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

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

$\hat{y}_i$ 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 \( a_0, a_1, \ldots \) 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.

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

All the parameters indicate a 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

Maximal Error ~ 1%

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
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:

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

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


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

$$\ln(P_{\text{Pa}}) = \frac{1}{T_{\text{K}}}$$
Modeling Vapor Pressure Data for Ethane by the Clapeyron Equation using Linear Regression

All the indicators show good, acceptable fit!

Maximal error in $\ln(P) \sim 80\%$ in $P \sim 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.

- Experimental and calculated values cannot be distinguished.

- Variance smaller by 2 orders of magnitude than Clapeyron
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 $T_C = 305.32 \text{ K}$, $P_C = 4.8720\times10^6 \text{ Pa}$

In order to obtain the model parameters using linear regression the following variables are defined:

- $\text{Tr} = T_K / 305.32$
- $\lnPr = \ln(P_{\text{Pa}} / 4872000)$
- $t = (1 - \text{Tr}) / \text{Tr}$
- $t15 = (1 - \text{Tr})^{1.5} / \text{Tr}$
- $t3 = (1 - \text{Tr})^3 / \text{Tr}$
- $t6 = (1 - \text{Tr})^6 / \text{Tr}$
Modeling Vapor Pressure Data for Ethane with the Wagner Equation using Multiple Linear Regression

\[ \text{Tr} = \frac{T_K}{305.32} \]

\[ \ln(\text{Pr}) = \ln(\text{P}_{\text{Pa}} / 4872000) \]

\[ t = \frac{1 - \text{Tr}}{\text{Tr}} \]

\[ t_{15} = \left(1 - \text{Tr}\right)^{1.5} / \text{Tr} \]

\[ t_3 = \left(1 - \text{Tr}\right)^3 / \text{Tr} \]

\[ t_6 = \left(1 - \text{Tr}\right)^6 / \text{Tr} \]

Maximal Error in \( \ln(\text{Pr}) \) ~ 0.6%

Note random residuals distribution in the entire data range