The relation between the normal distribution and least squares is rooted in the concept of maximum likelihood estimation (MLE). Let's walk through the details step by step.

Normal Distribution and Likelihood

First, consider a simple linear regression model where the observed data y_n is assumed to be normally distributed around a linear combination of the predictors $x_n^\top \theta$ with some variance σ^2. Mathematically, this can be written as:

$$y_n \sim \mathcal{N}(x_n^\top \theta, \sigma^2)$$

This means the probability density function for each y_n given x_n and θ is:

$$p(y_n \mid x_n, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_n - x_n^\top \theta)^2}{2\sigma^2}\right)$$

Log-Likelihood

To find the parameters θ that best fit the data, we maximize the likelihood function. The likelihood of the entire dataset $Y = (y_1, y_2, \ldots, y_N)$ given the predictors $X = (x_1, x_2, \ldots, x_N)$ and the parameters θ is the product of the individual probabilities:

$$p(Y \mid X, \theta) = \prod_{n=1}^{N} p(y_n \mid x_n, \theta)$$

To simplify the optimization, we typically maximize the log-likelihood instead of the likelihood itself. The log-likelihood function is:

$$\log p(Y \mid X, \theta) = \log\left(\prod_{n=1}^{N} p(y_n \mid x_n, \theta)\right) = \sum_{n=1}^{N} \log p(y_n \mid x_n, \theta)$$

Substituting the expression for $p(y_n \mid x_n, \theta)$:

$$\log p(y_n \mid x_n, \theta) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_n - x_n^\top \theta)^2}{2\sigma^2}\right)\right)$$

Simplifying the Log-Likelihood

Breaking down the log expression:

$$\log p(y_n \mid x_n, \theta) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{(y_n - x_n^\top \theta)^2}{2\sigma^2}\right)\right)$$

$$\log p(y_n \mid x_n, \theta) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_n - x_n^\top\theta)^2}{2\sigma^2}$$

Note that the first term $-\frac{1}{2}\log(2\pi\sigma^2)$ is a constant with respect to θ. Therefore, it can be absorbed into the constant term when we focus on the log-likelihood function:

$$\log p(y_n \mid x_n, \theta) = -\frac{(y_n - x_n^\top\theta)^2}{2\sigma^2} + \text{const}$$

Connecting to Least Squares

When we maximize the log-likelihood, it's equivalent to minimizing the negative log-likelihood. Dropping the constant term for optimization purposes:

$$-\log p(y_n \mid x_n, \theta) = \frac{(y_n - x_n^\top\theta)^2}{2\sigma^2} + \text{const}$$

Since σ^2 is also a constant (it represents the variance of the normal distribution), minimizing this expression with respect to θ is equivalent to minimizing the sum of squared residuals:

$$\sum_{n=1}^{N}(y_n - x_n^\top\theta)^2$$

This is the least squares objective function. Hence, maximizing the likelihood for normally distributed errors (under the assumptions of a linear model and constant variance) leads directly to minimizing the least squares cost function.