- *Semi-parametric regression:* It is a compromise between the above methods. $m(.)$ has some parameters to be estimated. This approach is more flexible than a fully parametric model.

1.2 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a method that determines values for the parameters of a model. In other words, we have some data points; we want to find the parameters of a curve that was most likely responsible for generating the data points that we observed. For example, suppose we observed several data points $x_1, x_2, x_3, ...$ with corresponding $y_1, y_2, y_3, ...$ and we want to know which curve is responsible for creating such data. One might decide that the Gaussian distribution is the model for the error between the y values and the regression curve.

The Gaussian distribution is characterized by two parameters μ and σ. The parameter μ is the mean or expectation of the distribution (and also its median and mode), while the parameter σ is its standard deviation. We use MLE to find these two parameters. The probability density of observing a single data point y_i that is generated from a Gaussian distribution is given by [42]:

$$P(y_i; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x_i - \mu)^2}{2\sigma^2}) \tag{1.3}$$

Under the assumption that each data point is generated independently of the others, the total probability density of $y_1, y_2, y_3, ...$ is obtained by multiplying all probability densities. To get a maximum likelihood estimator (MLE), we need to figure out the values of μ and σ that give the maximum value of total probability density. To do so, we find the partial derivative of the function and set it to zero, and then solve the resulting equations to get the parameters of interest [3]. MLE is usually asymptotically efficient, which means it has a smaller variance than alternatives for large samples of data.

1.3 Linear Regression

A simple form of parametric regression analysis is linear regression which assumes a linear form for $m(x)$ where the coefficients β's represent the relationship between

each predictor variable x_j and the response variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{1.4}$$

Equation (1.4) is the theoretical regression model; we want to obtain an estimated regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

To estimate coefficients ($\hat{\beta}_i$'s), Ordinary Least Squares (OLS) minimizes the sum of squared (L_2 norm) of the residuals.

$$\sum_{i=0}^{n} |y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik})|^2 = \sum_{i=0}^{n} |e_i|^2 \tag{1.5}$$

1.4 OLS is MLE for Normal Distribution

When data, more specifically error, follows a normal distribution with $\mu = 0$ and variance σ^2, i.e $\varepsilon \sim N(0, \sigma^2)$ and also, variable y is independent across observations which means given any data point (x_i, y_i), we can write down the probability density of seeing that data, under the model with parameters $\beta_0, \beta_1, ...\beta_n$ and σ^2 as:

$$\prod_{i=1}^{n} P(y_i | x_i; \beta_0, \beta_1, ..., \beta_k, \sigma^2) =$$
$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2}{2\sigma^2})$$

By taking the log of the equation above, we will have:

$$\log \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2}{2\sigma^2}) =$$
$$\sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2}{2\sigma^2}) =$$
$$\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=0}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2$$

This function is called log likelihood function. We pick the parameter values which maximize the log-likelihood, with respect to β.

$$\arg\max_{\beta}[\cancel{\frac{n}{2}\log 2\pi} - \cancel{n\log\sigma} - \frac{1}{2\sigma^2}\sum_{i=0}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2] =$$

$$\arg\max_{\beta}[-\sum_{i=0}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2]$$

$$= \arg\min_{\beta}[\sum_{i=0}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2]$$

By simplifying the equation above, we get to the same equation as in ordinary least squares regression (1.5). This indicates that OLS is equivalent to MLE; therefore, it will have the lowest variance among all linear unbiased estimators i.e. BLUE (Best Linear Unbiased Estimation) [26]. It is important to note that the statement previous is only valid under the following four assumptions:

1. The relationship between x and y is linear.
2. The error term has mean of zero ($\mu = 0$) and a constant variance ($\sigma^2 = c$).
3. Error terms are independent of the predictor and uncorrelated with each other.
4. All predictor variables are uncorrelated with the error term

1.5 The Generalized Error Distribution

There are numerous real-life applications where the data we have are not consistent with the hypothesis of normality. Often, collected data are not normally distributed and are contaminated by outliers, so least-squares estimates provide less accurate estimates even though they remain unbiased. Thus, the estimated solution provided by OLS differs from the true one when applied to non-normally distributed data [2].

Subbotin [42] has generalized the normal distributon to a law of errors' distribution that is much more flexible than the normal distribution and can describe a wider range of distributions than the normal distribution can. This is also known as the *family of normal distributions of order* p, the *exponential power function*