STAT 153 & 248 - Time Series
Lecture Four

Fall 2025, UC Berkeley

Aditya Guntuboyina

September 9, 2025

1 Bayesian Inference for Simple Linear Regression

We use the prior

$$\beta_0, \beta_1, \log\sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C)$$

for a large number C (the exact value of C will not matter in the following calculations). Note that as σ is always positive, we have made the uniform assumption on $\log\sigma$ (by the change of variable formula, the density of σ would be given by $f_\sigma(x) = f_{\log\sigma}(\log x)\frac{1}{x} = \frac{I\{-C<\log x<C\}}{2Cx} = \frac{I\{e^{-C}<x<e^C\}}{2Cx}$.

The joint posterior for all the unknown parameters β_0, β_1, σ is then given by (below we write the term "data" for $y_1, \ldots, y_n$):

$$f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma) \propto f_{y_1,\ldots,y_n|\beta_0,\beta_1,\sigma}(y_1,\ldots,y_n) f_{\beta_0,\beta_1,\sigma}(\beta_0,\beta_1,\sigma).$$

The two terms on the right hand side above are the likelihood:

$$f_{y_1,\ldots,y_n|\beta_0,\beta_1,\sigma}(y_1,\ldots,y_n) \propto \sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\beta_0-\beta_1x_i)^2\right) = \sigma^{-n}\exp\left(-\frac{S(\beta_0,\beta_1)}{2\sigma^2}\right),$$

and the prior:

$$\begin{aligned} f_{\beta_0,\beta_1,\sigma}(\beta_0,\beta_1,\sigma) &= f_{\beta_0}(\beta_0)f_{\beta_1}(\beta_1)f_\sigma(\sigma) \\ &\propto \frac{I\{-C<\beta_0<C\}}{2C}\frac{I\{-C<\beta_1<C\}}{2C}\frac{I\{e^{-C}<\sigma<e^C\}}{2C\sigma} \\ &\propto \frac{1}{\sigma}I\left\{-C<\beta_0,\beta_1,\log\sigma<C\right\}. \end{aligned}$$

Recall that $S(\beta_0,\beta_1)$ denotes the sum of squares:

$$S(\beta_0,\beta_1) := \sum_{i=1}^n(y_i-\beta_0-\beta_1x_i)^2.$$

We thus obtain

$$f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma) \propto \sigma^{-n-1}\exp\left(-\frac{S(\beta_0,\beta_1)}{2\sigma^2}\right)I\left\{-C<\beta_0,\beta_1,\log\sigma<C\right\}.$$

The above is the joint posterior over β_0, β_1, σ. The posterior over only the main parameters β_0, β_1 can be obtained by integrating (or marginalizing) the parameter σ.

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) = \int f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma)d\sigma$$
$$\propto I\{-C < \beta_0, \beta_1 < C\} \int_{e^{-C}}^{e^C} \sigma^{-n-1} \exp\left(-\frac{S(\beta_0,\beta_1)}{2\sigma^2}\right) d\sigma.$$

When C is large, the above integral can be evaluated from 0 to ∞ which gives

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) \propto I\{-C < \beta_0, \beta_1 < C\} \int_{0}^{\infty} \sigma^{-n-1} \exp\left(-\frac{S(\beta_0,\beta_1)}{2\sigma^2}\right) d\sigma.$$

The change of variable

$$s = \frac{\sigma}{\sqrt{S(\beta_0,\beta_1)}}$$

allows us to write the integral as

$$\int_0^\infty \sigma^{-n-1} \exp\left(-\frac{S(\beta_0,\beta_1)}{2\sigma^2}\right) d\sigma$$
$$= \left(\frac{1}{S(\beta_0,\beta_1)}\right)^{n/2} \int_0^\infty s^{-n-1} \exp\left(-\frac{1}{2s^2}\right) ds$$
$$\propto \left(\frac{1}{S(\beta_0,\beta_1)}\right)^{n/2}.$$

where, we are using the fact that $\int_0^\infty s^{-n-1} \exp(-1/(2s^2))ds$ is a constant (in the sense that it does not depend on β_0 and β_1).

The posterior density of (β_0, β_1) is thus

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) \propto I\{-C < \beta_0, \beta_1 < C\} \left(\frac{1}{S(\beta_0,\beta_1)}\right)^{n/2}. \tag{1}$$

In words, the posterior density is inversely proportional to $S(\beta_0, \beta_1)^{n/2}$. This implies that the posterior mode is just the least squares estimator $(\hat{\beta}_0, \hat{\beta}_1)$. It is nicer to write the posterior in the following equivalent form:

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) \propto \left(\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{S(\beta_0,\beta_1)}\right)^{n/2} I\{-C < \beta_0, \beta_1 < C\} \tag{2}$$

Note that (1) and (2) represent exactly the same density because the term $(S(\hat{\beta}_0, \hat{\beta}_1))^{n/2}$ does not depend on β_0, β_1 and is thus a constant.

Generally, the density (2) will be quite sharply concentrated around the least squares estimator $(\hat{\beta}_0, \hat{\beta}_1)$ especially when n is large. This is because, when (β_0, β_1) is such that $S(\beta_0, \beta_1)$ is large compared to $S(\hat{\beta}_0, \hat{\beta}_1)$, the quantity

$$\left(\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{S(\beta_0,\beta_1)}\right)^{n/2}$$

would be quite negligible because of the large power $n/2$. As a result, the posterior density $f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1)$ will be concentrated around those values of (β_0, β_1) for which $S(\beta_0, \beta_1)$

is quite close to $S(\hat{\beta}_0, \hat{\beta}_1)$. For example, suppose $n = 791$, and that (β_0, β_1) is such that $S(\beta_0, \beta_1) = (1.1)S(\hat{\beta}_0, \hat{\beta}_1)$. Then

$$\left(\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{S(\beta_0, \beta_1)}\right)^{n/2} = \left(\frac{1}{1.1}\right)^{395.5} \approx 4.26 \times 10^{-17}.$$

Such (β_0, β_1) will thus get negligible posterior probability. Even for (β_0, β_1) such that $S(\beta_0, \beta_1) = (1.01)S(\hat{\beta}_0, \hat{\beta}_1)$, we have

$$\left(\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{S(\beta_0, \beta_1)}\right)^{n/2} = \left(\frac{1}{1.01}\right)^{395.5} \approx 0.02$$

and so such (β_0, β_1) will also get fairly small posterior probability.

To sum up, when n is large, the posterior probability will be concentrated around those (β_0, β_1) for which $S(\beta_0, \beta_1)$ is very close to $S(\hat{\beta}_0, \hat{\beta}_1)$. Generally, this would imply that (β_0, β_1) would itself have to be close to $(\hat{\beta}_0, \hat{\beta}_1)$. For this reason, the indicator term in (2) has no effect when C is large. From now on, we shall drop this indicator term and refer to the Bayesian posterior as simply

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0, \beta_1) \propto \left(\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{S(\beta_0, \beta_1)}\right)^{n/2}. \tag{3}$$

It turns out that this represents a multivariate t-density, as well shall soon. Before that, let us first note that we have essentially the same formula in multiple linear regression as well.

2 Multiple Linear Regression

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

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \epsilon_i \qquad \text{with } \epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2). \tag{4}$$

In the time series context, the model (4) arises in the following ways:

1. **Regression with functions of time**: Suppose we want to fit a quadratic function of time to the data. We can do this via the model (4) with x_1 being the time variable, and x_2 being the squared time i.e., $x_{i1} = i$ and $x_{i2} = i^2$. Suppose we want to fit a simple sinusoidal function to the data. We can do this via (4) with $x_{i1} = \cos(2\pi i/12)$ and $x_{i2} = \sin(2\pi i/12)$.

2. **AutoRegression (AR)**: If we take $x_{ij} = y_{i-j}$, then we get the AR model:

   $$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_m y_{t-m} + \epsilon_t.$$

   The idea here is that we are using the m most recent values of the time series to predict the next observation. AR models are very commonly used and they work quite well for time series.

In Bayesian inference for (4), we work with the prior

$$\beta_0, \beta_1, \ldots, \beta_m, \log \sigma \overset{\text{i.i.d}}{\sim} \text{unif}(-C, C)$$

for a very large positive C. The joint posterior density of $\beta_0, \ldots, \beta_m, \sigma$ is then given by

$$\begin{aligned} &f_{\beta_0,\beta_1,\ldots,\beta_m,\sigma|\text{data}}(\beta_0, \beta_1, \ldots, \beta_m, \sigma) \\ &\propto \sigma^{-n-1} \exp\left(-\frac{S(\beta_0, \ldots, \beta_m)}{2\sigma^2}\right) I\left\{-C < \beta_0, \beta_1, \ldots, \beta_m, \log \sigma < C\right\}. \end{aligned}$$

where we use the notation

$$S(\beta_0, \beta_1, \ldots, \beta_m) := \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2$$

for the sum of squares.

The posterior over only the coefficient parameters β_0, β_1 can be obtained by integrating (or marginalizing) the parameter σ.

$$\begin{aligned} &f_{\beta_0,\beta_1,\ldots,\beta_m|\text{data}}(\beta_0, \beta_1, \ldots, \beta_m) \\ &= \int f_{\beta_0,\beta_1,\ldots,\beta_m,\sigma|\text{data}}(\beta_0, \beta_1, \ldots, \beta_m, \sigma) d\sigma \\ &\propto I\{-C < \beta_0, \beta_1, \ldots, \beta_m < C\} \int_{e^{-C}}^{e^C} \sigma^{-n-1} \exp\left(-\frac{S(\beta_0, \ldots, \beta_m)}{2\sigma^2}\right) d\sigma \\ &\approx I\{-C < \beta_0, \beta_1, \ldots, \beta_m < C\} \int_{0}^{\infty} \sigma^{-n-1} \exp\left(-\frac{S(\beta_0, \ldots, \beta_m)}{2\sigma^2}\right) d\sigma \\ &= I\{-C < \beta_0, \beta_1, \ldots, \beta_m < C\} \left(\frac{1}{S(\beta_0, \ldots, \beta_m)}\right)^{n/2} \int_{0}^{\infty} s^{-n-1} \exp\left(-\frac{1}{2s^2}\right) ds \\ &\propto I\{-C < \beta_0, \beta_1, \ldots, \beta_m < C\} \left(\frac{1}{S(\beta_0, \ldots, \beta_m)}\right)^{n/2} \\ &\approx \left(\frac{1}{S(\beta_0, \ldots, \beta_m)}\right)^{n/2} \propto \left(\frac{S(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)}{S(\beta_0, \beta_1, \ldots, \beta_m)}\right)^{n/2} \end{aligned}$$

where $\hat{\beta}_0, \ldots, \hat{\beta}_m$ denote the least squares estimators of $\beta_0, \ldots, \beta_m$ (i.e., $(\hat{\beta}_0, \ldots, \hat{\beta}_m)$ minimizes $S(\beta_0, \ldots, \beta_m)$ over all values of $\beta_0, \ldots, \beta_m$).

Our posterior density for $\beta_0, \ldots, \beta_m$ is thus:

$$f_{\beta_0,\beta_1,\ldots,\beta_m|\text{data}}(\beta_0, \beta_1, \ldots, \beta_m) \propto \left(\frac{S(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)}{S(\beta_0, \beta_1, \ldots, \beta_m)}\right)^{n/2}. \tag{5}$$

Now we will explain why this is a multivariate t-density.

3 Why is (5) a t-density?

If you go to the wikipedia page (`https://en.wikipedia.org/wiki/Multivariate_t-distribution`) for Multivariate t-distribution, it gives the following formula for the density:

$$f(x) := \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}\sqrt{\det\,\Sigma}}\left[\frac{1}{1+\frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)}\right]^{(\nu+p)/2}$$
$$\propto \left[\frac{1}{1+\frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)}\right]^{(\nu+p)/2}. \tag{6}$$

Their notation for this distribution is $t_p(\mu, \Sigma, \nu)$ where:

1. p denotes dimension of the vector x (this is a p-variate joint density)
2. μ is a $p \times 1$ vector called the location
3. Σ is a $p \times p$ matrix called the scale matrix
4. $\nu > 0$ denotes the degrees of freedom.

It turns out that (5) is a special case of (6) for some p, μ, Σ, ν. To see this, we need to first rewrite (5) using matrix notation which we do in the next section.

4 Matrix Notation for Multiple Linear Regression

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

This notation is used not just to write formulae for linear regression, but also in code. For example, the OLS function in `statsmodels` uses the syntax `sm.OLS(y, X).fit()` to fit the linear regression model, where y ($n \times 1$ vector) and X ($n \times (m+1)$ matrix) are defined above.

With this notation, one can write the sum of squares $S(\beta_0, \dots, \beta_m)$ as:

$$S(\beta) = S(\beta_0, \dots, \beta_m) = \|y - X\beta\|^2.$$

There are two important facts about $S(\beta)$:

1. **Fact 1**: the least squares estimator $\hat{\beta}$ is given by the formula:

$$\hat{\beta} = (X^TX)^{-1}X^Ty. \tag{7}$$

The proof of (7) is as follows. The gradient of $S(\beta)$ is given by

$$\begin{aligned}\nabla S(\beta) &= \nabla\left[\|y - X\beta\|^2\right] \\ &= \nabla\left[(y - X\beta)^T(y - X\beta)\right] \\ &= \nabla\left[y^Ty - \beta^TX^Ty - y^TX\beta + \beta^TX^TX\beta\right] = 2X^Ty - 2X^TX\beta.\end{aligned}$$

Because $\hat{\beta}$ minimizes $S(\beta)$, the gradient should equal zero when $\beta = \hat{\beta}$, and this leads to

$$X^T(y - X\hat{\beta}) = 0 \implies X^T X\hat{\beta} = X^T y \implies \hat{\beta} = (X^T X)^{-1} X^T y. \tag{8}$$

2. **Fact 2**: The following Pythagorean identity holds:

$$S(\beta) = S(\hat{\beta}) + \|X\beta - X\hat{\beta}\|^2 = S(\hat{\beta}) + (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}). \tag{9}$$

To prove (9), write

$$\begin{aligned} S(\beta) &= \|y - X\beta\|^2 \\ &= \|y - X\hat{\beta} + X\hat{\beta} - X\beta\|^2 \\ &= \|y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\beta\|^2 + 2\left\langle y - X\hat{\beta}, X\hat{\beta} - X\beta \right\rangle. \end{aligned}$$

The cross product is zero (leading to (9)) because:

$$\begin{aligned} \left\langle y - X\hat{\beta}, X\hat{\beta} - X\beta \right\rangle &= \left(X\hat{\beta} - X\beta\right)^T \left(y - X\hat{\beta}\right) \\ &= \left(\hat{\beta} - \beta\right)^T X^T(y - X\hat{\beta}) = \left(\hat{\beta} - \beta\right)^T \left(X^T y - X^T X\hat{\beta}\right) = 0 \end{aligned}$$

where we used (8).

Using (9), we can write the posterior density (2) as

$$\begin{aligned} f_{\beta|\text{data}}(\beta) &\propto \left(\frac{S(\hat{\beta})}{S(\beta)}\right)^{n/2} \\ &= \left(\frac{S(\hat{\beta})}{S(\hat{\beta}) + \left(\beta - \hat{\beta}\right)^T X^T X \left(\beta - \hat{\beta}\right)}\right)^{n/2} \\ &= \left(\frac{1}{1 + \left(\beta - \hat{\beta}\right)^T \frac{X^T X}{S(\hat{\beta})} \left(\beta - \hat{\beta}\right)}\right)^{n/2}. \end{aligned} \tag{10}$$

The above formula is a special case of (6) with

$$x = \beta, \quad p = m + 1, \quad \mu = \hat{\beta}, \quad \nu + p = n, \quad \frac{\Sigma^{-1}}{\nu} = \frac{X^T X}{S(\hat{\beta})}$$

or equivalently

$$x = \beta, \quad p = m + 1, \quad \mu = \hat{\beta}, \quad \nu = n - m - 1, \quad \Sigma = \frac{S(\hat{\beta})}{n - m - 1} \left(X^T X\right)^{-1}.$$

We thus have

$$\beta_0, \ldots, \beta_m \mid \text{data} \sim t_{m+1}\left(\hat{\beta}, \frac{S(\hat{\beta})}{n - m - 1} \left(X^T X\right)^{-1}, n - m - 1\right). \tag{11}$$

With the posterior density (11), one can do uncertainty quantification about the parameters $\beta_0, \beta_1, \ldots, \beta_m$. One can generate multiple samples from $t_{m+1}(\hat{\beta}, (S(\hat{\beta})/(n - m - 1))(X^T X)^{-1}, n - m - 1)$ and plot the resulting fitted values to visualize the uncertainty in the coefficients.