STAT 153 & 248 - Time Series
Lecture Four

Spring 2025, UC Berkeley

Aditya Guntuboyina

January 30, 2025

1 Bayesian Inference for Regression

We observe a time series $y_1, \dots, y_n$. We can fit a line to this data using the model:

$$y_t = \beta_0 + \beta_1 t + \epsilon_t \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2). \tag{1}$$

We can fit a more complicated trend function such as the cubic function to the data using the model:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2). \tag{2}$$

(1) and (2) are both examples of the multiple linear regression model. More generally, the multiple linear regression model is given by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \epsilon_i \qquad \text{with } \epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2). \tag{3}$$

There are m covariates here and x_{ij} is the i^{th} value of the j^{th} covariate. (1) is a special case of (3) with $m = 1$ and $x_{i1} = i$ for $i = 1, \dots, n$. (2) is a special case of (3) with $m = 3$ and $x_{i1} = i, x_{i2} = i^2, x_{i3} = i^3$. We shall assume that n is much larger than m (the case where n is comparable or even smaller to m is known as high-dimensional linear regression and we shall look at this later).

In Bayesian inference for (3), we work with the prior

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

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

$$f_{\beta_0, \beta_1, \sigma | \text{data}}(\beta_0, \beta_1, \dots, \beta_m, \sigma)$$
$$\propto \sigma^{-n-1} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2\right) I\left\{-C < \beta_0, \beta_1, \dots, \beta_m, \log \sigma < C\right\}.$$

The above is the joint posterior over $\beta_0, \beta_1, \dots, \beta_m \sigma$. The posterior over only the coefficient parameters β_0, β_1 can be obtained by integrating (or marginalizing) the parameter σ.

$$
\begin{aligned}
& f_{\beta_0,\beta_1,\ldots,\beta_m|\text{data}}(\beta_0,\beta_1,\ldots,\beta_m) \\
&= \int f_{\beta_0,\beta_1,\ldots,\beta_m,\sigma|\text{data}}(\beta_0,\beta_1,\ldots,\beta_m,\sigma)d\sigma \\
&\propto I\{-C < \beta_0,\beta_1,\ldots,\beta_m < C\} \int_{e^{-C}}^{e^C} \sigma^{-n-1} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2\right) d\sigma \\
&\approx I\{-C < \beta_0,\beta_1,\ldots,\beta_m < C\} \int_0^\infty \sigma^{-n-1} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2\right) d\sigma \\
&= I\{-C < \beta_0,\beta_1,\ldots,\beta_m < C\} \left(\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2\right)^{-n/2} \int_0^\infty s^{-n-1} \exp\left(-\frac{1}{2s^2}\right) ds \\
&\propto I\{-C < \beta_0,\beta_1,\ldots,\beta_m < C\} \left(\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2\right)^{-n/2} \\
&\approx \left(\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2\right)^{-n/2}.
\end{aligned}
$$

Using the notation

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

we can write

$$f_{\beta_0,\beta_1,\ldots,\beta_m|\text{data}}(\beta_0,\beta_1,\ldots,\beta_m) \propto \left(\frac{1}{S(\beta_0,\beta_1,\ldots,\beta_m)}\right)^{n/2}. \tag{4}$$

The mode of the above posterior is the least squares estimates $\hat{\beta}_0, \ldots, \hat{\beta}_m$ which minimize $S(\beta_0, \ldots, \beta_m)$ over all values of $\beta_0, \ldots, \beta_m$. (4) is equivalent to

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

Note that (4) and (5) represent exactly the same density because the term $(S(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_m)^{n/2}$ does not depend on $\beta_0,\beta_1,\ldots,\beta_m$ and is thus a constant.

The density (5) represents a multivariate t-distribution (see `https://en.wikipedia.org/wiki/Multivariate_t-distribution`). We demonstrate this below. It will be convenient to use the following vector-matrix notation here:

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

With this notation, one can check that

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

The following facts will be important:

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

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{6}$$

The proof of (6) is as follows. The gradient of $S(\beta)$ is given by

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

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

$$X^T(y - X\hat{\beta}) = 0 \implies X^T X\hat{\beta} = X^T y \implies \hat{\beta} = (X^T X)^{-1} X^T y. \tag{7}$$

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

$$S(\beta) = S(\hat{\beta}) + \|X\beta - X\hat{\beta}\|^2 = S(\hat{\beta}) + (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}). \tag{8}$$

To prove (8), 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 (8)) 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 (7).

Using (8), we can write the posterior density (5) as

$$\begin{aligned}f_{\beta|\text{data}}(\beta) &\propto \left(\frac{S(\hat{\beta})}{S(\beta)}\right)^{n/2} \\ &= \left(\frac{S(\hat{\beta})}{S(\hat{\beta}) + \left(\beta - \hat{\beta}\right)^T X^T X \left(\beta - \hat{\beta}\right)}\right)^{n/2} \\ &= \left(1 + (\beta - \hat{\beta})^T \frac{X^T X}{S(\hat{\beta})}(\beta - \hat{\beta})\right)^{-n/2}.\end{aligned} \tag{9}$$

The formula for the multivariate t-distribution will be reviewed next which will make clear that the above is an instance of the t-density.

2 Multivariate t-density

The multivariate t-density is obtained by changing the scale of a **multivariate** normal density. Let X have the p-variate normal distribution $N_p(\mu, \Sigma)$. This means that X is a

$p\times 1$ random vector, μ is a $p\times 1$ vector, Σ is a $p\times p$ (positive-definite) matrix and the density of X is equal to

$$x \mapsto \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

Let V be a chi-squared random variable with v degrees of freedom and assume that V and X are independent. Define

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

Note that X and T are both $p\times 1$ random vectors while V is a scalar. In other words, T is given by

$$\begin{pmatrix} T_1 \\ \cdot \\ \cdot \\ \cdot \\ T_d \end{pmatrix} = \begin{pmatrix} \mu_1 + \frac{X_1-\mu_1}{\sqrt{\frac{V}{v}}} \\ \cdot \\ \cdot \\ \cdot \\ \mu_d + \frac{X_d-\mu_d}{\sqrt{\frac{V}{v}}} \end{pmatrix}. \tag{10}$$

Note specifically that the scale change on each component is given by the same random variable V.

The distribution of this random vector T will be denoted by $t_{v,p}(\mu,\Sigma)$. Its density can be derived in the following way:

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

Observe that

$$T \mid V = x \sim N\left(\mu, \frac{v}{x}\Sigma\right)$$

so that

$$\begin{aligned} f_{T|V=x}(y) &= \frac{1}{(2\pi)^{p/2}\sqrt{\det(\frac{v}{x}\Sigma)}} \exp\left[-\frac{1}{2}(y-\mu)^T\left(\frac{v}{x}\Sigma\right)^{-1}(y-\mu)\right] \\ &= \frac{x^{p/2}}{(2\pi)^{p/2}v^{p/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{x}{2v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right) \end{aligned}$$

where we used $\det(\frac{v}{x}\Sigma) = (v/x)^p\det(\Sigma)$. As a result

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

The change of variable

$$t = x\left[1+\frac{1}{v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]$$

leads to

$$\begin{aligned} f_T(y) &\propto \frac{1}{\left[1+\frac{1}{v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]^{\frac{v+p}{2}}} \int_0^\infty t^{\frac{v+p}{2}-1}e^{-t/2}dt \\ &\propto \frac{1}{\left[1+\frac{1}{v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]^{\frac{v+p}{2}}}. \end{aligned}$$

Therefore the density corresponding to $t_{v,p}(\mu, \Sigma)$ distribution is proportional to

$$y \mapsto \frac{1}{\left[1 + \frac{1}{v}(y-\mu)^T \Sigma^{-1}(y-\mu)\right]^{\frac{v+p}{2}}}. \tag{11}$$

Note that, in the notation $t_{v,p}(\mu, \Sigma)$, v denotes degrees of freedom, p denotes dimension, μ and Σ denote the mean vector and covariance matrix of the corresponding normal random vector X.

When v is large, $t_{v,p}(\mu, \Sigma)$ is very close to $N_p(\mu, \Sigma)$. The following fact will be useful in the sequel.

Fact 2.1. *If $T \sim t_{v,p}(\mu, \Sigma)$ has components $T_1, \dots, T_p$, then, for each $j = 1, \dots, p$,*

$$T_j \sim t_{v,1}(\mu_j, \Sigma(j,j))$$

where μ_j is the j^{th} component of μ and $\Sigma(j,j)$ is the $(j,j)^{th}$ entry of Σ.

This fact follows directly from (10) (and the univariate definition of the t-density $t_{v,1}$) because

$$T_j = \mu_j + \frac{X_j - \mu_j}{\sqrt{\frac{V}{v}}}$$

and $X_j \sim N(\mu_j, \Sigma(j,j))$.

3 Back to the Bayesian Posterior in Linear Regression

Let us compare (9) and (11), and choose the parameters of the t-density so that (11) matches (9). First note that the dimension $p = m+1$ (as β has $m+1$ components). Matching the powers $(n/2)$ and $(p+v)/2$, we get

$$v = n - p = n - m - 1.$$

It is also clear that $\mu = \hat{\beta}$ and

$$\frac{1}{v}\Sigma^{-1} = \frac{X^T X}{S(\hat{\beta})} \implies \Sigma = \frac{S(\hat{\beta})}{v}(X^T X)^{-1} = \frac{S(\hat{\beta})}{n-m-1}(X^T X)^{-1}.$$

We have thus proved that

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

As we remarked in the frequentist treatment of the simple linear regression model, the quantity $S(\hat{\beta})/(n-m-1)$ is the frequentist unbiased estimator of σ^2. So we denote

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

With this notation, we get

$$\beta \mid \text{data} \sim t_{n-m-1,m+1}\left(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1}\right). \tag{12}$$

With the posterior density (12), one can do uncertainty quantification about the parameters $\beta_0, \beta_1, \ldots, \beta_m$. One can generate multiple samples from $t_{n-m-1,m+1}(\hat{\beta}, \hat{\sigma}^2(X^T X)^{-1})$ and plot the resulting fitted values to visualize the uncertainty in the coefficients. One can also use Fact 2.1 to deduce that

$$\beta_j \mid \text{data} \sim t_{n-m-1,1}(\hat{\beta}_j, \hat{\sigma}^2(X^T X)^{j+1,j+1}) \tag{13}$$

where $(X^T X)^{j+1,j+1}$ is the $(j+1)^{th}$ diagonal entry of $(X^T X)^{-1}$. These univariate t-densities describe the marginal uncertainty in the j^{th} coefficient β_j.

When n is large, the t-density (12) is approximately equal to the $N_{m+1}(\hat{\beta}, \hat{\sigma}^2(X^T X)^{-1})$. Further, when n is large, the distribution (13) will be close to the normal distribution $N(\hat{\beta}_j, \hat{\sigma}^2(X^T X)^{j+1,j+1})$. The quantity $\hat{\sigma}\sqrt{(X^T X)^{j+1,j+1}}$ is known as the standard error corresponding to β_j.