

STAT 238 - Bayesian Statistics

Lecture Fifteen

Spring 2026, UC Berkeley

Aditya Guntuboyina

25 Feb 2026

1 Linear Regression

Our next topic is (multiple) linear regression. Here, 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:

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

The default prior for linear regression is:

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

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

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

where we use the notation

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

for the sum of squares.

The posterior over only the coefficient parameters β_0, \dots, β_m can be obtained by integrat-

ing (or marginalizing) the parameter σ .

$$\begin{aligned}
& f_{\beta_0, \beta_1, \dots, \beta_m | \text{data}}(\beta_0, \beta_1, \dots, \beta_m) \\
&= \int f_{\beta_0, \beta_1, \dots, \beta_m, \sigma | \text{data}}(\beta_0, \beta_1, \dots, \beta_m, \sigma) d\sigma \\
&\propto I\{-C < \beta_0, \beta_1, \dots, \beta_m < C\} \int_{e^{-C}}^{e^C} \sigma^{-n-1} \exp\left(-\frac{S(\beta_0, \dots, \beta_m)}{2\sigma^2}\right) d\sigma \\
&\approx I\{-C < \beta_0, \beta_1, \dots, \beta_m < C\} \int_0^\infty \sigma^{-n-1} \exp\left(-\frac{S(\beta_0, \dots, \beta_m)}{2\sigma^2}\right) d\sigma \\
&= I\{-C < \beta_0, \beta_1, \dots, \beta_m < C\} \left(\frac{1}{S(\beta_0, \dots, \beta_m)}\right)^{n/2} \int_0^\infty s^{-n-1} \exp\left(-\frac{1}{2s^2}\right) ds \\
&\propto I\{-C < \beta_0, \beta_1, \dots, \beta_m < C\} \left(\frac{1}{S(\beta_0, \dots, \beta_m)}\right)^{n/2} \\
&\approx \left(\frac{1}{S(\beta_0, \dots, \beta_m)}\right)^{n/2} \propto \left(\frac{S(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m)}{S(\beta_0, \beta_1, \dots, \beta_m)}\right)^{n/2}
\end{aligned}$$

where $\hat{\beta}_0, \dots, \hat{\beta}_m$ denote the least squares estimators of β_0, \dots, β_m (i.e., $(\hat{\beta}_0, \dots, \hat{\beta}_m)$ minimizes $S(\beta_0, \dots, \beta_m)$ over all values of β_0, \dots, β_m).

Our posterior density for β_0, \dots, β_m is thus:

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

It turns out that this is a multivariate t -density. Here is some basic background on multivariate t -densities.

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):

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

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 $V \sim \chi_\nu^2$ (this is the chi-squared distribution with ν degrees of freedom) are independent. Then

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

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 3.1.

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 $a_0 + a_1 T_1 + a_2 T_2 + \dots + a_p T_p$ is also t -distributed. To see this, first write

$$a_0 + a_1 T_1 + \dots + a_p T_p = a_0 + a^T T$$

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

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

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

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

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 j th 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:

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

It turns out that (2) is a special case of (3) for some p, μ, Σ, ν . To see this, we need to first rewrite (2) using matrix notation which we do in the next section.

3 Matrix Notation for Multiple Linear Regression

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

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:

$$\hat{\beta} = (X^T X)^{-1} X^T y. \quad (5)$$

The proof of (5) is as follows. The gradient of $S(\beta)$ is given by

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

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

$$X^T (y - X\hat{\beta}) = 0 \implies X^T X \hat{\beta} = X^T y \implies \hat{\beta} = (X^T X)^{-1} X^T y. \quad (6)$$

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

$$S(\beta) = S(\hat{\beta}) + \|X\beta - X\hat{\beta}\|^2 = S(\hat{\beta}) + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}). \quad (7)$$

To prove (7), write

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

The cross product is zero (leading to (7)) because:

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

where we used (6).

Using (7), we can write the posterior density (2) as

$$\begin{aligned}
f_{\beta|\text{data}}(\beta) &\propto \left(\frac{S(\hat{\beta})}{S(\beta)} \right)^{n/2} \\
&= \left(\frac{S(\hat{\beta})}{S(\hat{\beta}) + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})} \right)^{n/2} \\
&= \left(\frac{1}{1 + (\beta - \hat{\beta})^T \frac{X^T X}{S(\hat{\beta})} (\beta - \hat{\beta})} \right)^{n/2}.
\end{aligned} \tag{8}$$

The above formula is a special case of (3) 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

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

We thus have

$$\beta_0, \dots, \beta_m \mid \text{data} \sim t_{m+1} \left(\hat{\beta}, \frac{S(\hat{\beta})}{n - m - 1} (X^T X)^{-1}, n - m - 1 \right). \tag{9}$$

With the posterior density (9), one can do uncertainty quantification about the parameters $\beta_0, \beta_1, \dots, \beta_m$. 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.

In (9), 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 (this will be a question in Homework three). The terminology **Residual Standard Error** is sometimes used for $\hat{\sigma}$.

With the notation for $\hat{\sigma}$, the posterior (9) becomes:

$$\beta_0, \dots, \beta_m \mid \text{data} \sim t_{m+1} \left(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1}, n - m - 1 \right). \tag{10}$$

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

$$\beta_j \mid \text{data} \sim t_1 \left(\hat{\beta}_j, \hat{\sigma}^2 (X^T X)^{j+1, j+1}, n - m - 1 \right) \tag{11}$$

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). Writing this density out, we have

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

which implies that

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

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

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

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:

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

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 .

The degrees of freedom corresponding to the t -density in (9) 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 (9) 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 (11) 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 .

3.1 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

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

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). □