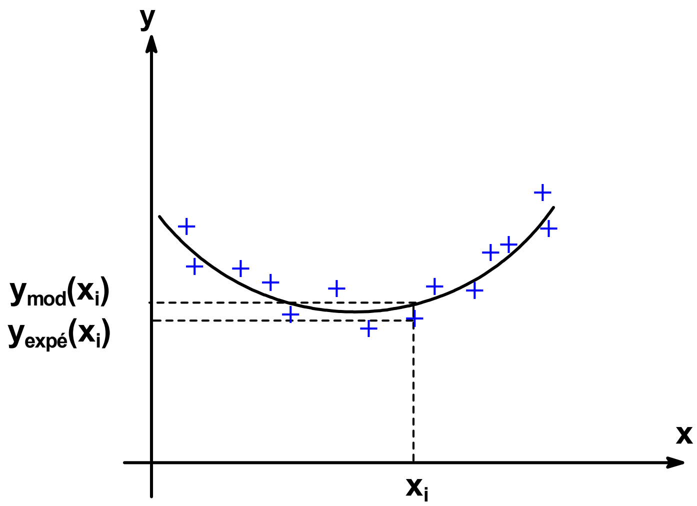
Régression polynomiale

Mise en œuvre de la regression parabolique :

Soit $\hat{y} = \hat{a}_2 x^2 + \hat{a}_1 x + \hat{a}_0$ l'équation de la parabole optimale pour N points.
Ce polynôme est obtenu par la méthode des moindres carrés .



Cette parabole est telle que la somme S des carrés des différences [yi(expé)-yi(modèle)] soit minimale.

$$S = \sum_i \left[y_{i(expé)} - \left(\hat{a}_2 x_i^2 + \hat{a}_1 x_i + \hat{a}_0 \right) \right]^2$$

$$\begin{cases} \dfrac{dS}{d\hat{a}_2} = -2\sum_i \left(x_i^2 y_i \right) + 2\hat{a}_2 \sum_i x_i^4 + 2\hat{a}_1 \sum_i x_i^3 + 2\hat{a}_0 \sum_i x_i^2 = 0 & (1) \\[2em] \dfrac{dS}{d\hat{a}_1} = -2\sum_i \left(x_i y_i \right) + 2\hat{a}_2 \sum_i x_i^3 + 2\hat{a}_1 \sum_i x_i^2 + 2\hat{a}_0 \sum_i x_i = 0 & (2) \\[2em] \dfrac{dS}{d\hat{a}_0} = -2\sum_i y_i + 2\hat{a}_2 \sum_i x_i^2 + 2\hat{a}_1 \sum_i x_i + 2\hat{a}_0 N = 0 & (3) \end{cases}$$

On en tire :

$$
\begin{cases}
\hat{a}_2 \sum\limits_i x_i^4 + \hat{a}_1 \sum\limits_i x_i^3 + \hat{a}_0 \sum\limits_i x_i^2 = \sum\limits_i \left(x_i^2 y_i\right) & (1') \\
\hat{a}_2 \sum\limits_i x_i^3 + \hat{a}_1 \sum\limits_i x_i^2 + \hat{a}_0 \sum\limits_i x_i = \sum\limits_i \left(x_i y_i\right) & (2') \\
\hat{a}_2 \sum\limits_i x_i^2 + \hat{a}_1 \sum\limits_i x_i + \hat{a}_0 N = \sum\limits_i y_i & (3')
\end{cases}
$$

soit sous forme matricielle :

$$
\begin{vmatrix} \hat{a}_2 \\ \hat{a}_1 \\ \hat{a}_0 \end{vmatrix} \times
\begin{vmatrix}
\sum\limits_i x_i^4 & \sum\limits_i x_i^3 & \sum\limits_i x_i^2 \\
\sum\limits_i x_i^3 & \sum\limits_i x_i^2 & \sum\limits_i x_i \\
\sum\limits_i x_i^2 & \sum\limits_i x_i & N
\end{vmatrix}
=
\begin{vmatrix} \hat{a}_2 \\ \hat{a}_1 \\ \hat{a}_0 \end{vmatrix} \times |M| =
\begin{vmatrix} \sum\limits_i x_i^2 y_i \\ \sum\limits_i x_i y_i \\ \sum\limits_i y_i \end{vmatrix}
$$

dont on tire :

$$
\hat{a}_2 = \frac{\begin{vmatrix}
\sum\limits_i \left(x_i^2 y_i\right) & \sum\limits_i x_i^3 & \sum\limits_i x_i^2 \\
\sum\limits_i \left(x_i y_i\right) & \sum\limits_i x_i^2 & \sum\limits_i x_i \\
\sum\limits_i y_i & \sum\limits_i x_i & N
\end{vmatrix}}{\begin{vmatrix}
\sum\limits_i x_i^4 & \sum\limits_i x_i^3 & \sum\limits_i x_i^2 \\
\sum\limits_i x_i^3 & \sum\limits_i x_i^2 & \sum\limits_i x_i \\
\sum\limits_i x_i^2 & \sum\limits_i x_i & N
\end{vmatrix}}
\qquad
\hat{a}_1 = \frac{\begin{vmatrix}
\sum\limits_i x_i^4 & \sum\limits_i \left(x_i^2 y_i\right) & \sum\limits_i x_i^2 \\
\sum\limits_i x_i^3 & \sum\limits_i \left(x_i y_i\right) & \sum\limits_i x_i \\
\sum\limits_i x_i^2 & \sum\limits_i y_i & N
\end{vmatrix}}{\begin{vmatrix}
\sum\limits_i x_i^4 & \sum\limits_i x_i^3 & \sum\limits_i x_i^2 \\
\sum\limits_i x_i^3 & \sum\limits_i x_i^2 & \sum\limits_i x_i \\
\sum\limits_i x_i^2 & \sum\limits_i x_i & N
\end{vmatrix}}
$$

$$\hat{a}_0 = \frac{\begin{vmatrix} \sum_i x_i^4 & \sum_i x_i^3 & \sum_i \left(x_i^2 y_i\right) \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i \left(x_i y_i\right) \\ \sum_i x_i^2 & \sum_i x_i & \sum_i y_i \end{vmatrix}}{\begin{vmatrix} \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^2 & \sum_i x_i & N \end{vmatrix}}$$

$$D = \sum_i x_i^4 \left(N \sum_i x_i^2 - \left[\sum_i x_i \right]^2 \right) - \sum_i x_i^3 \left(N \sum_i x_i^3 - \sum_i x_i^2 \sum_i x_i \right) + \sum_i x_i^2 \left(\sum_i x_i \sum_i x_i^3 - \left[\sum_i x_i^2 \right]^2 \right)$$

$$\hat{a}_2 = \frac{\sum_i x_i^2 y_i \left(N \sum_i x_i^2 - \left[\sum_i x_i \right]^2 \right) - \sum_i x_i^3 \left(N \sum_i x_i y_i - \sum_i x_i \sum_i y_i \right) + \sum_i x_i^2 \left(\sum_i x_i y_i \sum_i x_i - \sum_i y_i \sum_i x_i^2 \right)}{D}$$

$$\hat{a}_1 = \frac{\sum_i x_i^4 \left(N \sum_i x_i y_i - \sum_i x_i \sum_i y_i \right) - \sum_i x_i^2 y_i \left(N \sum_i x_i^3 - \sum_i x_i^2 \sum_i x_i \right) + \sum_i x_i^2 \left(\sum_i y_i \sum_i x_i^3 - \sum_i x_i y_i \sum_i x_i^2 \right)}{D}$$

$$\hat{a}_0 = \frac{\sum_i x_i^4 \left(\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i \right) - \sum_i x_i^3 \left(\sum_i x_i^3 \sum_i y_i - \sum_i x_i^2 \sum_i x_i y_i \right) + \sum_i x_i^2 y_i \left(\sum_i x_i \sum_i x_i^3 - \left(\sum_i x_i^2 \right)^2 \right)}{D}$$

Variances de a_2, a_1 et a_0

Les variances sont les termes diagonaux de la matrice V :

$$V = \begin{vmatrix} var(\hat{a}_2) & cov(\hat{a}_2,\hat{a}_1) & cov(\hat{a}_2,\hat{a}_0) \\ cov(\hat{a}_1,\hat{a}_2) & var(\hat{a}_1) & cov(\hat{a}_1,\hat{a}_0) \\ cov(\hat{a}_0,\hat{a}_2) & cov(\hat{a}_0,\hat{a}_1) & var(\hat{a}_0) \end{vmatrix}$$

Cette matrice V est telle que : $|V| = s_y^2 |M|^{-1}$ avec :

$$M = \begin{vmatrix} \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^2 & \sum_i x_i & N \end{vmatrix}$$

$$\mathrm{Det}|M| = \sum_i x_i^4 \begin{vmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & N \end{vmatrix} - \sum_i x_i^3 \begin{vmatrix} \sum_i x_i^3 & \sum_i x_i \\ \sum_i x_i^2 & N \end{vmatrix} + \sum_i x_i^2 \begin{vmatrix} \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^2 & \sum_i x_i \end{vmatrix}$$

$$= \sum_i x_i^4 \left[N\sum_i x_i^2 - \left(\sum_i x_i \right)^2 \right] - \sum_i x_i^3 \left[N\sum_i x_i^3 - \sum_i x_i \sum_i x_i^2 \right] + \sum_i x_i^2 \left[N\sum_i x_i^3 \sum_i x_i - \left(\sum_i x_i^2 \right)^2 \right]$$

Exercice RegPoly1 :

Effectuer une régression parabolique sur les données relatives à l'indice de réfraction n de mélanges eau-glucose : $n=a_0+a_1x+a_2x^2$ (x désigne le titre (fraction) massique exprimé en % de glucose) pour déterminer les valeurs de $\hat{a}_0$, $\hat{a}_1$ et $\hat{a}_2$.

On utilisera la fonction SOMMEPROD() d'EXCEL pour calculer les expressions du type : $\sum_i x_i^4$, $\sum_i x_i^3$,......., $\sum_i x_i^2 y_i$.

La matrice $M = \begin{vmatrix} \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^2 & \sum_i x_i & N \end{vmatrix}$

sera inversée en utilisation la fonction INVERSEMAT d'EXCEL.

Déterminer les valeurs σ_{a_0} , σ_{a_1} et σ_{a_2} des écarts-types de a_0, a_1 et a_2 .

σ_{a_0} , σ_{a_1} et σ_{a_2} sont les éléments diagonaux de $\sigma_y^2 \times |M|^{-1}$.

- En déduire $a_0 = a_{0MC} \pm \Delta a_{0MC}$

  $a_1 = a_{1MC} \pm \Delta a_{1MC}$

  $a_2 = a_{2MC} \pm \Delta a_{2MC}$

avec un intervalle de confiance de 95%.

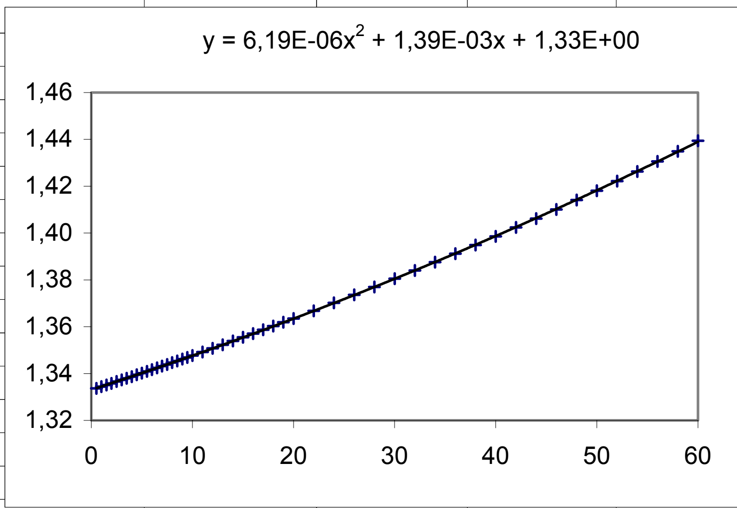
- La quantité $\frac{a_2 - \hat{a}_2}{\sigma_{a_2}}$ suit une loi de Student à N-3 degrés de liberté. On pourra tester l'hypothèse $a_2 = 0$ pour savoir si a_2 a une contribution significative dans la régression. Dans l'exemple envisagé ici :

$\hat{a}_2 = 6{,}19\times10^{-6}$
$\sigma_{a_2} = 5{,}85\times10^{-8}$

Sous l'hypothèse $a_2 = 0$, $\frac{a_2 - \hat{a}_2}{\sigma_{a_2}} = \frac{-6{,}19\times10^{-6}}{5{,}85\times10^{-8}}$ = -106. La probabilité d'observer une valeur supérieure à 106 en valeur absolue pour une variable de Student à 50-3=47 degrés de liberté est très faible , ce qui amène à rejeter l'hypothèse de nullité de a_2.

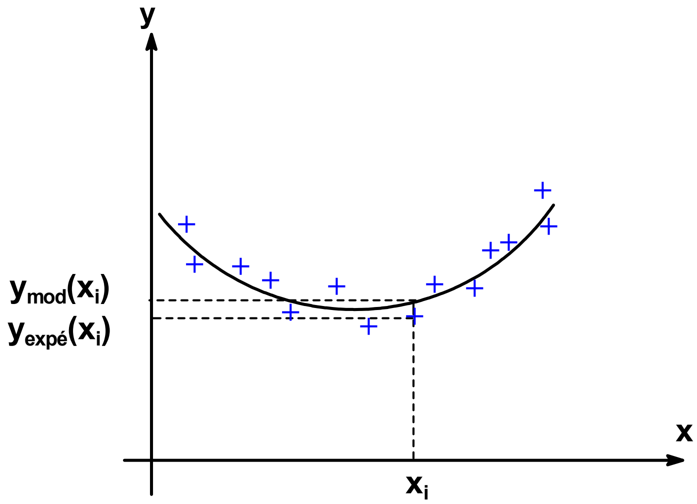
	A	B	C	D	E	F	G	H
1	**Indice de réfraction d'une sol aq de glucose $n=a_2*x^2+a_1*x+a_0$**							
2								
3		8,47E+07	1,75E+06	3,95E+04	**S**			
4	**M=**	1,75E+06	3,95E+04	1,08E+03	**5,58E-07**	**a_2**	**a_1**	**a_0**
5		3,95E+04	1,08E+03	50		6,19E-06	1,394E-03	1,33
6					**σ(y)**	**$\sigma(a_2)$**	**$\sigma(a_1)$**	**$\sigma(a_0)$**
7		**2,88E-07**	-1,60E-05	1,19E-04	**1,09E-04**	5,85E-08	3,37E-06	3,41E-05
8	**M^{-1}=**	-1,60E-05	**9,54E-04**	-7,95E-03				
9		1,19E-04	-7,95E-03	**9,80E-02**		**1,06E+02**		
10								
11								
12	**Sx**	**Sy**	**Sx4**	**Sx3**	**Sx2**			
13	1080,00	68,40	8,47E+07	1,75E+06	3,95E+04			
14	**D**	**Sxy**	**Sx2y**	N				
15	2,81E+12	1,51E+03	5,56E+04	50				
16	x	y						**ycalculé**
17	0,5000	1,3337						**1,3338**
18	1,0000	1,3344						**1,3345**
19	1,5000	1,3351						**1,3352**
20	2,0000	1,3358						**1,3359**
21	2,5000	1,3366						**1,3366**
22	3,0000	1,3373						**1,3373**
23	3,5000	1,3380						**1,3381**
24	4,0000	1,3387						**1,3388**
25	4,5000	1,3395						**1,3395**
26	5,0000	1,3402						**1,3402**
27	5,5000	1,3409						**1,3410**
28	6,0000	1,3417						**1,3417**
29	6,5000	1,3424						**1,3424**
30	7,0000	1,3432						**1,3432**
31	7,5000	1,3439						**1,3439**
32	8,0000	1,3447						**1,3446**
33	8,5000	1,3454						**1,3454**
34	9,0000	1,3462						**1,3461**
35	9,5000	1,3469						**1,3469**
36	10,0000	1,3477						**1,3477**

Quotient $a_2/\sigma(a_2)$
Ici la valeur de ce quotient justifie une regression parabolique plutot que linéaire



Régression polynomiale : Mise en œuvre de la régression cubique

Soit $y = a_3x^3+a_2x^2+a_1x+a_0$ l'équation utilisée pour estimer les valeurs de y_i pour N points.
Les a_i sont déterminés par la méthode des moindres carrés .



La somme S des carrés des différences [yi(expé)-yi(modèle)] doit être minimale.

$$S = \sum_{i=1}^{N} \left[y_{i(expé)} - \left(a_3x_i^2 + a_2x_i^2 + a_1x_i + a_0 \right) \right]^2$$

Les valeur des a_i sont telles que $\dfrac{dS}{da_i} = 0$

On en déduit :

$$\begin{cases} a_3\sum_i x_i^6 + a_2\sum_i x_i^5 + a_1\sum_i x_i^4 + a_0\sum_i x_i^3 = \sum_i \left(x_i^3 y_i\right) & (1') \\ a_3\sum_i x_i^5 + a_2\sum_i x_i^4 + a_1\sum_i x_i^3 + a_0\sum_i x_i^2 = \sum_i \left(x_i^2 y_i\right) & (2') \\ a_3\sum_i x_i^4 + a_2\sum_i x_i^3 + a_1\sum_i x_i^2 + a_0\sum_i x_i = \sum_i \left(x_i y_i\right) & (3') \\ a_3\sum_i x_i^3 + a_2\sum_i x_i^2 + a_1\sum_i x_i + a_0 N = \sum_i y_i & (4') \end{cases}$$

soit sous forme matricielle :

$$\begin{vmatrix} a_3 \\ a_2 \\ a_1 \\ a_0 \end{vmatrix} \times \begin{vmatrix} \sum_i x_i^6 & \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 \\ \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i & N \end{vmatrix} = \begin{vmatrix} a_3 \\ a_2 \\ a_1 \\ a_0 \end{vmatrix} \times |M| = \begin{vmatrix} \sum_i x_i^3 y_i \\ \sum_i x_i^2 y_i \\ \sum_i x_i y_i \\ \sum_i y_i \end{vmatrix}$$

On en déduit :

$$a_3 = \frac{\begin{vmatrix} \sum_i x_i^3 y_i & \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 \\ \sum_i x_i^2 y_i & \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i y_i & \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i y_i & \sum_i x_i^2 & \sum_i x_i & N \end{vmatrix}}{\begin{vmatrix} \sum_i x_i^6 & \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 \\ \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i & N \end{vmatrix}}$$

$$a_2 = \frac{\begin{vmatrix} \sum_i x_i^6 & \sum_i x_i^3 y_i & \sum_i x_i^4 & \sum_i x_i^3 \\ \sum_i x_i^5 & \sum_i x_i^2 y_i & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^4 & \sum_i x_i y_i & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^3 & \sum_i y_i & \sum_i x_i & N \end{vmatrix}}{\begin{vmatrix} \sum_i x_i^6 & \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 \\ \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i & N \end{vmatrix}}$$

$$a_1 = \frac{\begin{vmatrix} \sum_i x_i^6 & \sum_i x_i^5 & \sum_i x_i^3 y_i & \sum_i x_i^3 \\ \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^2 y_i & \sum_i x_i^2 \\ \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i y_i & \sum_i x_i \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i y_i & N \end{vmatrix}}{\begin{vmatrix} \sum_i x_i^6 & \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 \\ \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i & N \end{vmatrix}}$$

$$a_0 = \frac{\begin{vmatrix} \sum_i x_i^6 & \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 y_i \\ \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 y_i \\ \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i y_i \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i & \sum_i y_i \end{vmatrix}}{\begin{vmatrix} \sum_i x_i^6 & \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 \\ \sum_i x_i^5 & \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i & N \end{vmatrix}}$$