CURVE FITTING AND DATA REGRESSION

Numerical methods in chemical engineering

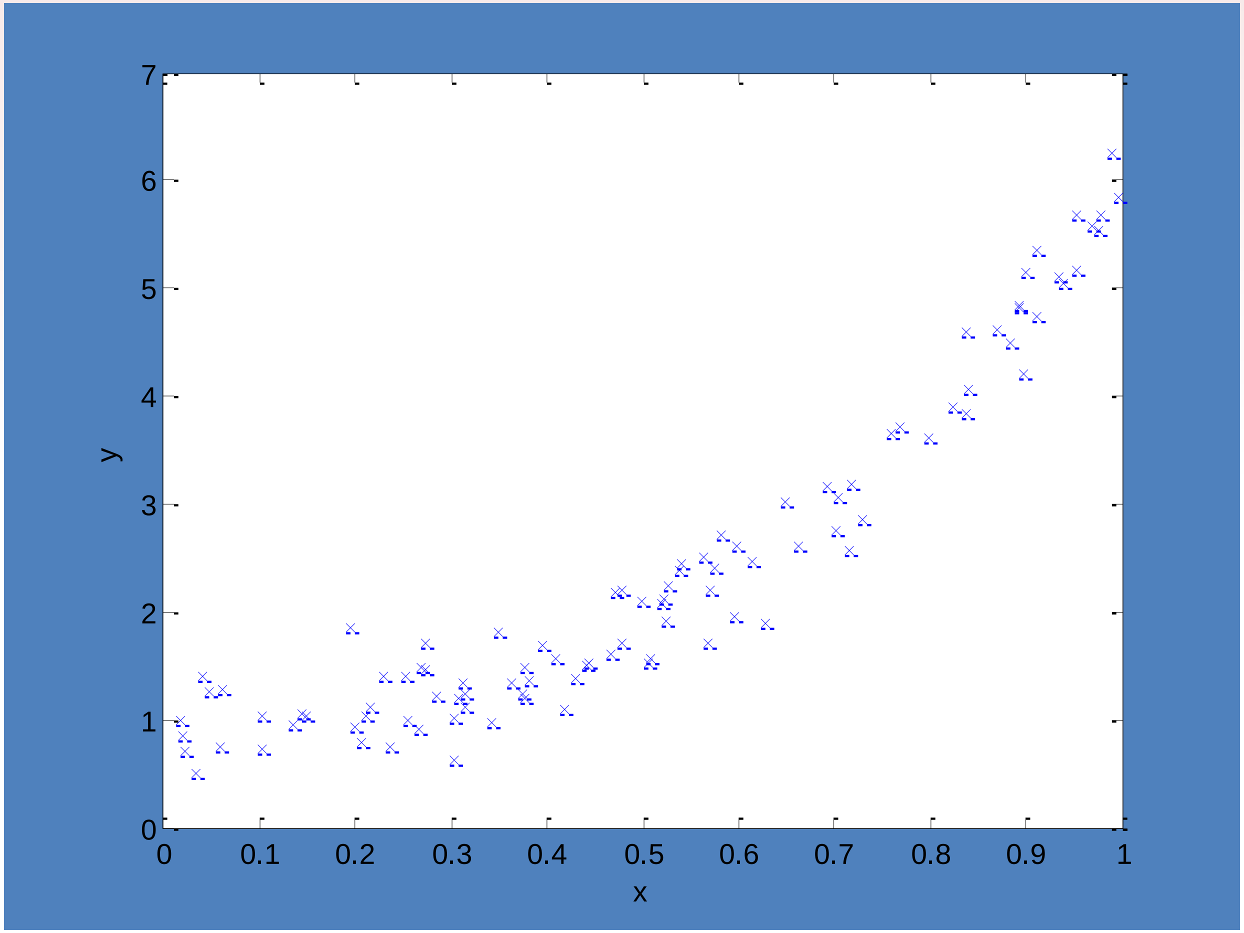
Dr. Edwin Zondervan

OVERVIEW

- We are going to fit measurements to models today.
- You will also learn what R^2 actually means

FITTING MODELS TO DATA

y is the measured variable



x is the controlled variable

HOW TO FIT A MODEL TO THE DATA

- We would like to fit the following model to the data:

$$\hat{y} = a_1 + a_2 x + a_3 x^2 + a_4 x^3 \quad \textbf{(9-1)}$$

- First step: If we have N data points, we could write the model as the product of a matrix and a vector:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$$

(9-2)

$$\hat{y} = Xa \quad \textbf{(9-3)}$$

X is called the design matrix and a are the fit parameters.

RESIDUALS

- Second step: work out the residuals for each data point:

$$d_i = (y_i - \hat{y}_i) \qquad \textbf{(9-4)}$$

- Third step: Work out the sum of squares of the residuals:

$$\sum_i (d_i)^2 = \sum_i (y_i - \hat{y}_i)^2 \qquad \textbf{(9-5)}$$

$$\sum_i (d_i)^2 = d \bullet d = d^T \times d = (y_i - \hat{y}_i)^T (y_i - \hat{y}_i) \qquad \textbf{(9-6)}$$

MINIMIZING THE SUM OF SQUARES

- Choose the parameter vector such that the sum of squares of the residuals is minimized; the partial derivative with respect to each parameter should be set to zero:

$$\frac{\partial}{\partial a_j}[(y-(Xa)^T(y-Xa)]=0 \qquad \textbf{(9-7)}$$

MINIMIZING THE SUM OF SQUARES

$$\frac{\partial}{\partial a_j}[(y^T-(Xa)^T)(y-Xa)]=0 \Leftrightarrow$$

$$\frac{\partial}{\partial a_j}[(y^T-X^Ta^T)(y-Xa)]=0 \Leftrightarrow$$

$$(y^T-X^Ta^T)X\frac{\partial}{\partial a_j}[(a)]+\frac{\partial}{\partial a_j}[(a)^T]X^T(y-Xa)=0 \Leftrightarrow \tag{9-8}$$

$$(y^T-X^Ta^T)Xe_j+e_j{}^TX^T(y-Xa)=0$$

$$(y^T-(Xa)^T)Xe_j+e_j{}^TX^T(y-Xa)=$$

$$(y-(Xa))^TXe_j+e_j{}^TX^T(y-Xa)=0 \Leftrightarrow$$

$$(Xe_j)^T(y-Xa)+e_j{}^TX^T(y-Xa)=0 \Leftrightarrow$$

$$e_j{}^TX^T(y-Xa)+e_j{}^TX^T(y-Xa)=0 \Leftrightarrow$$

$$e_j{}^T(X^T(y-Xa)+X^T(y-Xa))=0 \Leftrightarrow$$

$$2X^T(y-Xa)=0 \Leftrightarrow$$

$$X^Ty=X^TXa$$

$$a=(X^TX)^{-1}X^Ty \tag{9-9}$$

USING MATLAB FOR LLSQ

- If we have the same number of data points as fit parameters, we can solve the system:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^{\,2} & x_1^{\,3} \\ 1 & x_2 & x_2^{\,2} & x_2^{\,3} \\ 1 & x_3 & x_3^{\,2} & x_3^{\,3} \\ 1 & x_4 & x_4^{\,2} & x_4^{\,3} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \tag{9-10}$$

- As `a = X\y`

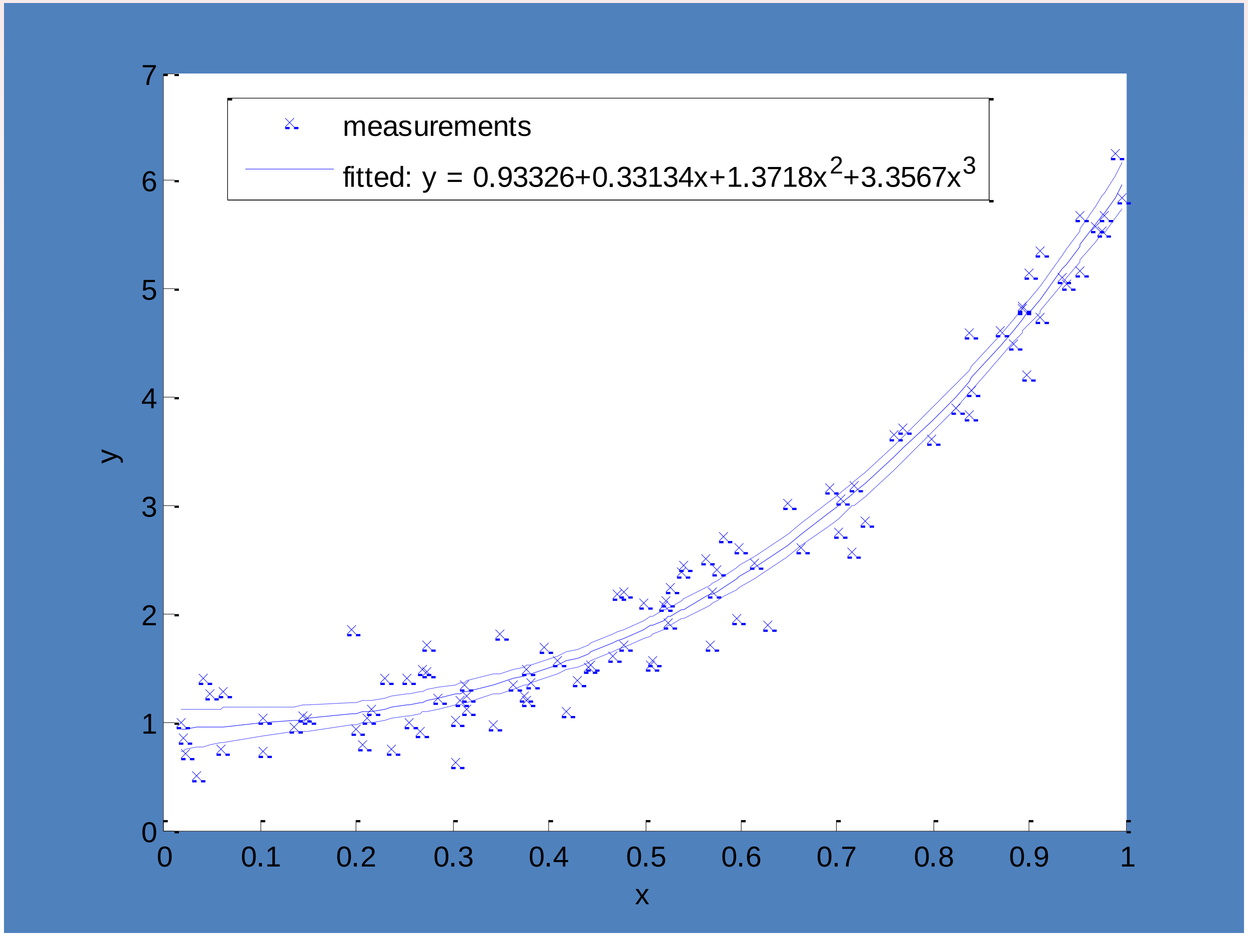
USING MATLAB FPR LLSQ

- If there are more data points (N>4), we can write an analogue, but maybe a consistent solution does not exist (the system is over specified).

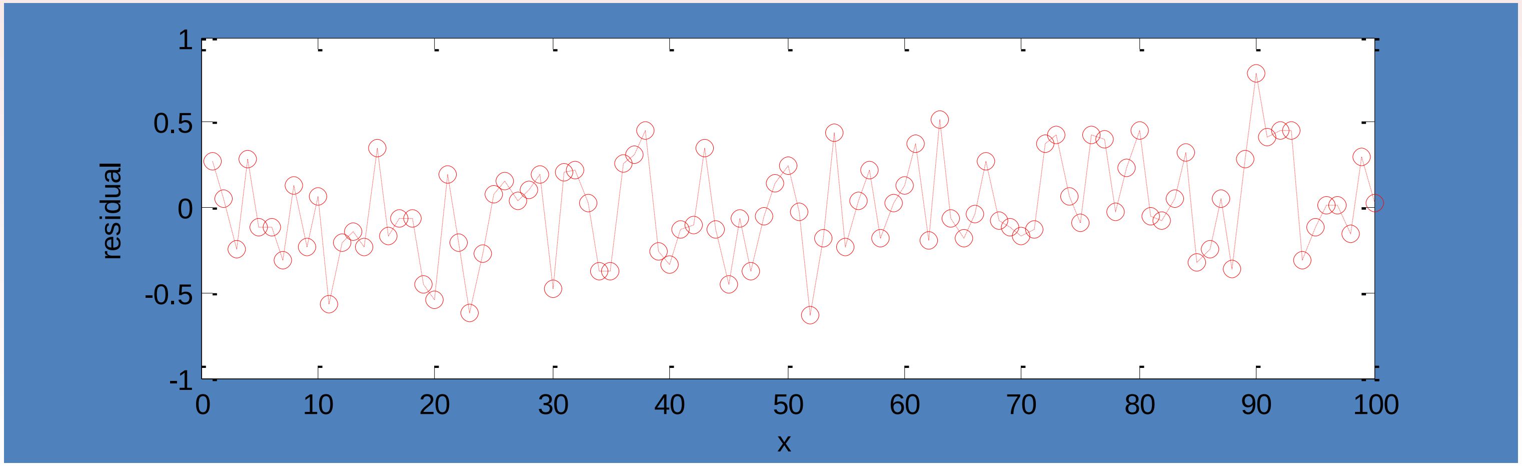
- However, matlab will find values for the vector a which minimize $\|y\text{-}aX\|^2$, so i.e. a least squares fit.

FIT TO OUR PROBLEM

```
N=length(x);

X(:,1) = ones(N,1);

X(:,2) = x;

X(:,3) = x.^2;

X(:,4) = x.^3;


A =X\y;
```



HOW GOOD IS THE MODEL?



- For a model to make sense the data points should be scattered randomly around the model predictions, the mean of the residuals d should be zero.

$$d_i = (y_i - \hat{y}_i)$$

- It's always good to check if the residuals are not correlated with the measured values, if that is the case, it can indicate that your model is wrong.

REGRESSION COEFFICIENTS

- Variance measured in the data (y) is:

$$\sigma_y^2 = \frac{1}{N}\sum_i (y_i - \bar{y})^2 \qquad \textbf{(9-11)}$$

- Variance of the residuals is:

$$\sigma_{error}^2 = \frac{1}{N}\sum_i (d_i)^2 \qquad \textbf{(9-12)}$$

- Variance in the model is:

$$\sigma_{\mathrm{model}}^2 = \frac{1}{N}\sum_i (\hat{y}_i - \bar{\hat{y}}_i)^2 \qquad \textbf{(9-13)}$$

REGRESSION COEFFICIENTS

- Given that the error is uncorrelated we can state that:

$$\sigma_y^2 = \sigma_{error}^2 + \sigma_{\text{model}}^2 \qquad \textbf{(9-14)}$$

$$R^2 = \frac{\sigma_{\text{model}}^2}{\sigma_y^2} = 1 - \frac{\sigma_{error}^2}{\sigma_y^2} \qquad \textbf{(9-15)}$$

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

(9-16)

SSE: Sum of errors squared

SSR: Sum of squares (data)

SST: Total sum of squares (model)

STATISTICAL ANALYSIS

- An uncorrellated error (mean will be zero) → SSE, SST and SSR will have χ^2-distributions and the ratios will have an *F*-distribution. If SSR/SSE is large, the model is good!
- There is a change that the model is rubbish, but that SSR/SSE will yield a good value, Analysis of Variance (ANOVA) will be a good tool to calculate the probability of such a thing happening!

BACK TO THE EXAMPLE

- Stats:
  - N = 100
  - SSE = 8.1031
  - SST = 232.5490
  - SSR = 224.4459
  - $R^2 = 0.9652$

Source	Deg. Of freedom	Sum of squares	F-value
Regression	K = 4	SSR = 224.44	F = 657.84
Residual	N-K-1= 95	SSE = 8.103	
Total	99	SST = 232.55	

F>657 means: **very unlikely!!!**

F = (SSR/4)/(SSE/95)

CONFIDENCE LIMITS FOR THE FIT PARAMETERS

- Using the t-distribution, the confidence limits for the fit parameters can be set as:

(9-17)

$$a_j - t\frac{\sigma^2_{error}}{\nu}[(X^T X)^{-1}]_{j,j} < a_j < a_j + t\frac{\sigma^2_{error}}{\nu}[(X^T X)^{-1}]_{j,j}$$

Degree of freedom

Is the j-th diagonal element of a the symmetric matrix

Must be looked up from a statistical table

CONFIDENCE LIMITS FOR THE PREDICTED POINTS

- A confidence interval for each predicted value is given by:

$$\hat{y}_i - t\frac{\sigma_{error}}{\nu}\sqrt{[X(X^T X)^{-1}X^T]}_{j,j} < \hat{y}_j < \hat{y}_i + t\frac{\sigma_{error}}{\nu}\sqrt{[X(X^T X)^{-1}X^T]}_{j,j}$$

(9-18)

SUMMARY

- We have seen how fit parameters of a model can fitted to a data set, using the linear least squares method.
- We found out how to calculate the regression coefficients and how to perform a statistical analysis of the model using ANOVA.
- We also postulated expressions for the confidence limits for the fit parameters and the predicted points