverse-gamma distribution

$$\rho(\sigma^2) \propto (\sigma^2)^{-\frac{v_0}{2}-1} \exp\left(-\frac{v_0 s_0^2}{2\sigma^2}\right).$$

In the notation introduced in the inverse-gamma distribution article, this is the density of an $\textbf{Inv-Gamma}(a_0, b_0)$ distribution with $a_0 = \frac{v_0}{2}$ and $b_0 = \frac{1}{2} v_0 s_0^2$ with v_0 and s_0^2 as the prior values of v and s^2, respectively. Equivalently, it can also be described as a scaled inverse chi-squared distribution, $\textbf{Scale-inv-}\chi^2(v_0, s_0^2)$.

Further the conditional prior density $\rho(\boldsymbol{\beta}|\sigma^2)$ is a normal distribution,

$$\rho(\boldsymbol{\beta}|\sigma^2) \propto (\sigma^2)^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Lambda}_0 (\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right).$$

In the notation of the normal distribution, the conditional prior distribution is $\mathcal{N}\left(\boldsymbol{\mu}_0, \sigma^2 \boldsymbol{\Lambda}_0^{-1}\right)$.

Posterior distribution

With the prior now specified, the posterior distribution can be expressed as

$$\begin{aligned}\rho(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) &\propto \rho(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)\rho(\boldsymbol{\beta}|\sigma^2)\rho(\sigma^2) \\ &\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)(\sigma^2)^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_0(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right)(\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2}\right)\end{aligned}$$

With some re-arrangement,[1] the posterior can be re-written so that the posterior mean $\boldsymbol{\mu}_n$ of the parameter vector $\boldsymbol{\beta}$ can be expressed in terms of the least squares estimator $\hat{\boldsymbol{\beta}}$ and the prior mean $\boldsymbol{\mu}_0$, with the strength of the prior indicated by the prior precision matrix $\boldsymbol{\Lambda}_0$

$$\boldsymbol{\mu}_n = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0)^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0).$$

To justify that $\boldsymbol{\mu}_n$ is indeed the posterior mean, the quadratic terms in the exponential can be re-arranged as a quadratic form in $\boldsymbol{\beta} - \boldsymbol{\mu}_n$.[2]

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_0(\boldsymbol{\beta}-\boldsymbol{\mu}_0) = (\boldsymbol{\beta}-\boldsymbol{\mu}_n)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\boldsymbol{\Lambda}_0)(\boldsymbol{\beta}-\boldsymbol{\mu}_n) + \mathbf{y}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\mu}_n^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\boldsymbol{\Lambda}_0)\boldsymbol{\mu}_n + \boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0.$$

Now the posterior can be expressed as a normal distribution times an inverse-gamma distribution:

$$\rho(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto (\sigma^2)^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_n)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\boldsymbol{\Lambda}_0)(\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right)(\sigma^2)^{-\frac{n+2a_0}{2}-1} \exp\left(-\frac{2b_0 + \mathbf{y}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\mu}_n^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\boldsymbol{\Lambda}_0)\boldsymbol{\mu}_n + \boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0}{2\sigma^2}\right.$$

Therefore, the posterior distribution can be parametrized as follows.

$$\rho(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto \rho(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})\rho(\sigma^2|\mathbf{y}, \mathbf{X}),$$

where the two factors correspond to the densities of $\mathcal{N}\left(\boldsymbol{\mu}_n, \sigma^2\boldsymbol{\Lambda}_n^{-1}\right)$ and $\textbf{Inv-Gamma}\left(a_n, b_n\right)$ distributions, with the parameters of these given by

$$\begin{aligned}&\boldsymbol{\Lambda}_n = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0), \quad \boldsymbol{\mu}_n = (\boldsymbol{\Lambda}_n)^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0), \\ &a_n = a_0 + \frac{n}{2}, \qquad b_n = b_0 + \frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^{\mathrm{T}}\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n).\end{aligned}$$

This can be interpreted as Bayesian learning where the parameters are updated according to the following equations.

$$\begin{aligned}&\boldsymbol{\mu}_n = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0)^{-1}(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0)^{-1}(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \mathbf{X}^{\mathrm{T}}\mathbf{y}), \\ &\boldsymbol{\Lambda}_n = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0), \\ &a_n = a_0 + \frac{n}{2}, \\ &b_n = b_0 + \frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^{\mathrm{T}}\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n).\end{aligned}$$

Model evidence

The model evidence $p(\mathbf{y}|m)$ is the probability of the data given the model m. It is also known as the marginal likelihood, and as the *prior predictive density*. Here, the model is defined by the likelihood function $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma)$ and the prior distribution on the parameters, i.e. $p(\boldsymbol{\beta}, \sigma)$. The model evidence captures in a single number how well such a model explains the observations. The model evidence of the Bayesian linear regression model presented in this section can be used to compare competing linear models by Bayesian model comparison. These models may differ in the number and values of the predictor variables as well as in their priors on the model parameters. Model complexity is already taken into account by the model evidence, because it marginalizes out the parameters by integrating $p(\mathbf{y}, \boldsymbol{\beta}, \sigma|\mathbf{X})$ over all possible values of $\boldsymbol{\beta}$ and σ.

$$p(\mathbf{y}|m) = \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma)\, p(\boldsymbol{\beta}, \sigma)\, d\boldsymbol{\beta}\, d\sigma$$

This integral can be computed analytically and the solution is given in the following equation.[3]

$$p(\mathbf{y}|m) = \frac{1}{(2\pi)^{n/2}} \sqrt{\frac{\det(\boldsymbol{\Lambda}_0)}{\det(\boldsymbol{\Lambda}_n)}} \cdot \frac{b_0^{a_0}}{b_n^{a_n}} \cdot \frac{\Gamma(a_n)}{\Gamma(a_0)}$$

Here Γ denotes the gamma function. Because we have chosen a conjugate prior, the marginal likelihood can also be easily computed by evaluating the following equality for arbitrary values of $\boldsymbol{\beta}$ and σ.

$$p(\mathbf{y}|m) = \frac{p(\boldsymbol{\beta}, \sigma|m)\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma, m)}{p(\boldsymbol{\beta}, \sigma|\mathbf{y}, \mathbf{X}, m)}$$

Note that this equation is nothing but a re-arrangement of Bayes theorem. Inserting the formulas for the prior, the likelihood, and the posterior and simplifying the resulting expression leads to the analytic expression given above.

Other cases

In general, it may be impossible or impractical to derive the posterior distribution analytically. However, it is possible to approximate the posterior by an approximate Bayesian inference method such as Monte Carlo sampling[4] or variational Bayes.

The special case $\boldsymbol{\mu}_0 = \mathbf{0}, \boldsymbol{\Lambda}_0 = c\mathbf{I}$ is called ridge regression.

A similar analysis can be performed for the general case of the multivariate regression and part of this provides for Bayesian estimation of covariance matrices: see Bayesian multivariate linear regression.

See also

- Bayes linear statistics
- Regularized least squares
- Tikhonov regularization
- Spike and slab variable selection
- Bayesian interpretation of kernel regularization

Notes

1. The intermediate steps of this computation can be found in O'Hagan (1994) at the beginning of the chapter on Linear models.
2. The intermediate steps are in Fahrmeir et al. (2009) on page 188.
3. The intermediate steps of this computation can be found in O'Hagan (1994) on page 257.
4. Carlin and Louis(2008) and Gelman, et al. (2003) explain how to use sampling methods for Bayesian linear regression.

References

- Box, G. E. P.; Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Wiley. ISBN 0-471-57428-7.
- Carlin, Bradley P.; Louis, Thomas A. (2008). *Bayesian Methods for Data Analysis, Third Edition*. Boca Raton, FL: Chapman and Hall/CRC. ISBN 1-58488-697-8.
- Fahrmeir, L.; Kneib, T.; Lang, S. (2009). *Regression. Modelle, Methoden und Anwendungen* (Second ed.). Heidelberg: Springer. doi:10.1007/978-3-642-01837-4 (https://doi.org/10.1007%2F978-3-642-01837-4). ISBN 978-3-642-01836-7.
- Fornalski K.W.; Parzych G.; Pylak M.; Satuła D.; Dobrzyński L. (2010). "Application of Bayesian reasoning and the Maximum Entropy Method to some reconstruction problems". *Acta Physica Polonica A*. **117** (6): 892–899. doi:10.12693/APhysPolA.117.892 (https://doi.org/10.12693%2FAPhysPolA.117.892).
- Fornalski, Krzysztof W. (2015). "Applications of the robust Bayesian regression analysis". *International Journal of Society Systems Science*. **7** (4): 314–333. doi:10.1504/IJSSS.2015.073223 (https://doi.org/10.1504%2FIJSSS.2015.073223).
- Gelman, Andrew; Carlin, John B.; Stern, Hal S.; Rubin, Donald B. (2003). *Bayesian Data Analysis, Second Edition*. Boca Raton, FL: Chapman and Hall/CRC. ISBN 1-58488-388-X.
- Goldstein, Michael; Wooff, David (2007). *Bayes Linear Statistics, Theory & Methods*. Wiley. ISBN 978-0-470-01562-9.
- Minka, Thomas P. (2001) *Bayesian Linear Regression* (http://research.microsoft.com/~minka/papers/linear.html), Microsoft research web page
- Rossi, Peter E.; Allenby, Greg M.; McCulloch, Robert (2006). *Bayesian Statistics and Marketing*. John Wiley & Sons. ISBN 0470863676.
- O'Hagan, Anthony (1994). *Bayesian Inference*. Kendall's Advanced Theory of Statistics. **2B** (First ed.). Halsted. ISBN 0-340-52922-9.
- Sivia, D.S.; Skilling, J. (2006). *Data Analysis - A Bayesian Tutorial* (Second ed.). Oxford University Press.
- Walter, Gero; Augustin, Thomas (2009). "Bayesian Linear Regression—Different Conjugate Models and Their (In)Sensitivity to Prior-Data Conflict" (http://epub.ub.uni-muenchen.de/11050/1/tr069.pdf) (PDF). *Technical Report Number 069, Department of Statistics, University of Munich*.

External links

- Bayesian estimation of linear models (R programming wikibook). Bayesian linear regression as implemented in R.