18. de los Reyes, M.F., F.L. de los Reyes, M. Hernandez, and L. Raskin, "Quantification of Gordona amarae Strains in Foaming Activated Sludge and Anaerobic Digester Systems with Oligonucleotide Hybridization Probes," *Appl. Environ. Microbiol.*, **(64)**, p. 2503, (1998)
19. E-C Apparatus Corp., Holbrook, NY
20. Promega, Inc., Madison, WI
21. Cel-Line Associates, New Field, NJ
22. Oerther, D.B., J. Pernthaler, A. Schramm, R. Amann, and L. Raskin, "Monitoring Precursor 16S rRNA of Acinetobacter spp. in Activated Sludge Wastewater Treatment Systems," *Appl. Environ. Microbiol.*, **(66)**, p. 2154 (2000)
23. Type FF, Cedar Grove, NJ
24. Nikon Instruments, Inc., Melville, NY
25. Diagnostic Instruments, Inc. Sterling Heights, MI ❐

ChE letter to the editor

To The Editor:

This letter is motivated by the paper "An Undergraduate Course in Applied Probability and Statistics" that appeared in the Spring 2002 issue of *Chemical Engineering Education.*[1] Probability and statistics are difficult subjects to teach to engineering students, and Professor Fahidy is to be congratulated on his efforts in this area.

In this letter we would like to refer to the discussion and examples related to regression analysis. Professor Fahidy discusses in detail the use of numeric information (such as error variance, confidence intervals, correlation coefficient, etc.) for regression analysis, but does not mention graphic information (residual plots) and physical insight for regression analysis. Using the examples presented by Professor Fahidy,[1] we would like to demonstrate the importance of including graphical information and physical arguments in the regression analysis.

Let us refer first to Example 4 in the paper. In this example, the integral method of rate data analysis is used for a (supposedly) first-order reaction. Nonlinear regression can be used for finding the reaction rate coefficient (k) using concentration (Y) versus time (t) data, on the regression model $Y = \exp(-kt)$. Alternatively, this equation can be linearized to yield $\ln Y = -kt$, where linear regression can be applied. The results of the linear and nonlinear regression that were obtained using POLYMATH 5.1 are shown in the first two columns of Table 1. Note that these results are different from what is presented in [1], but they are correct and were confirmed by the author of the original article.[2] Looking at the numerical information presented in Table 1 (parameter values, confidence intervals, correlation coefficients, and variances) leads to the conclusion that there is no significant difference between linear and nonlinear regression for determining k (the variances are almost the same, contrary to what is argued in [1]). The same information may also lead to the conclusion that the model fits the data reasonably well. This conclusion, however, is contradicted by the residual plot shown in Figure 1. The residuals are not randomly distributed around a zero value. This may indicate either lack of fit of the model, or that the underlying assumption of a random error distribution for the dependent variables is incorrect.

TABLE 1
Regression Results for Example 4 in Reference 1

Reaction Order	*1st Order*	*1st Order*	*0th Order*	*2nd Order*
Model	$\log Y = -k*t$	$Y = \exp(-k*t)$	$Y = Y_0 + k*t$	$1/Y = 1/Y_0 + k*t$
k (value)	0.0039888	0.0038126	-0.0042162	0.0059893
95% Conf. Interval	±0.0011009	±0.0010816	0.0015209	±0.0059893
Y_0 (or $1/Y_0$, value)	-	-	1.0329275	0.9365288
95% Conf. interval	-	-	±0.586582	±0.1012594
R^2	0.7620164	0.7770319	00.8362884	0.7757433
Variance (based on Y)	0.0023055	0.002271	0.0018759	0.0021994

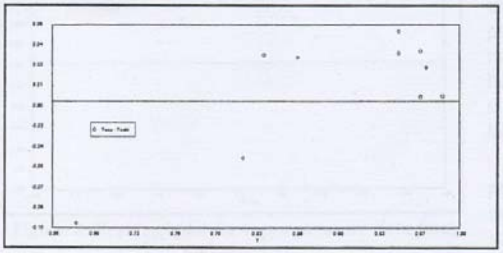

Figure 1. *Residual plot for Example 4 in Fahidy paper.*[1]

Physical insight can suggest alternative regression models, but more information regarding the reaction involved is needed. Since no such information is available, we will assume a homogeneous reaction, just for the sake of the demonstration. Assuming 0th order reaction or 2nd order reaction yields the models shown in the third and fourth columns of Table 1, respectively. The numeric information presented in the Table points on the 0th order reaction as the most appropriate one (smallest variance value—note that in order to be on a unique scale, all the variance calculations must be based on Y). The residual plot for the 0th order reaction is not significantly different, however, from that shown in Figure 1; thus, this model is not supported by the residual plot either.

The conclusion from proper analysis of this example is that the data available are insufficient (in quality, quantity, or both) to determine in any certainty the order of the reaction it represents. To obtain a more definite result, additional measurements must be made.

In Example 5, a linear model $Y = a + bx$ is fitted to data of mean fuel consumption rate (Y) versus vehicle mass (x). The numerical results that were obtained for this example, using POLYMATH, are: parameter values (including 95% confidence intervals) a=-0.8695975±2.0733031;

Continued on page 277

Letter to the Editor

Continued from page 262.

b=8.5164364±1.5315505; the error variance s^2=0.467503; and correlation coefficient R^2=0.953603. Professor Fahidy advises not to put too much faith in the linear regression model, in spite of the relatively large R^2 value, because of the extremely wide confidence intervals on the parameter a. The fairly random distribution of the residuals (see Figure 2) suggests, however, that the linear model may be the correct one. Furthermore, both physical considerations (fuel consumption should be zero for a zero mass vehicle) and the wide confidence intervals on the free parameter a, indicate that the model can be improved by setting the free parameter at zero. Indeed, carrying out the regression while setting a=0 yields: b=7.892916±0.3599903; s^2=0.4641509, and R^2=0.9481781. Thus, this model is now acceptable, even with respect to the confidence interval values.

One of Professor Fahidy's objectives in presenting this example was to warn against accepting relatively large R^2 values as proof of good linear relationship between the dependent and independent variables. The limitations of the R^2 statistics in this respect can be most strikingly demonstrated using residual plots. Shacham, et al.,[3] for example, fitted vapor pressure data of 1-propanol with the two-parameter Clapeyron equation. This regression yields the values: R^2=0.9998818 and s^2=1.659E-05 (based on log P). Such a high value of R^2 can be interpreted as a perfect fit. But the residual plot (seen in Figure 3) shows that the vapor pressure data set exhibits a curvature, which is not predicted by the Clapeyron equation. Indeed, using the four-parameter Riedel equation for representation of the same data yields: R^2=1; s^2=1.327E-09 and randomly distributed residuals.

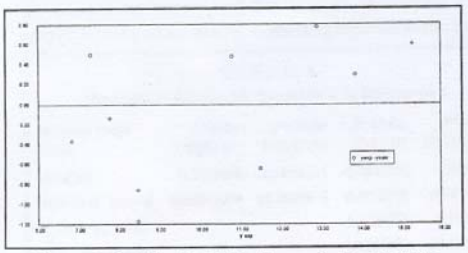

Figure 2. *Residual plot for Example 5 in Fahidy paper.*[1]

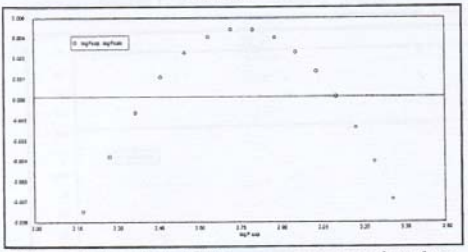

Figure 3. *Residual plot for vapor pressure data from Reference 3.*
Regression model: log P = 7.6380342-1622.8666/T

The last example, given in the Appendix of the article deals with a linear model for representing coded effectiveness indicators versus catalysts containing various coded platinum mass units. Analysis of this example shows that if the free parameter, a, is set at zero (as suggested by the wide confidence intervals on a and physical considerations) the linear model is appropriate to represent the data with ß=1.6437659±0.0845917, R^2=0.8860414, and s^2=0.8508906.

We can conclude that teaching statistical analysis of data and regression models is very important, but interpretation of numeric statistical indicators must be complemented by graphical analysis and consideration of the physical nature of the model in order to arrive at the correct conclusions.

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

References

1. T.Z., "An Undergraduate Course in Applied Probability and Statistics," *Chem. Eng. Ed.*, **36**(2), 170 (2002)
2. Fahidy, T.Z., Personal communication (2002)
3. Shacham, M., N. Brauner, and M.B. Cutlip, "Replacing the Graph Paper with Interactive Software in Modeling and Analysis of Experimental Data," *Comp. Appl. Eng. Ed.*, **4**(1), 241 (1996) ❐

Author's Response

I am delighted at Professor Shacham's interest in my paper. I also fully concur with the argument that the residual plots are an important and integral part of regression analysis. This is now standard textbook material, and I do routinely discuss this subject in my course. Although my intention was to keep the article from being too long, in retrospect I should have spent a paragraph or two on residual analysis, and I regret the omission.

In Example 4 it was stated that the reaction mechanism was first-order irreversible, but perhaps not strongly enough to imply an *a priori* knowledge of non-statistical origin, so that 0^{th} and 2^{nd} order models are beyond consideration. With limited data and given a physically correct model, the method that provides regression parameters relating data to model with the smallest error variance may be acceptable in lack of something better, even if the residual plot does not show randomness of a desired degree. The quest for additional measurements is almost universal in the case of limited-size data.

My views about R2 versus confidence intervals for true regression parameters do not fully coincide with the respondents', but may I point out the redundancy of seven-digit values, computer printouts notwithstanding. An R2=0.8860414 is not more meaningful than R2=0.89

Thomas Z. Fahidy