Comparing the Bayesian Linear Regression Model to Ordinary Least Squares

Joe Zhang and Christian Paravalos

Winter 2024

Abstract

Linear regression is often used to model relationships between a response variable and one or more explanatory variables. In traditional settings, the ordinary least squares method is used to estimate the model. In this paper, our goal is to develop a linear model by estimating the model parameters under a Bayesian framework and compare the model against the traditional ordinary least squares model using a dataset.

Introduction

To model the relationship between a response variable and one or more explanatory variables, the simplest way to estimate a model is to assume that there is a linear relationship between the response variable and the predictor variables. The formulation of the model is

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

where X is an $n \times k$ matrix with the observations of the explanatory variables, $\boldsymbol{y}$ is an $n \times 1$ vector of observations of the response variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of model parameters, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors with each observation, n is the number of observations, and k is the number of parameters. The goal is to estimate $\boldsymbol{\beta}$ and then construct and evaluate the model from the estimated values. In other words, we have

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The most well-known way is to use the ordinary least squares method to estimate the parameters. The goal of this paper is to compare the ordinary least squares method with the Bayesian method of estimating parameters.

Ordinary Least Squares Regression Revisited

Estimation of Model Parameters

Consider a multiple linear regression model

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

assume that the random errors have mean 0, variance σ^2, are uncorrelated, and they follow a normal distribution. In other words, we assume that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$, where $\mathbf{0}$ is the zero vector, and I is the identity matrix.

To estimate the coefficients, the following quantity should be minimized:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta})$$

We need to take derivatives of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. We get that

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T\boldsymbol{y} + 2X^T X\boldsymbol{\beta}$$

If we set this equation to 0 and we solve for $\boldsymbol{\beta}$, we get

$$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}$$

If we multiply both sides by the inverse of $X^T X$, assuming that it is invertible, then the least squares estimator of $\boldsymbol{\beta}$ is

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

The least squares estimator of $\boldsymbol{\beta}$ has the following properties:

- $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
- $Var[\hat{\boldsymbol{\beta}}] = \sigma^2 (X^T X)^{-1}$

Estimation of Variance

In the frequentist approach, we estimate σ^2 using the residual sum of squares. We have

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^T \boldsymbol{e}$$

where $\hat{y}_i$ are the fitted values of y_i using the least squares model and $e_i = y_i - \hat{y}_i$ are the residuals. $\boldsymbol{e}$ is the vector of the residuals. If we simplify further, we get

$$SS_{res} = \boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}^T X^T \boldsymbol{y}$$

The residual sum of squares has $n - k$ degrees of freedom. The residual mean square is

$$MS_{res} = \frac{SS_{res}}{n - p}$$

We can show that

$$\mathbb{E}[MS_{res}] = \sigma^2$$

Therefore, an unbiased estimator for σ^2 is given by

$$\hat{\sigma^2} = MS_{res} = \frac{\boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}^T X^T \boldsymbol{y}}{n - k}$$

Inference on model parameters

After obtaining the estimates of the model parameters, we need to make inference on them. The main goal of inference is to determine if all the explanatory variables are significant to the model. Sometimes when there are too many variables, the accuracy of predictions using the model can decrease significantly. Making inferences can help decide if some variables should not be included in the model.

We have assumed that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$. This consequently leads to that $\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$. Under this assumption, we can also get that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

From this, we can also deduce that

$$\hat{\beta_j} \sim N(\beta_j, \sigma^2 (X^T X)^{-1}_{jj})$$

for $j = 1, 2, ..., k$ and $(X^T X)^{-1}_{jj}$ is the j-th diagonal element of the matrix $(X^T X)^{-1}$.

If we standardize $\hat{\beta}$, we get that

$$\frac{\hat{\beta_j} - \beta_j}{\sigma \sqrt{(X^T X)^{-1}_{jj}}} \sim N(0, 1)$$

In practice, we do not know the value of σ^2, which we will need to use in order to make inference on the model parameters. Therefore, the best way is to use the estimate for σ^2, which is MS_{res}. However, if we replace σ^2 with MS_{res}, we get

$$\frac{\hat{\beta_j} - \beta_j}{\sqrt{MS_{res} (X^T X)^{-1}_{jj}}} \sim t(n - k)$$

If we want to test the null hypothesis $H_0 : \beta_j = \beta_{j_0}$ against the hypothesis $H_1 : \beta_j \neq \beta_{j_0}$, we have that under H_0 that

$$t_0 = \frac{\hat{\beta_j} - \beta_{j_0}}{\sqrt{MS_{res} (X^T X)^{-1}_{jj}}} \sim t(n - k)$$

t_0 is the test statistic for the hypothesis $\beta_j = \beta_{j_0}$. We reject H_0 if $t_0 > t_{\alpha/2}(n-k)$, where α is the significance level and $t_{\alpha/2}(n-k)$ is the $100(1-\alpha/2)$th quantile of the Student t distribution with $n-k$ degrees of freedom. We can also compute the p-value, which is

$$p\text{-value} = P(|T| > t_0) = 2P(T > t_0)$$

and we reject H_0 if it is less than α.

Another way to test the hypothesis H_0 against H_1 is to construct confidence intervals for the model parameters β_j. We can deduce that a $100(1 - \alpha)\%$ confidence interval for β_j is

$$\hat{\beta_j} \pm t_{\alpha/2}(n - k) \cdot \sqrt{MS_{res} (X^T X)^{-1}_{jj}}$$

From the confidence interval, we can reject H_0 if β_{j_0} is contained in the interval. Otherwise, we accept H_0.

Prediction

The purpose of constructing a model between a response variable and several explanatory variables is to make predictions of future values of the response given certain values of the explanatory variables. To predict $\boldsymbol{y}$ at a given X, we simply use the fitted model

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$$

To construct a confidence interval for a prediction, we have that

Bayesian Linear Regression

Now, we turn to the Bayesian approach to estimate a linear regression model between a response variable and several explanatory variables. The form of the model is the same as in the ordinary least squares framework

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

Estimation of Parameters

We make the same assumptions as in the ordinary least squares framework, where $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 I)$. We need to estimate $k+1$ parameters: the k model coefficients, and σ^2.

Likelihood Function

With the assumption, we have that

$$\boldsymbol{y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$$

Therefore, the likelihood function is

$$f(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left[\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})\right]$$

Choosing a prior distribution

To estimate $\boldsymbol{\beta}$ and σ^2, we treat them as random variables and we make a prior assumption on their distribution. The assumption we will use are

$$\boldsymbol{\beta}|\sigma^2 \sim N_k(\boldsymbol{m}, \sigma^2 V)$$

and

$$\sigma^2 \sim IG(a, b)$$

where $\boldsymbol{m}$ is a $k \times 1$ vector of real numbers, V is a $k \times k$ matrix of real numbers, and a and b are both real numbers.

Then, we get that the prior for $(\boldsymbol{\beta}, \sigma^2)$ is

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

where

$$g(\boldsymbol{\beta}|\sigma^2) = \frac{1}{(2\pi\sigma^2)^{k/2}}|V|^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{m})^T V^{-1}(\boldsymbol{\beta} - \boldsymbol{m})\right)$$

and

$$g(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} \exp\left(-\frac{b}{\sigma^2}\right)$$

Computation of posterior distribution

Now that we have the likelihood function and the prior probability distribution of the parameters, we can now compute the posterior distribution of the parameters. We let $h(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, X)$ represent the posterior distribution function:

$$\begin{aligned}
h(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, X) &\propto f(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2) g(\boldsymbol{\beta}, \sigma^2) \\
&\propto f(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2) g(\boldsymbol{\beta}|\sigma^2) g(\sigma^2) \\
&\propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})\right) (2\pi\sigma^2)^{-\frac{k}{2}} |V|^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{m})^T V^{-1} (\boldsymbol{\beta} - \boldsymbol{m})\right) \\
&\times \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp\left(-\frac{b}{\sigma^2}\right) \\
&\propto (\sigma^2)^{-\frac{n}{2} - \frac{k}{2} - (a+1)} \exp\left(-\frac{1}{2\sigma^2}[(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{m})^T V^{-1}(\boldsymbol{\beta} - \boldsymbol{m}) + 2b]\right) \\
&\propto (\sigma^2)^{-\frac{n}{2} - \frac{k}{2} - (a+1)} \exp\left(-\frac{1}{2\sigma^2} A\right)
\end{aligned}$$

We now want to rewrite the equation above by simplifying the quantity inside the exponential function. We do the following:

$$\begin{aligned}
A &= (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{m})^T V^{-1}(\boldsymbol{y} - \boldsymbol{m}) + 2b \\
&= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\beta} - \boldsymbol{\beta}^T X^T\boldsymbol{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T V^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^T V^{-1}\boldsymbol{m} - \boldsymbol{m}^T V^{-1}\boldsymbol{\beta} + \boldsymbol{m}^T V^{-1}\boldsymbol{m} + 2b \\
&= \boldsymbol{\beta}^T(X^T X + V^{-1})\boldsymbol{\beta} - \boldsymbol{\beta}^T(X^T\boldsymbol{y} + V^{-1}\boldsymbol{m}) + (\boldsymbol{m}^T V^{-1}\boldsymbol{m} + 2b + \boldsymbol{y}^T\boldsymbol{y}) - (\boldsymbol{y}^T X + \boldsymbol{m}^T V^{-1})\boldsymbol{\beta}
\end{aligned}$$

We define the following quantities:

$$\Lambda = (X^T X + V^{-1})^{-1}$$

and

$$\boldsymbol{\mu} = (X^T X + V^{-1})^{-1}(X^T\boldsymbol{y} + V^{-1}\boldsymbol{m})$$

where Λ is a $k \times k$ matrix and $\boldsymbol{\mu}$ is a $k \times 1$ vector.

Then we can write

$$\begin{aligned}
A &= \boldsymbol{\beta}^T\Lambda^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^T\Lambda^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\Lambda^{-1}\boldsymbol{\beta} + \boldsymbol{m}^T V^{-1}\boldsymbol{m} + 2b + \boldsymbol{y}^T\boldsymbol{y} \\
&= (\boldsymbol{\beta} - \boldsymbol{\mu})^T\Lambda^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}) - \boldsymbol{\mu}^T\Lambda^{-1}\boldsymbol{\mu} + \boldsymbol{m}^T V^{-1}\boldsymbol{m} + 2b + \boldsymbol{y}^T\boldsymbol{y}
\end{aligned}$$

Now, we go back to the posterior distribution and we can write:

$$h(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2} - \frac{k}{2} - (a+1)} \exp\left(-\frac{1}{2\sigma^2} A\right)$$

Application

Conclusion

References

Note: These are not cited in proper format yet.

https://en.wikipedia.org/wiki/Bayesian_linear_regression

https://gregorygundersen.com/blog/2020/02/04/bayesian-linear-regression/

https://www.researchgate.net/publication/333917874_Bayesian_Linear_Regression#pf18