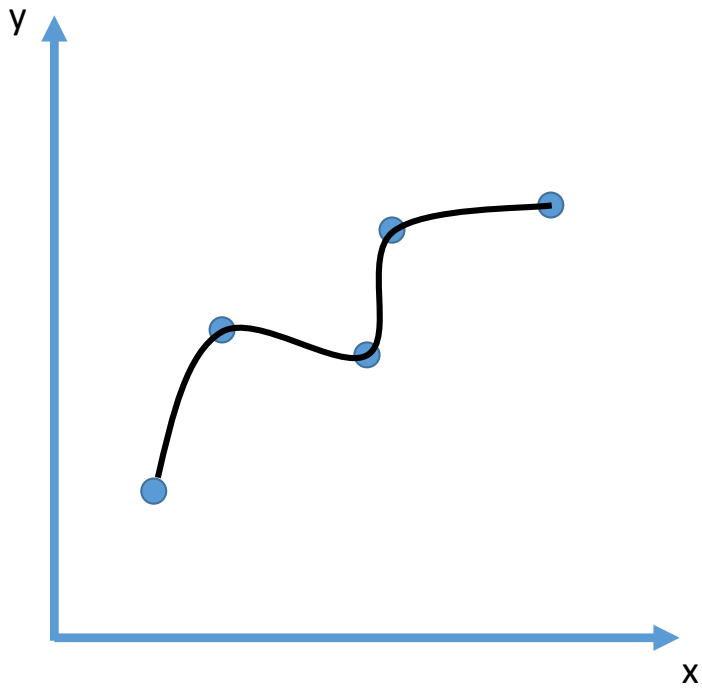
Least Squares Regression

Lecture 6

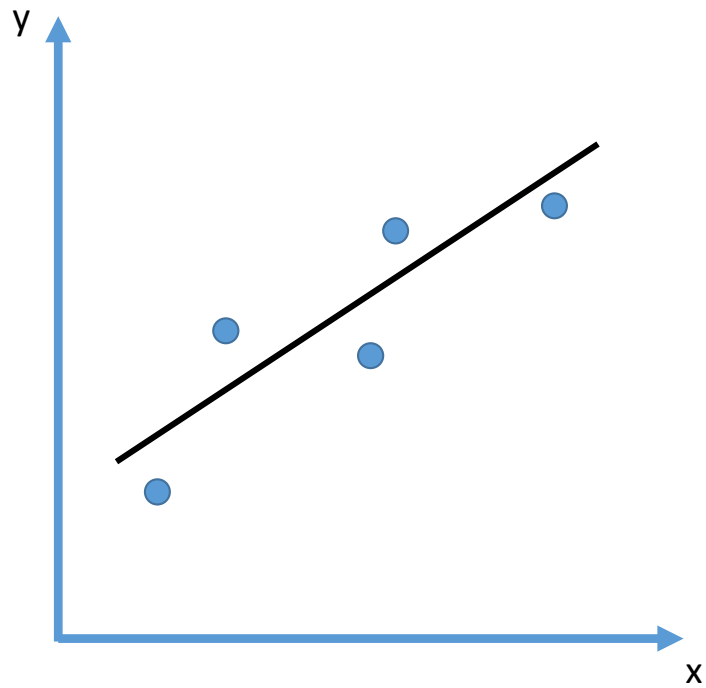
Dr. Görkem Saygılı

Department of Biomedical Engineering
Ankara University

Numerical Methods, 2017-2018 Fall



(a) Fitting a Polynomial



(b) Least Squares Regression

Until this lecture, we have seen how to fit a polynomial between data samples by using different methods as shown above in (a).

We do not always want our fit to pass through all the data samples:

- There may be noise in the data
- We might want a general model that represents the characteristics of the data.

Hence, rather than fitting a polynomial as in (a), we may want to fit a line as in (b).

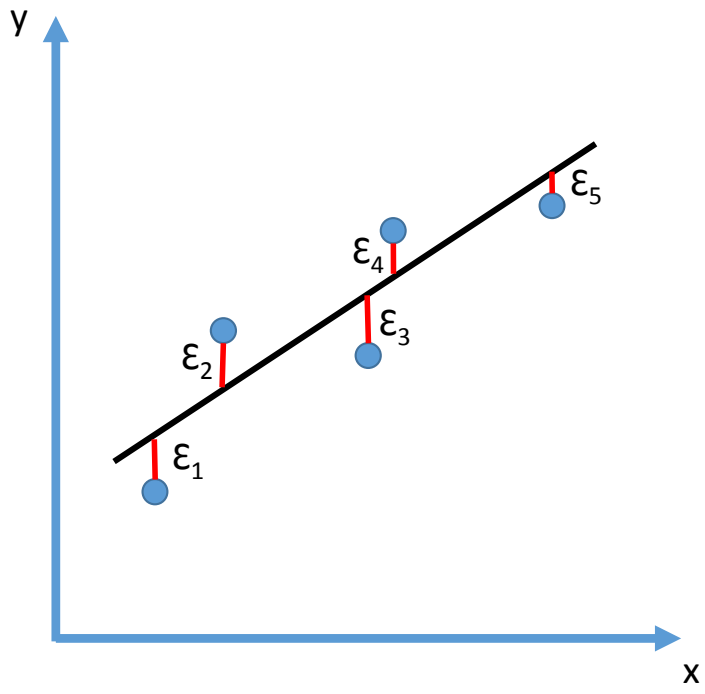
The line is fit by minimizing the squared distance of the samples to the estimated line:

$$\hat{y}_i = \theta_0 + \theta_1 x_i \tag{1}$$

$$\epsilon_i = y_i - \hat{y}_i$$

$$\epsilon = \sum \epsilon_i^2 \tag{2}$$

where $\hat{y}_i$ in Eqn. 1 is the value of the fitted line at x_i. For each sample, error is calculated by calculating the sum of squared difference between the samples and the fitted line as in Eqn. 2 and as shown in the figure below:



Total error:

$$\epsilon = \sum(\epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 + \epsilon_4^2 + \epsilon_5^2)$$

Since we define our error function, the rest is to minimize the error with respect to the parameters of our fitted function. To do that, we take the derivative of the error with respect to each parameter and equalize it to zero:

$$\frac{\partial \epsilon}{\partial \theta_0} = \frac{\partial \sum (y_i - \theta_0 - \theta_1 x_i)^2}{\partial \theta_0} = 0 \tag{3}$$

$$\frac{\partial \epsilon}{\partial \theta_1} = \frac{\partial \sum (y_i - \theta_0 - \theta_1 x_i)^2}{\partial \theta_1} = 0 \tag{4}$$

Let total number of samples be N. From Eqn. 3:

$$
\begin{aligned}
-2\sum(y_i - \theta_0 - \theta_1 x_i) &= 0 \\
\sum y_i &= \sum \theta_0 + \sum \theta_1 x_i \\
\sum y_i &= N\theta_0 + \theta_1 \sum x_i \\
\theta_0 &= \frac{\sum y_i - \theta_1 \sum x_i}{N}
\end{aligned} \tag{5}
$$

From Eqn. 4:

$$
\begin{aligned}
&-2\sum x_i(y_i - \theta_0 - \theta_1 x_i) = 0 \\
&\sum x_i y_i = \theta_0 \sum x_i + \theta_1 \sum x_i^2 \\
&\sum x_i y_i = \frac{\sum x_i \sum y_i - \theta_1 (\sum x_i)^2}{N} + \theta_1 \sum (x_i^2) \\
&N\sum x_i y_i = \sum x_i \sum y_i - \theta_1 (\sum x_i)^2 + N\theta_1 \sum x_i^2 \\
&N\sum x_i y_i - \sum x_i \sum y_i = \theta_1 (N\sum x_i^2 - (\sum x_i)^2 \\
&\theta_1 = \frac{N(\sum x_i y_i) - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2} \qquad (6)
\end{aligned}
$$

Hence, we can calculate least squares regression line by using the following equations:

$$\boxed{\begin{aligned}\theta_1 &= \frac{N(\sum x_i y_i) - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2} \\ \theta_0 &= \frac{\sum y_i - \theta_1 \sum x_i}{N}\end{aligned}}$$

We first calculte θ_1 and then we calculate θ_0.