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)

Regression and Analysis of "Heat of Hardening" Data

No.	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.7
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

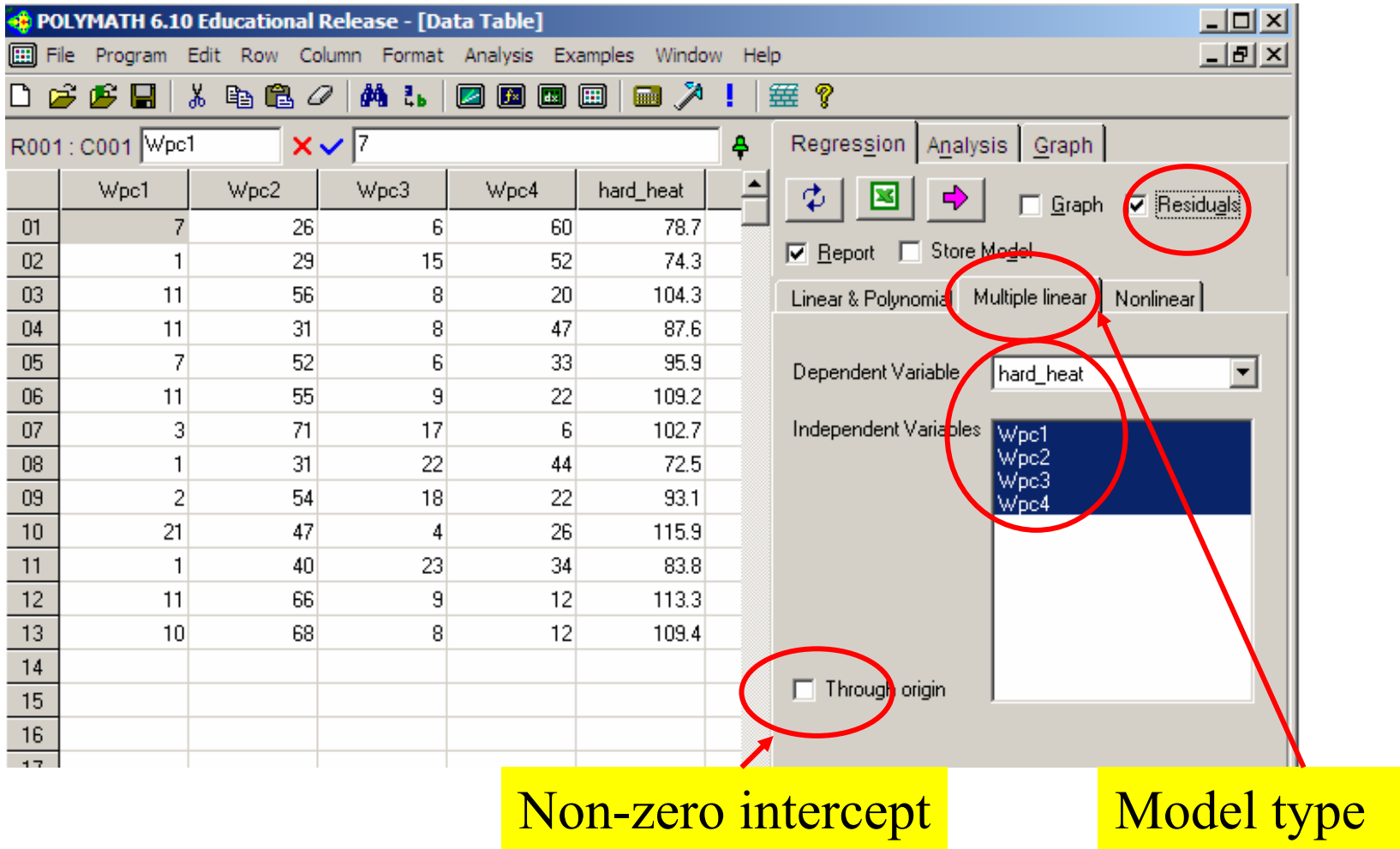
Woods *et al*(1932) investigated the integral heat of hardening of cement as a function of composition. The independent variables represent *weight percent* of the clinker compounds: x_1-tricalcium aluminate ($3CaO \cdot Al_2O_3$), x_2-tricalcium silicate ($3CaO \cdot SiO_2$), x_3-tetracalcium alumino-ferrite ($4CaO \cdot Al_2O_3 \cdot Fe_2O_3$), and x_4- β-dicalcium silicate ($3CaO \cdot SiO_2$). The dependent variable, y is the total heat evolved (in calories per gram cement) in a 180-day period.

Regression and Analysis of "Heat of Hardening" Data

No.	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.7
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Calculate the coefficients of a linear model representation of y as function of x_1, x_2, x_3, and x_4, calculate the variance and the correlation coefficient R^2 and the 95% confidence intervals.
Prepare a residual plot.
Consider the cases when the model includes and does not include a free parameter.

“Heat of Hardening” Data – Regression and Analysis by Polymath



"Heat of Hardening" Data –Analysis of the Linear Model that Includes a Free Parameter

File Edit Window Help

POLYMATH Report — Heat of hardening of portland cement
Multiple linear regression — 11-2007-יול

Model: hard_heat = a0 + a1*Wpc1 + a2*Wpc2 + a3*Wpc3 + a4*Wpc4

Variable	Value	95% confidence
a0	60.89893	161.6172
a1	1.562729	1.717796
a2	0.5265023	1.669402
a3	0.1125452	1.740721
a4	-0.1266181	1.635414

General
Number of independent variables = 4
Regression including a free parameter
Number of observations = 13

Statistics

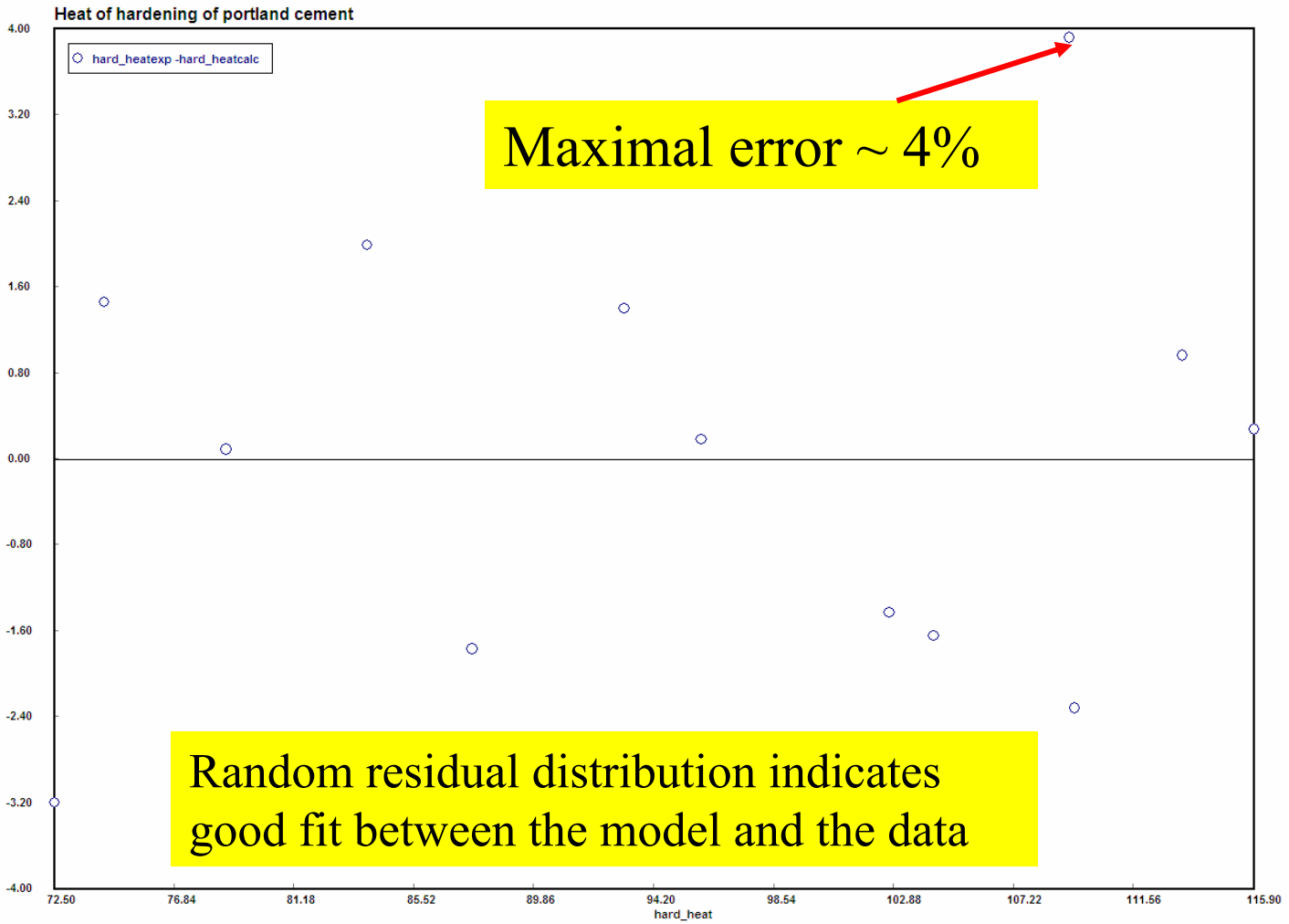
R^2	0.9823245
R^2adj	0.9734867
Rmsd	0.5322918
Variance	5.985442

Highly unstable model. All 95% confidence intervals larger in absolute value than the respective parameters.

R^2 value close to 1, indicates good fit between model and data. May be misleading occasionally.

Rmsd and variance values used for comparison between different models

"Heat of Hardening" Data – Residual Plot of the Linear Model that Includes a Free Parameter



"Heat of Hardening" Data – Demonstration of the Harmful Effect of the Instability

Model: hard_heat = a0 + a1*Wpc1 + a2*Wpc2 + a3*Wpc3 + a4*Wpc4

Variable	Value	95% confidence
a0	60.89893	161.6172
a1	1.562729	1.717796
a2	0.526503	1.669402
a3	0.112545	1.740721
a4	-0.12662	1.635414
Statistics		
R^2	0.982325	
Variance	5.985442	

Last data point removed

Removal of the last data point causes substantial change in all the parameter values.

In case of the parameter a_4 even its sign is changed

Variable	Value	95% confidence
a0	36.20362	170.1033
a1	1.783367	1.784379
a2	0.802752	1.770434
a3	0.328399	1.804248
a4	0.12209	1.720564
Statistics		
R^2	0.983897	
Variance	5.746289	

"Heat of Hardening" Data – A Stable Model is Obtained After Removing the Free Parameter

Model: hard_heat = a1*Wpc1 + a2*Wpc2 + a3*Wpc3 + a4*Wpc4

Variable	Value	95% confidence
a1	2.189177	0.4182687
a2	1.154136	0.1082325
a3	0.753295	0.3601112
a4	0.488545	0.093483
Statistics		
R^2	0.980656	
Variance	5.822523	

Last data point removed

Variable	Value	95% confidence
a1	2.151451	0.4078854
a2	1.17869	0.1115862
a3	0.703614	0.3563326
a4	0.487699	0.0901595
Statistics		
R^2	0.983314	
Variance	5.209989	

Correlation of Heat Capacity Data for Ethane

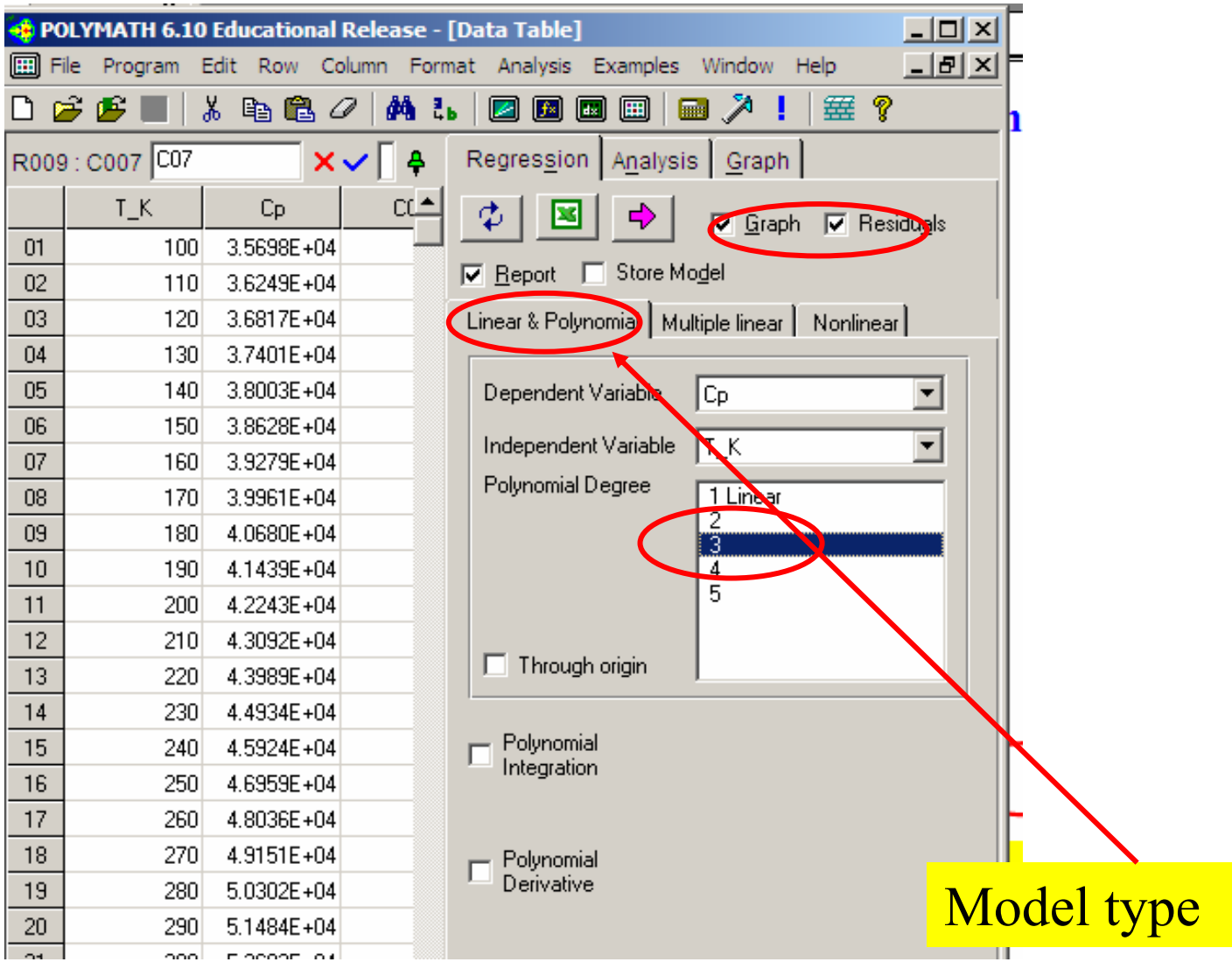
A polynomial has to be fitted to heat capacity data provided by Ingham et al*. This data set includes 41 data points in the temperature range of 100 K – 400 K.
The degree of the polynomial:

$$C_p = a_0 + a_1 T + a_2 T^2 + a_3 T^3 + \ldots$$

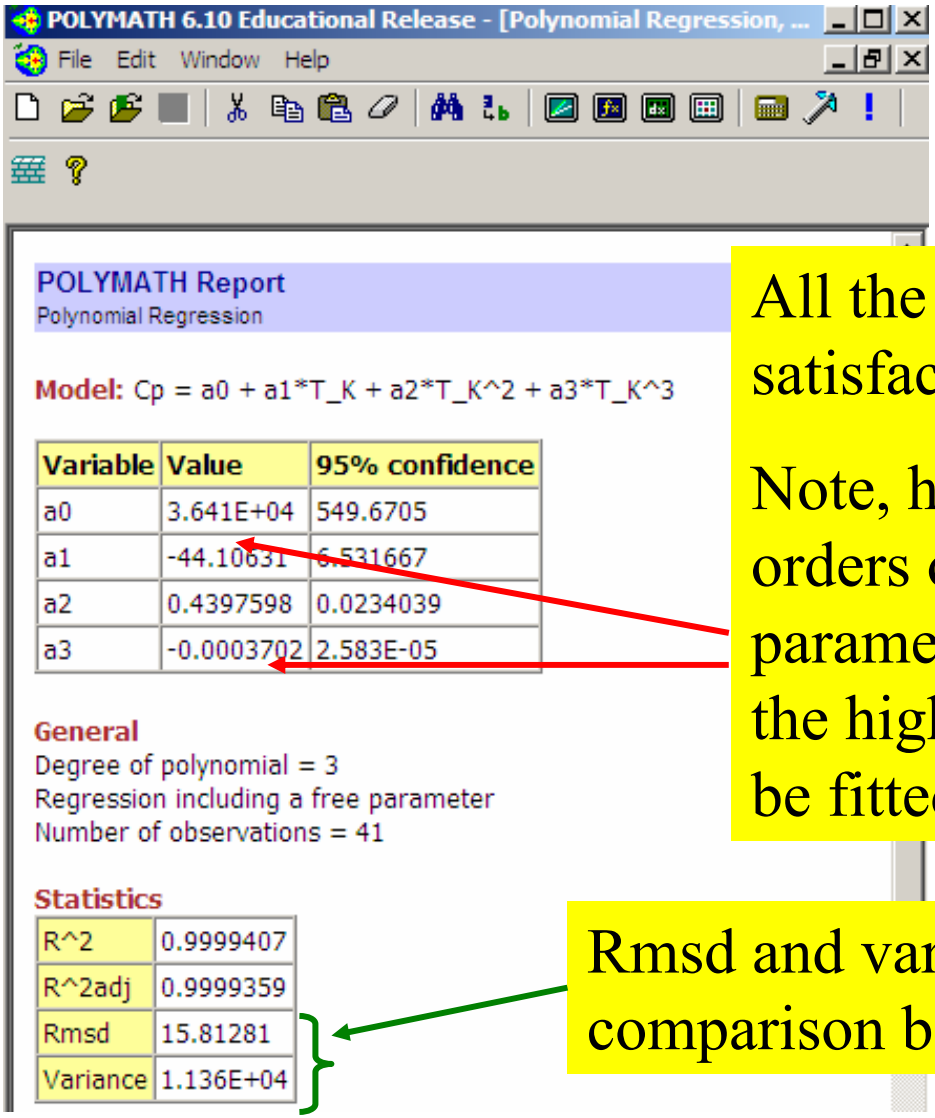
where C_p is the heat capacity in J/kg-mol·K, T is the temperature in K, and a_0, a_1,... are the regression model parameters. The set of parameters which best represents the data, has to be found.
The goodness of fit should be determined based on the variance, the correlation coefficient (R^2), the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

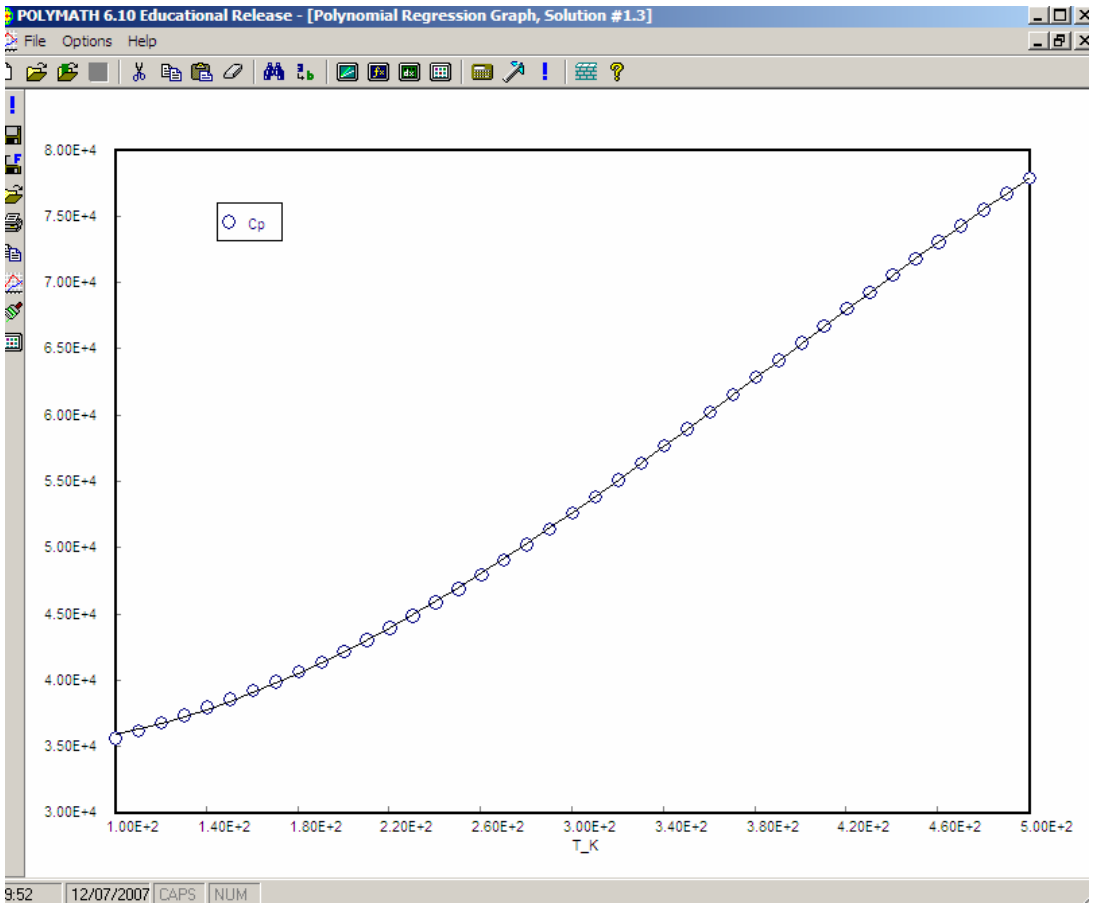


All the parameters indicate satisfactory model.

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

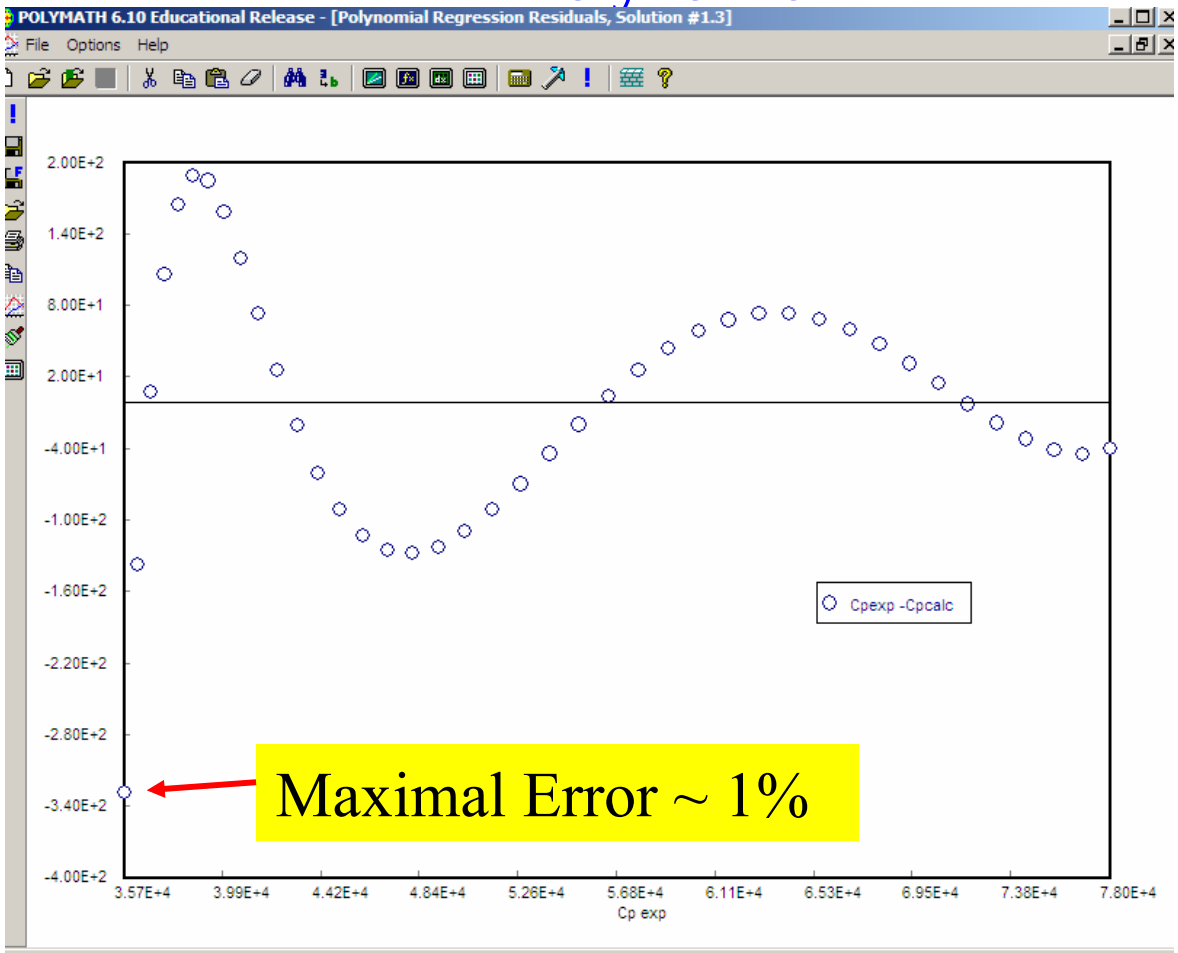
Rmsd and variance values used for comparison between different models

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]

File Program Edit Row Column Format Analysis Examples Win

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]

File Edit Window Help

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

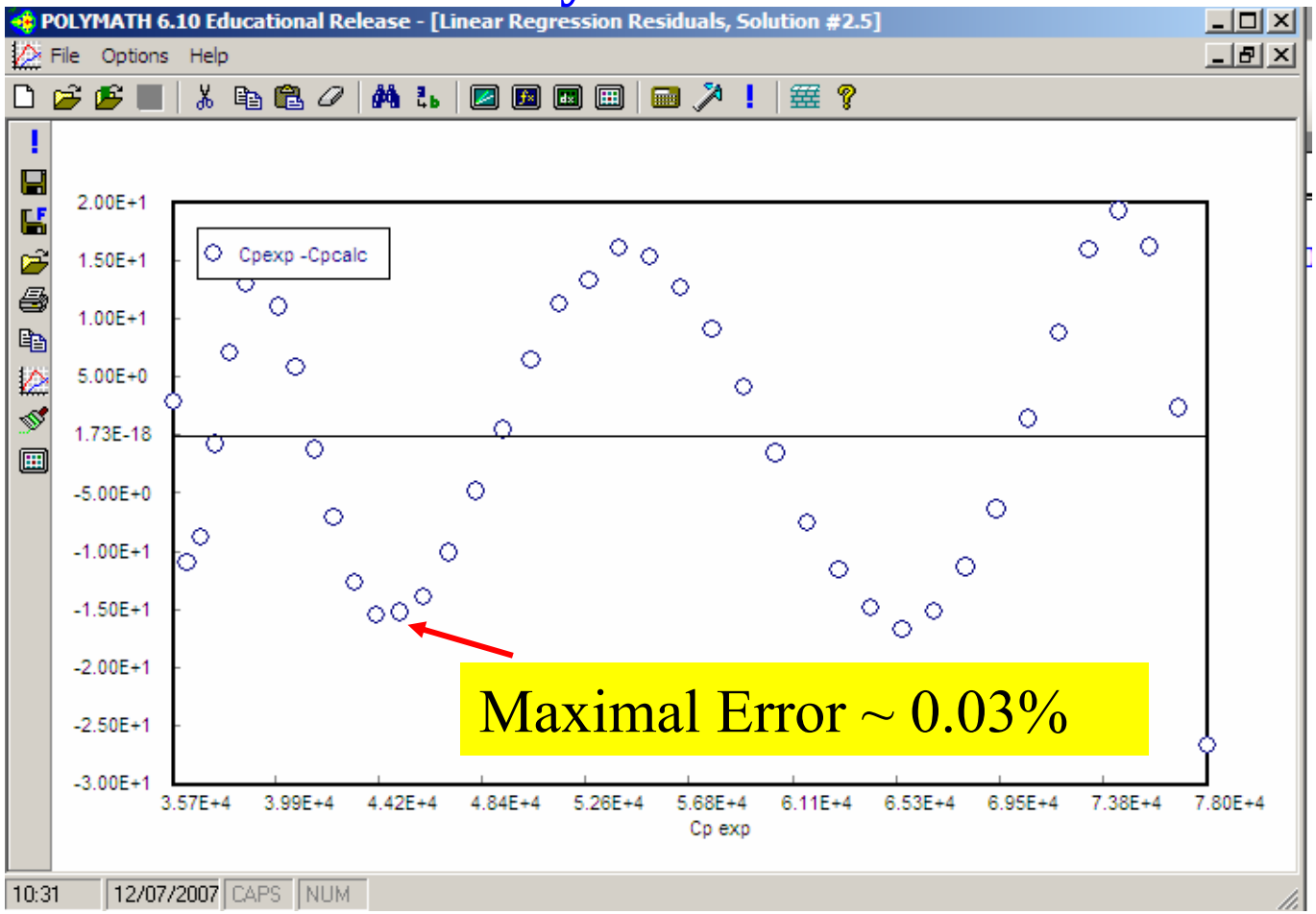
General
Degree of polynomial = 5
Regression including a free parameter
Number of observations = 41

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

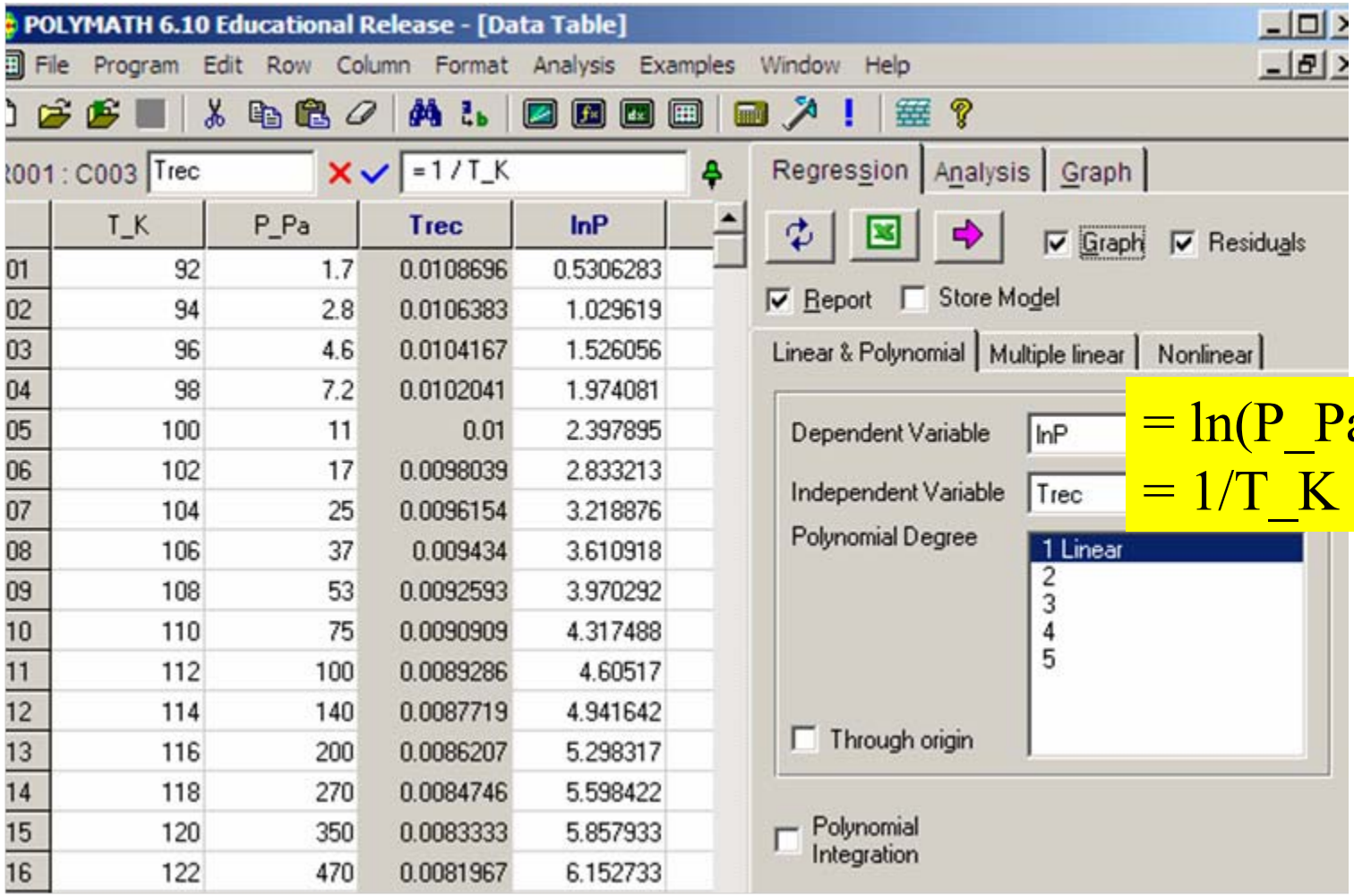
The Clapeyron equation is a two parameter equation:

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

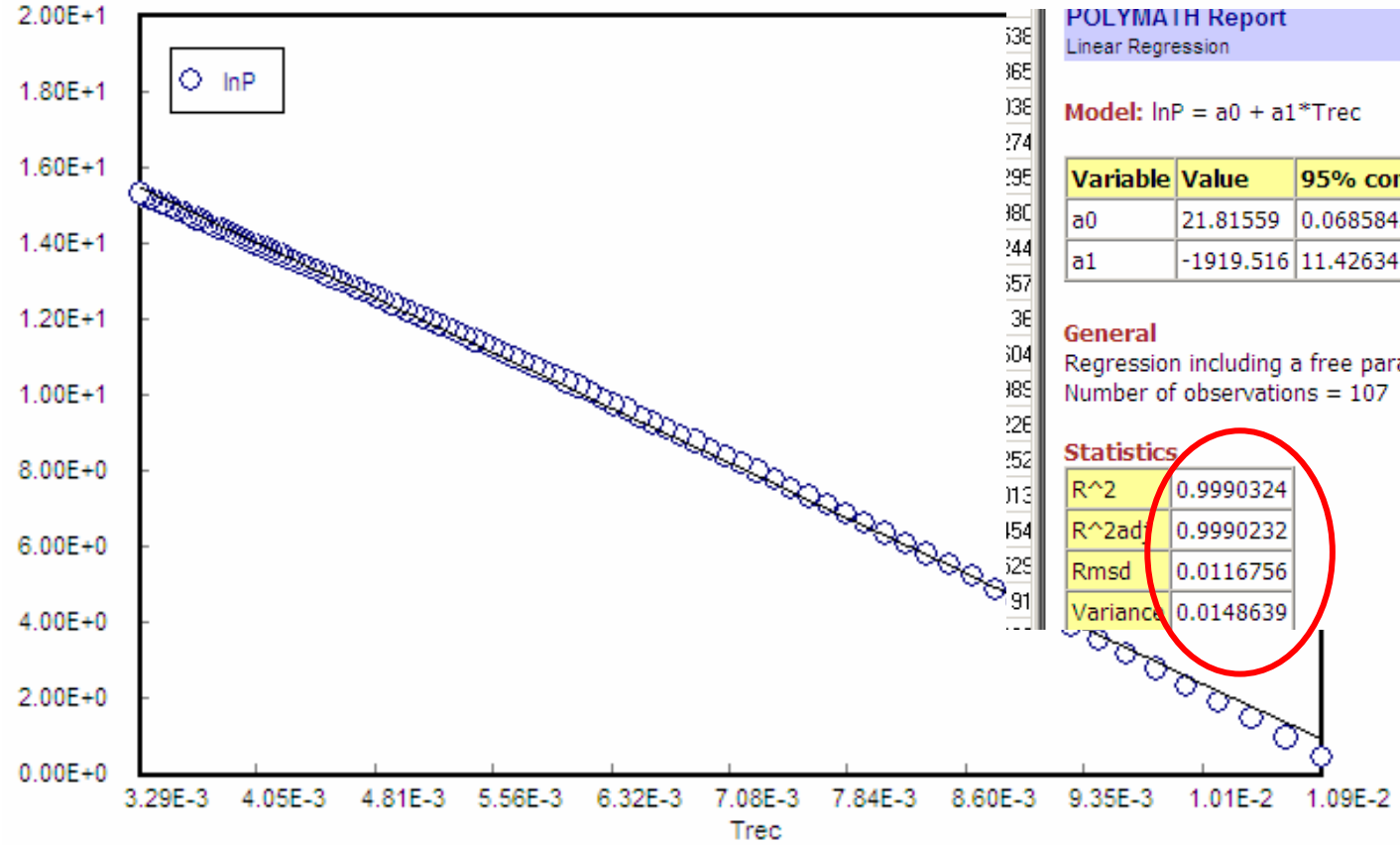
where P is the vapor pressure (Pa), T – temperature (K) , A and B are parameters

*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



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



POLYMATH Report
Linear Regression

Model: lnP = a0 + a1*Trec

Variable	Value	95% confidence
a0	21.81559	0.0685843
a1	-1919.516	11.42634

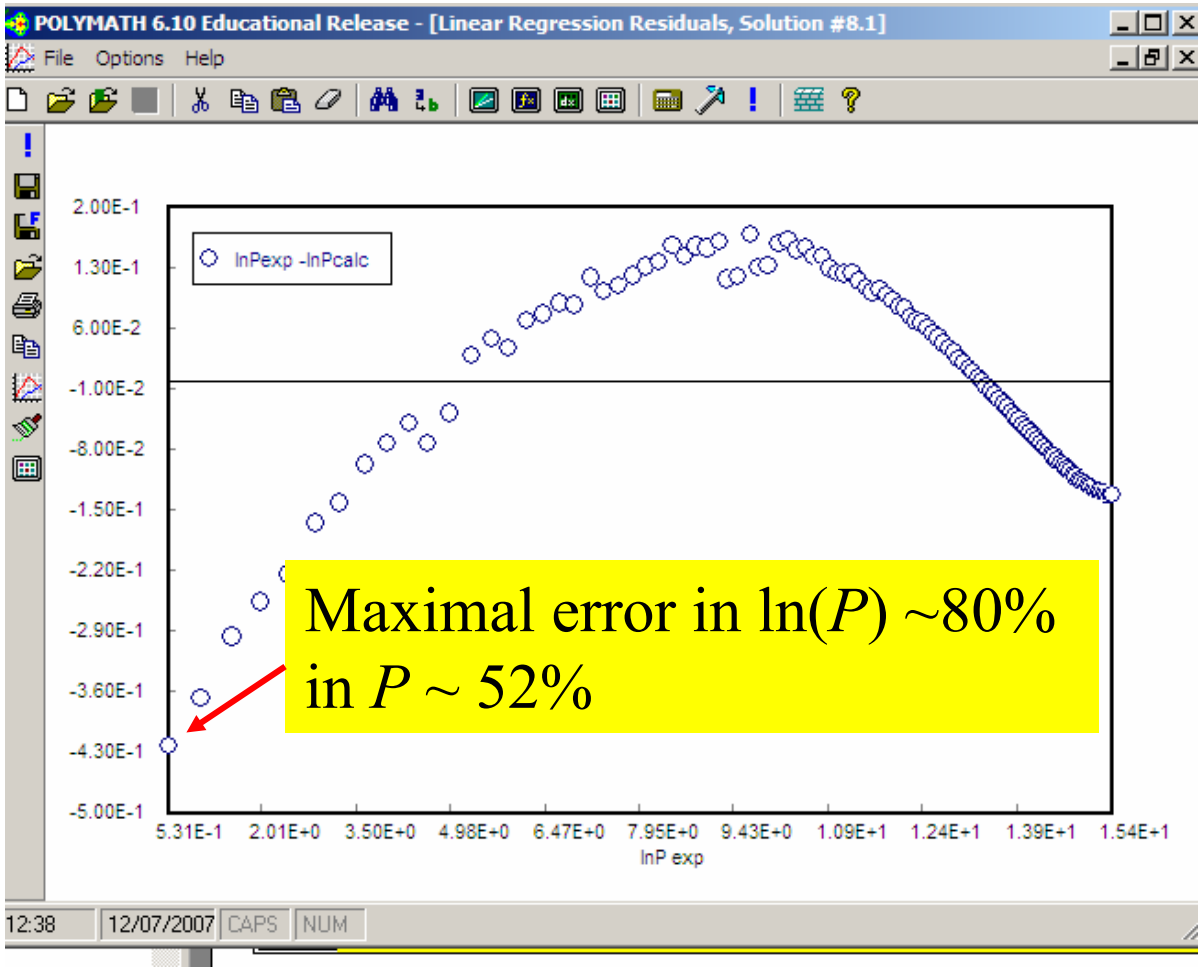
General
Regression including a free parameter
Number of observations = 107

Statistics

R^2	0.9990324
R^2adj	0.9990232
Rmsd	0.0116756
Variance	0.0148639

All the indicators show good, acceptable fit!

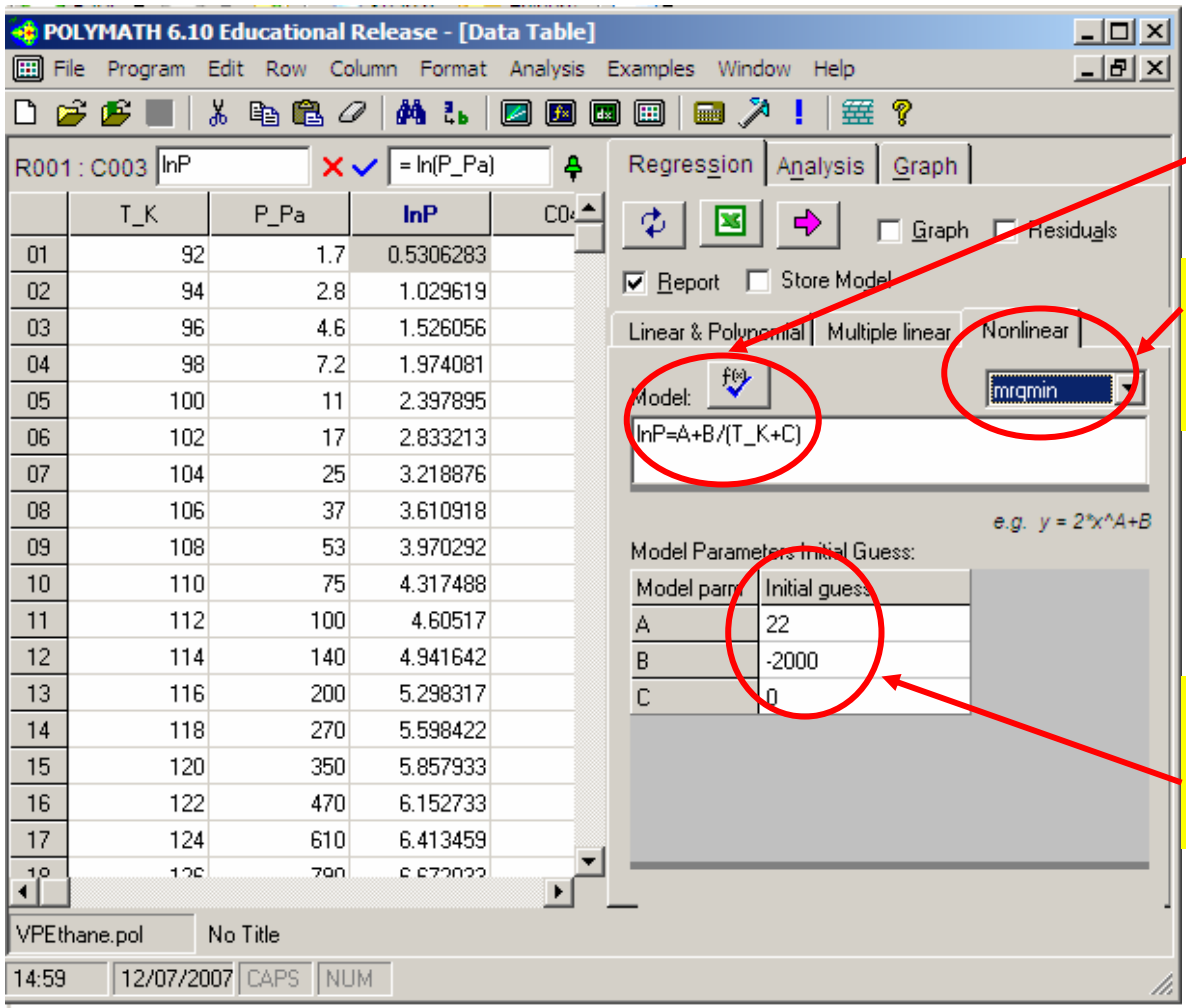
Modeling Vapor Pressure Data for Ethane by the Clapeyron Equation using Linear Regression



The residual plot reveals large unexplained curvature in the data

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

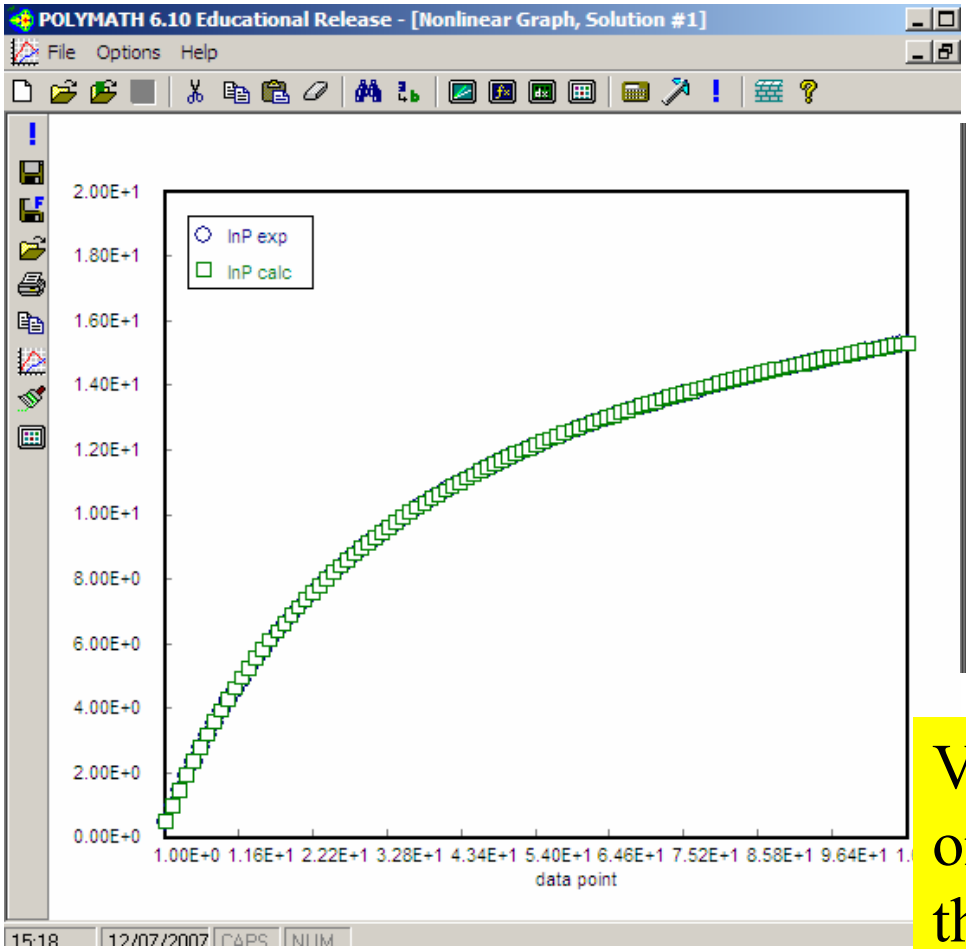
$$\ln P = A + \frac{B}{T + C}$$



Model type and solution algorithm

Initial guess from Clapeyron eqn.

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



Model: lnP = A+B/(T_K+C)

Variable	Initial guess	Value	95% confidence
A	22.	20.79896	0.0208029
B	-2000.	-1583.811	6.234624
C	0	-13.88808	0.2627736

Nonlinear regression settings
Max # iterations = 64
Tolerance = .0001

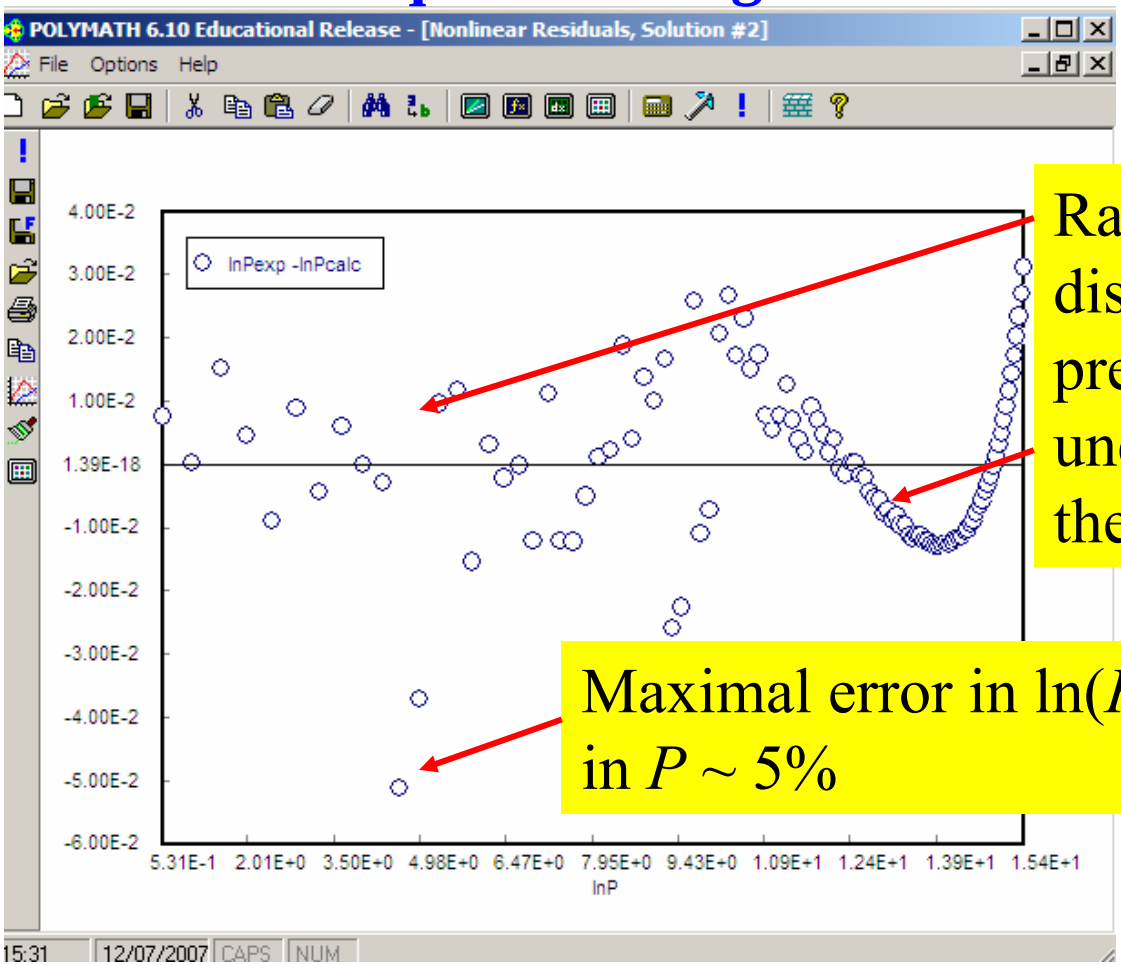
Precision

R^2	0.9999885
R^2adj	0.9999883
Rmsd	0.0012724
Variance	0.0001782
Chi-Sq	1.853625

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

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

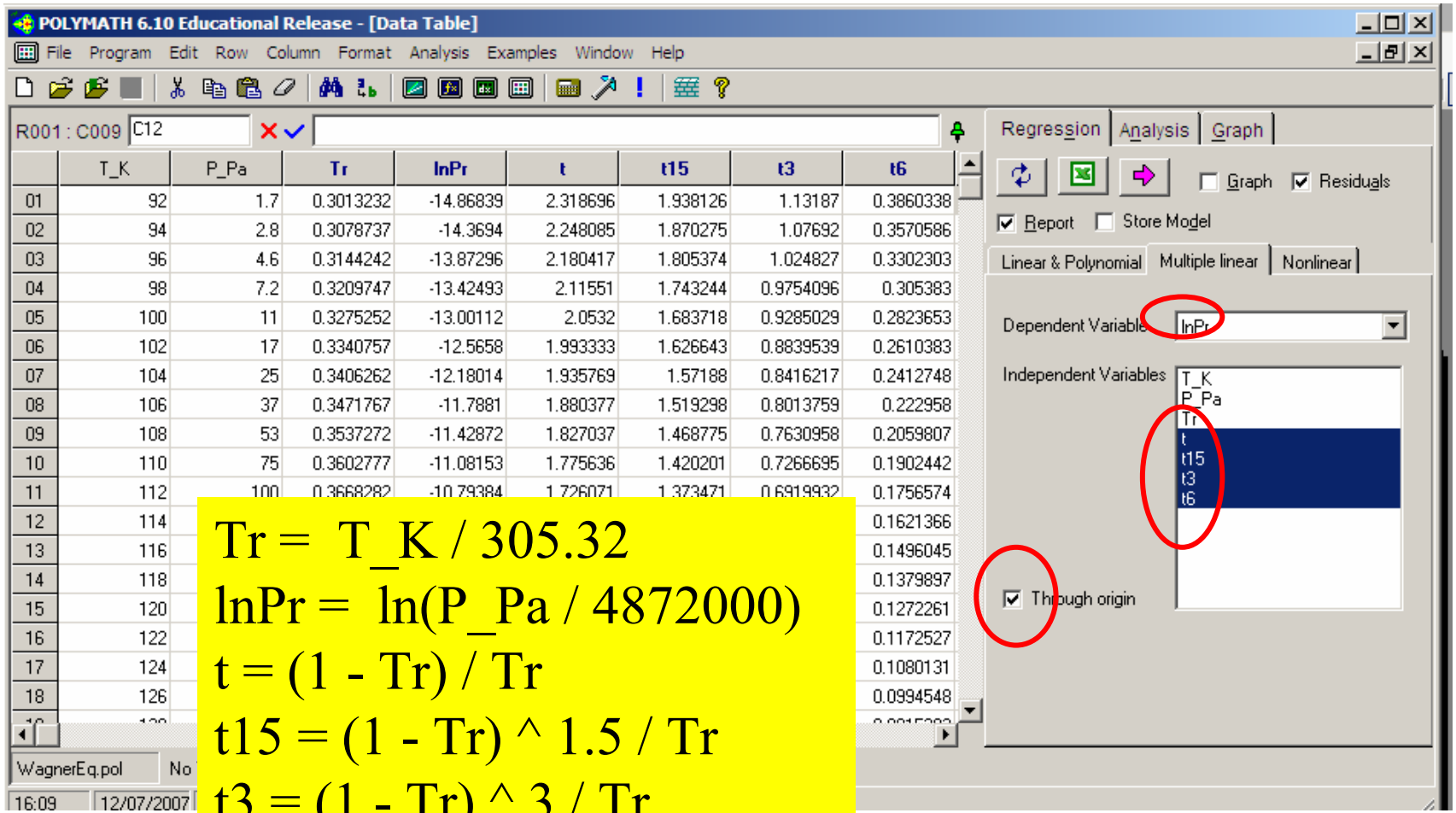
Where $T_R = T/T_C$ is the reduced temperature $P_R = P/P_C$ is the reduced pressure and $\tau = 1 - T_R$.

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)

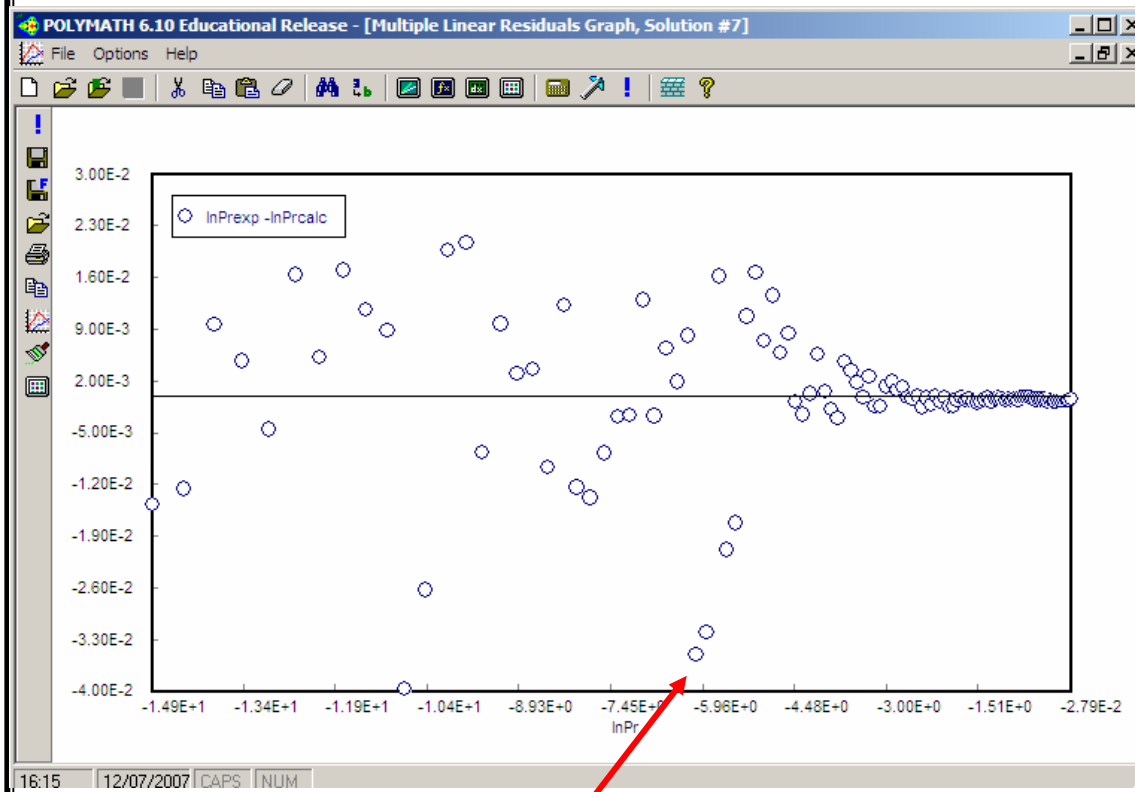
t = (1 - Tr) / Tr

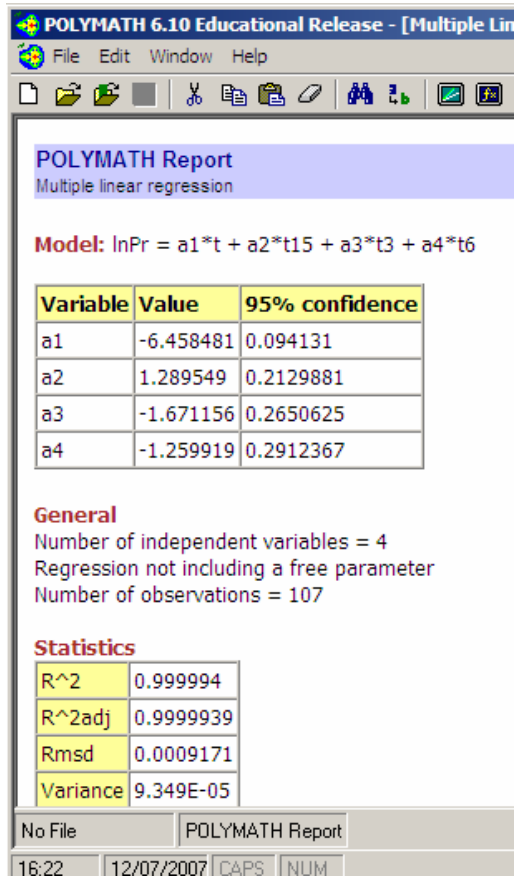
t15 = (1 - Tr) ^ 1.5 / Tr

t3 = (1 - Tr) ^ 3 / Tr

t6 = (1 - Tr) ^ 6 / Tr

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





Maximal Error in ln(Pr) ~ 0.6%

Note random residuals distribution in the entire data range