

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 $x_{1i}, x_{2i}, \dots, x_{ni}$, multiple linear regression attempts to find the “best” values of the parameters a_0, a_1, \dots, a_n for the equation

$$\hat{y}_i = a_0 + a_1x_{1,i} + a_2x_{2,i} + \dots + a_nx_{n,i}$$

\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

$$\hat{y}_i = a_0 + a_1x_i + a_2x_i^2 + \dots + a_nx_i^n$$

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:

$$C_p = a_0 + a_1T + a_2T^2 + \dots + a_nT^n$$

where C_p is the heat capacity in J/kg-mol·K, T is the temperature in K, and a_0, a_1, \dots 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

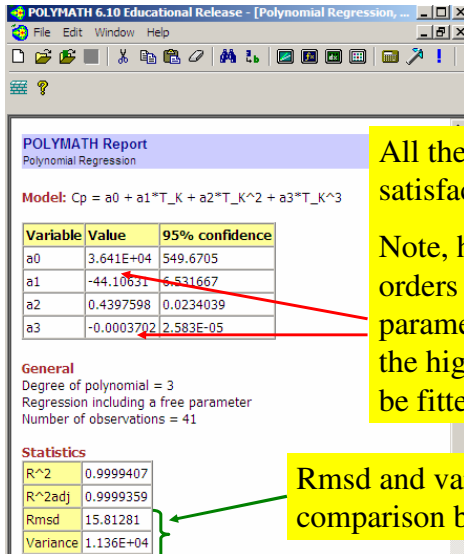
The screenshot shows the POLYMATH 6.10 Educational Release interface. On the left, a data table is displayed with columns for T_K and Cp. The regression settings on the right are as follows:

- Regression: Report, Store Model
- Analysis: Graph, Residuals
- Model type: Linear & Polynomial, Multiple linear, Nonlinear
- Dependent Variable: Cp
- Independent Variable: T_K
- Polynomial Degree: 3 (selected)
- Through origin
- Polynomial Integration
- Polynomial Derivative

A red arrow points from the text "Model type" in a yellow box to the "Linear & Polynomial" radio button. Another red circle highlights the "3" in the Polynomial Degree list.

	T_K	Cp
01	100	3.5698E+04
02	110	3.6249E+04
03	120	3.6817E+04
04	130	3.7401E+04
05	140	3.8003E+04
06	150	3.8628E+04
07	160	3.9279E+04
08	170	3.9961E+04
09	180	4.0680E+04
10	190	4.1439E+04
11	200	4.2243E+04
12	210	4.3092E+04
13	220	4.3989E+04
14	230	4.4934E+04
15	240	4.5924E+04
16	250	4.6959E+04
17	260	4.8036E+04
18	270	4.9151E+04
19	280	5.0302E+04
20	290	5.1484E+04

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

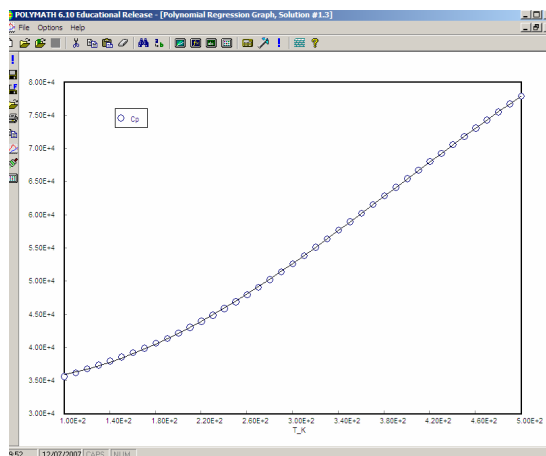


All the parameters indicate satisfactory model.

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

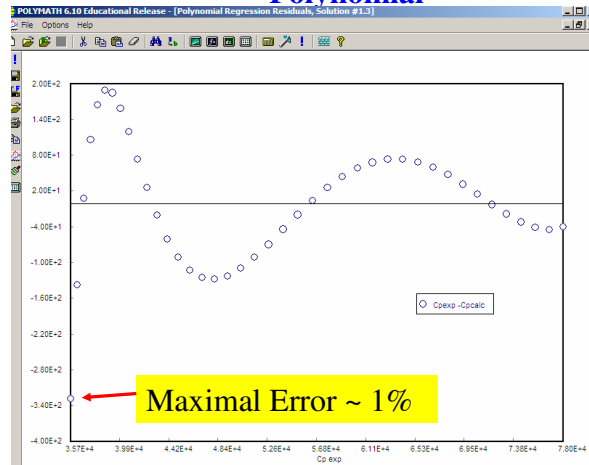
Rmsd and variance values used for comparison between different models

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]

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

	T_K	Cp	Tstd	C04	C05
01	100	3.5690E+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	-0.918262		
11	200	4.2243E+04	-0.834783		
12	210	4.3092E+04	-0.751305		
13	220	4.3989E+04	-0.667827		
14	230	4.4934E+04	-0.584348		

Column Statistics

Column name : T_K
 Number of rows : 41
 Sum : 1.235E+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: $C_p = a_0 + a_1 \cdot T_{std} + a_2 \cdot T_{std}^2 + a_3 \cdot T_{std}^3 + a_4 \cdot T_{std}^4 + a_5 \cdot T_{std}^5$

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

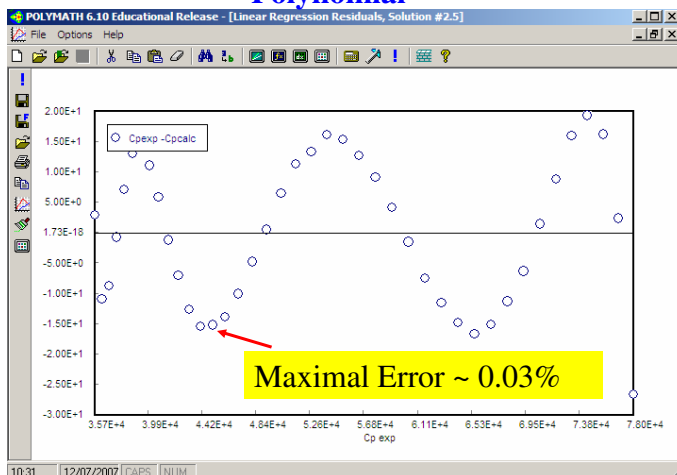
Degree of polynomial = 5
 Regression including a free parameter
 Number of observations = 41

Statistics

R ²	0.9999992
R ² adj	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 triple 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:

$$\ln P = A + \frac{B}{T} \quad \text{where } P \text{ is the vapor pressure (Pa), } T \text{ – temperature (K), } A \text{ and } B \text{ are parameters}$$

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

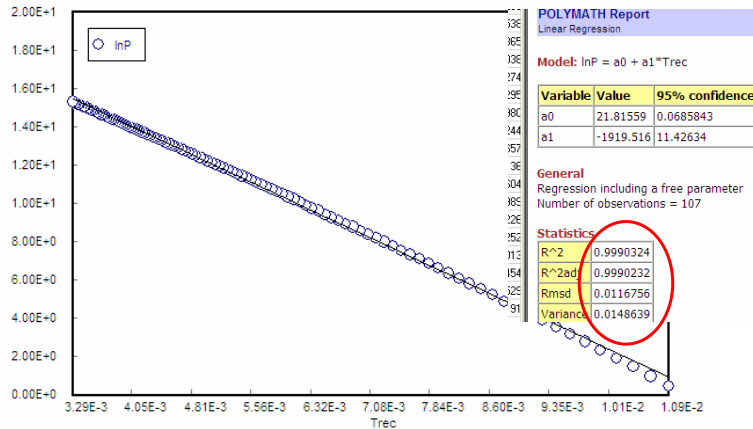
Modeling Vapor Pressure Data for Ethane by the Clapeyron Equation using Linear Regression

The screenshot shows the POLYMATH 6.10 Educational Release interface. The main window displays a data table with the following columns: T_K, P_Pa, Trec, and lnP. The regression settings are configured as follows:

- Dependent Variable: lnP = ln(P_Pa)
- Independent Variable: Trec = 1/T_K
- Polynomial Degree: 1 Linear
- Through origin:
- Polynomial Integration:

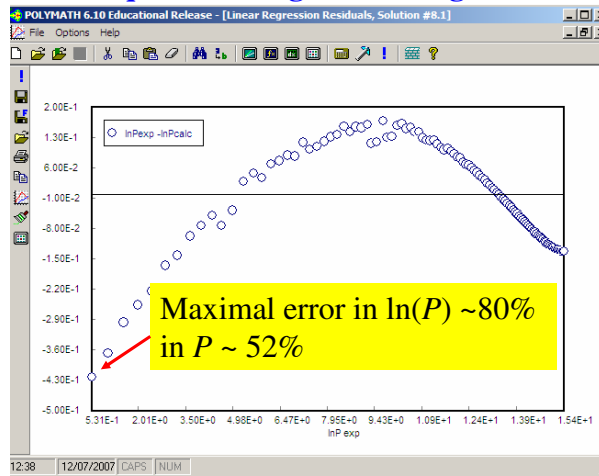
	T_K	P_Pa	Trec	lnP
01	92	1.7	0.0108696	0.5306283
02	94	2.8	0.0106383	1.029619
03	96	4.6	0.0104167	1.526056
04	98	7.2	0.0102041	1.974081
05	100	11	0.01	2.397895
06	102	17	0.0098039	2.833213
07	104	25	0.0096154	3.218876
08	106	37	0.009434	3.610918
09	108	53	0.0092593	3.970292
10	110	75	0.0090909	4.317488
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



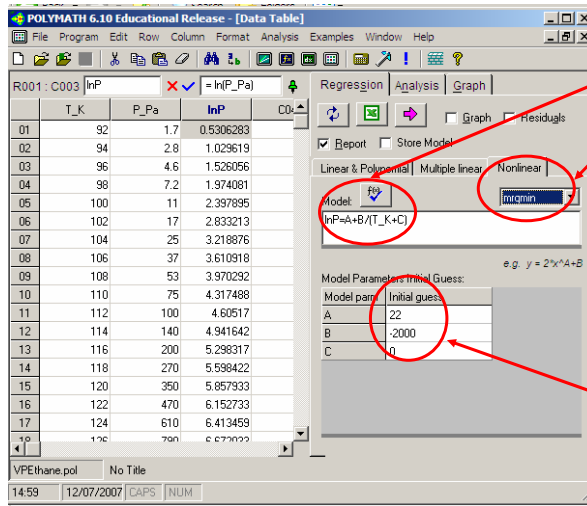
All the indicators show good, acceptable fit!

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



The residual plot reveals large unexplained curvature in the data

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

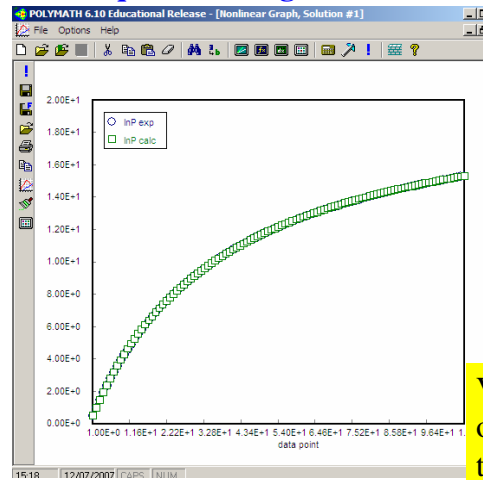


$$\ln P = A + \frac{B}{T + C}$$

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: $\ln P = A + B / (T + C)$

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

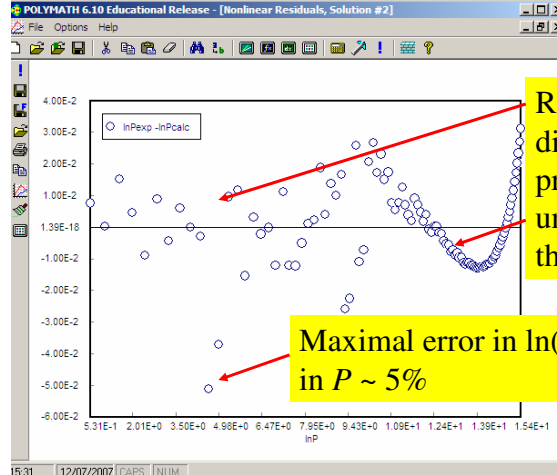
Precision

R^2	0.9999885
R^2adj	0.9999883
Rmsd	0.0012724
Variance	0.0001782
Chi-Sq	1.453625

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 $\ln(P) \sim 1\%$
in $P \sim 5\%$

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:

$$\begin{aligned} Tr &= T_K / 305.32 \\ \ln Pr &= \ln(P_Pa / 4872000) \\ t &= (1 - Tr) / Tr \\ t15 &= (1 - Tr)^{1.5} / Tr \\ t3 &= (1 - Tr)^3 / Tr \\ t6 &= (1 - Tr)^6 / Tr \end{aligned}$$

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

The screenshot shows the POLYMATH 6.10 Educational Release interface. The main window displays a data table with columns: T_K, P_Pa, Tr, lnPr, t, t15, t3, t6. The regression settings on the right are configured for a multiple linear regression with the dependent variable 'lnPr' and independent variables 'T_K', 'P_Pa', 'Tr', 't', 't15', 't3', and 't6'. The 'Through origin' checkbox is checked.

	T_K	P_Pa	Tr	lnPr	t	t15	t3	t6
01	92	1.7	0.3013232	-14.86839	2.318636	1.938126	1.13187	0.3860338
02	94	2.8	0.3078737	-14.3694	2.248085	1.870275	1.07652	0.3570986
03	96	4.6	0.3144242	-13.87296	2.180417	1.805374	1.024827	0.3302303
04	98	7.2	0.3209747	-13.42493	2.11551	1.743244	0.9754096	0.305383
05	100	11	0.3275252	-13.00112	2.0532	1.683718	0.9285029	0.2823653
06	102	17	0.3340757	-12.5658	1.993333	1.626643	0.8839539	0.2610383
07	104	25	0.3406262	-12.18014	1.935769	1.57198	0.8416217	0.2412748
08	106	37	0.3471767	-11.7881	1.880377	1.519298	0.8013759	0.222958
09	108	53	0.3537272	-11.42872	1.827037	1.468755	0.7630958	0.2059807
10	110	75	0.3602777	-11.08153	1.775636	1.420201	0.7266895	0.1902442
11	112	100	0.3668282	-10.79384	1.736701	1.373471	0.6919932	0.1756574
12	114							0.1621365
13	116							0.1498045
14	118							0.1379897
15	120							0.1272261
16	122							0.1172527
17	124							0.1080131
18	126							0.0994548

$$Tr = T_K / 305.32$$

$$\ln Pr = \ln(P_Pa / 4872000)$$

$$t = (1 - Tr) / Tr$$

$$t15 = (1 - Tr)^{1.5} / Tr$$

$$t3 = (1 - Tr)^3 / Tr$$

$$t6 = (1 - Tr)^6 / Tr$$

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

The screenshot shows the POLYMATH 6.10 Educational Release interface with a residuals graph and a report. The residuals graph plots lnPr - lnPrCalc against lnPr, showing a random distribution of points around zero. The report on the right provides the model equation and statistical results.

POLYMATH Report
Multiple linear regression

Model: $\ln Pr = a1*t + a2*t15 + a3*t3 + a4*t6$

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 ² adj	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