DATA MINING LECTURE 6

Linear Regression
Logistic Regression
Neutral Networks
A Little Bit of Deep Learning

CSC177
Dr. Victor Chen

Regression Problem

- The problem of predicting continuous values is called regression problem

- General approach: find a continuous function that models the continuous points.

Linear Regression with one input

A simple regression has the coefficient β and the constant α. The equation is then:

$$y = α + β * x$$

where α is   *y*-intercept and β is slope
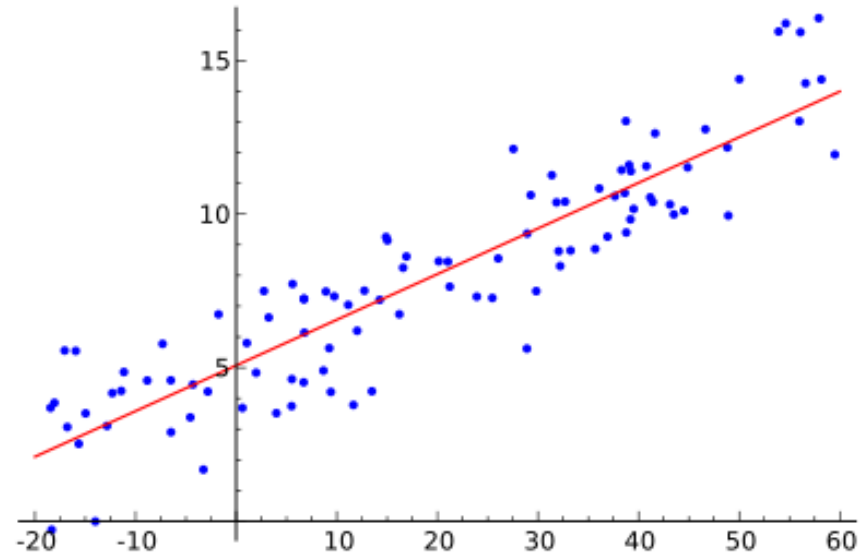
Linear Regression with more than one input

A multiple regression is in the following form:

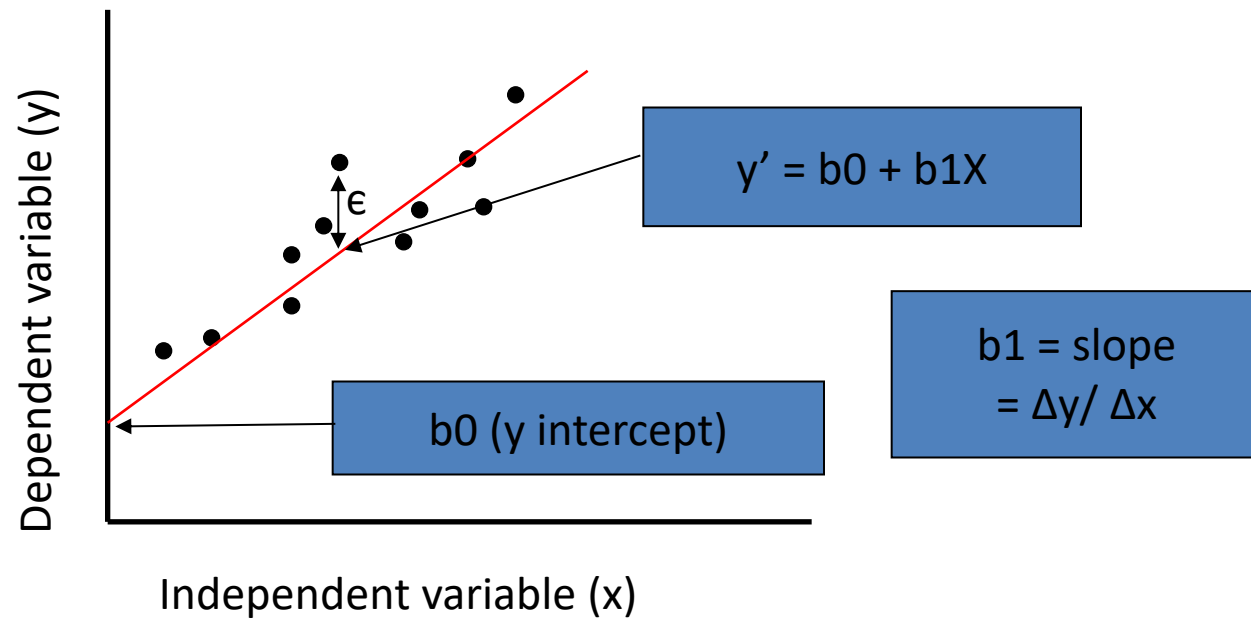$$y = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_k * X_k$$

where k is the number of variables, or parameters.
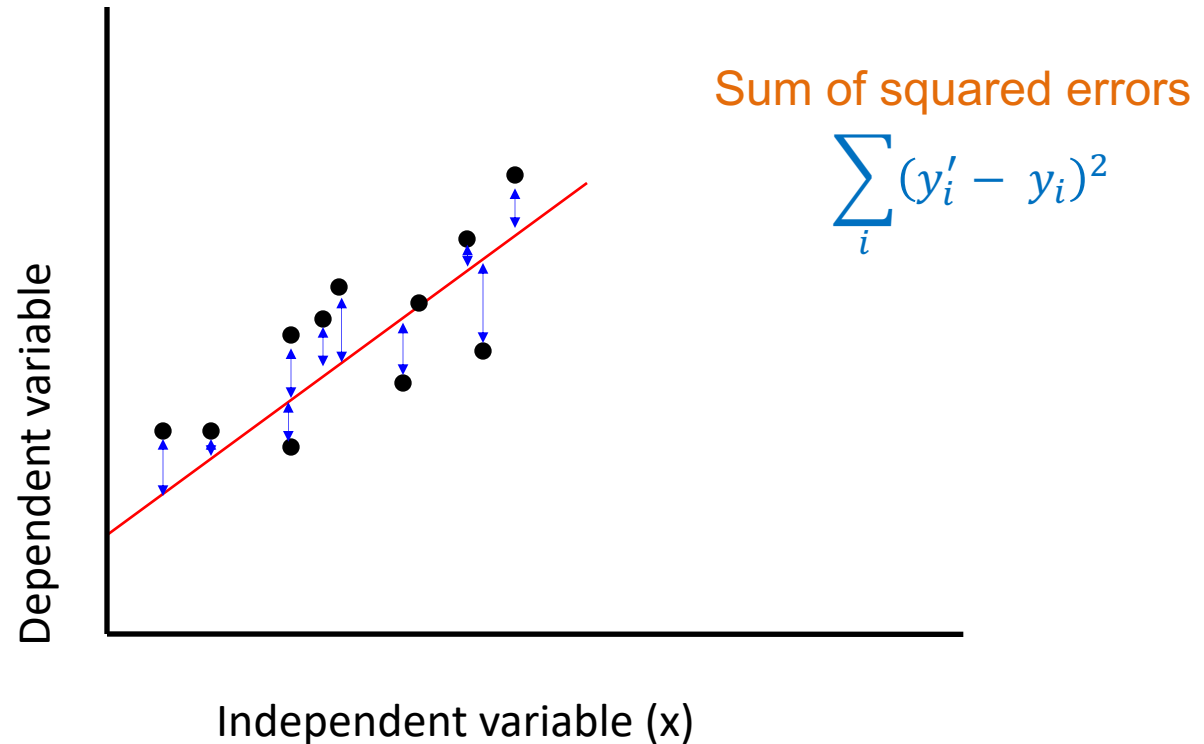
Linear regression for 2D data

- Given a dataset of the form $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, find a linear continuous function that minimizes the Sum of Squares of Error (SSE)
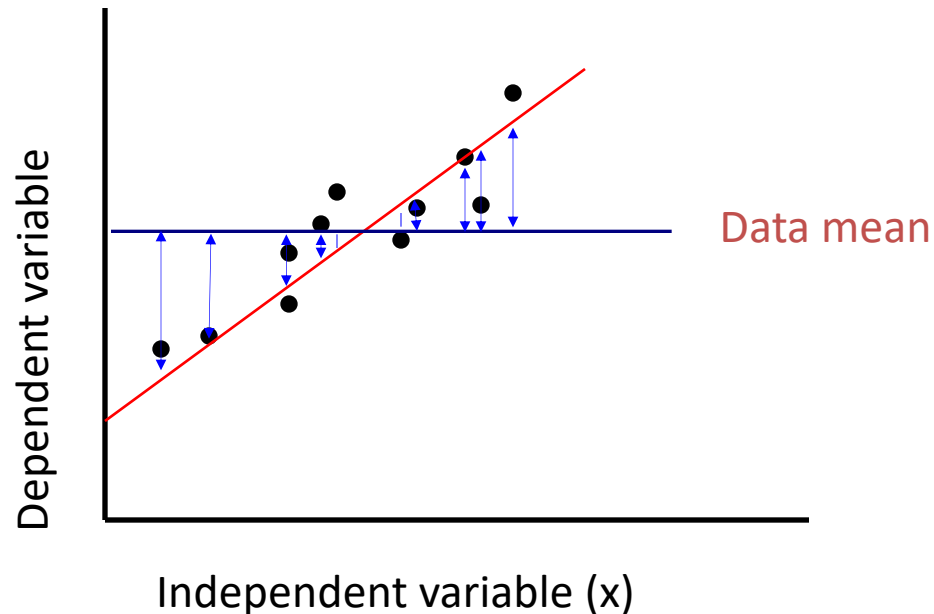
Linear Regression

Sum of Squares of Error (SSE)

Sum of squared errors

$$\sum_i (y_i' - y_i)^2$$

Dependent variable

Independent variable (x)

Sum of Squares of Error (SSE) is the sum of all the squared errors.
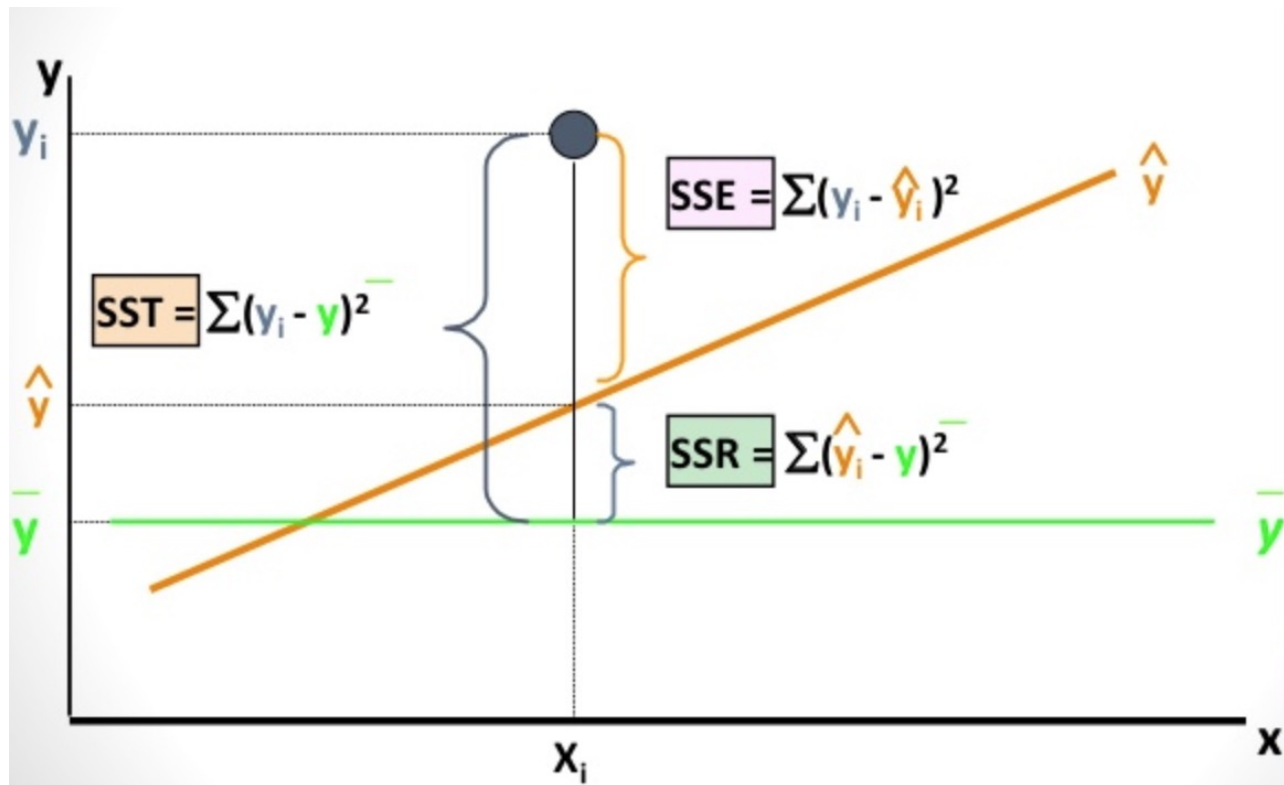
Sum of Squares of Regression (SSR)

Sum of Squares of Regression (SSR) is the sum of the squared differences between each prediction and the mean of data.

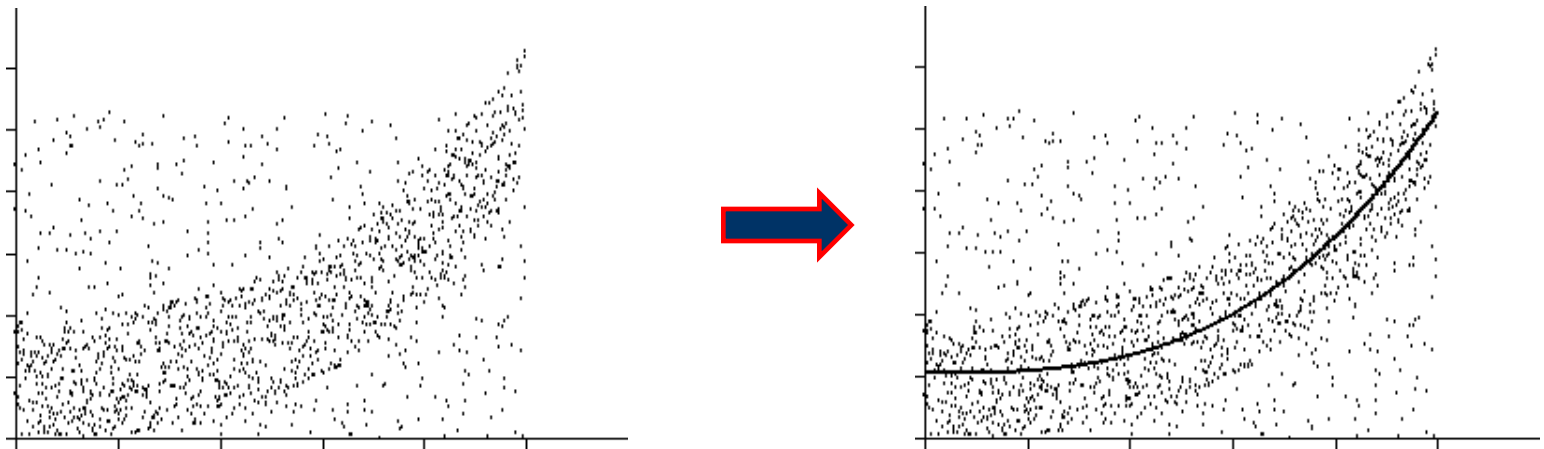# Sum of Squares of Total (SST)

SST = SSE + SSR

R-Squared score

- R-Squared score can be used as a single summary number to measure the quality of linear regression model

- The value of R^2 can range between 0 and 1.

- **The higher R^2, the more accurate the regression model is.**

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Nonlinear functions can also be fit as regressions

Any nonlinear continuous functions can also be fit as regressions, including power, Logarithmic, Exponential, and Logistic.