Considering Numerical Error Propagation in Modeling and Regression of Data

Neima Brauner, Mordechai Shacham
Tel-Aviv University/Ben-Gurion University of the Negev

Abstract

The use of user-friendly interactive regression software enables undergraduate engineering students to reach a high level of sophistication in regression, correlation and analysis of data. In order to interpret correctly the results, the students must be familiar with potential causes for poor fits in correlations, should be able to recognize a poor correlation and improve it if possible. They should also be aware of the practical consequences of using a correlation which has no statistical validity.

In this paper, the harmful effects of numerical error propagation (resulting from collinearity among the independent variables) are explained and demonstrated. Simple methods for minimizing such error propagation in polynomial regression are introduced. This material can be presented, for example, as part of 3rd year undergraduate mathematical modeling and numerical methods course.

Introduction

Realistic modeling and accurate correlation of experimental data are essential to sound engineering design. Many of the statistical techniques for analyzing the accuracy of the correlations have been known for several decades (see, for example, Draper and Smith, 1981, Himmelblau, 1970, Bates and Watts, 1988 and Noggle, 1993). But, until recently, those techniques have not been utilized in a significant level in undergraduate engineering education. One of the main reasons for not utilizing those techniques was that statistical tests usually yield numbers (variance, standard deviation, correlation coefficient, etc.). The meaning of these numbers can be easily misinterpreted if the statistical theory and the assumptions made in developing the tests are not well understood.

The emergence of software packages with interactive regression and statistical analysis capabilities (such as POLYMATH, MATLAB, MATHEMATICA, EXCEL) which provides both numerical and graphical output changes the situation. These software packages enable undergraduate engineering students, with moderate statistical background, to carry out rigorous regression and statistical analysis of data. They are able to select the most appropriate correlation model and test its statistical validity using residual and confidence region plots. They can analyze the quality and precision of the laboratory data by plotting one independent variable versus the others to detect hidden collinearity that may exist among the variables.

Shacham et al (1996) had described a set of lectures and exercises that is used to introduce freshman engineering students to the basics of data modeling and analysis using interactive software packages. This material is included in an introductory computing course or as part of an introductory engineering course. The introduction to data modeling and analysis (described by Shacham et al, 1996) includes the following subjects:

1. Basic statistical concepts.
2. Discrimination between real experimental data and smoothed interpolated data.
3. Using residual plots and confidence intervals for selecting the most appropriate model.
4. The dangers of extrapolation, in particular, when a non-theory based model is used.

This introductory material is very helpful to the students for modeling and analyzing their own data. However, they may need more advanced material when dealing with, for example, models containing large numbers of parameters.

In this paper, more advanced material related to regression is presented. The discussion includes models which are comprised of a sum of functions of the same independent variable (as in polynomial regression). Various effects of the interdependency between these functions are described and demonstrated.

The material presented is taught to third year, undergraduate Chemical Engineering students at the Ben Gurion University as part of a mathematical modeling and numerical methods' course.

The calculations involved in solving the examples presented have been carried out using the POLYMATH 4.0 (Shacham and Cutlip, 1996) and MATLAB (MathWorks, 1992) packages, but other similar packages can be used for this purpose.

Linear Regression with Models Comprising of Functions of One Independent Variable

Let us assume that there is a set of N data points of a dependent variable (measured variable, such as vapor pressure, viscosity, heat capacity, etc.), y_i versus an independent variable (controlled variable, such as temperature,. concentration, pressure) x_i, $i = 1, 2, \cdots N$. A regression model comprised of a linear combination of n different functions of the independent variable is considered. Thus, the regressors are $x_1 = f_1(x)$, $x_2 = f_2(x) \ldots x_n = f_n(x)$. For instance, in polynomials $x_1 = x^0$, $x_2 = x, \cdots x_n = x^{n-1}$.

A linear model fitted to the data is of the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \cdots + \beta_n x_{ni} + \in_i \qquad (1)$$

where $\beta_0, \beta_1, \ldots \beta_n$ are the parameters of the model and $\in_i$ is a measurement error in y_i. It is assumed that $\in_i$ is independently and identically (i.i.d.) distributed. The vector of estimated parameters $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_0, \hat{\beta}_1 \mathrm{K}\ \hat{\beta}_n)$ is usually calculated using the least squares error approach, by minimizing the following function:

$$S^2 = \sum_{i=1}^{N} \left[y_i - \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \Lambda + \beta_n x_{ni} \right) \right]^2 \qquad (2)$$

If the parameters appear in a linear expression (as in eq. (1)), the minimization can be carried out by solving a set of simultaneous linear algebraic equations (the normal equation):

$$\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y} \qquad (3)$$

The columns of $\mathbf{X}$ are: $\mathbf{x}_0 = \mathbf{1}$, $\mathbf{x}_1$, $\mathbf{x}_2 \cdots \mathbf{x}_n$ and $\mathbf{X}^T \mathbf{X} = \mathbf{A}$ is the normal matrix.

To check the goodness of the fit between the observed y_i and estimated $\hat{y}_i$ values of the dependent variable, they can be plotted versus x_i (when there is a single independent variable) or versus i, the point number (when there are several independent variables). The distance between the observed and estimated values can serve as an indication for the quality of the fit. These distances can be amplified using a "residual plot". In the residual plot, the model error (residual) $\hat{\varepsilon}_i$ is plotted usually versus y_i, where:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \qquad (4)$$

A random distribution of the residuals around zero indicates that the model correctly represent the particular set of data. A definite trend or pattern in the residual plot may indicate either a lack of fit of the model, or that the assumption of random error distribution for the dependent variable is incorrect. In some cases (for example, in cases where the value of the dependent variable changes by several orders of magnitude over the range of interest) the relative error is distributed normally. The relative error is defined as:

$$\hat{\varepsilon}_{ir} = \frac{\hat{\varepsilon}_i}{y_i} \qquad (5)$$

The appropriate transformation, which results in minimization of the relative error in a regression, is taking the logarithm of both sides of the model equation. It should be emphasized, however, that when the variables are transformed, the residual plot must be constructed using the transformed form of the dependent variable, in order to account for the change in the error distribution introduced by the transformation.

A numerical indicator for the quality of the fit which is used most frequently is the square of standard error of the estimate, which represents the sample variance, and is given by:

$$s^2 = \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2}{N - n - 1} \qquad (6)$$

Thus, the sample variance is the sum of squares of errors divided by the degrees of freedom (where the number of parameters, n+1, is subtracted from the number of data points, N) and is a measure for the variability of the actual y values from the predicted $\hat{y}$ values. Smaller variance indicates a better fit of the model to the data.

Confidence intervals (in particular the 95% confidence interval) on the parameter values can be very useful indicators of the fit between the model and the data. A better fit and more precise data lead to narrow confidence intervals, while a poor model fit and/or imprecise data result in wide confidence intervals. Furthermore, confidence intervals which are larger (in absolute value) than the respective parameters themoften indicate that the model contains too many parameters. Confidence interval is defined by:

$$\hat{\beta}_j - t(\nu, \alpha) s\sqrt{a_{jj}} \le \beta_j < \hat{\beta}_j + t(\nu, \alpha) s\sqrt{a_{jj}} \tag{7}$$

where t (ν, α) is the statistical t distribution corresponding to ν degrees of freedom ($\nu = N-(n+1)$) and a desired confidence level, α and s is the standard error of the estimate. If the number of data points is large enough, the value of t approaches a constant value (for $\nu > 15$, $t \sim 2$ for the 95% confidence interval). Therefore, t (ν, α) is often omitted. The term $s\sqrt{a_{jj}}$ is called the standard error of the estimate of parameter β.

One of the assumptions of the least squares error approach is that there is no error in the independent variables. This is rarely true, however. Most reports on experimental measurements include estimated error in the independent variable. If such an estimate is not included, a lower limit on the error can be estimated from the number of decimal digits in which the data are reported. (For example if the temperature is reported with one digit after the decimal point in degrees K, then the error is at least $\pm 0.05K$). Thus, the true value of an independent variable can be represented by:

$$x_i = \tilde{x}_i + \delta x_i \tag{8}$$

where $\tilde{x}_i$ is the expected measured value and δx_i is the error (or uncertainty) in its value. The error in the independent variable is, of course, carried over to its functions. The error in the different functions can be estimated from:

$$\left|\delta x_{ji}\right| \le \left|\frac{\partial f_j}{\partial x}\right|_{x=x_i} \left|\delta x_i\right| \tag{9}$$

where δx_{ji} is the estimated error in the jth function of x_i.

The errors in $\mathbf{x}$ and its functions are carried over to the normal matrix and to the $\mathbf{X}^T\mathbf{y}$ term in eq. (3). These errors can propagate during the solution process of eq. (3) yielding inaccurate and unstable parameter estimates. Denoting $\mathbf{b} = \mathbf{X}^T\mathbf{y}$, the errors in the calculated parameter values, $\delta\hat{\boldsymbol{\beta}}$ are bounded by (p. 176 in Dahlquist et al, 1974):

$$\kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} \geq \frac{\|\delta\hat{\boldsymbol{\beta}}\|}{\|\hat{\boldsymbol{\beta}}+\delta\hat{\boldsymbol{\beta}}\|} \tag{10}$$

where κ (**A**) is the condition number of the normal matrix and δ**A** is the matrix of errors in **A**. A similar equation relates the error in **b**, δ**b**, to the error $\delta\hat{\boldsymbol{\beta}}$:

$$\frac{\|\delta\hat{\boldsymbol{\beta}}\|}{\|\hat{\boldsymbol{\beta}}\|} \leq \kappa(\mathbf{A})\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \tag{11}$$

The condition number is the ratio of the largest to the smallest eigenvalue of **A**. If the condition number is large, the δ**A** and δ**b** are amplified considerably in the calculation of $\hat{\boldsymbol{\beta}}$, yielding very poor estimate of the true parameter values. Large condition numbers result from linear dependency (collinearity) between the various (supposedly) independent variables. The normal matrix which has large condition number is called "ill-conditioned".

The problems of collinearity has been extensively discussed in the literature (see, for example, Shacham and Brauner (1997), Brauner and Shacham (1998)). In the next three sections, the practical results of collinearity, ill-conditioning and numerical error propagation will be demonstrated using a vapor pressure correlation example.

Vapor Pressure Models and Data

There are several models which are being extensively used for correlating vapor pressure versus temperature data. Three types of equations will be used in this work. The first is a polynomial:

$$\ln P = \beta_0 + \beta_1 T + \beta_2 T^2 \Lambda\ \beta_{n+1} T^n \tag{12}$$

where T is temperature (K), P is pressure ($k\,Pa$) and n is the order of the polynomial.

The precision and stability of polynomial regression can often be increased by transforming the data. The following transformations (which were investigated by Shacham and Brauner (1997)) will be used. Normalization of the temperature:

$$\iota_i = T_i / T_{\max} \tag{13}$$

where $T_{\max}$ is the largest temperature value in the data set. The w transformation defined by:

$$w_i = (T_i - T_{\min})/(T_{\max} - T_{\min}) \tag{14}$$

which yields values in the rage $0 \leq w_i \leq 1$, and the z transformation

$$z_i = \frac{2T_i - T_{\max} - T_{\min}}{T_{\max} - T_{\min}} \tag{15}$$

which yields variable distribution in the range of $-1 \le z_i \le 1$.

The Riedel equation is a four parameter equation, which is used in several slightly different forms. The following definition is used in this work

$$\ln \pi = \beta_0 + \frac{\beta_1}{\tau} + \beta_2 \ln \tau + \beta_3 \tau^2 \tag{16}$$

where π is the normalized pressure $\pi_i = P_i / P_{\max}$. When only the first two terms of the equation are used, the model reduces to what is known as Clapeyron's equation.

The Wagner (1973) equation is considered the most accurate equation, with the smallest number of constants, for correlating vapor pressure data in a wide range of temperatures (between the triple point and the critical temperatures). While the number of terms and the exponents of the different terms in Wagner's equation may change, the most widely used form of this equation is:

$$\ln P_R = \frac{1}{P_R}\left[\beta_0 (1 - T_R) + \beta_1 (1 - T_R)^{1.5} + \beta_2 (1 - T_R)^3 + \beta_3 (1 - T_R)^6\right] \tag{17}$$

where $T_R = T/T_c$ is the reduced temperature, $P_R = P/P_c$ is the reduced pressure, T_c is the critical temperature (K) and P_c is the critical pressure ($k\,P_a$).

To enable discrimination between possible different sources of inaccuracy in a. regression model, 'exact' data generated by Wagner's equation will be used rather than real experimental data. This way, the precision of the data can be controlled by introducing randomly distributed error (noise) to the variables.

Table 1 shows the critical constants and the Wagner equation coefficients for Toluene, the substance which is used in this study. Three different data sets were generated using the Wagner equation with the constants given in Table 1. In each set, 21 data points were generated in the vicinity of the normal boiling point. The three sets are covering temperature ranges of 100K, 50K and 20K. Minimal and maximal values of T, P, ι and π for the three data sets are also shown in Table 2. In Figure 1, the vapor pressure of Toluene in the range of 310K-590K as calculated from Wagner's equation is displayed. It can be seen that the vapor pressure changes by more than four orders of magnitude over this temperature range. The temperature ranges corresponding to the three data sets are also marked on the figure. These data sets will be regressed using the various vapor pressure models.

Table 1: Critical Properties and the Wagner Equation Constants for Toluene (McGarry (1983)).

Melting point temperature[1]		178.15K
Normal boiling point temperature[1]		383.75K
Critical temperature		591.72K
Critical pressure		4106.45kPa
Wagner constants	β_0	-7.28607
	β_1	1.38091
	β_2	-2.83433
	β_3	-2.79168

[1] Reference: Weast (1978)

Table 2: Minimal and Maximal Values for the Vapor Pressure Data Sets.

	Data Set 1 Range = 100K	Data Set 2 Range = 50K	Data Set 3 Range = 20K
$T_{\max}(K)$	433.75	408.75	393.75
$T_{\min}(K)$	333.75	358.75	373.75
$\iota_{\max}$	1.0	1.0	1.0
$\iota_{\min}$	0.769452	0.877676	0.949206
$P_{\max}(kPa)$	347.812	195.922	133.381
$P_{\min}(kPa)$	18.9721	46.9463	75.5311
$\pi_{\max}$	1.0	1.0	1.0
$\pi_{\min}$	0.0545471	0.239617	0.566279

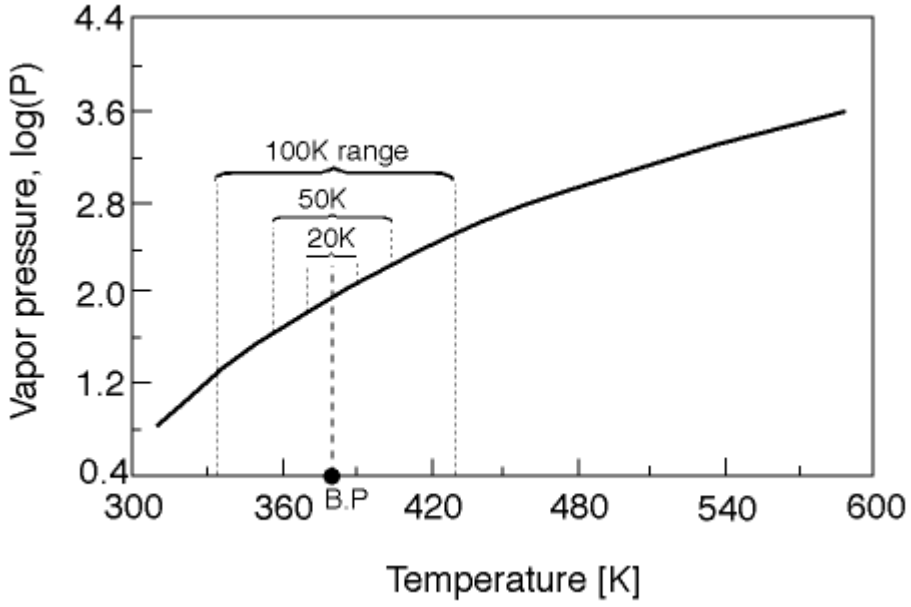

Figure 1: Vapor pressure of toluene in the 310*K* - 590*K* temperature range.

Fitting Polynomials to the Data

An equation in the form of polynomial is often used for representing the relationship between an independent variable and a dependent variable. The resulting correlation is an empirical model, which often lacks a theoretical basis. Some of the pitfalls of using such a model for correlating vapor pressure data were discussed by Shacham et al (1996). Recently, theoretical aspects of the stability of parameters estimation as function of range and precision of the data and the number of terms in the polynomial have been extensively studied (Shacham and Brauner, 1997). The vapor pressure data can be used to demonstrate the practical results predicted by these theoretical studies.

The model of eq. (2) with data set 1 (range of 100*K*) were used in all polynomial regression studies. The calculations were carried out using MATLAB (double precision computation with approximately 16 decimal digits accuracy) where eq. (3) is explicitly solved for $\hat{\boldsymbol{\beta}}$. Table 3 shows the coefficients, variance and condition numbers of the $\mathbf{X^T X}$ matrix for polynomials up to the 6th orders when the regression is done on *T* and *P* without any normalization or transformation. It can be seen that the variance decreases from 5.746×10^{-3} for a linear dependency to 1.6522×10^{-9} for 4th order polynomial of *T*. Increasing the order of the polynomial further causes a sharp increase of the variance to 2.2565×10^{-2} for the 6th order polynomial. The standard errors of the parameters (not shown) indicate a similar trend. They are smaller by about two orders of magnitude then the parameter values for polynomials up to the 4th order. For the 5th order polynomial (and higher orders), the standard errors become larger then the parameter values.

Table 3: Coefficients, Variance and Condition Number in Polynomial Regression of the Vapor Pressure Data

Order / cons	1	2	3	4	5	6
β_0	-6.53471E+00	-1.94056E+01	-3.46712E+01	-5.12182E+01	-6.85307E+01	-5.60273E+01
β_1	2.88523E-2	9.63520E-02	2.16591E-01	3.90455E-01	6.17904E-01	4.19561E-01
β_2		-8.79474E-05	-4.02445E-04	-1.08567E-03	-2.27844E-03	-9.77902E-04
β_3			2.73179E-07	1.46326E-06	4.58435E-06	2.31157E-08
β_4				-7.75297E-10	-4.85019E-09	4.15379E-09
β_5					2.12358E-12	-7.32026E-12
β_6						4.14629E-15
s^2	5.74600E-03	4.05780E-05	2.33390E-07	1.65220E-09	4.21080E-06	2.25650E-02
$\kappa(\mathbf{A})$	2.39540E+07	7.25390E+14	2.84320E+22	5.82150E+25	3.35990E+31	3.09540E+34

The practical significance of such variations in the variances and the standard errors can be seen in Figures 2 and 3. Figure 2 shows the relative error in the calculated pressure values for 1st, 2nd and 6th order polynomials versus the "experimental" pressure. For the 1st order polynomial, there is a clear trend of the error and it reaches a maximum value of about 16%. For the 2nd order polynomial, the error is much smaller and it hardly reaches 1%. Errors in the 3rd and 4th order polynomials cannot even be observed in the scale of Figure 2. For the 6th order polynomial, there is again a clear trend in the error and the maximal error reaches 30%. The same trend is even more pronounced when looking at the errors in the calculated derivative values (Figure 3). The maximal error in the derivative values using the straight line correlation reaches only about 50%, it reduces to 0.1% for the 4th order polynomial, whereas with the 6th order polynomial it increases up to 65% error. Thus, using a too high order polynomial (often called overcorrelation) yields even poorer results in the dependent variable values and its derivative values then using a polynomial of a too low order.

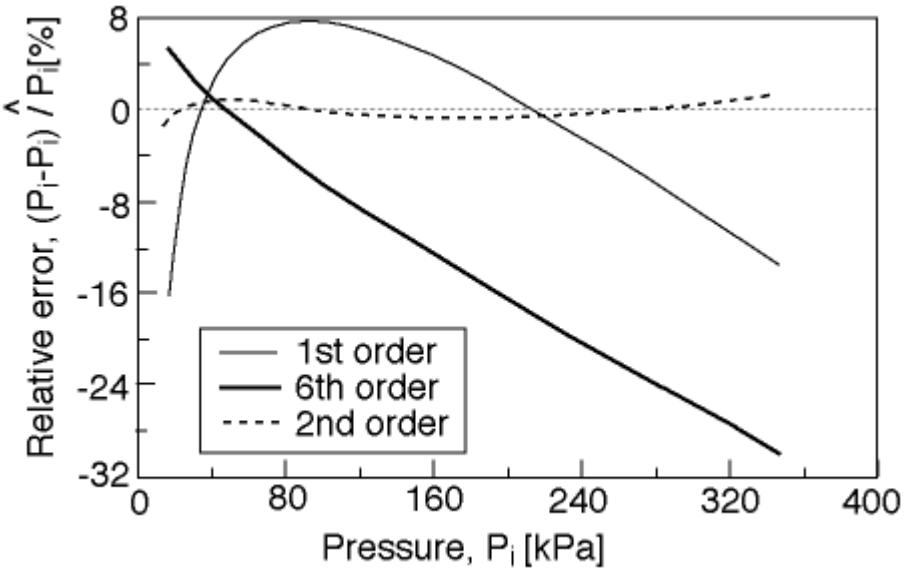

Figure 2: Relative error in the calculated pressure values (polynomial regression).

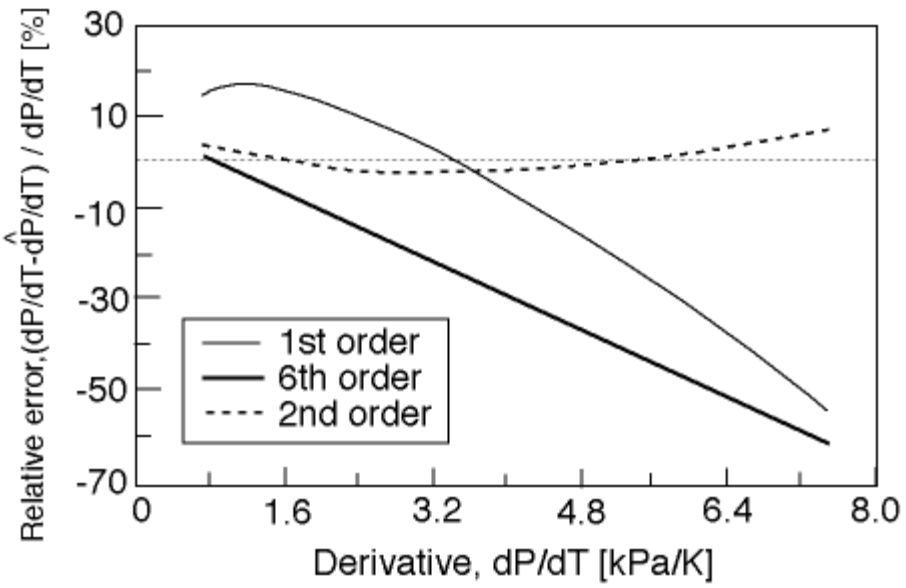

Figure 3: Relative error in calculated pressure derivatives (polynomial regression).

In trying to determine what causes the poor results obtained with high order polynomials, the condition number of the normal matrix (shown in Table 3) can provide valuable information. The value of the condition number is 2.3954×10^{7} for the 1st order polynomial and it increases to 3.0854×10^{34} for the 6th order polynomial. The MATLAB program, which relies on the condition number in determining the conditioning of the normal matrix, starts issuing warning messages "Matrix is close to singular or badly scaled. Results may be inaccurate" already for the 3rd order polynomial.

Does the 4th order polynomial represent the limit of precision for this correlation? To answer this question, regressions have been carried out using the various transformations of the temperature as defined in eqs. (13), (14) and (15). The results of these regressions are summarized in Fig. 4, which shows the variance as function of the polynomial order for the various transformations. With no-transformation or with a simple normalization, a. steady decrease of the variance up to the 4th order polynomial is obtained, whereby each additional term in the polynomial reduces the variance by about two orders of magnitude. Starting at the 5th order polynomial, the normal matrix becomes ill-conditioned, which causes a sharp increase in the variance. With the *w* and *z* transformations, however, the same rate of decrease of the variance extends to the 6th order polynomial. For the *w* transformation, the variance increase starts using a 7th order polynomial, while for the *z* transformation, the variance reaches a steady minimal level at the 8th order polynomial. Thus, using the *z*-transformation enables reducing the variance of the best fit to 5.0107×10^{-16} using the 8th order polynomial from the minimal value of 1.6522×10^{-9} obtained with the 4th order polynomial and the untransformed original data.

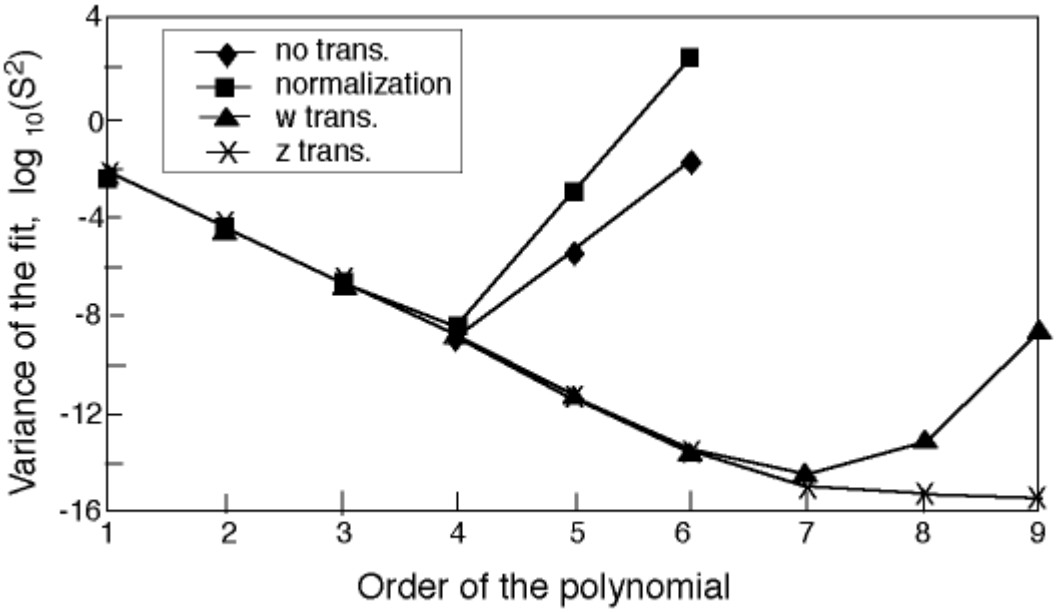

Figure 4: Variance in polynomial regression using various transformations of the temperature data.

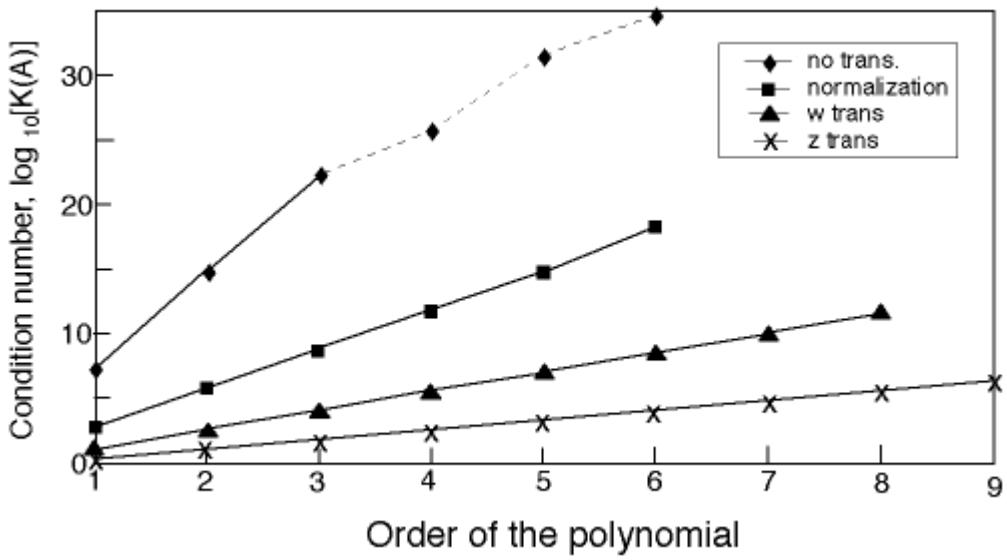

Figure 5: Condition numbers in polynomial regression using various transformations of the temperature data.

Figure 5 shows the logarithm of the condition number as function of the polynomial order for the various transformation. The logarithm of the condition number increases linearly as the polynomial order increases. (The only exception is the case of "no transformation" where numerical error propagation prevents obtaining accurate values for the condition numbers for

high order polynomials). The condition numbers are minimal for the *z*-transformation and increase in the following order: *w*-transformation, normalization and no transformation. The reduction of the condition number corresponding to a particular polynomial order when using the *w* and *z* transformation (instead of normalization) is reflected in the lower variance values achieved with these transformations. It is to be noted, however, that the reduction in the condition number when using normalized data instead of the original data, has no significant effect on the order of the best-fit polynomial and its variance.

Fitting Clapeyron's and Riedel's Equations to the Data

Clapeyron's equation (the first two terms in eq. (5)) and Riedel's equation are both frequently used for modeling of vapor pressure data. To investigate the effects of the range and precision of the independent variable data, both equations were fitted using the three data sets shown in Table 2. The data was generated using Wagner's equation, but for this study, a random and normally.distributed error was introduced into the temperature data. Three cases were tested with the following error levels: $|\delta T| = 0.0005$, 0.05 and 0.5*K*. Using the highest precision ($\delta T = 0.0005K$) and the widest temperature range (100*K*) data with Riedel's model yields a very accurate correlation. The parameter values obtained are $\beta_0 = 14.848$, $\beta_1 = -15.994$, $\beta_2 = -8.9698$, $\beta_4 = 1.1456$, the variance is 3.04×10^{-10} and the standard errors of the parameters are smaller by three orders of magnitude than the parameter values. The resultant maximal error in the calculated pressure value is 0.00487 *kPa*. For the same data set, Clapeyron's equation yields a much worse fit. The resultant parameter values are $\beta_0 = 9.7118$, $\beta_1 = -9.6941$, the variance is 1.1589×10^{-4} and the maximal error in the calculated pressure is 6.2335 *kPa*. Thus, the use of the four parameter Riedel's equations adds approximately three more decimal digits to the accuracy of the correlation.

To observe the effects of the range and precision of the temperature data on the accuracy of the correlation, the various data sets were regressed using Clapeyron's and Riedel's models. Figure 6 shows the variance of the fit obtained with Clapeyron's equation as function of the range with the error level δT as parameter. It can be seen that for the high precision temperature data ($\delta T = 0.0005K$), the variance is 1.1589×10^{-4} for the 100*K* range. It reduces to 6.8956×10^{-6} for the 50*K* temperature range and further to 1.7399×10^{-7} for the 20*K* range. This is the trend that could have been expected in case the variance results from a lack of fit of a model (due to a limited number of parameters). In such a case, the model can better represent the data in a narrower range of temperature and the variance is expected to decrease accordingly. However, the trend is different for the low precision data ($\delta T = 0.5K$). For the 100*K* range, the variance is 3.0519×10^{-4}, it reduces to 1.29×10^{-4} for the 50*K* range and increases to 2.256×10^{-4} for the 20*K* range. For this low precision data, numerical error propagation (resulting from collinearity) is intensified with reducing the range and for the range of 20*K*, it dominates over a better model fit which could have been achieved in a narrower range.

Figure 7 shows the variance of the Riedel correlation as function of the temperature range with the error level δT as parameter. It can be seen that for all the δT's studied, reducing the range from 50*K* to 20*K* affects an of the variance. For the low precision data, Riedel's model yields even a larger variance than that of Clapeyron's equation.

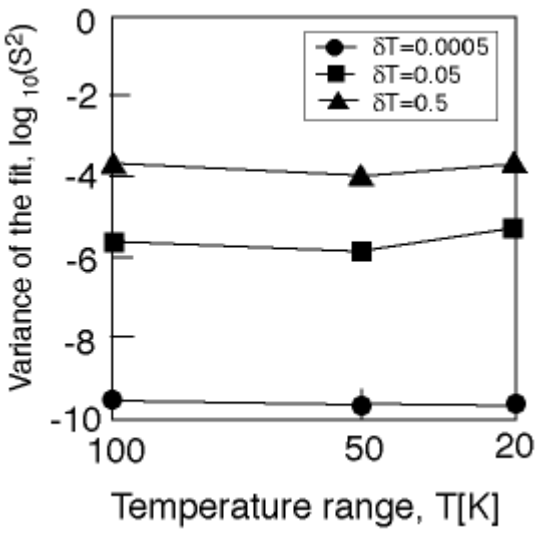

Figure 6: Change of the variance as function of range for Clapeyron's equation.

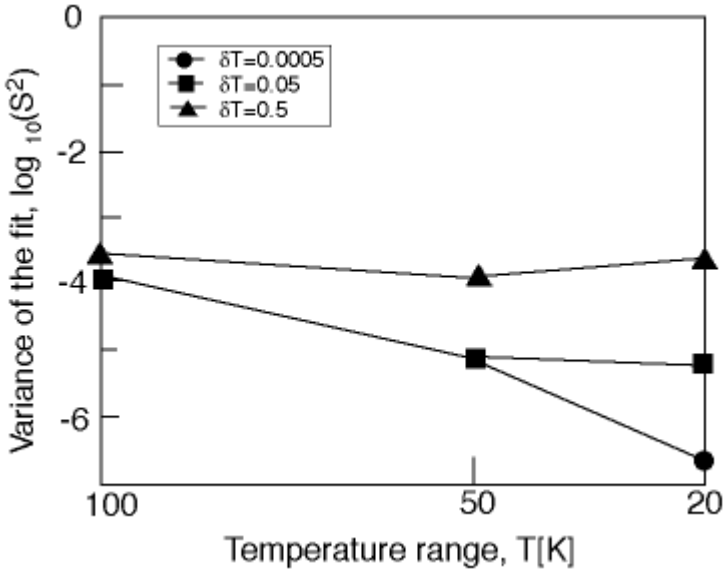

Figure 7: Change of the variance as function of range for Riedel's equation.

This study clearly demonstrates that the optimal number of terms/parameters in a model is a function of the range and precision of the independent variable data. When reducing the range improves the accuracy of the correlation considerably, the use of a different model (consists of more terms/parameters) should be considered. But, when range reduction does not improve the accuracy, adding more terms to the model will probably lead to over-correlation. The harmful effects of over-correlation have been demonstrated in the previous section with reference to the polynomial regression.

Conclusions

Numerical error propagation, resulting from collinearity between various functions of the same independent variable, can severely limit the accuracy of a correlation. Because of the complex theory involved, demonstrating to undergraduate students the causes and practical significance of such inaccuracies and identifying methods to improve the accuracy can represent a great challenge. In this paper, the graphical and interactive capabilities of two numerical computation packages (MATLAB and POLYMATH) have been used to demonstrate these topics.

It was shown that in polynomial regression, the highest order of polynomial to be used is often limited by numerical error propagation. Disregarding this limit can yield large errors in the calculated values of the dependent variable and even more severe errors in its derivatives. The condition number of the normal matrix provides information regarding the rate of the error propagation, but this information can be inaccurate and sometimes even misleading.

Collinearity in polynomial regression can be significantly reduced by using data transformations, in particular the z-transformation, which transform the independent variable data to the [-1, +1] range. To determine whether it is collinearity that limits the highest order polynominal and the accuracy of the fit, it is advisable to repeat the regression using z-transformation of the independent variable data. If the transformed data allows fitting by a higher order polynomial, collinearity is a limiting factor, otherwise other reasons for the low accuracy must be sought.

Using the two-parameter Clapeyron equation and the four-parameter Riedel equation, it was shown that there is not a single model which is superior in representing a particular type of data. For correlating low precision and/or narrow range data of vapor pressure, the Clapeyron equation is more appropriate than the Riedel equation. Whereas, for high precision and wide range data, Riedel's equation is more appropriate. If reducing the range results in a significant reduction of the variance, adding more terms/parameters to the model can improve the accuracy of the correlation. But, adding more terms when range reduction does not lead to a better fit can result in over-correlation.

References

1. Bates DM and Watts DG (1988), "Nonlinear Regression Analysis and its Application", John Wiley & Sons, New York.
2. Brauner N and Shacham M (1998), "The Role of Range and Precision of the Independent Variable in Regression of Data", AIChE J., 44(3), 603-611.
3. Dahlquist A, Björk A and Anderson N (1974), "Numerical Methods", Prentice-Hall, Englewood Cliffs, N.J.
4. Draper NR and Smith H (1981), "Applied Regression Analysis", John Wiley & Sons, New York.
5. Himmelblau DM (1970), "Process Analysis by Statistical Methods", John Wiley & Sons, New York.
6. Math Works, Inc. (1992), "*The Student Edition of MATLAB*"; Prentice-Hall, Englewood Cliffs, N.J.
7. McGarry J (1983), "Correlation and Prediction of the Vapor Pressures of Pure Liquids over Large Pressure Ranges", Ind. Eng. Chem. Process. Des. Dev, 22, 313.
8. Noggle JH (1993), "Practical Curve Fitting and Data Analysis", Prentice-Hall, Englewood Cliffs, N.J.

9. Shacham M and Brauner N (1997), "Minimizing the Effects of Collinearity in Polynomial Regression", Ind. & Chem. Res. 36, (10), 4405-4412.
10. Shacham M, Brauner N and Cutlip MB (1996), "Replacing Graph Paper with Interactive Software in Modeling and Analysis of Experimental Data". Comp. Appl. Eng. Edu. 4(3), 241-251.
11. Shacham M and Cutlip MB (1996), "POLYMATH 4.0 User's Manual", CACHE corporation, Austin, TX.
12. Wagner W (1973) "New Vapor Pressure Measurements for Argon and Nitrogen and a New Method for Establishing Rational Vapor Pressure Equations", Cryogenics 13, 470.
13. Weast RC (Ed) (1975), "Handbook of Chemistry and Physics", 56th Ed, CRC Press, Ohio.

NEIMA BRAUNER received her BSc and MSc from the Technion, Israel Institute of Technology, and her PhD from the University of Tel-Aviv. She is currently a Professor in the Fluid Mechanics and Heat Transfer Department. She teaches courses in Mass and Heat Transfer and Process Control. Her main research interests include hydrodynamics and transport phenomena in two-phase systems and applied regression analysis.

MORDECHAI SHACHAM is a Professor in the Chemical Engineering Department at the Ben Gurion University of the Negev, Beer-Sheva, Israel. He received his BSc and DSc from the Technion, Israel Institute of Technology. His research interests include applied numerical and statistical methods, computer-aided instruction, chemical process simulation, design and optimization.