Multiple Linear and Polynomial Regression with Statistical Analysis

Given a set of data of measured (or observed) values of a dependent variable: y_i versus n independent variables x_{1i}, x_{2i}, … x_{ni}, multiple linear regression attempts to find the "best" values of the parameters a_0, a_1, …a_n for the equation

$$\hat{y}_i = a_0 + a_1 x_{1,i} + a_2 x_{2,i} + \ldots + a_n x_{n,i}$$

$\hat{y}_i$ is the calculated value of the dependent variable at point i. The "best" parameters have values that minimize the squares of the errors

$$S = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

In polynomial regression there is only one independent variable, thus

$$\hat{y}_i = a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_n x_i^n$$

Multiple Linear and Polynomial Regression with Statistical Analysis

Typical examples of multiple linear and polynomial regressions include correlation of temperature dependent physical properties, correlation of heat transfer data using dimensionless groups, correlation of non-ideal phase equilibrium data and correlation of reaction rate data.

The software packages enable high precision correlation of the data, however statistical analysis is essential to determine the ***quality of the fit*** (how well the regression model fits the data) and the ***stability of the model*** (the level of dependence of the model parameters on the particular set of data).

The most important indicators for such studies are the ***residual plot*** (quality of the fit) and ***95% confidence intervals*** (stability of the model)

Correlation of Heat Capacity Data for Ethane

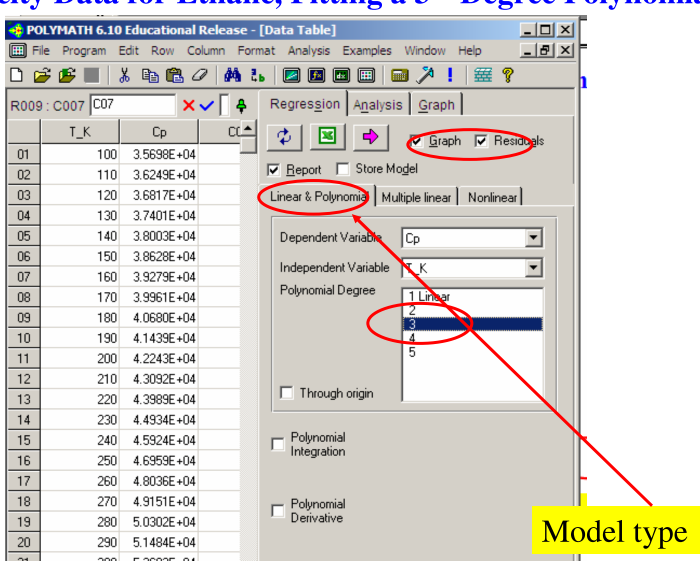
A polynomial has to be fitted to heat capacity data provided by Ingham et al*. This data set includes 41 data points in the temperature range of 100 K – 400 K.
The degree of the polynomial:

$$C_p = a_0 + a_1 T + a_2 T^2 + \ldots + a_n T^n$$

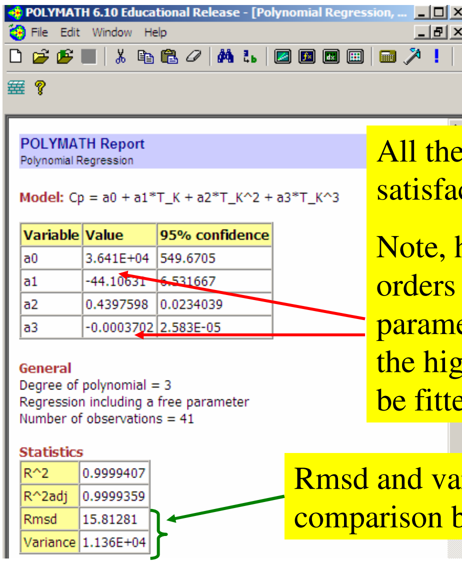
where C_p is the heat capacity in J/kg-mol·K, T is the temperature in K, and a_0, a_1,... are the regression model parameters, which best represents the data, has to be found.
The goodness of fit should be determined based on the variance, the correlation coefficient (R^2), the confidence intervals of the parameters, and the residual plot.

Ingham, H.; Friend, D.G.; Ely, J.F.; "Thermophysical Properties of Ethane"; *J. Phys. Ref. Data* 1991, 20, 275

Heat Capacity Data for Ethane, Fitting a 3rd Degree Polynomial



Heat Capacity Data for Ethane, Fitting a 3rd Degree Polynomial


POLYMATH 6.10 Educational Release - [Polynomial Regression, ...

File Edit Window Help

POLYMATH Report

Polynomial Regression

Model: Cp = a0 + a1*T_K + a2*T_K^2 + a3*T_K^3

Variable	Value	95% confidence
a0	3.641E+04	549.6705
a1	-44.10631	6.531667
a2	0.4397598	0.0234039
a3	-0.0003702	2.583E-05

General

Degree of polynomial = 3

Regression including a free parameter

Number of observations = 41

Statistics

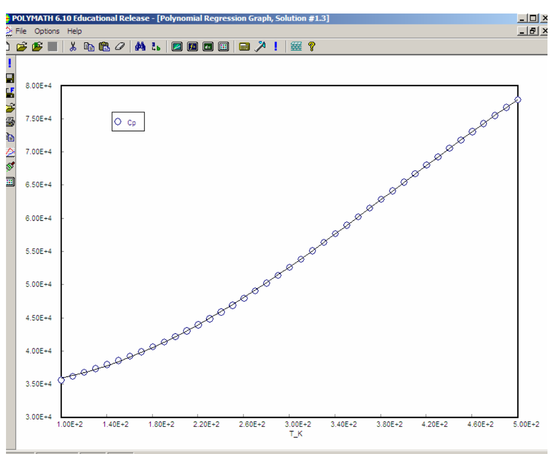
R^2	0.9999407
R^2adj	0.9999359
Rmsd	15.81281
Variance	1.136E+04

All the parameters indicate satisfactory model.

Note, however, the differences of orders of magnitude between the parameter values. This may limit the highest degree of polynomial to be fitted

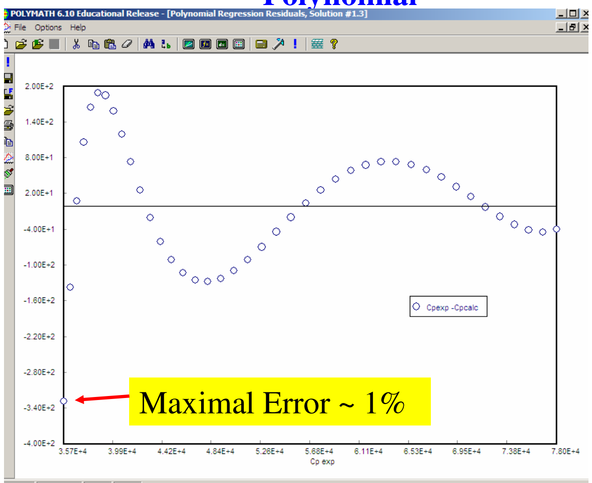
Rmsd and variance values used for comparison between different models

Heat Capacity Data for Ethane, Calculated (3rd Degree Polynomial) and Experimental Values



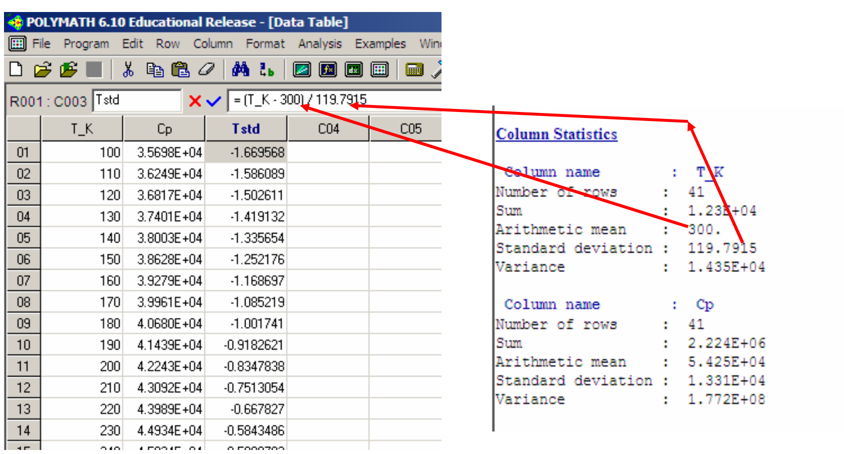
On the scale of the entire range of the C_p data the fit seems to be excellent

Heat Capacity Data for Ethane, Residual Plot for the 3rd Degree Polynomial



High resolution residual plot shows oscillatory behavior which is not explained by the 3rd degree polynomial

Heat Capacity Data for Ethane, Defining Standardized Temperature Values for High Order Polynomial Fitting



POLYMATH 6.10 Educational Release - [Data Table]

R001 : C003 Tstd = (T_K - 300) / 119.7915

	T_K	Cp	Tstd	C04	C05
01	100	3.5698E+04	-1.669568		
02	110	3.6249E+04	-1.586089		
03	120	3.6817E+04	-1.502611		
04	130	3.7401E+04	-1.419132		
05	140	3.8003E+04	-1.335654		
06	150	3.8628E+04	-1.252176		
07	160	3.9279E+04	-1.168697		
08	170	3.9961E+04	-1.085219		
09	180	4.0680E+04	-1.001741		
10	190	4.1439E+04	-0.9182621		
11	200	4.2243E+04	-0.8347838		
12	210	4.3092E+04	-0.7513054		
13	220	4.3989E+04	-0.667827		
14	230	4.4934E+04	-0.5843486		

Column Statistics

```
 Column name        :  T_K
Number of rows      :  41
Sum                 :  1.23E+04
Arithmetic mean     :  300.
Standard deviation  :  119.7915
Variance            :  1.435E+04

 Column name        :  Cp
Number of rows      :  41
Sum                 :  2.224E+06
Arithmetic mean     :  5.425E+04
Standard deviation  :  1.331E+04
Variance            :  1.772E+08
```

Heat Capacity Data for Ethane, Fitting a 5rd Degree Polynomial

POLYMATH 6.10 Educational Release - [Polynomial Regression, Solution #1]

POLYMATH Report
Polynomial Regression

Model: Cp = a0 + a1*Tstd + a2*Tstd^2 + a3*Tstd^3 + a4*Tstd^4 + a5*Tstd^5

Variable	Value	95% confidence
a0	5.268E+04	7.53794
a1	1.461E+04	17.87844
a2	1812.29	16.18934
a3	-1041.693	24.12796
a4	-112.7446	6.199
a5	125.1056	7.262645

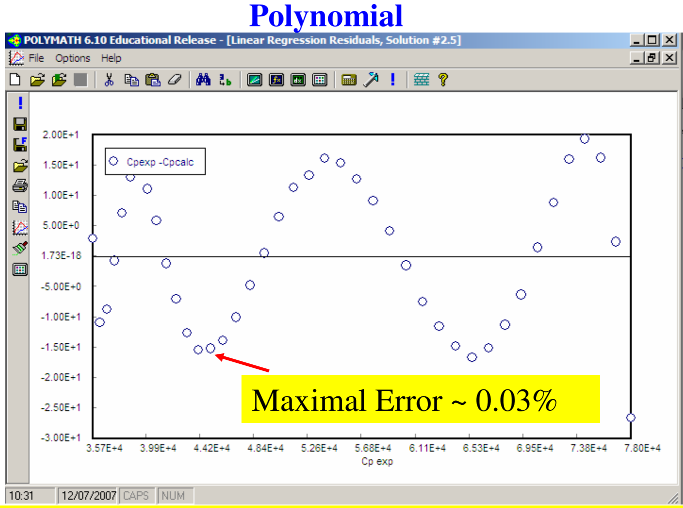
General
Degree of polynomial = 5
Regression including a free parameter
Number of observations = 41

Statistics

R^2	0.9999992
R^2adj	0.9999991
Rmsd	1.834451
Variance	161.6262

Using standardized values yields model parameters of similar magnitude, enables fitting higher order polynomials and improves considerably all the statistical indicators

Heat Capacity Data for Ethane, Residual Plot of a 5th Degree Polynomial



Using standardized independent variable values enables fitting polynomials with ***precision higher than justified by the experimental error***.

Modeling Vapor Pressure Data for Ethane

A vapor pressure data set provided by Ingham et al* includes 107 data points in the temperature range of 92 K – 304 K. This temperature range covers almost completely the range between the tripe point temperature (= 90.352 K) and the critical temperature (T_C = 305.32 K).
The temperature dependence of the vapor pressure should be modeled by the Clapeyron, Antoine and Wagner equations

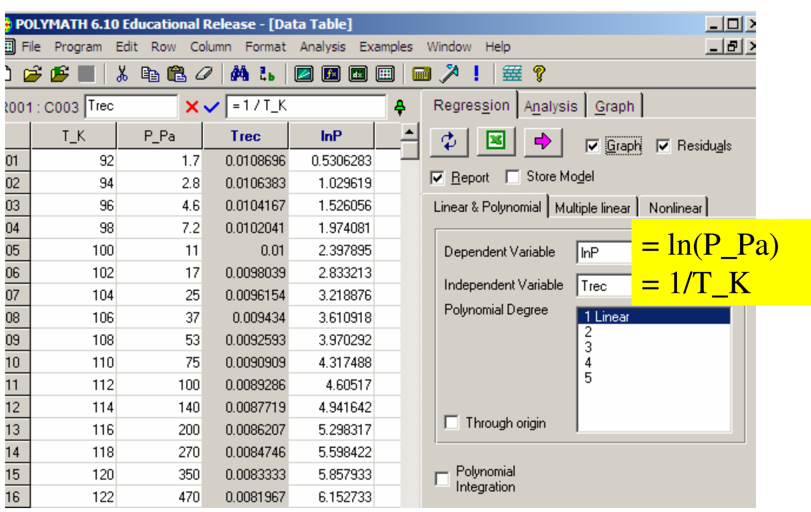
The Clapeyron equation is a two parameter equation:

$$\ln P = A + \frac{B}{T}$$

where P is the vapor pressure (Pa), T – temperature (K) , A and B are parameters

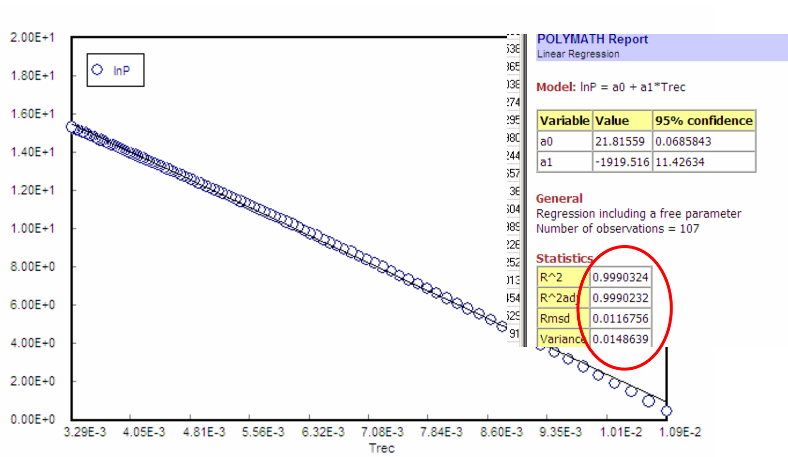
*Ingham, H.; Friend, D.G.; Ely, J.F.; "Thermophysical Properties of Ethane"; *J. Phys. Ref. Data* 1991, 20, 275

Modeling Vapor Pressure Data for Ethane by the Clapeyron Equation using Linear Regression



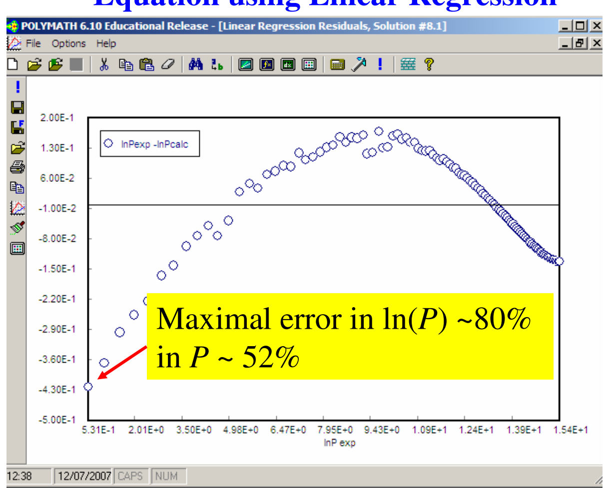
	T_K	P_Pa	Trec	lnP
01	92	1.7	0.0108696	0.5306283
02	94	2.8	0.0106383	1.029619
03	96	4.6	0.0104167	1.526056
04	98	7.2	0.0102041	1.974081
05	100	11	0.01	2.397895
06	102	17	0.0098039	2.833213
07	104	25	0.0096154	3.218876
08	106	37	0.009434	3.610918
09	108	53	0.0092593	3.970292
10	110	75	0.0090909	4.317488
11	112	100	0.0089286	4.60517
12	114	140	0.0087719	4.941642
13	116	200	0.0086207	5.298317
14	118	270	0.0084746	5.598422
15	120	350	0.0083333	5.857933
16	122	470	0.0081967	6.152733

Modeling Vapor Pressure Data for Ethane by the Clapeyron Equation using Linear Regression



All the indicators show good, acceptable fit!

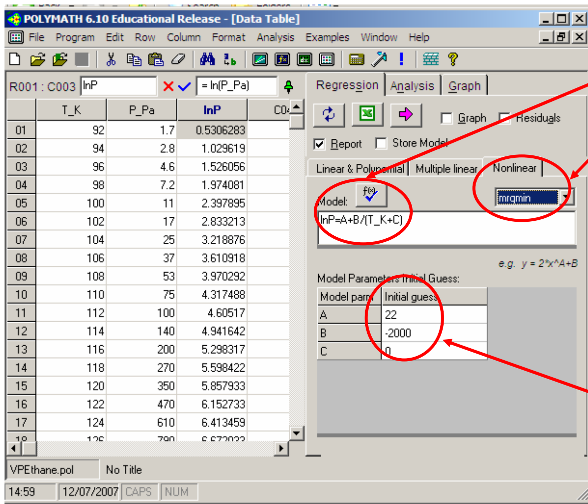
Modeling Vapor Pressure Data for Ethane by the Clapeyron Equation using Linear Regression



Maximal error in ln(*P*) ~80% in *P* ~ 52%

The residual plot reveals large unexplained curvature in the data

Modeling Vapor Pressure Data for Ethane by the Antoine Equation using Non-linear Regression

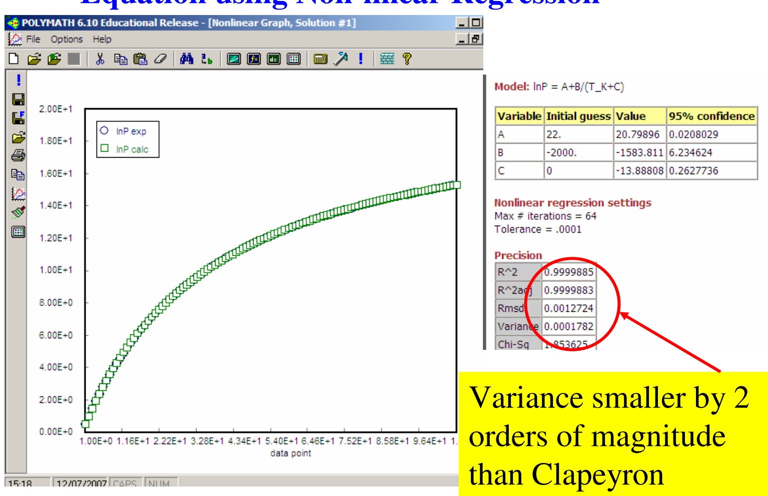


$$\ln P = A + \frac{B}{T + C}$$

Model type and solution algorithm

Initial guess from Clapeyron eqn.

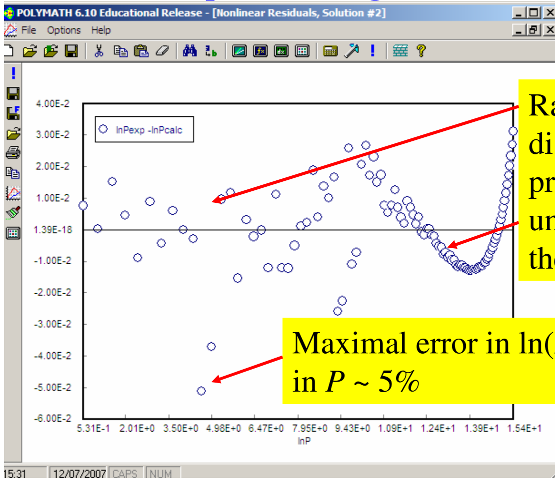
Modeling Vapor Pressure Data for Ethane by the Antoine Equation using Non-linear Regression



Variance smaller by 2 orders of magnitude than Clapeyron

Experimental and calculated values cannot be distinguished.

Modeling Vapor Pressure Data for Ethane by the Antoine Equation using Non-linear Regression



Random residual distribution in the low pressure range, unexplained curvature in the high pressure range

Maximal error in ln(*P*) ~1% in *P* ~ 5%

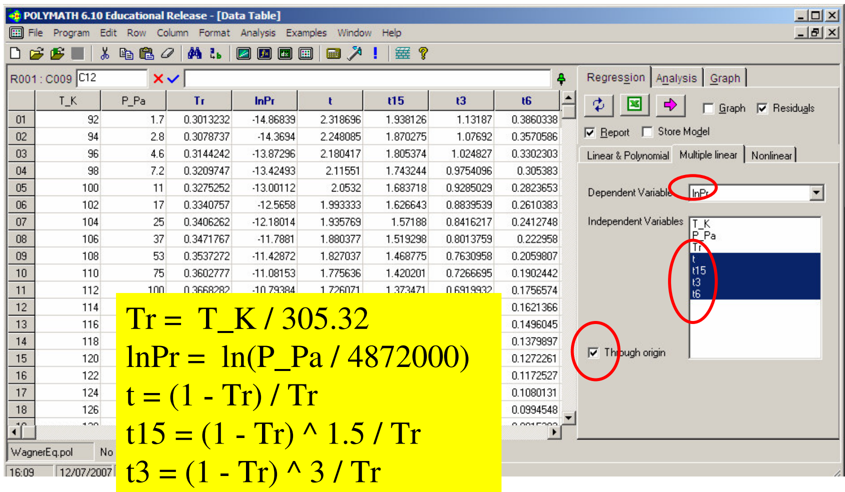
Modeling Vapor Pressure Data for Ethane with the Wagner Equation

$$\ln P_R = \frac{a\tau + b\tau^{1.5} + c\tau^3 + d\tau^6}{T_R}$$

Where $T_R = T/T_C$ is the reduced temperature $P_R = P/P_C$ is the reduced pressure and $\tau = 1 - T_R$.
For ethane T_C = 305.32 K, P_C =4.8720E+06 Pa
In order to obtain the model parameters using linear regression the following variables are defined:

Tr = T_K / 305.32
lnPr = ln(P_Pa / 4872000)
t = (1 - Tr) / Tr
t15 = (1 - Tr) ^ 1.5 / Tr
t3 = (1 - Tr) ^ 3 / Tr
t6 = (1 - Tr) ^ 6 / Tr

Modeling Vapor Pressure Data for Ethane with the Wagner Equation using Multiple Linear Regression



Tr = T_K / 305.32

lnPr = ln(P_Pa / 4872000)

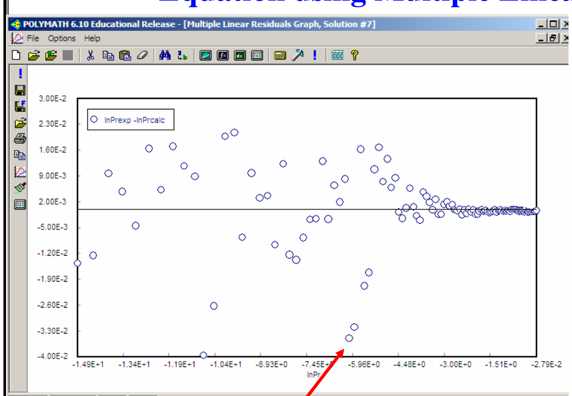
t = (1 - Tr) / Tr

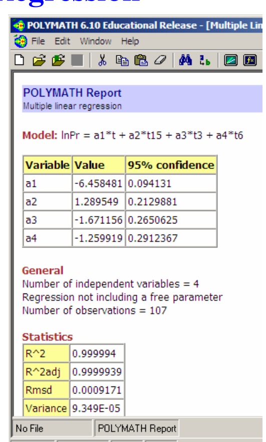
t15 = (1 - Tr) ^ 1.5 / Tr

t3 = (1 - Tr) ^ 3 / Tr

t6 = (1 - Tr) ^ 6 / Tr

Modeling Vapor Pressure Data for Ethane with the Wagner Equation using Multiple Linear Regression





POLYMATH Report

Multiple linear regression

Model: lnPr = a1*t + a2*t15 + a3*t3 + a4*t6

Variable	Value	95% confidence
a1	-6.458481	0.094131
a2	1.289549	0.2129881
a3	-1.671156	0.2650625
a4	-1.259919	0.2912367

General

Number of independent variables = 4

Regression not including a free parameter

Number of observations = 107

Statistics

R^2	0.999994
R^2adj	0.9999939
Rmsd	0.0009171
Variance	9.349E-05

Maximal Error in ln(Pr) ~ 0.6%

Note random residuals distribution in the entire data range