STAT 153 & 248 - Time Series
Lecture Five

Fall 2025, UC Berkeley

Aditya Guntuboyina

September 11, 2025

1 Posterior t-density in Multiple Linear Regression

In the last lecture, we saw the following formula for the posterior distribution in multiple linear regression:

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

Recall in multiple regression: the data is $(x_{i1}, \dots, x_{im}, y_i), i = 1, \dots, n$, and the model is:

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

For doing calculations in regression, we use the matrix notation:

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

In terms of this notation, we can rewrite the model equation (2) as:

$$y = X\beta + \epsilon.$$

In (1), $S(\hat{\beta})$ denotes the sum of squares evaluated at the least squares estimator. It is the smallest possible value of the sum of squares, and it is also known as the Residual Sum of Squares (RSS).

(1) represents the joint density of $(\beta_0, \dots, \beta_m)$ given the observed data. It turns out that the posterior distribution of each individual β_j is also given by a t-density. This follows from properties of the multivariate t-density which we go over next.

2 t-density

The formula for the density corresponding to the t-distribution $t_p(\mu, \Sigma, \nu)$ is (see (`https://en.wikipedia.org/wiki/Multivariate_t-distribution`):

$$
\begin{aligned}
f(x) &:= \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}\sqrt{\det\Sigma}} \left[\frac{1}{1+\frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)}\right]^{(\nu+p)/2} \\
&\propto \left[\frac{1}{1+\frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)}\right]^{(\nu+p)/2}. \qquad (3)
\end{aligned}
$$

Here:

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.

Here is some more information about the t-density (3):

1. **Connection to the Multivariate Normal Density**: The most important term in the formula (3) is $(x-\mu)^T\Sigma^{-1}(x-\mu)$. This exact term also appears in the multivariate normal density. If $X \sim N(\mu, \Sigma)$, then the density of X is given by:

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

This suggests that the t-density is closely related to the multivariate normal density. Here is the connection. Suppose $X \sim N_p(\mu, \Sigma)$ and $V \sim \chi^2_\nu$ (this is the chi-squared distribution with ν degrees of freedom) are independent. Then

$$T := \mu + \frac{X-\mu}{\sqrt{V/\nu}} \sim t_p(\mu, \Sigma, \nu). \qquad (4)$$

Thus, in the notation $t_p(\mu, \Sigma, \nu)$, ν denotes degrees of freedom, p denotes dimension, μ and Σ denote the mean vector and covariance matrix of the corresponding normal random vector X. For completeness, we include a proof of (4) in Section 4.

2. **Individual Components as well as Linear Combinations of Components of T are also t-distributed**: Suppose $T \sim t_p(\mu, \Sigma, \nu)$ and the components of T are $T_1, \dots, T_p$. Then each individual component T_j is also t-distributed. Also every linear combination $a_0 + a_1T_1 + a_2T_2 + \dots + a_pT_p$ is also t-distributed. To see this, first write

$$a_0 + a_1T_1 + \dots + a_pT_p = a_0 + a^TT$$

where a is the $p \times 1$ vector with components $a_1, \dots, a_p$. Using the formula (4), we can write

$$a_0 + a^TT = (a_0 + a^T\mu) + \frac{(a_0 + a^TX) - (a_0 + a^T\mu)}{\sqrt{V/\nu}}$$

Because $a_0 + a^TX \sim N(a_0 + a^T\mu, a^T\Sigma a)$, the same fact (4) applied to this case gives:

$$a_0 + a^TT \sim t_1(a_0 + a^T\mu, a^T\Sigma a, \nu).$$

In particular, this implies that for each $j = 1, \dots, p$,

$$T_j \sim t_1(\mu_j, \Sigma(j,j), \nu)$$

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

3. **When ν is large, t is very close to normal**: This can intuitively be seen by noting that when ν is large, the term $(x-\mu)^T\Sigma^{-1}(x-\mu)/\nu$ is small so that

$$1 + \frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu) \approx \exp\left(\frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

where we used the observation that $1 + z \approx e^z$ when z is small. Thus the t-density (3) for large ν becomes approximately:

$$\exp\left(-\frac{\nu+p}{2\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \approx \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

because $\frac{\nu+p}{\nu} \approx 1$ when ν is large. This gets us the normal density:

$$t_p(\mu, \Sigma, \nu) \approx N_p(\mu, \Sigma) \qquad \text{if } \nu \text{ is large.}$$

In our regression case, the degrees of freedom 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:

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

3 Back to Regression

Let us get back to

$$\beta_0, \dots, \beta_m \mid \text{data} \sim t_{m+1}\left(\hat{\beta}, \frac{S(\hat{\beta})}{n-m-1}\left(X^TX\right)^{-1}, n-m-1\right). \tag{5}$$

The quantity $S(\hat{\beta})/(n-m-1)$ is the **frequentist unbiased** estimator for σ^2, so we denote it by $\hat{\sigma}^2$:

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

$\hat{\sigma}$ can also be justified as a Bayesian estimator of σ (See Question 5 (e) of Homework One). The terminology **Residual Standard Error** is sometimes used for $\hat{\sigma}$.

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

$$\beta_0, \dots, \beta_m \mid \text{data} \sim t_{m+1}\left(\hat{\beta}, \hat{\sigma}^2\left(X^TX\right)^{-1}, n-m-1\right). \tag{6}$$

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

$$\beta_j \mid \text{data} \sim t_1\left(\hat{\beta}_j, \hat{\sigma}^2(X^TX)^{j+1,j+1}, n-m-1\right) \tag{7}$$

where $(X^TX)^{j+1,j+1}$ is the $(j+1)^{th}$ diagonal entry of $(X^TX)^{-1}$ (note that we are using the $(j+1)$th diagonal entry of X^TX because β_j is the $(j+1)$th component of β). Writing this density out, we have

$$f_{\beta_j|\text{data}}(\beta_j) \propto \left(\frac{1}{1+\frac{1}{n-m-1}\frac{(\beta_j-\hat{\beta}_j)^2}{\hat{\sigma}^2(X^TX)^{j+1,j+1}}}\right)^{n/2}$$

which implies that

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

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

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

which is same as:

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

This interval:

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

is called the $100(1-\alpha)\%$ Bayesian Credible interval for β_j. It **exactly coincides** with the frequentist $100(1-\alpha)\%$ confidence interval for β_j.

When $n-m-1$ is large, the t-density (6) is approximately equal to the $N_{m+1}(\hat{\beta}, \hat{\sigma}^2(X^TX)^{-1})$. Further, when $n-m-1$ is large, the distribution (7) will be close to the normal distribution $N(\hat{\beta}_j, \hat{\sigma}^2(X^TX)^{j+1,j+1})$. The quantity $\hat{\sigma}\sqrt{(X^TX)^{j+1,j+1}}$ is known as the standard error corresponding to β_j.

4 Proof of (4)

Proof of (4). Start with the formula:

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

Observe that

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

so that

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

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

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

The change of variable

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

leads to

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

which proves (4). □

5 Nonlinear Regression

We shall next start our discussion on models in which certain parameters appear in a nonlinear fashion. A simple example is:

$$y_t = \beta_0 + \beta_1 t + \beta_2 \mathrm{ReLU}(t-c) + \epsilon_t \tag{8}$$

with $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0,\sigma^2)$. Here $\mathrm{ReLU}(t-c) = (t-c)_+$ equals 0 if $t \le c$ and equals $t-c$ if $t \ge c$. We can also write

$$\mathrm{ReLU}(t-c) = (t-c)_+ = (t-c)I\{t > c\} = \max(t-c, 0).$$

$(\cdot)_+$ is also called the positive part function, or, the ramp function.

The model (8) says that for times $t \le c$, the slope of the regression line is β_1, while for $t > c$, the slope changes to $(\beta_1 + \beta_2)$. We shall refer to (8) as the 'Change of Slope' model. An alternative name for this model is "Broken-stick regression". This is because the function

$$t \mapsto \beta_0 + \beta_1 t + \beta_2 \mathrm{ReLU}(t-c)$$

resembles a broken stick.

The unknown parameters for this model are $c, \beta_0, \beta_1, \beta_2$ as well as σ. The unknown parameter c makes (8) a nonlinear regression model. If c were known, then (8) would be a linear regression model:

$$y = X_c\beta + \epsilon \tag{9}$$

with

$$X_c = \begin{pmatrix} 1 & 1 & \mathrm{ReLU}(1-c) \\ 1 & 2 & \mathrm{ReLU}(2-c) \\ 1 & 3 & \mathrm{ReLU}(3-c) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & n & \mathrm{ReLU}(n-c) \end{pmatrix} \text{ and } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \text{ and } \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{pmatrix}$$

5.1 Estimation of $c, \beta_0, \beta_1, \beta_2, \sigma$

Least squares again is the most basic estimation procedure. The sum of squares is given by:

$$S(\beta_0, \beta_1, \beta_2, c) := \sum_{t=1}^{n} \left(y_t - \beta_0 - \beta_1 t - \beta_2 \text{ReLU}(t-c)\right)^2 .$$

We need to minimize this over all the four variables $\beta_0, \beta_1, \beta_2, c$. Using matrix notation, we can write

$$S(\beta, c) = \|y - X_c \beta\|^2.$$

If we fix c, then it is easy to minimize $S(\beta, c)$ over β. This is the same as linear regression and the minimizing β is given by:

$$\hat{\beta}_c := (X_c^T X_c)^{-1} X_c^T y,$$

and the smallest value of $S(\beta, c)$ for fixed c is $S(\hat{\beta}_c, c)$ which is just the RSS in the multiple linear regression with fixed c. We use the notation:

$$RSS(c) = S(\hat{\beta}_c, c) = \min_{\beta} S(\beta, c).$$

The least squares estimates of β and c can therefore be found in the following way:

1. Fix a finite set of possible values of c. In this change of slope model, it is reasonable to assume that $c \in \{1, \dots, n\}$.
2. For each value of c in the chosen set, calculate $\hat{\beta}_c$ and define $RSS(c) = S(\hat{\beta}_c, c)$.
3. Take $\hat{c}$ to be the value of c which minimizes $RSS(c)$.
4. Take $\beta = \hat{\beta}_{\hat{c}}$.

We shall revisit this and also look at uncertainty quantification in this model next week .