STAT 153 & 248 - Time Series
Lecture Three

Spring 2025, UC Berkeley

Aditya Guntuboyina

January 28, 2025

1 Bayesian Inference in Simple Linear Regression

We observe data $(x_1, y_1), \dots, (x_n, y_n)$. In the linear regression model, it is assumed that $x_1, \dots, x_n$ are fixed deterministic values, and that the response values $y_1, \dots, y_n$ satisfy the model equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{with } \epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2).$$

Another way of writing the model is:

$$y_i \overset{\text{independent}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2).$$

There are three parameters in this model: β_0, β_1 and σ^2.

In Bayesian inference, the first step is to select a prior for the unknown parameters β_0, β_1, σ. A reasonable prior reflecting ignorance is

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

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 $y_1, \dots, y_n$):

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

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

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

and the prior:

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

We thus obtain

$$
\begin{aligned}
&f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma) \\
&\propto \sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2\right) I\left\{-C<\beta_0,\beta_1,\log\sigma<C\right\}.
\end{aligned}
$$

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.

$$
\begin{aligned}
f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) &= \int f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma)d\sigma \\
&\propto I\{-C<\beta_0,\beta_1<C\}\int_{e^{-C}}^{e^C}\sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2\right)d\sigma.
\end{aligned}
$$

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

$$
f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1)\propto I\{-C<\beta_0,\beta_1<C\}\int_0^{\infty}\sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2\right)d\sigma.
$$

The change of variable

$$
s=\frac{\sigma}{\sqrt{\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2}}
$$

allows us to write the integral as

$$
\begin{aligned}
&\int_0^{\infty}\sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2\right)d\sigma \\
&=\left(\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2\right)^{-n/2}\int_0^{\infty}s^{-n-1}\exp\left(-\frac{1}{2s^2}\right)ds\propto\left(\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2\right)^{-n/2}.
\end{aligned}
$$

The posterior density of (β_0,β_1) is thus

$$
f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1)\propto I\{-C<\beta_0,\beta_1<C\}\left(\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2\right)^{-n/2}.
$$

Using the notation

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

we write

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

In most regression problems, the least squares criterion $S(\beta_0,\beta_1)$ will take large values (for example, in the US population dataset, the smallest possible value of $S(\beta_0,\beta_1)$ is of the order of billions). This would mean that $\left(\frac{1}{S(\beta_0,\beta_1)}\right)^{n/2}$ would be very small for all values of β_0,β_1 (of course, the normalizing constant in front of (1) would then have to be quite large). In order to not deal with such small values, it makes sense to rewrite the posterior density as:

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

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 $n = 791$ (as in the US population dataset), 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

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

A more precise understanding of the posterior density can be obtained by noting its connection to the multivariate t-density. Before looking at this connection, let us briefly recall t-densities.

1.1 t-densities

We first look at the univariate case.

1.1.1 Univariate t-density

The t-density is obtained by changing the scale of a normally distributed random variable through an independent chi-squared distributed random variable. More precisely, suppose X has the $N(\mu, \sigma^2)$ distribution. First write

$$X = \mu + (X - \mu).$$

Now consider an independent random variable V such that

$$V \sim \chi_v^2.$$

Recall that χ_v^2 is the same as the Gamma$(v/2, 1/2)$ distribution so that

$$f_V(x) \propto x^{\frac{v}{2}-1} e^{-x/2} I\{x > 0\}.$$

We now change the scale of X using V to create a new random variable T by

$$T := \mu + \frac{X - \mu}{\sqrt{\frac{V}{v}}}. \tag{4}$$

The distribution of T will be denoted by $t_v(\mu.\sigma^2)$ (here v is known as the degrees of freedom). The density of T can be derived as follows:

$$f_T(y) = \int_0^\infty f_{T|V=x}(y) f_V(x) dx.$$

Observe now that

$$T \mid V = x \quad = \mu + \frac{X - \mu}{\sqrt{\frac{x}{v}}} \sim N\left(\mu, \sigma^2 \frac{v}{x}\right)$$

so that

$$f_{T|V=x}(y) = \frac{\sqrt{x}}{\sqrt{2\pi}\sigma\sqrt{v}} \exp\left(-\frac{x}{2\sigma^2 v}(y-\mu)^2\right).$$

As a result

$$\begin{aligned}
f_T(y) &= \int_0^\infty f_{T|V=x}(y) f_V(x) dx \\
&\propto \int_0^\infty \frac{\sqrt{x}}{\sqrt{2\pi}\sigma\sqrt{v}} \exp\left(-\frac{x}{2\sigma^2 v}(y-\mu)^2\right) x^{\frac{v}{2}-1} e^{-x/2} dx \\
&\propto \int_0^\infty x^{\frac{v}{2}-\frac{1}{2}} \exp\left(-\frac{x}{2}\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)\right) dx.
\end{aligned}$$

The change of variable

$$t = x\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)$$

now leads to

$$f_T(y) \propto \frac{1}{\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)^{\frac{v+1}{2}}} \int_0^\infty t^{\frac{v}{2}-1} e^{-t/2} dt \propto \frac{1}{\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)^{\frac{v+1}{2}}}.$$

Therefore the density corresponding to the $t_v(\mu, \sigma^2)$ distribution is proportional to

$$y \mapsto \frac{1}{\left(1 + \frac{(y-\mu)^2}{v\sigma^2}\right)^{\frac{v+1}{2}}}.$$

It is useful to note that when the degrees of freedom v is large, the distribution $t_v(\mu, \sigma^2)$ is very close to the normal distribution $N(\mu, \sigma^2)$. There are many ways of seeing this. One way is to note that the mean and variance of $V \sim \chi_v^2$ are given by v and $2v$ respectively. This implies that

$$\mathbb{E}\left(\frac{V}{v}\right) = 1 \quad \text{and} \quad \text{var}\left(\frac{V}{v}\right) = \frac{2v}{v^2} = \frac{2}{v}.$$

Thus when v is large, the random variable $\frac{V}{v}$ has mean 1 and very small variance so that $\frac{V}{v}$ will be very close to 1 with very high probability. As a result, the scale change by $\sqrt{V/v}$ in (4) has little effect so that T will have the same distribution as $X \sim N(\mu, \sigma^2)$.