Replacing Graph Paper with Interactive Software in Modeling and Analysis of Experimental Data

MORDECHAI SHACHAM,[1] NEIMA BRAUNER,[2] and MICHAEL B. CUTLIP[3]

[1]*Department of Chemical Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel,* [2]*School of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel, and* [3]*Department of Chemical Engineering, University of Connecticut, Storrs, Connecticut 06269*

ABSTRACT

The techniques presented in introductory engineering textbooks for modeling and analysis of experimental data nowadays are essentially the same as the ones that were presented over 35 years ago. Considerable potential now exists for dramatic improvements in data correlation and analysis because of the introduction of personal computers along with user-friendly interactive software which performs linear and nonlinear regressions and yields standard statistical results. This article presents some basic statistical concepts which are required to understand the results obtained from a regression package and demonstrates, using an example of vapor pressure correlation the proper technique for modeling and analysis of experimental data. Our experience and student performance has indicated that a 1st-year course for engineering students can effectively introduce students to the correlation and the modeling of experimental data. This capability can be given to students during two lectures and a 1-hour computer laboratory period plus an appropriate assignment, provided that an interactive regression package (Polymath or EZfit, for example) or a spreadsheet program with multiple linear regression capabilities is available.

INTRODUCTION

Chemical engineering students have been traditionally taught modeling and analysis of experimental data in the first introductory course of chemical engineering. Anderson and Wenzel [1], in their classic book from 1961, entitled this subject ''Presentation and correlation of data.'' Topics under this title included fitting a straight line to data on rectangular, semilogarithmic, and logarithmic coordinates; interpolation using Lagrange polynomial; and the method of least-squares regression for a second-order polynomial. In the more recent textbook of Felder and Rousseau [2], the subjects covered in this area still fit a straight line on rectangular, semilogarithmic, and logarithmic coordinates and least-squares regression for a straight line.

The question arises as to whether the methods that were taught 35 years ago are still sufficient for the data analysis needs of a chemical engineering student or a practicing chemical engineer today. The same question is probably valid for most other engineering disciplines.

Over the years, many books have been published

Computer Applications in Engineering Education, Vol. 4(3) 241–251 (1996)

CCC 1061-3773/96/030241-11

which deal with a statistical approach to modeling and analysis of experimental data. These books have been published by statisticians as well as chemical engineers [3–6]. The methods presented in these books have failed to reach most engineering students and practicing engineers because they require an extensive knowledge of statistics, which only a specialized group of engineers possesses. The emergence of interactive regression programs for personal computers with both graphic and numeric output of the results brought advanced capabilities for data analysis and modeling into the reach of all engineering students and practicing engineers. Representative software includes the regression program in the Polymath package [7], the EZfit program [6], and the linear regression option in spreadsheet programs such as Excel [8]. The appropriate time for introducing students to the use of such programs and the extent of the desirable theoretical background material that should be included in such an introduction deserves further discussion.

Engineering students typically start doing laboratory experiments requiring data analysis in the 1st year of their studies, so they should be introduced to the basic concepts of data modeling and analysis as early as possible. This material can be included in an introductory computing course which is given to freshman engineering students in several schools, or it can be given as part of an introductory chemical engineering course.

It is challenging to decide what to include in this introduction to modeling and analysis of experimental data. Faculty should assume that the students do not have previous knowledge of statistics, but this introduction should not replace the traditional statistics course. On the other hand, regression programs cannot be used in a ''cookbook'' manner, because such a use often leads to wrong conclusions. The students must understand how the statistical software works and appreciate the meaning and importance of different statistical indicators.

In this article we present material that, in our opinion (and experience), can provide an introduction to the use of interactive programs for data modeling and analysis to 1st-year engineering students. The time framework for this introduction includes two lectures, 1 computer laboratory hour, and an extensive data analysis assignment. This introductory material will provide students with a basis for advancing from the era of graph paper to that of computerized modeling and analysis of experimental data.

THE OBJECTIVES OF DATA MODELING AND ANALYSIS

Generally data modeling and analysis is applied to either experimental data from the laboratory or tabulated data from the literature. This is accomplished by postulating a particular form for a model and fitting parameters to the model by regression of the data. Obviously the objective is to find the model which best represents the data.

In the past, tabulated data were often used in process calculations. The introduction of computers and more rigorous calculations into the process design and analysis area has made it imperative to fit a model to the data. Such a model can represent the data in a more compact, easier to use, and often more accurate form.

There are, in general, three types of uncertainties associated with the process of trying to fit a model to a set of data: (1) The validity of the model representation of the physical phenomena is not known. (2) As a result of experimental (or other) error, the true value of a particular variable at a particular point is not known. (3) It is not known how well the sample set of data represents the complete set (full population), if such were available.

There are many statistical tests available to judge the reliability of a regressed model and compare various models in light of the uncertainties involved. A comprehensive discussion, review, and demonstration of this subject is presented, for example, in Himmelblau [4]. But there are actually a few statistical indicators on which the comparison of the different models can be based and which are usually sufficient for carrying out statistical analysis. Further analysis will often be of diminishing value.

For statistical analysis, it is important to understand the difference between smoothed data as opposed to unaltered and untreated experimental data. Tabulated data published in handbooks (such as Perry's [11], for example) are often smoothed and interpolated, and as such they do not provide a true representation of the experimental error involved.

In the next section, the basic statistical concepts which should be included in an introductory course are described. An example, which includes regression of both experimental and smoothed data for vapor pressure of a particular substance, will be used to demonstrate the proposed approach for data modeling and analysis.

BASIC STATISTICAL CONCEPTS

The student (or practicing engineer) must be familiar with a few basic statistical concepts to be able to select from among several models and estimate the uncertainty of using the selected model.

Let us assume that there is a set of N data points of a dependent variable y_i versus x_{1i}, $x_{2i} \cdots x_{ni}$, where $x_1, x_2 \cdots x_n$ are n independent variables. A particular model to be fitted to the data is of the form

$$y_i = g(x_{1i}, x_{2i} \cdots x_{ni}, \beta_1, \beta_2 \cdots \beta_m) \quad (1)$$

where $\beta_1, \beta_2 \cdots \beta_m$ are the parameters of the model. The least-squares error approach is most often used to find the parameters of Equation (1).

The statistical assumption behind the least-squares error method for parameter estimation is that the measured value of the dependent variable has a deterministic and a stochastic part. The stochastic part is often denoted as an error, ϵ_i. Thus, Equation (1) can be rewritten

$$y_i = g(x_{1i}, x_{2i} \cdots x_{ni}, \beta_1, \beta_2 \cdots \beta_m) \pm \epsilon_i \quad (1a)$$

It is further assumed that the origin of ϵ_i is measurement error which is randomly distributed.

An infinite number of measurements would be required to obtain the true values of the parameters $\beta_1, \beta_2 \cdots \beta_m$. Because a sample always contains a finite number of measurements, the calculated parameters are always approximations for the true values. They are denoted with a circumflex. Thus, $\hat{\beta}_1, \hat{\beta}_2 \cdots \hat{\beta}_m$ are the calculated values of the parameters and $\hat{y}_i$ is the estimate for the dependent variable y_i.

In the least-squares error approach, the estimates $\hat{\beta}_1, \hat{\beta}_2 \cdots \hat{\beta}_m$ are found so that they minimize the following function (squares of errors):

$$F = \sum_{i=1}^{N} [y_i - g(x_{1i}, x_{2i} \cdots x_{ni}, \beta_1, \beta_2 \cdots \beta_m)]^2 \quad (2)$$

The particular mathematic technique of finding the set of parameter values that minimizes the function F depends on the form of the function $g(\underline{x}_i, \underline{\beta})$.

If the parameters appear in a linear expression in the function g, the minimization can be carried out by solving a set of simultaneous linear algebraic equations (the normal equations). Representation of the minimization of the sum of squares as a system of linear equations, for the cases of linear and polynomial regression, and the solution of the system are described in detail in several introductory chemical engineering textbooks (for example, [1,2]), and will not be repeated here. As part of the discussion on linear regression, it should be noted that several nonlinear models can be transformed to linear ones by transformation of variables. Such transformations are also discussed in the introductory textbooks.

Nonlinear regression (when the parameters appear in nonlinear expressions in $g(\underline{x}_i, \underline{\beta})$) requires the application of more advanced numeric methods. This subject is usually included in more advanced numeric methods courses [9].

An assessment of the quality of the fit of a particular model and a comparison between different models is based on graphic and numeric information.

Graphic Information

The observed (y_i) and estimated ($\hat{y}_i$) values of the dependent variable can be plotted versus x_i (if there is a single independent variable) or versus i, the point number (if 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) $\hat{\epsilon}_i$ is plotted usually versus y_i, where

$$\hat{\epsilon}_i = y_i - \hat{y}_i \quad (3)$$

A random distribution of the residuals around zero indicates that the model correctly represents 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 assumed error distribution for the data (random error distribution for the dependent variable) is not correct. In such cases, the use of statistical indicators to evaluate the validity of the model for the particular phenomena based on the available data is not justified. When the source of the nonrandomness in the residual distribution is a lack of fit of the model, appropriate modification and upgrade of the model can eliminate the problem. In other cases, data transformation can be beneficial. Many forms of data weighting and transformation can be covered only in advanced statistics courses, but some simple forms are routinely used in data analysis. For example, in cases where the value of the dependent variable changes by several orders of magnitude over the range of interest, often the relative error is distributed normally. The relative error is defined as

$$\tilde{\epsilon}_i = \frac{\hat{\epsilon}_i}{y_i} \quad (4)$$

An appropriate transformation, which results in minimization of the relative error in the regression, is taking the logarithm of both sides of the model equation.

It should be emphasized that when the variables are transformed the residual plot must be constructed using the transformed form of the independent variable, to account for the change in the error distribution introduced by the transformation.

Numeric Information

The most frequently used numeric indicator of the quality of the fit is the standard error of the estimate which represents the sample variance, and is given by

$$S^2 = \frac{\Sigma_{i=1}^{N} (y_i - \hat{y}_i)^2}{N - m} \tag{5}$$

Thus, the sample variance is the sum of squares of errors divided by the degrees of freedom (where the number of parameters, m, 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 means a better fit of the model to the data. It should be emphasized that when the sample variance is used for the comparison of different models, the same independent variable (transformed or nontransformed) should be used in Equation (5) for all models. The variance is an unscaled variable which can take on any value from zero to infinity. Consequently the variance alone cannot be used to indicate the goodness of fit between the data and the respective model.

When fitting a straight line to data of y versus x, most software will provide as an output the correlation coefficient (R^2). The correlation coefficient represents the ratio between the sum of squares about the mean due to regression to the total sum of squares (about the mean), and is obtained by

$$R^2 = \frac{\Sigma_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\Sigma_{i=1}^{N} (y_i - \bar{y})^2} \tag{6}$$

Hence, R^2 measures the proportion of variation about the mean value $\bar{y}$ that is explained by the independent variables in the regression model. In case of simple linear regression (single independent variable), the correlation coefficient ($r = R$) can be calculated also from the following equation:

$$r = \frac{\Sigma_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i (x_i - \bar{x})^2}\sqrt{\Sigma_i (y_i - \bar{y})^2}} \tag{6a}$$

where $\bar{x} = \Sigma_{i=1}^{N} x_i/N$ and $\bar{y} = \Sigma_{i=1}^{N} y_i/N$ are the sample mean of the independent and dependent variables, respectively. The correlation coefficient ($-1 \leq r \leq 1$) indicates the strength of the correlation between x and y. If r is close to -1 or 1, there is a strong correlation between the variables, whereas a value close to zero indicates a weak or no correlation. Because the correlation coefficient is a scaled variable, it is often used to indicate how well a straight line represents 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 model fit and more precise data lead to narrow confidence intervals, while a poor model fit and/or imprecise data cause wide confidence intervals. Furthermore, confidence intervals which are larger (in absolute value) than the respective parameters themselves often indicate that the model contains too many parameters.

Calculation of the confidence intervals requires knowledge of some advanced statistical concepts (such as t distribution), so its discussion can be deferred to a course in statistics. In an introductory engineering course, the explanation that the confidence interval represents uncertainty in the calculated parameter values is probably sufficient. It should be emphasized, however, that the validity of the calculated confidence intervals is based on the assumption of random distribution of the residuals. Violation of this assumption renders the calculated confidence intervals doubtful.

The maximum errors (absolute and relative) have no statistical significance, but they can give an indication of the error that can be expected in the worst case, and therefore should be included in the comparison of different models. If some models involve transformation of variables, it is important to calculate the errors based on the same variables (as is also necessary in the case of the variance).

In the next section, an example involving modeling of vapor pressure will be presented and different models will be compared based on the graphic and numeric indicators outlined in this section.

VAPOR PRESSURE DATA AND MODELS

Table 1 shows vapor pressure data for 1-propanethiol, in the low-pressure range (1 to 760 mmHg) as published by Stull in 1947 [10]. The same data appear in several editions of *Chemical Engineers Handbook* (the last being the 6th edition, from 1984 [11]). Table 2 shows vapor pressure data for the same substance, 1-propanethiol, in a slightly different pressure range (149.41 up to 2026

Table 1 Vapor Presure Data for 1-Propanethiol (from Stull [10])

	Temperature (°C)	Pressure (mmHg)
1	−56.0	1
2	−36.3	5
3	−26.3	10
4	−15.4	20
5	−3.2	40
6	4.6	60
7	15.3	100
8	31.5	200
9	49.2	400
10	67.4	760

mmHg). These data are from Osborn and Douslin [12].

There are several obvious differences between the data in Tables 1 and 2. In Table 1, temperatures at even pressures are provided. Because pressure is the dependent variable, it is evident that the data in this table are not experimental but have been smoothed and interpolated. According to the number of significant digits provided, the error in the temperature can be at least 0.05°C and the error in the pressure is ±0.5 mmHg.

In view of the number of significant digits provided, the data in Table 2 appear to be real experimental data of higher precision. Judging merely from the number of significant digits, the error in the temperature is ±0.0005°C and the error in the pressure is 0.005 mmHg.

It is essential to find out whether the data represent either real measured values or smoothed and interpolated values. While the objective in both cases is to fit the best possible model to the data, only real measured data allow determination of the uncertainty associated with the use of the model. Statistical indicators obtained from smoothed data will show the model's deviation, not from the originally measured values but from the processed data. The uncertainty introduced by smoothing and interpolation is usually unknown. Furthermore, the smoothing process is usually carried out using a particular model. Fitting the same model to the data may falsely indicate excellent agreement between the model and the data.

In this respect, it is interesting to read what Stull [10] wrote about the method used for obtaining the data in Table 1. Vapor pressure data from different sources were combined using the following technique:

> A 56 × 38 inch Cox chart was used together with a set of map tacks of different colors. All the information on a given compound was plotted using different colors to represent the work of different individuals. With compounds that had been much worked on, it was easy to see which of the points did not fall on the median line. This median line was actually a taut thread placed so that it touched the data in which one had the most confidence. . . . By choice and elimination the thread was placed (under slight tension) so that it fit the points consistently, and the temperature values were "read back" at predetermined pressures and copied onto yellow cards. (Stull [10], p. 517)

Obviously, it is impossible to determine what is the uncertainty introduced by such a treatment of the data. Consulting the source of the data in Table 2 [12] confirms that these data indeed represent experimental results.

It is not always as evident, as in Table 1, that the data do not represent originally measured values. The source of the data must always be consulted to determine whether the data were not altered in a way that can make statistical indicators for evaluating the uncertainty in the calculated values of the model parameters meaningless.

The following models (equations) are compared with regard to their ability to represent the vapor pressure data of Tables 1 and 2:

1. The two-parameter Clapeyron equation:

$$\log(P) = A + \frac{B}{T} \qquad (7)$$

where P is the vapor pressure (mmHg), T is the temperature (°K), and A and B are param-

Table 2 Vapor Pressure Data for 1-Propanethiol (from Osborn and Douslin [12])

	Temperature (°C)	Pressure (mmHg)
1	24.275	149.41
2	29.563	187.57
3	34.891	233.72
4	40.254	289.13
5	45.663	355.22
6	51.113	433.56
7	56.605	525.86
8	62.139	633.99
9	67.719	760.00
10	73.341	906.06
11	79.004	1074.60
12	84.710	1268.00
13	90.464	1489.10
14	96.255	1740.80
15	102.088	2026.00

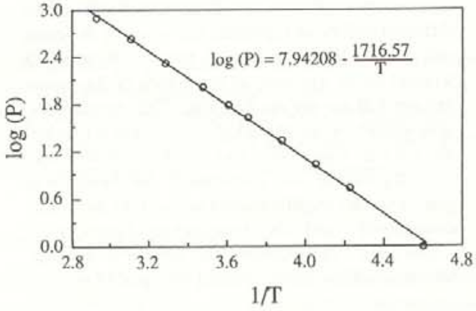

Figure 1 Plot of data points (Stull's [10] data) and calculated straight line for the Clapeyron equation.

eters to be estimated by regression of the experimental data.

2. The three-parameter Antoine equation:

$$\log(P) = A + \frac{B}{t + C} \qquad (8)$$

where t is the temperature (°C) and A, B, and C are parameters.

3. A polynomial with up to six parameters:

$$P = a_0 + a_1t + a_2t^2 + \cdots + a_5t^5 \quad (9)$$

All calculations were carried out using the "Polynomial, Multiple Linear, and Nonlinear Regression" program of the Polymath 4.0 package [7].

THE CLAPEYRON EQUATION

The Clapeyron equation can be linearized by defining $x = 1/T$ and $y = \log(P)$. Using this transformation, the plot of y versus x should give a straight line. This is demonstrated in several introductory chemical engineering textbook (e.g., [1]), usually by plotting P versus $1/T$ on a semilogarithmic paper.

Figures 1 and 2 show the plot of the data points, the calculated straight line, and the calculated parameters of Equation (7) for the data of Stull [10] and Osborn and Douslin [12], respectively. It can be seen that in both cases the experimental data are aligned nicely along a straight line. The calculated parameter values differ, however, in the second significant digit for the two data sets. The source of the discrepancy is the use of two different samples with different accuracies for correlating the same physical property.

The pertinent numeric results are summarized in Table 3. Some of the indicators show that the fit between the Clapeyron equation and the data is not very good for either of the data sets. The variance (based on the pressure itself) is 198.7 for the less accurate data set and 110.76 for the more accurate data set. Similarly, the maximum errors in the pressure are 37.07 and 30.57 mmHg for the less and more accurate data sets, respectively.

Because the accuracy of the data is not reflected and has a little effect on the accuracy of the correlation, it can be concluded that this particular model is limited in its ability to represent vapor pressure data. This conclusion is further supported by examination of the residual plots of the two data sets, shown in Figures 3 and 4. For both sets of data, the residuals are not randomly distributed, but show a clear pattern. Both data sets exhibit a curvature which is not predicted by the Clapeyron equation.

The curvature in both residual plots nullifies the apparent straight line representations in Figures 1 and 2 and also the values of the correlation coefficient (r) in Table 3 (very close to -1). The residual plot indeed magnifies the deviation between the model and the data points, making it easier to distinguish between an appropriate and an inappropriate model. The value of the correlation coefficient r, which is very close to -1, demonstrates a major limitation which may exist in some of the widely used statistical indicators. One should not rely on a single indicator but consider the consistency of several relevant indicators in data analysis.

THE ANTOINE EQUATION

The Antoine equation, given by Equation (8), is nonlinear, and the most appropriate treatment would

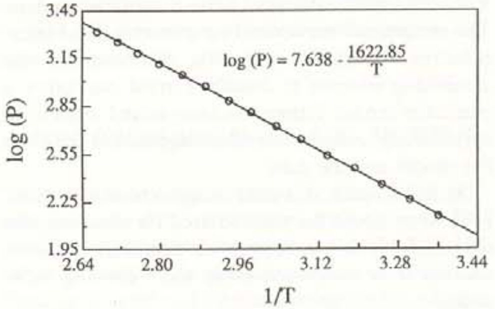

Figure 2 Plot of data points (Osborn and Douslin's [12] data) and calculated straight line for the Clapeyron equation.

Table 3 Numeric Results for the Clapeyron Equation

	Stull [10] Data		Osborn and Douslin [12] Data	
	Value	95% Confidence Interval	Value	95% Confidence Interval
A	7.94208	0.111465	7.638	0.0317435
B	−1716.57	29.9813	−1622.85	10.5741
Variance*	4.09×10^{-4}		1.66×10^{-5}	
r*	−0.99977		−0.999944	
Variance†	198.7		110.759	
Max. abs. error†	37.07		30.57	
Max. rel. error†	8.9%		1.69%	

*Based on $y = \log(P)$.
†Based on $y = P$.

be to use nonlinear regression for fitting parameters to this equation. However, it is preferable to postpone the discussion on nonlinear regression from the introductory course to a more advanced numeric analysis or statistics course. For the introductory engineering course, the Antoine equation can be linearized by multiplying both sides of the equation by $t + C$ and rearranging as follows:

$$\log(P) = A + \frac{(AC + B)}{t} - C\,\frac{\log(P)}{t} \quad (10)$$

In this equation, the independent variables are $\log(P)/t$ and $1/t$, and the dependent variable is $\log(P)$. Multiple linear regression can be used to calculate the parameters A, $(AC + B)$, and $(-C)$.

It should be noted that P, the original dependent variable, appears both as a dependent and independent variable in Equation (10). The inclusion of the dependent variable in the righthand side of Equation (10) violates one of the statistical assumptions on which the regression error analysis is based. This limitation must, of course, be pointed out to the students. But in most cases (as in this example) the linearization does not cause significant errors which would be apparent in the residual plot.

Table 4 shows the numeric results, and Figures 5 and 6 the residual plots, when Equation (10) is regressed using the data of Stull [10] and Osborn and Douslin [12]. A comparison of the results in Tables 3 and 4 shows that the combination of accurate data (Osborn and Douslin's) and a more appropriate model yields a very significant improvement in all of the statistical indicators. The variance (based on P) is reduced by more than four orders of magnitude (from ~110 in Table 3 to ~0.0085 in Table 4). The maximum relative and absolute errors are reduced by two orders of magnitude. Similar reduction results in the confidence interval for the calculated parameters values. The residual plot (in Figure 6) shows a random distribution of the errors, and the absolute value of the error based on $\log(P)$ is reduced by one order of magnitude

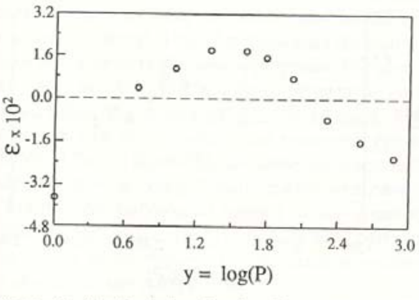

Figure 3 Residual plot for the Clapeyron equation (Stull's [10] data).

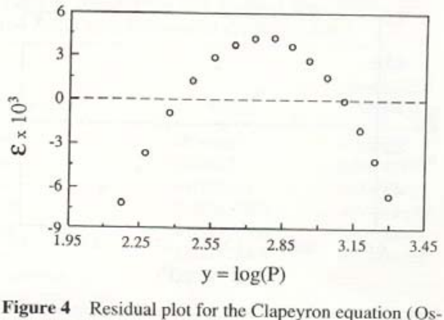

Figure 4 Residual plot for the Clapeyron equation (Osborn and Douslin's [12] data).

Table 4 Numeric Results for the Antoine Equation

	Stull [10] Data		Osborn and Douslin [12] Data	
	Value	95% Confidence Interval	Value	95% Confidence Interval
A	7.01572	0.176281	6.9311	0.00551491
AC + B	393.556	12.6924	373.323	0.654306
C	234.871	7.60125	224.802	0.358564
Variance*	1.276×10^{-3}		9.735×10^{-8}	
Variance†	89.96		0.008561	
Max. abs. error†	24.89		0.2504	
Max. rel. error†	3.27%		0.019%	

*Based on $y = \log(P)$.
†Based on $y = P$.

compared to that obtained with the Clapeyron equation.

When using the more appropriate model with the less accurate data [10], the improvement is not nearly as significant. Comparison of Table 3 and 4 shows that the variance based on P, as well as the maximal relative and absolute errors are reduced only by a factor of ~2. There is no appreciable change in the confidence intervals of the parameters. While the pattern of the residual plot (Fig. 5) is not as clear as with the Clapeyron equation (Fig. 3), there is still an observable trend of increasing residuals with increasing value of $\log(P)$.

POLYNOMIALS

An equation in the form of polynomial is often used for representing the relation between one independent variable and one dependent variable. The resulting correlation is an empiric model, which lacks a theoretic basis. The use of such a model for the vapor pressure data can demonstrate some of the advantages and the disadvantages of using empiric models.

In the selection of the order of polynomial which best represents the data, the residual plot and numeric values of the variance and confidence intervals on the parameter values are the most important indicators. Table 5 shows the variances of polynomials of orders 1–5 which have been fitted to the data of P versus t. It can be seen that for Stull's [10] data, there is a continuous, significant decrease of the variance from the first and up to the fifth-order polynomial, indicating that the fifth-order polynomial is probably the best from among the ones tested. Table 6 shows the parameter values (including the 95% confidence intervals) for the fifth-order polynomial for this case. It can be seen that all of the confidence intervals are smaller than the respective parameter values; thus, all of the parameters are significantly different from zero. The residual plot (not shown) indicates random distribution of the error. Comparing the pertinent results in Tables 4 to 6 shows that the variance (based on P) for the fifth-order polynomial is smaller by two or-

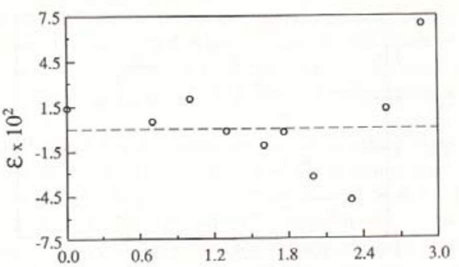

Figure 5 Residual plot for the linearized Antoine equation (Stull's [10] data).

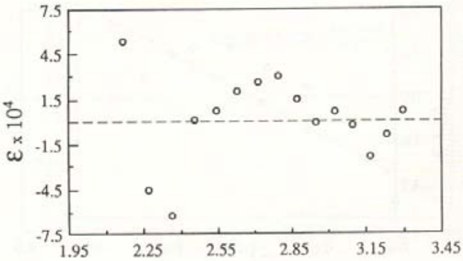

Figure 6 Residual plot for the linearized Antoine equation (Osborn and Douslin's [12] data).

Table 5 Variance of Polynomials of Different Orders for Representing Vapor Pressure Data

	Variance	
Order of Polynomial	Stull [10] Data	Osborn and Douslin [12] Data
1	17801.7	24875.6
2	2024.17	394.2
3	93.32	0.81
4	1.18	0.0021
5	0.1545	0.00205

ders of magnitude compared to that obtained with the Antoine equation, and the maximum absolute error is smaller by more than one order of magnitude. Thus, the polynomial fits the data better than the Antoine equation. One should remember, of course, that the fifth-order polynomial has six adjustable parameters; thus, the better fit is achieved by using a more cumbersome and complicated model. Obviously the number of data points required to obtain significant parameter values for the six-parameter model is larger.

The second column in Table 5 shows the variances of the different order polynomials for Osborn and Douslin's [12] data. It can be seen that for these data, there is a continuous and significant decrease of the variance with increasing order of the polynomial, up to the fourth order. While the variance for the fifth-order polynomial is slightly better than that of the fourth, the difference seems to be of little significance. This conclusion is further enforced by looking at the parameter values of the two polynomials in Table 7. It can be seen that, while for the fourth-order polynomial all the parameter values are significantly different from zero, for the last parameter of the fifth-order polynomial (a_5) the confidence interval is twice the parameter value itself, indicating that using zero value for this parameter can be as good as using any other value inside the confidence interval. Thus, for these data the fourth-order polynomial is the most appropriate. The residual plot for the fourth-order polynomial (not shown) indicates a random error distribution. The comparison of the variances and maximum errors (Tables 4 and 7) shows that for these more accurate data, the errors are only slightly smaller when using a fourth-order polynomial instead of the Antoine equation. This slight improvement in data correlation comes at the expense of two extra adjustable parameters in the polynomial.

It should be emphasized that an empiric model (such as a polynomial) can represent the data (often very well, as in this case) only locally: namely, inside the interval where the measurements were taken. The use of such a model for extrapolation, or for studying asymptotic behavior of the independent variable, can be completely misleading.

We can try, for example, to use the various models with the parameter values obtained with the Osborn and Douslin [12] data to extrapolate to lower temperatures (covered by Stull's [10] data). For $T = -26.3$°C, the reported vapor pressure is 10 mmHg. The Clapeyron equation yields $P = 11.58$ mmHg and the Antoine equation yields $P = 9.16$ mmHg. But the fourth-order polynomial (with the parameter values shown in Table 7), which represents the data excellently inside the interval where the measurements were taken, yields negative vapor pressure (-4.0 mmHg) when extrapolating.

CLASSROOM IMPLEMENTATION

The students in the Chemical Engineering Department at the Ben Gurion University of the Negev are introduced to the use of the "Polynomial, Multiple Linear, and Nonlinear Regression" program of the Polymath [7] package in the 1st year of their studies as part of an "Introduction to Personal Computers" course. Two hours of lecture are spent on introducing them to the basic concepts (roughly, the material included in the "Basic Statistical Concepts" section of this article), and 1 additional hour is spent in the computer lab, familiarizing the students with the technical details of the Polymath regression program. For practice with data analysis, every student is given a different set of physical and thermodynamic data (heat capacity, viscosity, heat of vaporization, vapor pressure, and thermal conductivity) for various substances. They are asked to calculate the parameters of different models representing the

Table 6 Parameters and Maximal Error Values for the Fifth-Order Polynomial (Stull [10] Data)

Parameter	Value	95% Confidence Interval
$a0$	47.4619	0.620243
$a1$	2.44657	0.0434703
$a2$	0.0521241	0.00122225
$a3$	0.000610835	4.44444e-05
$a4$	4.63332e-06	3.6769e-07
$a5$	2.03111e-08	9.6309e-09
Variance	0.1545	
Max. abs. error	0.53	
Max. rel. error	2.36%	

Table 7 Parameters and Maximal Error Values for Fourth- and Fifth-Order Polynomials (Osborn and Douslin [12] Data)

	Fourth Order		Fifth Order	
Parameter	Value	95% Confidence Interval	Value	95% Confidence Interval
$a0$	44.1152	1.36474	45.8882	3.8466
$a1$	2.65097	0.1049	2.47299	0.375874
$a2$	0.048802	0.00280121	0.0554638	0.0137958
$a3$	0.000762749	3.11485e-05	0.000645722	0.000239316
$a4$	3.6009e-06	1.22945e-07	4.57329e-06	1.97535e-06
$a5$			−3.08024e-09	6.24507e-09
Variance	0.002107		0.002057	
Max. abs. error	0.068		0.0825	
Max. rel. error	0.0156%		0.00845%	

data and then compare the models following the guidelines in the vapor pressure example shown here. After accomplishing this assignment, they submit a report summarizing their findings.

Reading their reports, it is evident that a brief introduction to an interactive regression program, as presented here, enables 1st-year engineering students to

1. Discriminate and appreciate the difference between true experimental data and smoothed or interpolated data.
2. Calculate the parameters of models requiring linear or polynomial regression.
3. Use transformations to linearize nonlinear models.
4. Apply the relevant statistical indicators: variance, confidence intervals, and residual plots to check the appropriateness of a model and compare the quality of representation of the same data using different models.

CONCLUSIONS

Considerable potential now exists for dramatic improvements in the correlation and analysis of experimental data because of the introduction of personal computers along with user-friendly interactive software which performs linear regressions and yields standard statistical results.

Our experience and student performance have indicated that a 1st-year course for engineering students can effectively introduce students to the correlation and modeling of experimental data. This capability can be given to students during two lectures and a 1-hour computer laboratory period plus an appropriate assignment.

Basic instruction to 1st-year students prior to the use of regression software should include the basic statistical terms of variance, least-squares, correlation coefficients, confidence intervals, and residual plots. Students should be exposed to the differences between experimental data and smoothed or interpolated data. Experience with data sets, such as presented here, can be helpful in learning to employ various indicators properly when making comparisons among various models and assessing model applicability.

As engineering students progress through their educational program, the ability to manipulate and correlate data sets can be used in a variety of courses and student capabilities can be enhanced through subsequent statistics and numerical analysis coursework.

The definitions, explanations, and examples set forth in this article can be used to provide freshman engineering students with data sets which help to advance their capabilities beyond the basic use of various forms of graph paper, which has been the standard for data correlation for much too long.

REFERENCES

[1] L. B. Anderson and L. A. Wenzel, *Introduction to Chemical Engineering*, McGraw-Hill, New York, 1961.

[2] R. M. Felder and R. W. Rousseau, *Elementary Principles of Chemical Processes*, Wiley, New York, 1986.

[3] N. R. Draper and H. Smith, *Applied Regression Analysis*, Wiley, New York, 1981.

[4] D. M. Himmelblau, *Process Analysis by Statistical Methods*, Wiley, New York, 1970.

[5] D. M. Bates and D. G. Watts, *Nonlinear Regression Analysis and Its Application,* Wiley, New York, 1988.

[6] J. H. Noggle, *Practical Curve Fitting and Data Analysis,* Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[7] M. Shacham and M. B. Cutlip, *Polymath 4.0 User's Manual,* CACHE Corporation, Austin, Texas, 1995.

[8] Microsoft Corporation, Microsoft Excel, Vers. 5, *User's Guide,* Microsoft Corporation, Redmond, Washington, 1993.

[9] A. Constantinides, *Applied Numerical Methods with Personal Computers,* McGraw-Hill, New York, 1987.

[10] D. R. Stull, "Vapor pressure of pure substances organic compounds," *Ind. Eng. Chem.,* Vol. 39, No. 4, 1947, pp. 517–540.

[11] R. H. Perry, D. W. Green, and J. D. Maloney, eds., *Perry's Chemical Engineers Handbook,* McGraw-Hill, New York, 1984.

[12] A. G. Osborn and D. R. Douslin, "Vapor pressure relations of 36 sulfur compounds present in petroleum," *J. Chem. Eng. Data,* Vol. 11, 1966, pp. 502–507.

BIOGRAPHIES

Mordechai Shacham is a professor in and head of 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 methods, computer-aided instruction, and chemical-process simulation, design, and optimization. He is a coauthor of the POLYMATH numerical software package.

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 an associate professor in the Fluid Mechanics and Heat Transfer Department and she serves as the president of the Israel Institute of Chemical Engineers. She teaches courses in Mass and Heat Transfer and Process Control. Her main research interests include two-phase flows and transport phenomena in thin films.

Michael B. Cutlip is a professor of chemical engineering at the University of Connecticut, where he has served since 1968 at every academic level including a nine-year term as department head. He received his BChE and MS from The Ohio State University and his PhD from the University of Colorado. He is coauthor of the POLYMATH numerical computation package with Mordechai Shacham and is the immediate past president of the CACHE Corporation (Computer Aids for Chemical Engineering Education). He is currently involved with the production of a CD-ROM for chemical engineering students, which is available through CACHE. His laboratory research encompasses chemical reaction engineering with a current interest in photocatalytic oxidation of ozone-depleting chemicals and the electrocatalytic processes in fuel cells.