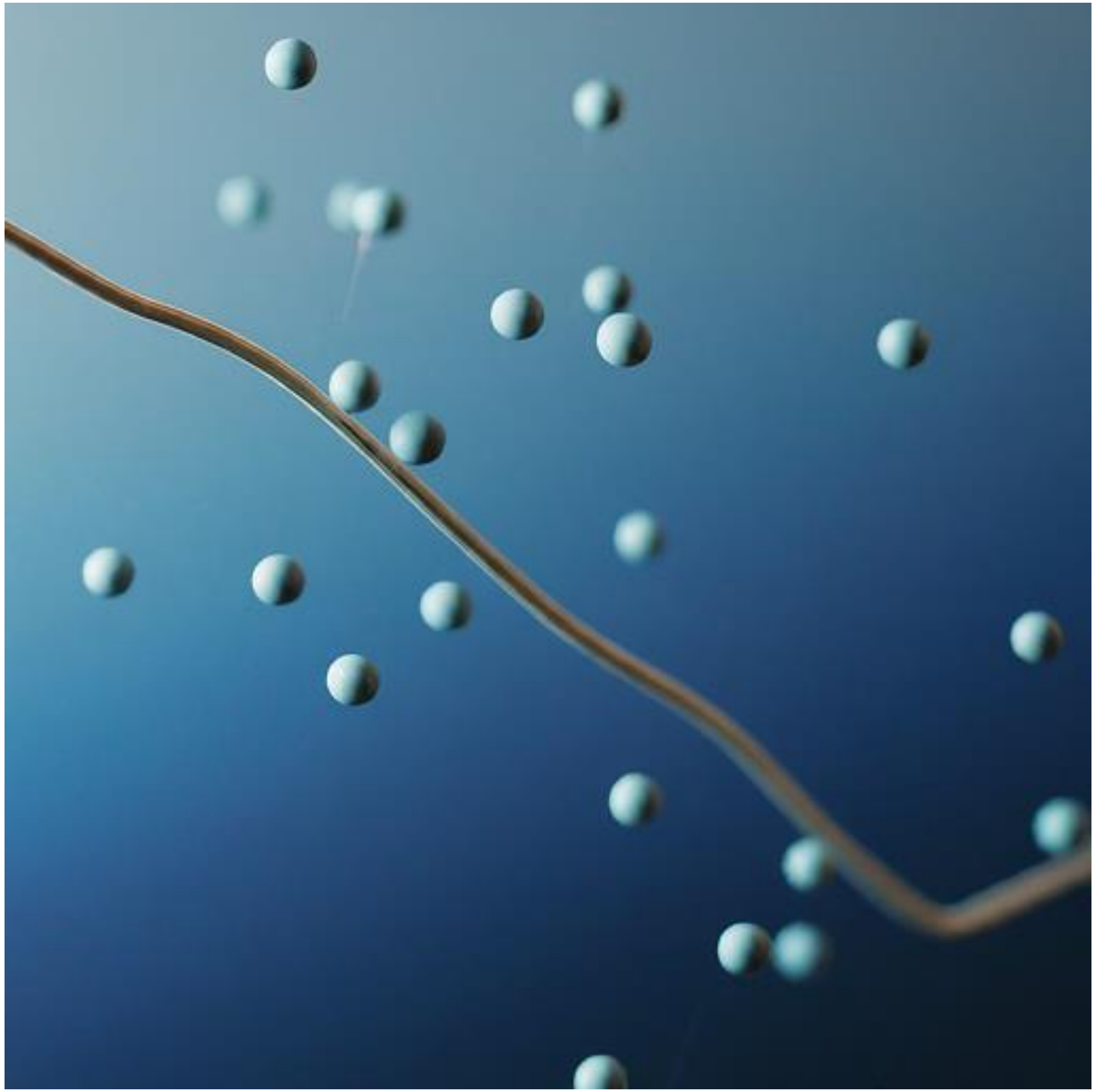
PH 3120 – Computational Physics Laboratory I

Regression and Interpolation

Dr. E. M. D. Siriwardane

Department of Physics

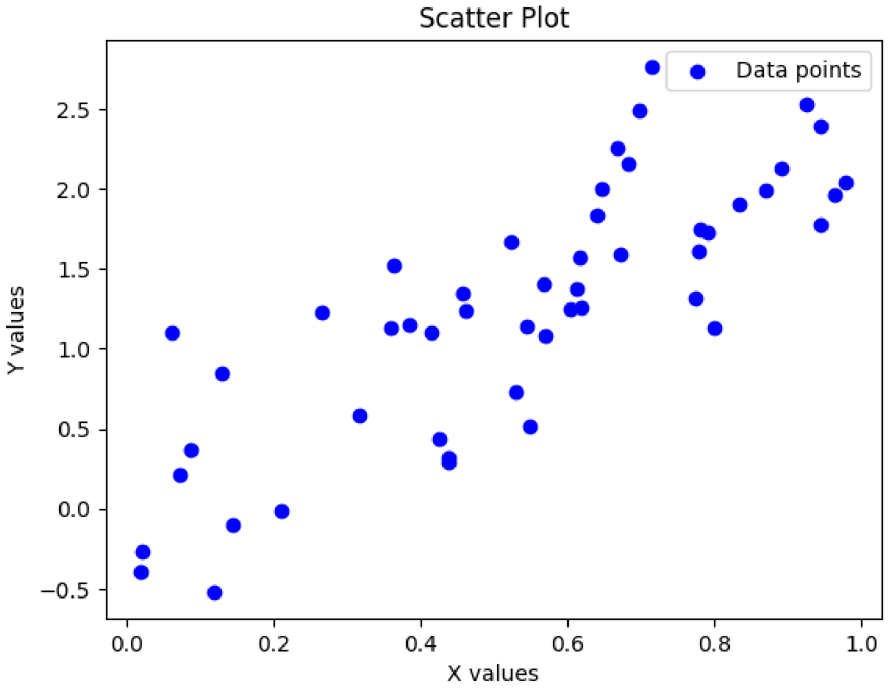
University of Colombo



Regression

- Regression is a statistical method used to understand the relationship between variables.
- It allows you to model the relationship between a dependent variable (often called the response variable) and one or more independent variables (predictors or features).
- There are several types of regression, each suited to different types of data and research questions.
- In this laboratory, we will focus on
    - Least Square Regression - Linear Regression
    - Polynomial Regression

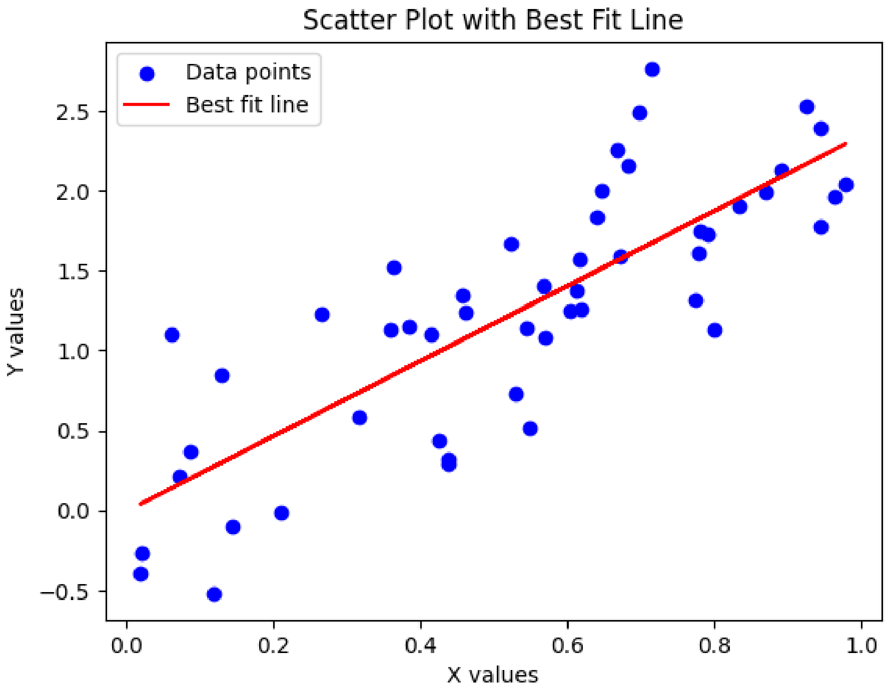
Regression



- The scatter plot consists of individual points plotted on a two-dimensional graph, where each point represents a pair of values from two variables:
  - X values: The independent variable, plotted along the horizontal axis.
  - Y values: The dependent variable, plotted along the vertical axis.

- The objective of regression is to model the relationship between a dependent variable and one or more independent variables.

Regression



- The best fit line or curve is a straight line or a curve that best represents the data points on the scatter plot.
- The figure shows the best fit line found using least square regression.

Applications of Regression

- **Identify Patterns**: Determine whether and how the dependent variable changes as the independent variable(s) change.
- **Quantify Relationships**: Quantify the strength and direction (positive or negative) of relationships between variables.
- **Predict Values for New X Values**: Use the regression model to make predictions about the dependent variable based on new values of the independent variable(s).
- **Forecasting**: Provide estimates for future observations, which is especially useful in fields like finance, economics, and engineering.

Least Squares Regression

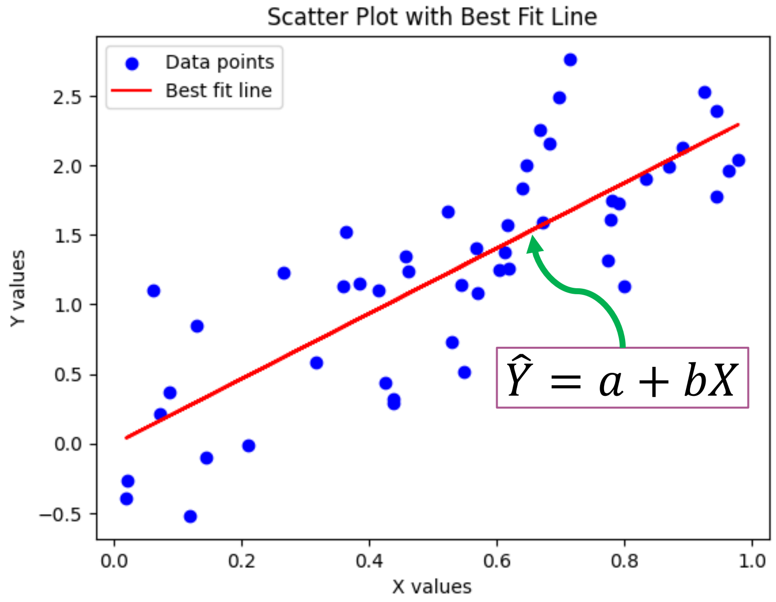
- Least Squares Regression is a method used to determine the best-fit line or model for a given set of data by minimizing the sum of the squares of the residuals (the differences between observed and predicted values).
- This technique is fundamental in statistical modeling and machine learning, especially for linear regression analysis.

Least Squares Regression

- **Linear Relationship**: The basic assumption is that there is a linear relationship between the dependent variable Y and the independent variable(s) X.
- **Model Equation**: For simple linear regression, the model can be expressed as:

$$Y = a + bX + \epsilon :$$

- Y is the dependent variable.
- X is the independent variable.
- a is the intercept.
- b is the slope.
- ϵ is the error term (residual).



$\hat{Y}$: Dependent variable values on the best fit line

Least Squares Regression

The goal is to find the values of a and b that minimize the Residual Sum of Squares (RSS) between the observed values (Y_i) and the predicted values ($\hat{Y}_i$):

$$RSS = \sum_i^n \left(Y_i - \hat{Y}_i\right)^2 = \sum_i^n (Y_i - (a + bX_i))^2$$

where:

- n is the number of observations.
- Y_i is the observed value.
- $\hat{Y}_i = a + bX_i$ is the predicted value.

Least Squares Regression

To find a and b, we take the partial derivatives of RSS with respect to a and b, set them to zero, and solve for a and b.

$$\frac{\partial RSS}{\partial a} = -2\sum_{i}^{n}\big(Y_i - (a + bX_i)\big) = 0$$

$$\frac{\partial RSS}{\partial b} = -2\sum_{i}^{n}X_i\big(Y_i - (a + bX_i)\big) = 0$$

By solving the two simultaneous equations

$$b = \frac{n\sum_i^n X_iY_i - \sum_i^n X_i \sum_i^n Y_i}{n\sum_i^n Xi^2 - \left(\sum_i^n Xi\right)^2}$$

$$a = \bar{Y} - b\bar{X}$$

$\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively.

Error Analysis

R-squared (R^2): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

$$\boxed{R^2 = 1 - \frac{\sum_i^n \left(Y_i - \hat{Y}_i\right)^2}{\sum_i^n (Y_i - \bar{Y})^2}}$$

- Y_i is the observed value.
- $\hat{Y}_i$ is the predicted value
- $\bar{Y}$ is the average value of observed values

Mean Absolute Error (MAE) : MAE is calculated as the average of the absolute differences between the predicted values ($\bar{Y}_i$) and the actual values (Y_i):

$$\boxed{MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \bar{Y}_i|}$$

Polynomial Regression

For a polynomial regression of degree d, the model can be expressed as:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \epsilon$$

Y is the dependent variable.

X is the independent variable.

$\beta_0, \beta_1, \ldots, \beta_d$ are the coefficients.

ϵ is the error term (residual).

Polynomial Regression

Consider n number of data points

X_i	Y_i
X_1	Y_1
X_2	Y_2
X_3	Y_3
X_4	Y_4
X_5	Y_5
$\vdots$	$\vdots$
X_n	Y_n

Polynomial Regression

Transform the Independent Variables: For polynomial regression, we transform the original variable X into polynomial features.

For example, if you have a polynomial of degree 2, the independent variable will be $[1, X, X^2]$

Set Up the Design Matrix: The design matrix $\boldsymbol{X}$ for a polynomial of degree d will include d columns corresponding to each polynomial term. There is an additional column for constant 1.

For a degree d polynomial, the design matrix will look like:

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^d \\ 1 & X_2 & X_2^2 & \cdots & X_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^d \end{bmatrix} \qquad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix}$$

Polynomial Regression

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_d X_i^d + \epsilon_i$$

The residual or the error term can be rewritten using matrices as follows

$$\epsilon = \boldsymbol{Y} - \boldsymbol{X}\beta$$

RSS can be written as follows

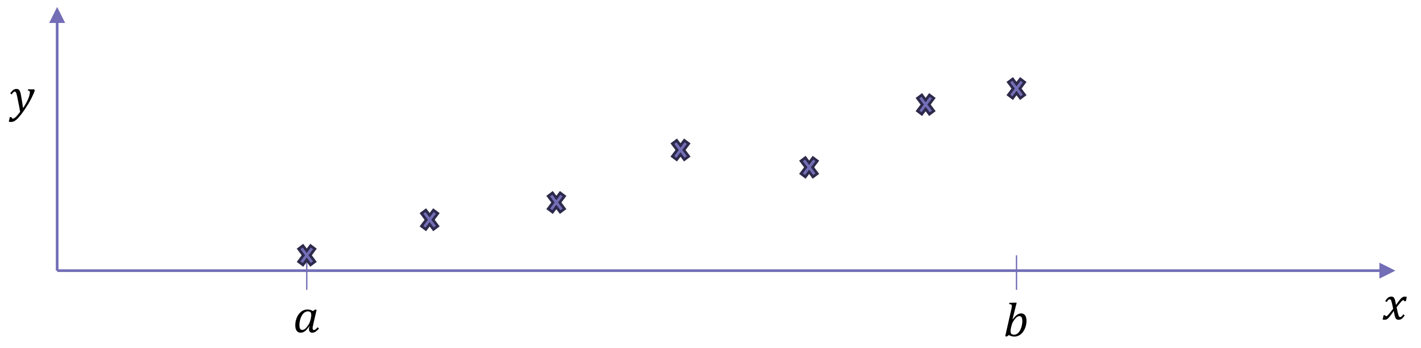
$$RSS = \epsilon^T \epsilon = (\boldsymbol{Y} - \boldsymbol{X\beta})^T (\boldsymbol{Y} - \boldsymbol{X\beta})$$

To find the coefficients β that minimize the RSS, we take the derivative of RSS with respect to β and set it to zero:

$$\frac{\partial RSS}{\partial \beta} = \frac{\partial}{\partial \beta} (\boldsymbol{Y} - \beta \boldsymbol{X})^T (\boldsymbol{Y} - \beta \boldsymbol{X}) = 0 \longrightarrow \beta = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T Y}$$

Interpolation

- Interpolation is a method used to estimate unknown values that fall between known values.
- In other words, interpolation involves constructing new data points within the range of a discrete set of known data points.
- This is commonly used in numerical analysis, data science, and various fields of engineering and science where the data points are discrete, and a continuous function is needed to approximate the data.

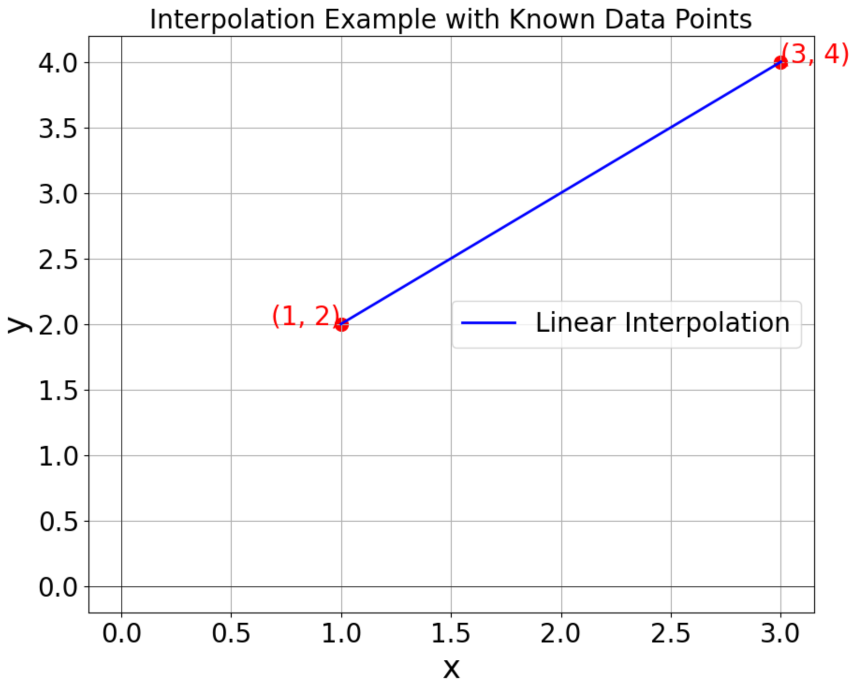


In this example, we attempt to find a function represents the data points within $[a, b]$interval.

Linear Interpolation

- Linear interpolation involves connecting two adjacent known data points with a straight line.
- It is the simplest form of interpolation.

Given two known points (x_0, y_0) and (x_1, y_1), the linear interpolation formula for a point x is:



$$(x_0, y_0) \equiv (1,2)$$

$$(x_1, y_1) \equiv (3,4)$$

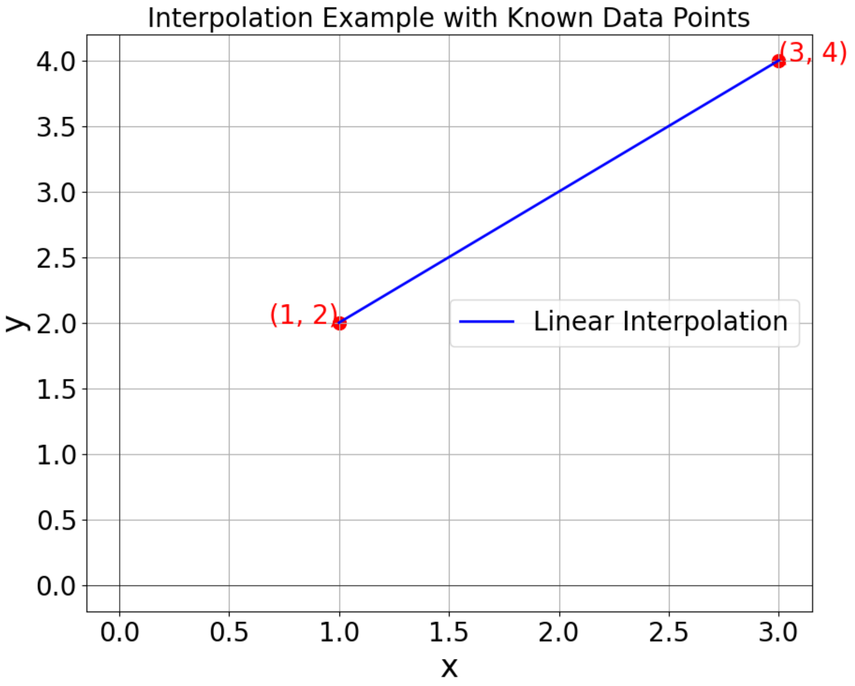
Linear Interpolation

Using the slope of the straight line

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0}$$

For the example

$$\frac{y - 2}{x - 1} = \frac{4 - 2}{3 - 2} \longrightarrow y = 2x$$



Lagrange Interpolation

Lagrange Interpolation constructs a polynomial $P(x)$ of degree $n-1$ that passes through n given data points (x_i, y_i).

The Lagrange polynomial $P(x)$ is given by:

$$\boxed{P(x) = \sum_{i=0}^{n-1} y_i L_i(x)}$$

where $L_i(x)$ are the Lagrange basis polynomials defined as:

$$\boxed{L_i(x) = \prod_{\substack{0 \le j \le n-1 \\ j \ne i}} \frac{x - x_j}{x_i - x_j}}$$

Lagrange Interpolation

Example

$(x_0, y_0) = (1,2), (x_1, y_1) = (2,3), (x_2, y_2) = (3,5)$

Calculate the basis polynomials:

$$\boxed{L_i(x) = \prod_{\substack{0 \le j \le n-1 \\ j \ne i}} \frac{x - x_j}{x_i - x_j}}$$

$$L_0(x) = \frac{(x-2)(x-3)}{(1-2)(1-3)} = \frac{(x-2)(x-3)}{2}$$

$$L_1(x) = \frac{(x-1)(x-3)}{(2-1)(1-3)} = \frac{-(x-1)(x-3)}{1}$$

$$L_0(x) = \frac{(x-1)(x-2)}{(3-1)(3-2)} = \frac{(x-1)(x-2)}{2}$$

Lagrange Interpolation

Example

Form the Lagrange polynomial:

$$\boxed{P(x) = \sum_{i=0}^{n-1} y_i L_i(x)}$$

$$(x_0, y_0) = (1,2), (x_1, y_1) = (2,3), (x_2, y_2) = (3,5)$$

$$P(x) = 2.\,L_0(x) + 3.\,L_1(x) + 5.\,L_2(x)$$

$$P(x) = 2.\frac{(x-2)(x-3)}{2} + 3.\frac{-(x-1)(x-3)}{1} + 5.\frac{(x-1)(x-2)}{2}$$