Theorem 1. *Under Gaussian assumption, linear regression amounts to least square.*

Let us assume that the target variable y and the inputs are related as given below

$$y^i = \theta^T x^{(i)} + \epsilon^{(i)}$$

where $\epsilon^{(i)}$ is the error term which accounts for the features that the equation can't take care of, or some random noise. Let us also assume that $\epsilon^{(i)}$'s are distributed IID accoding to a Gaussian distribution with mean o and variance σ^2. So,

$$\epsilon^{(i)} \sim N(0, \sigma^2)$$

i.e. the density of $\epsilon^{(i)}$ is given by

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

This implies that

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

The notation "$p(y^{(i)}|x^{(i)};\theta)$" indicates that this is the distribution of $x^{(i)}$ given $y^{(i)}$ and parametrized by θ. Now, as θ is not a random variable, it shouldn't be conditioned on. We can also write the distribution of $y^{(i)}$ as $y^{(i)}|x^{(i)} \sim N(\theta^T x^{(i)}, \sigma^2)$.

Given X, i.e. the feature matrix, and θ, we are looking for the distribution of $y^{(i)}$'s. The probability of the data is given by "$p(\vec{y}|X;\theta)$". This is typically viewed as a function of $\vec{y}$ (and perhaps X) for a particular θ. To explicitly view this as a function of θ, we shall call it the **likelihood** function, and express it as

$$L(\theta) = L(\theta; X, \vec{y}) = p(\vec{y}|X;\theta)$$

By independence assumption on $\epsilon^{(i)}$'s, we can also write this as

$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

For this probabilistic model relating the $y^{(i)}$'s and $x^{(i)}$'s we need to find a reasonable way of choosing the parameters θ. The principal of **maximum likelihood** says that we should choose θ so as to make the data as high probability as possible, i.e. we have to maximize $L(\theta)$ w.r.t θ.

Instead of $L(\theta)$, we choose to maximize the **log likelihood** $l(\theta)$:

$$\ell(\theta) = \log L(\theta) = \log \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)$$

$$= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} . \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

Hence, maximizing $\ell(\theta)$ becomes same as minimizing

$$\frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2,$$

which is our least-squares cost function.