Computational Techniques
Module 5: Regression and Interpolation

Dr. Niket Kaisare
Department of Chemical Engineering
Indian Institute of Technology - Madras

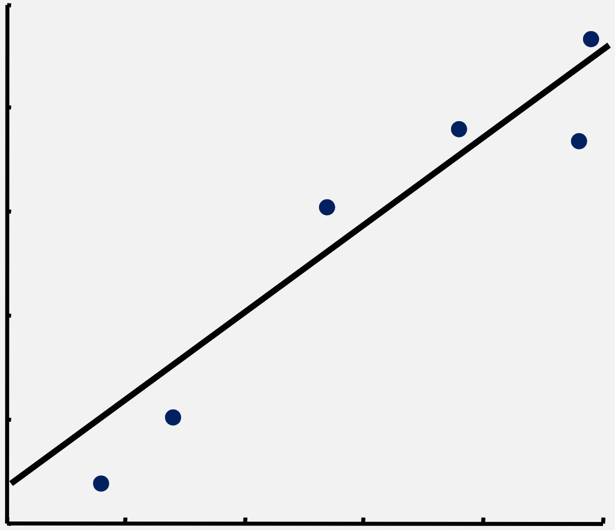
Regression Example

- Given the following data:

x	0.8	1.4	2.7	3.8	4.8	4.9
y	0.69	1.00	2.02	2.39	2.34	2.83

Regression:

Obtain a straight line that best fits the data



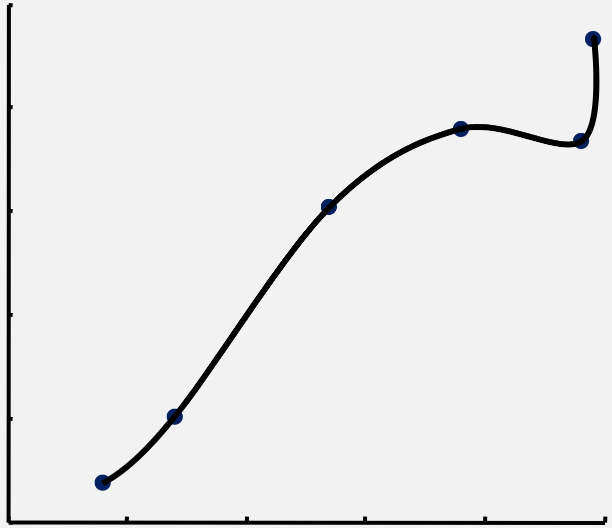
Interpolation Example

- Given the following data:

x	0.8	1.4	2.7	3.8	4.8	4.9
y	0.69	1.00	2.02	2.39	2.34	2.83

Interpolation:

"Join the dots" and find a curve passing through the data.



Regression vs. Interpolation

x	0.8	1.4	2.7	3.8	4.8	4.9
y	0.69	1.00	2.02	2.39	2.34	2.83

- In regression, we are interested in fitting a chosen function to data

$$y = 0.45 + 0.47x$$

- In interpolation, given finite amount of data, we are interested in obtaining new data-points within this range.

$$\text{At } x = 2.0,\ y = 1.87$$

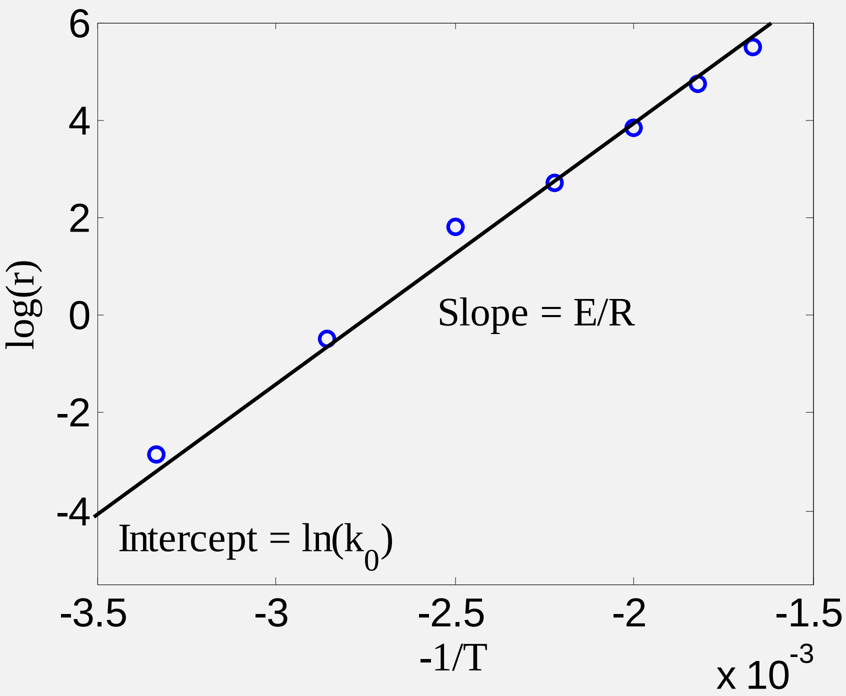
Example: Kinetic Rate Constants (Regression)

- Experiments with conversion measured at various temperatures

$$r = k_0 e^{-E/(RT)} c_a$$

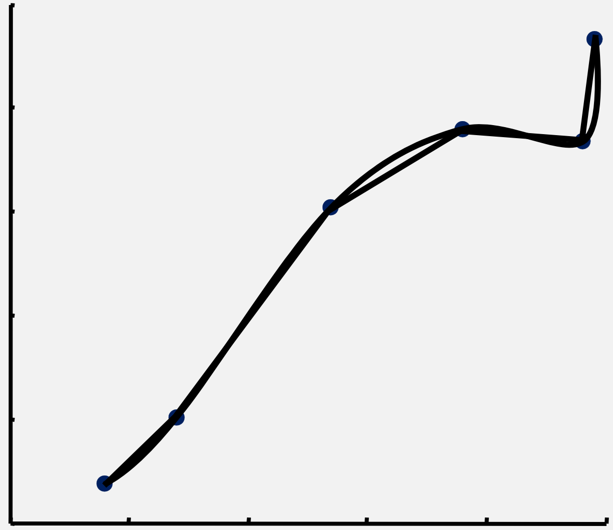
$$\Downarrow$$

$$\ln(r) = \ln(k_0) + \frac{E}{R}\left(\frac{-1}{T}\right) + \ln(c_a)$$



Example: Viscosity of Oil (Interpolation)

- Viscosity of lubricant oil was measured between -20 to 200 degrees Celcius in steps of 20 °C.
- Interpolation is used if viscosity is desired at an intermediate temperature



General Setup

- Let x be an *independent* variable and y be a *dependent* variable

- Given the data:

$$(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_N, y_N)$$

Find parameters θ to get a "best-fit" curve

$$y = f(x; \theta)$$

Regression vs. Interpolation

Regression

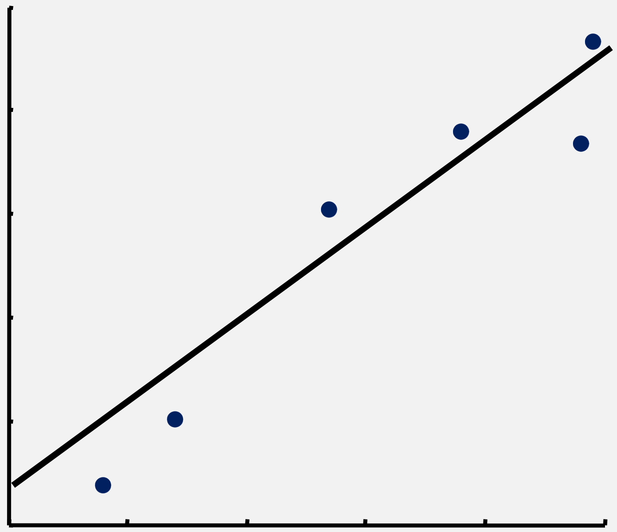
- Choose a function form for $f(x;\theta)$
- For a given θ, obtain the values $\hat{y}_i$ from the model
- The best θ minimizes the error $\|y_i - \hat{y}_i\|$

Interpolation

- Various *standard* function forms exist
- The interpolating function passes through all the points
- Can be used to "fill-in" the data at new points

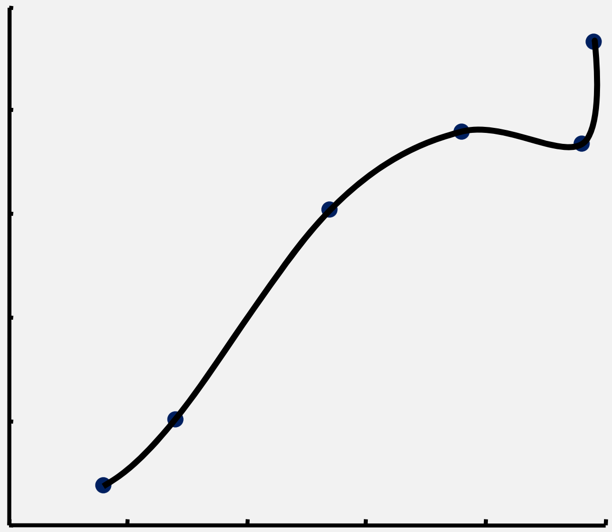
Regression vs. Interpolation

- "Curve Fitting"



- Obtain a functional form to fit the data

- "Joining the dots"



- Obtain the value of y at intermediate point

Outline: Regression

- Linear Regression in One Variable
- Linear Regression in Multiple Variables
- Polynomial Regression
- Analysis and Extension
- Non-Linear Regression

Linear Regression: One Variable

- Model: $y = f(x;\theta)$
- Actual Data: $(x_1, y_1),\ (x_2, y_2),\ \cdots,(x_N, y_N)$
- Prediction: $\hat{y}_i = f(x_i;\theta)$
- Errors: $e_i = y_i - \hat{y}_i$

$$\Downarrow$$

$$y_i = \underbrace{f(x_i;\theta)}_{\hat{y}_i} + e_i$$

- Mean / Variance:

$$\bar{x} = \sum x_i \Big/ N \qquad s_x = \sum (x_i - \bar{x})^2 \Big/ (N-1) \qquad \sigma = \sqrt{s_x}$$

Extension to Multi-Variables

- Let $x_1, x_2, \ldots, x_n$ be n variables.
  Let there be N data points for each:

$$
\begin{aligned}
&\left(x_{11}, x_{21}, \cdots, x_{n1}; y_1\right),\\
&\left(x_{12}, x_{22}, \cdots, x_{n2}; y_2\right),\\
&\vdots\\
&\left(x_{1N}, x_{2N}, \cdots, x_{nN}; y_N\right)
\end{aligned}
$$

⇓ Obtain $\boldsymbol{\theta}$ for

$$y = f\left(\mathbf{x}; \boldsymbol{\theta}\right)$$

Linear Regression (multi-variable)

- Data $(x_i, u_i, w_i; y_i)$
- Model $y = a_0 + a_1 x + a_2 u + a_3 w$
- Error $e_i = y_i - (a_0 + a_1 x_i + a_2 u_i + a_3 w_i)$

- Least Squares Criterion

$$\min_{a_0, a_1, a_2, a_3} \sum_{i=1}^{N} \left(\underbrace{y_i - (a_0 + a_1 x_i + a_2 u_i + a_3 w_i)}_{e_i} \right)^2$$

Linear Regression (alternate)

$$\underbrace{\begin{bmatrix} 1 & x_1 & u_1 & w_1 \\ 1 & x_2 & u_2 & w_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & u_N & w_N \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}}_{\Phi} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{Y}$$

⇓ Least Squares

$$\Phi = \left(X^T X\right)^{-1} X^T Y$$

Polynomial / Functional Regression

- Example: Specific heat as a function of T
  - Methane: $c_p = 85.8 + 1.126\text{e-}2\ T - 2.1141\text{e-}6\ T^2$

- Example: Antoine's vapor pressure relationship
  - $\ln\left(p_{sat}\right) = a - \dfrac{b}{T + c}$

Polynomial / Functional Regression

Model:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^2 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 \end{bmatrix}$$

Model:

$$y = a_0 + a_1 \ln(x) + \frac{a_2}{x} + a_3 x$$

$$X = \begin{bmatrix} 1 & \ln(x_1) & 1/x_1 & x_1 \\ 1 & \ln(x_2) & 1/x_2 & x_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \ln(x_N) & 1/x_N & x_N \end{bmatrix}$$

Outline: Interpolation

- Polynomial fitting and limitations
- Lagrange interpolating polynomials
- Newton's methods
- Spline interpolation

Lagrange Polynomials

$$P_i = \frac{(x - x_1)(x - x_2)\ldots(x - x_{i-1})\ (x - x_{i+1})\ldots(x - x_N)}{(x_i - x_1)(x_i - x_2)\ldots(x_i - x_{i-1})\ (x_i - x_{i+1})\ldots(x_i - x_N)}$$

$$= \prod_{j \neq i} \left(\frac{x - x_j}{x_i - x_j} \right)$$

The interpolating polynomial becomes

$$f(x) = y_1 P_1 + y_2 P_2 + \ldots + y_N P_N$$

Newton's Divided Differences

$$y[i+1,i] = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}$$

$$y[i+2,i+1,i] = \frac{y[i+2,i+1] - y[i+1,i]}{x_{i+2} - x_i}$$

$$y[i+3,i+2,i+1,i] = \frac{y[i+3,i+2,i+1] - y[i+2,i+1,i]}{x_{i+3} - x_i}$$

$$\vdots$$

$$\underbrace{\hspace{10em}}$$

$$c_0 = y_1, \;\; c_1 = y[2,1], \;\; c_2 = y[3,2,1], \;\; c_3 = y[4,3,2,1], \ldots$$

$$f(x) = c_0 + c_1(x - x_1) + c_2(x - x_1)(x - x_2) + \ldots$$