

STAT 238 - Bayesian Statistics

Lecture Sixteen

Spring 2026, UC Berkeley

Aditya Guntuboyina

02 March 2026

1 Linear Regression

In the last lecture, we discussed Bayesian inference for multiple linear regression. The setting is as follows: we have one response variable y and m covariates x_1, \dots, x_m ($m = 1$ corresponds to simple linear regression). We observe data on n instances or subjects for all these variables: $(y_i, x_{i1}, \dots, x_{im})$ for $i = 1, \dots, n$. The multiple linear regression model (with normal errors) is given by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} + \epsilon_i \quad \text{with } \epsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2). \quad (1)$$

Using the matrix vector notation:

$$y = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 1 & x_{n1} & \dots & x_{nm} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_m \end{pmatrix} \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdot \\ \cdot \\ \hat{\beta}_m \end{pmatrix}$$

the model can be written as:

$$y = X\beta + \epsilon \quad \text{where } \epsilon \sim N(0, \sigma^2 I_n).$$

Below we review standard frequentist and Bayesian approaches to linear regression. Both lead to identical answers.

1.1 Frequentist Approach

Here the focus is on the least squares estimate which is given by:

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

To perform inference, the key is to compute the distribution of $\hat{\beta}$. The dependence of the above on the X matrix is quite complicated. On the other hand, the dependence on y is linear. If we assume that X is fixed (non-random), then it is easy to write the distribution of $\hat{\beta}$. Indeed because $y \sim N(X\beta, \sigma^2 I_n)$, it is easy to check that

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

This implies that

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^T X)^{j+1, j+1}) \quad \text{for each } j = 0, 1, \dots, m,$$

where $(X^T X)^{j+1, j+1}$ is the $(j+1, j+1)$ -th diagonal entry of $(X^T X)^{-1}$. The above is equivalent to:

$$\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{(X^T X)^{j+1, j+1}}} \sim N(0, 1). \quad (2)$$

Because σ is unknown, it is natural to replace it with a suitable estimate $\hat{\sigma}$. The standard estimate used is the residual standard error:

$$\hat{\sigma} := \sqrt{\frac{S(\hat{\beta})}{n - m - 1}}.$$

where $S(\beta) = \|y - X\beta\|^2$ is the sum of squares and $S(\hat{\beta})$ is the smallest possible value of the sum of squares. Replacing σ by $\hat{\sigma}$ in (2) will change the distribution. In fact, it can be shown that $N(0, 1)$ will be replaced by the standard t distribution with $n - m - 1$ degrees of freedom:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}}} \sim t_{n-m-1}. \quad (3)$$

This can be used to derive confidence intervals for β_j . If $t_{n-m-1, \alpha/2}$ is the point beyond which the t -distribution (with $n - m - 1$ degrees of freedom) assigns probability $\alpha/2$, then

$$\mathbb{P} \left\{ -t_{n-m-1, \alpha/2} \leq \frac{\beta_j - \hat{\beta}_j}{\hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}}} \leq t_{n-m-1, \alpha/2} \right\} = 1 - \alpha$$

which is same as:

$$\mathbb{P} \left\{ \hat{\beta}_j - \hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}} t_{n-m-1, \alpha/2} \leq \beta_j \leq \hat{\beta}_j + \hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}} t_{n-m-1, \alpha/2} \right\} = 1 - \alpha$$

This interval:

$$\left[\hat{\beta}_j - \hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}} t_{n-m-1, \alpha/2}, \hat{\beta}_j + \hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}} t_{n-m-1, \alpha/2} \right] \quad (4)$$

is the $100(1 - \alpha)\%$ confidence interval for β_j .

The quantity $\hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}}$ is known as the standard error corresponding to β_j .

1.2 Bayesian Approach

The default prior for linear regression is:

$$\beta_0, \beta_1, \dots, \beta_m, \log \sigma \stackrel{\text{i.i.d.}}{\sim} \text{unif}(-\infty, \infty).$$

The likelihood is given by:

$$\propto \sigma^{-n} \exp \left(-\frac{1}{2\sigma^2} \|y - X\beta\|^2 \right). \quad (5)$$

To obtain this likelihood, we can assume that X is fixed (and that $\epsilon \sim N(0, \sigma^2 I_d)$). Or we can assume that ϵ is independent of X and that the distribution of X does not depend on the parameters β, σ . There could be other assumptions on ϵ, X which can lead to the same (or very similar) likelihood (see Section 2).

Under this prior and likelihood, we calculated (see last lecture notes) that the posterior density of β is given by:

$$\beta_0, \dots, \beta_m \mid \text{data} \sim t_{m+1} \left(\hat{\beta}, \frac{S(\hat{\beta})}{n-m-1} (X^T X)^{-1}, n-m-1 \right). \quad (6)$$

Note that, in (6), the quantity $S(\hat{\beta})/(n-m-1)$ is the **frequentist unbiased** estimator for σ^2 , which is denoted by $\hat{\sigma}^2$. $\hat{\sigma}$ can also be justified as a Bayesian estimator of σ (this will be a question in Homework three).

With the notation for $\hat{\sigma}$, the posterior (6) becomes:

$$\beta_0, \dots, \beta_m \mid \text{data} \sim t_{m+1} \left(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1}, n-m-1 \right). \quad (7)$$

By one of the facts mentioned about the multivariate t -distribution, the posterior of each individual β_j is also t :

$$\beta_j \mid \text{data} \sim t_1 \left(\hat{\beta}_j, \hat{\sigma}^2 (X^T X)^{j+1, j+1}, n-m-1 \right) \quad (8)$$

where $(X^T X)^{j+1, j+1}$ is the $(j+1)^{\text{th}}$ diagonal entry of $(X^T X)^{-1}$ (note that we are using the $(j+1)$ th diagonal entry of $X^T X$ because β_j is the $(j+1)$ th component of β). This implies that

$$\frac{\beta_j - \hat{\beta}_j}{\hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}}} \mid \text{data} \sim \text{univariate standard } t \text{ with } n-m-1 \text{ d.f.} \quad (9)$$

This can be used to obtain uncertainty intervals for β_j . If $t_{n-m-1, \alpha/2}$ is the point beyond which the t -distribution (with $n-m-1$ degrees of freedom) assigns probability $\alpha/2$, then

$$\mathbb{P} \left\{ -t_{n-m-1, \alpha/2} \leq \frac{\beta_j - \hat{\beta}_j}{\hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}}} \leq t_{n-m-1, \alpha/2} \mid \text{data} \right\} = 1 - \alpha$$

which is same as:

$$\mathbb{P} \left\{ \hat{\beta}_j - \hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}} t_{n-m-1, \alpha/2} \leq \beta_j \leq \hat{\beta}_j + \hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}} t_{n-m-1, \alpha/2} \mid \text{data} \right\} = 1 - \alpha$$

This interval:

$$\left[\hat{\beta}_j - \hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}} t_{n-m-1, \alpha/2}, \hat{\beta}_j + \hat{\sigma} \sqrt{(X^T X)^{j+1, j+1}} t_{n-m-1, \alpha/2} \right] \quad (10)$$

is the $100(1 - \alpha)\%$ Bayesian Credible interval for β_j .

From (4) and (10), it is clear that the Bayesian and frequentist $100(1 - \alpha)\%$ intervals for β_j coincide.

1.3 When n is large

The degrees of freedom corresponding to the t -density in (6) is $n - m - 1$ where n is the number of observations, and m is the number of covariates. Thus **if $n - m - 1$ is large**, then

the posterior distribution (which is actually t) is approximately normal:

$$t_{m+1} \left(\hat{\beta}, \frac{S(\hat{\beta})}{n-m-1} (X^T X)^{-1}, n-m-1 \right) \approx N_{m+1} \left(\hat{\beta}, \frac{S(\hat{\beta})}{n-m-1} (X^T X)^{-1} \right).$$

In other words, when $n-m-1$ is large, the t -density (6) is approximately equal to the $N_{m+1}(\hat{\beta}, \hat{\sigma}^2(X^T X)^{-1})$. Further, when $n-m-1$ is large, the distribution (8) will be close to the normal distribution $N(\hat{\beta}_j, \hat{\sigma}^2(X^T X)^{j+1,j+1})$. In such cases, $t_{n-m-1, \alpha/2}$ can be replaced by $z_{\alpha/2}$ in the intervals.

1.4 Bayesian vs Frequentist

We mentioned above that the Bayesian uncertainty interval for β_j exactly coincides with the standard frequentist confidence interval. But the Bayesian reasoning that led to the interval differs fundamentally from the frequentist reasoning. The Bayesian reasoning is based on the fact (9) while frequentist reasoning is based on (3).

The only difference between (3) and (9) is in the conditioning. In (9), the random variable is β_j and we are computing its distribution conditional on the observed data. In contrast, in the frequentist statement (3), the random variable is $\hat{\beta}_j$ (which is a function of y_1, \dots, y_n) and we are computing its distribution the parameters β, σ are fixed. Note that X is assumed to be fixed here.

The fact that both Bayesian and frequentist reasoning lead to the same interval is a matter of coincidence. By changing the setting slightly, this coincidence can be broken. One such example involves AR models and is discussed next.

2 AutoRegression

We observe a time series $\{y_t, t = 1, \dots, n\}$. The AutoRegressive Model of order 1 (in short, AR(1)) is given by:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t \quad \text{where } t = 2, \dots, n. \quad (11)$$

In the above equation, y_1 does not appear on the right hand side, so we can treat y_1 as a constant. The equation (11) is simply a linear regression model with covariate given by $x_t = y_{t-1}$. Frequentist and Bayesian inference now give different answers (although in practice the answers will be quite similar) and with very different justifications. This is explained below.

2.1 Frequentist Inference

We start with the least squares estimate of $\hat{\beta}$:

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

where

$$y = \begin{pmatrix} y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & y_1 \\ 1 & y_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & y_{n-1} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

Unlike in the previous case, we cannot compute the distribution of $\hat{\beta}$ assuming that X is fixed. This is because X involves depends on y_1, \dots, y_{n-1} . Instead, we need to use other arguments. For example, note that

$$\hat{\beta}_1 = \frac{\sum_{t=2}^n (y_t - \bar{y})(x_t - \bar{x})}{\sum_{t=2}^n (x_t - \bar{x})^2} = \frac{\sum_{t=2}^n (y_t - \bar{y}_{2:n})(y_{t-1} - \bar{y}_{1:(n-1)})}{\sum_{t=2}^n (y_{t-1} - \bar{y}_{1:(n-1)})^2}$$

where $\bar{y} = \bar{y}_{2:n} := (y_2 + \dots + y_n)/(n-1)$ and $\bar{x} = \bar{y}_{1:(n-1)} := (y_1 + \dots + y_{n-1})/(n-1)$. The distribution of $\hat{\beta}_1$ is actually quite complicated. Asymptotic arguments can be used to derive a limiting normal distribution. For the full argument, see, for example, Shumway and Stoffer [1, Subsection B.3] which uses a dependent Central Limit Theorem.

2.2 Bayesian Inference

For Bayesian inference, we need to write the prior and likelihood. The prior will obviously be the same as in usual linear regression:

$$\beta_0, \beta_1, \log \sigma \stackrel{\text{i.i.d}}{\sim} \text{unif}(-\infty, \infty).$$

For the likelihood, we write:

$$\begin{aligned} \text{Likelihood} &= f_{y_1, \dots, y_n | \beta, \sigma}(y_1, \dots, y_n) \\ &= f_{y_1 | \beta, \sigma}(y_1) \prod_{t=2}^n f_{y_t | y_1, \dots, y_{t-1}, \beta, \sigma}(y_t) \\ &= f_{y_1 | \beta, \sigma}(y_1) \prod_{t=2}^n f_{\beta_0 + \beta_1 y_{t-1} + \epsilon_t | y_1, \dots, y_{t-1}, \beta, \sigma}(y_t) \\ &= f_{y_1 | \beta, \sigma}(y_1) \prod_{t=2}^n f_{\epsilon_t | y_1, \dots, y_{t-1}, \beta, \sigma}(y_t - \beta_0 - \beta_1 y_{t-1}). \end{aligned}$$

We can now assume that ϵ_t is independent of y_1, \dots, y_{t-1} for each $t = 2, \dots, n$. This gives

$$\begin{aligned} \text{Likelihood} &= f_{y_1 | \beta, \sigma}(y_1) \prod_{t=2}^n f_{\epsilon_t}(y_t - \beta_0 - \beta_1 y_{t-1}) \\ &= f_{y_1 | \beta, \sigma}(y_1) \prod_{t=2}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(y_t - \beta_0 - \beta_1 y_{t-1})^2}{2\sigma^2}\right). \end{aligned}$$

For $f_{y_1 | \beta, \sigma}(y_1)$, the simplest thing is to assume that it does not depend on the parameters β, σ so that this term can be ignored. This assumption leads to

$$\begin{aligned} \text{Likelihood} &= f_{y_1 | \beta, \sigma}(y_1) \prod_{t=2}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(y_t - \beta_0 - \beta_1 y_{t-1})^2}{2\sigma^2}\right) \\ &\propto \sigma^{-(n-1)} \exp\left(-\frac{1}{2\sigma^2} \|y - X\beta\|^2\right). \end{aligned} \tag{12}$$

This is exactly the same likelihood as in usual linear regression (see (5)) except that n in (5) is now replaced by $n-1$ which is the number of observations in the regression model (11). Because the likelihood is of the same form as in (5), the posterior of β follows the same calculation as in usual linear regression and we obtain:

$$\beta_0, \beta \mid \text{data} \sim t_2\left(\hat{\beta}, \frac{S(\hat{\beta})}{n-3} (X^T X)^{-1}, n-3\right)$$

This is the same as (6) with $n - 1$ replacing n and $m = 2$.

Therefore, Bayesian inference gives the same posterior t -distribution in AutoRegression as in usual linear regression. Unlike frequentist AutoRegression, there is no need for asymptotics assuming $n \rightarrow \infty$. This example is useful for highlighting the differences between Bayesian and frequentist inference.

References

- [1] Shumway, R. H. and Stoffer, D. S. (2017) *Time Series Analysis and Its Applications: With R Examples*. Springer, 4th edition.