The object of this column is to enhance our readers' collections of interesting and novel problems in chemical engineering. Problems of the type that can be used to motivate the student by presenting a particular principle in class, or in a new light, or that can be assigned as a novel home problem, are requested, as well as those that are more traditional in nature and that elucidate difficult concepts. Manuscripts should not exceed fourteen double-spaced pages and should be accompanied by the originals of any figures or photographs. Please submit them to Professor James O. Wilkes (e-mail: wilkes@umich.edu), Chemical Engineering Department, University of Michigan, Ann Arbor, MI 48109-2136.

DATA ANALYSIS MADE EASY WITH DATAFIT

JAMES R. BRENNER
Florida Institute of Technology • Melbourne, FL 32901

Shortly after starting as an assistant professor, I realized that quite a few of our students were unable to analyze laboratory data at a level consistent with that expected when I had worked in industry. Having been put in charge of the Florida Institute of Technology's introductory chemical engineering course and its materials science and engineering laboratory course, I decided that a strong emphasis on data analysis would be added to each of these courses in order to satisfy ABET's requirement regarding the ability of students to analyze data.

Most departments emphasize spreadsheet calculations and plotting of data in Microsoft Excel as part of their introductory chemical engineering course. Experience in our department has shown that unless sufficient time is spent on data analysis instruction such that spreadsheet calculations, plotting, and curve fitting become second nature, such skills are either forgotten or are never learned properly.

We have incorporated DataFit from Oakdale Engineering[1] throughout the entire curriculum at Florida Tech, beginning with ChE 1102, an eight-week, one-day-per-week, two-hour, one-credit-hour, second-semester Introduction to Chemical Engineering course in a hands-on computer classroom. The syllabus for CHE 1102 is shown in Table 1. The examples

James R. Brenner *received his B.S. degree from The University of Delaware and M.S. and Ph.D. degrees from The University of Michigan. After a postdoc at Argonne National Laboratory and industrial experience at Westinghouse Savannah River Company, he became an assistant professor of chemical engineering at Florida Institute of Technology. His research interests are in hydrogen purification and sensing, electronic noses, and nanoporous materials.*

chosen, shown in parentheses, are selected so as to be consistent with concepts that students learn concurrently in other courses. DataFit also has become commonly used in our Physical Chemistry Lab and Materials Science and Engineering Lab courses, as well as in several courses in other engineering departments. Our experience at Florida Tech is that students retain data analysis concepts best when such concepts are formally taught to them in this short course and then periodically reinforced throughout their academic career. Several examples covered in weeks three through eight will be discussed here.

TABLE 1
Data Analysis Curriculum

1) Statistics and Confidence Intervals
2) Introduction to Plotting and Calculations in Excel
3) y = ax + b Fitting in DataFit (Pressure Transducer Calibration)
4) y = ax Requires User-Defined Models (Hygrometer Calibration)
5) Semi-Log Functions (First-Order Rate Laws - Felder & Rousseau 2.34)
6) Plotting and Curve Fitting of Power-Law Functions (Crystal Growth - Felder and Rousseau 2.37)
7) Nonlinear Functions (Vapor Pressures)
8) Curve-Fitting in 3-D (Rate Laws with 2 Reactants)

An introduction to basic statistics is included in nearly all introductory ChE courses and will not be discussed in this article. Students in ChE 1102 cover basic statistics during the first week of the course and get constant reinforcement of these concepts through the use of DataFit.[1] The second half of ChE 1102 consists of problems that require Polymath- or Excel-based solutions to either sets of linear and nonlinear algebraic equations or numeric integration, as suggested by Clough.[2]

All Excel and DataFit files are available at <http://my.fit.edu/~jbrenner/dataanalysispaper1.zip>.

Experience in our department has shown that unless sufficient time is spent on data analysis instruction such that spreadsheet calculations, plotting, and curve fitting become second nature, such skills are either forgotten or are never learned properly.

SOLVING PROBLEMS WITH DATAFIT

Problem 1. Calibration of a Pressure Transducer

Following the introduction to basic statistics, the first problem that I assign students is the calibration of a 0-250 psig Span Instruments NTT-204 (now Millipore) pressure transducer against a 0-1000 psia Paroscientific pressure transmitter. In addition to being useful for teaching students how to make plots with error bars and determine the difference between absolute and gauge pressures, it provides a relatively simple problem for studying linear regression with DataFit. The repeatability and lack of drift of Paroscientific pressure transmitters is even superior to that of a deadweight tester that was calibrated at NIST.[3] The repeatability of the quartz oscillator that the Paroscientific pressure transmitters use is certainly within the quoted 0.01% of full-scale precision (*i.e.*, 0.1 psia fixed error for a 1000-psia transmitter). Span Instruments' pressure transducers output a signal that ranges from 4-20 milliamps to within 0.08 milliamps.

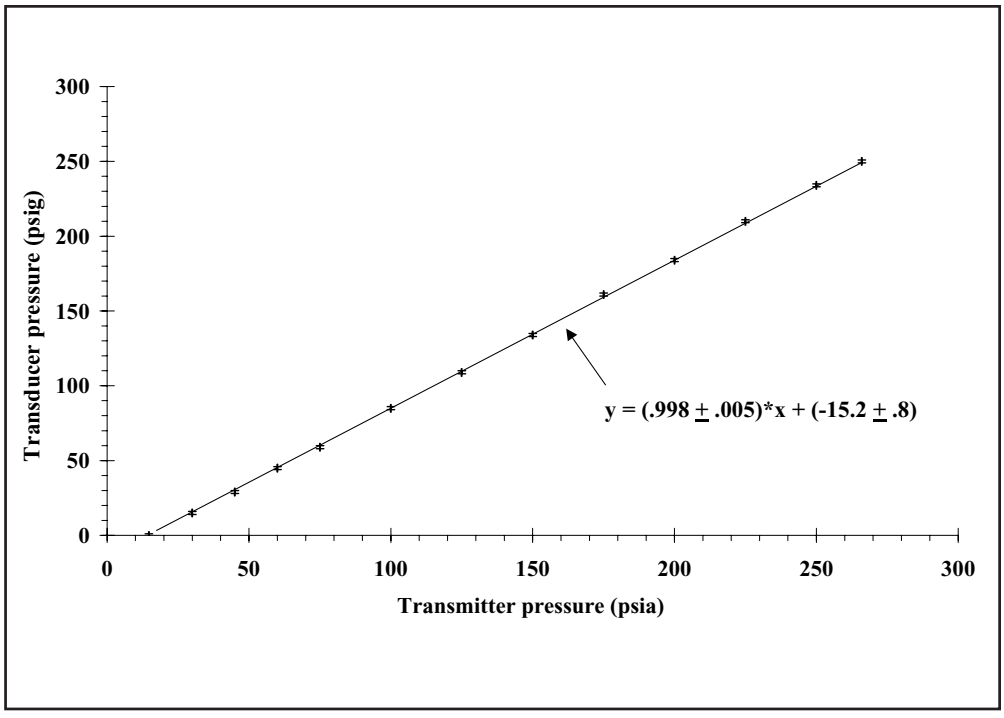

Figure 1 . *Calibration of a Span Instruments pressure transducer against a NIST-traceable Paroscientific pressure transmitter.*

After having the students prepare a plot of the data shown in Figure 1, including error bars, the students copy and paste the data into

DataFit, click on the Solve Regression option, click on OK, select the y = ax + b option, and let DataFit do the work for them. After clicking on Results Detailed, the Fit Information output is obtained (Table 2). Included in the output are the residual sum of squares (RSS), which is the sum of the squares of the differences between the calculated values of Y, the pressure in psig as determined by the pressure transducer, and the corresponding experimental values. Also evaluated are the commonly seen R2 correlation parameter, as well as several more-advanced goodness-of-fit parameters. Most importantly, the 68%, 90%, 95%, and 99% confidence intervals are conveniently tabulated. This is an excellent opportunity to reinforce basic statistics, most notably the Gaussian distribution, which is typically taught at the beginning of ChE 1102.

Problem 2. Calibration of a Hygrometer

The second problem that I assign is Problem 2.32 from Felder and Rousseau's textbook.[4] This problem involves the correlation of a signal from a hygrometer versus the mass fraction of water in the inlet stream to the hygrometer. For this problem, first ask students to do the y = ax + b fit as described in the previous section. The 95% confidence intervals on the slope, a, and the intercept, b, are as follows: a = 470 ± 20; b = 0 ± 2, at the appropriate number of significant figures; proper use of significant figures is an extremely difficult concept to get students to consistently apply. Then ask them whether the intercept, b, is mathematically significant (*i.e.*, nonzero within the 95% confidence interval). They should answer that b is not mathematically significant at the 95% confidence level. Out of a sample of 100 students asked over the last five years as part of an in-class exercise, only 50% have answered correctly to this question; 25% of students replied "don't know." This is a surprisingly difficult concept to master that requires consistent reinforcement throughout ChE 1102. Yet, of the same sample of students, 98% replied correctly to a similar question during hourly and final exams.

Once the students have realized that b is unnecessary, it is time to teach them how to create a user-defined model in DataFit, as y = ax is not one of the built-in models (one of DataFit's few shortcomings). This can be done by returning to DataFit's main menu and clicking on the Define User Model option under the Solve menu. The user defines a Model ID, (which I defined as "Linear, no intercept," in this case). The user also inputs the Model Definition, in this case Y = a*x. Mathematical functions in DataFit, such as multiplication and exponentiation, work in the same way as Excel.

In many cases, including this one and all cases where the fitting is of a linear function, initial estimates are unneces-

TABLE 2
Fit Information for Pressure Transducer Calibration

DataFit version 6.1.10
Results from project
"F:\brenner\datafit\pcalib.dft"
Equation ID: a*x+b

Number of observations = 13
Number of missing observations = 0
Solver type: Nonlinear
Nonlinear iteration limit = 2000
Diverging nonlinear iteration limit =10
Number of nonlinear iterations performed = 1
Residual tolerance = 0.0000000001

Sum of Residuals = 4.08562073062058E-14
Average Residual = 3.14278517740044E-15
Residual Sum of Squares (Absolute) = 5.20799168906741
Residual Sum of Squares (Relative) = 5.20799168906741
Standard Error of the Estimate = 0.688079784556427
Coefficient of Multiple Determination (R^2) = 0.99994096
Proportion of Variance Explained = 99.994096%
Adjusted coefficient of multiple determination (Ra^2) = 0.9999355927
Durbin-Watson statistic = 2.88469613789683

Regression Variable Results

Variable	Value	Standard Errort-ratio	Prob(t)
a	0.998001779	0.002312177	431.6287071 0
b	-15.1762779	0.359917526	-42.1659876 0

68% Confidence Intervals

Variable	Value	68% (+/-)	Lower Limit	Upper Limit
a	0.998001779	0.002408132	0.995593648	1.000409911
b	-15.1762779	0.374854103	-15.551132	-14.8014238

90% Confidence Intervals

Variable	Value	90% (+/-)	Lower Limit	Upper Limit
a	0.998001779	0.004152438	0.993849342	1.002154217
b	-15.1762779	0.646375884	-15.8226538	-14.529902

95% Confidence Intervals

Variable	Value	95% (+/-)	Lower Limit	Upper Limit
a	0.998001779	0.005089101	0.992912679	1.00309088
b	-15.1762779	0.792178474	-15.9684564	-14.3840995

99% Confidence Intervals

Variable	Value	99% (+/-)	Lower Limit	Upper Limit
a	0.998001779	0.007181158	0.990820622	1.005182937
b	-15.1762779	1.117831851	-16.2941098	-14.0584461

Variance Analysis

Source	DF	Sum of Square	Mean Square	F Ratio	Prob(F)
Regression	1	88206.02278	88206.02278	186303.3408	0
Error	11	5.207991689	0.47345379		
Total	12	88211.23077			

sary, but they become critical when doing some nonlinear fitting. The default values of each of the curve-fit parameters are unity in all cases. I look at this as one of DataFit's very few design flaws. When one goes through a Taylor series expansion, terms involving higher-order parameters are supposed to be corrections to the previous terms, meaning that the *product* of the curve-fit coefficient multiplying a high-order term and that higher-order term (*i.e.*, d*x3) should be less than those of previous terms. Without some exceptional physical justification, it would be difficult to throw out constant, linear, or quadratic terms and keep a cubic term.

After manually assigning initial estimates and/or constraints on the curve-fit coefficients, clicking OK, clicking Solve Regression, and OK again, the user will need to locate his or her user-defined model in the list of models. After locating your recently defined model, click on Solve, click OK, and then click on Results Detailed to return to the Fit Information screen once again. The models are ranked by the RSS, and so the Fit Information that pops up first is the one with the lowest RSS, not the one for the most recent fit. By clicking on the uppermost dialog box to locate the user-defined model, one will get the Fit Information associated with the user-defined model, "Linear, no intercept." Interestingly, scrolling down to the 95% confidence interval shows that the confidence interval for the one-parameter model (a = 473 ± 8) is narrower than the slope from the two-parameter model (a = 470 ± 20).

Problem 3. Fitting Water Vapor Pressures to the Clausius-Clapeyron and Antoine Equations

Fitting water vapor-pressure data to the Clausius-Clapeyron equation is challenging for underclassmen, but usually can be done successfully if the previous examples have been worked out in class or for homework. This problem, along with the follow-up fitting of the same data to the Antoine equation, typically is either the final in-class or homework problem that students are asked to solve during CHE 1102. Data for the vapor pressure of water is tabulated in Appendix B.3 of Felder and Rousseau.[4] The Clausius-Clapeyron equation is as follows, and requires conversion of temperatures into Kelvin:

$$\log_{10} P = A - \frac{B}{T} \qquad (1)$$

At this point in the course, the students know that they should plot pressure on a logarithmic scale on the y-axis and reciprocal temperature on the x-axis. Students are asked to plot 1,000/T so that the values on the x-axis are between a more aesthetically pleasing 0 and 10, to estimate the slope (-B) and the intercept (A) graphically, to use DataFit to determine A and B, and finally to superimpose the curve fit (the solid line) on top of the experimental points (Figure 2).

The Clausius-Clapeyron equation is a reasonably good fit of the vapor pressure of water data from 0 to 60 °C, but one can see that there is a systematic deviation from linearity at low temperature and pressure. By graphically extrapolating

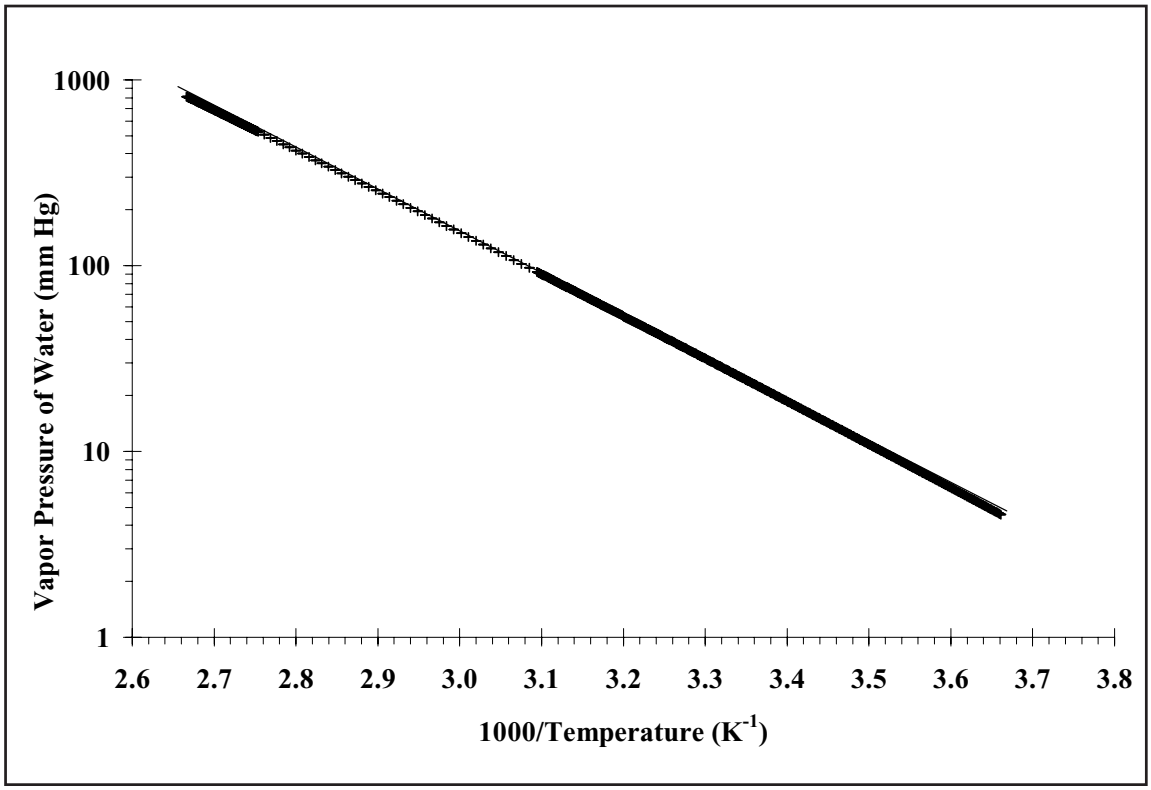

Figure 2. *Clausius-Clapeyron plot for water vapor pressures.*[4]

a straight line through the portion of the data that appears to be linear, one can estimate the slope (-B) as -2200 and the intercept (A) as 109 from Figure 2. Interestingly, there are slight differences in the DataFit estimates of the curve fit parameters, depending on whether the logarithm of the pressure data and the inversion of the temperature data are taken before curve fitting in DataFit or not (Table 2). In the case where the data are not so linearized before entry into DataFit and then a nonlinear model is generated in DataFit, the points at low vapor pressures are de-emphasized relative to the other points.

If one tries to fit the Antoine equation for water vapor pressures either below 60 °C or above 60 °C, in either case if one does not manually change the default parameter guesses of unity, DataFit's "solution" will require more iterations than the default number of iterations, which is 250.

$$\log_{10} P = A - \frac{B}{(T+C)} \qquad (2)$$

This problem can be changed using Edit Preferences. I have changed the default number of iterations permanently to 2,000.

The problem with using the results for A and B from the Clausius-Clapeyron equation as initial guesses for A and B for the Antoine equation fit is that the Antoine equation requires temperatures to be in degrees Celsius instead of in Kelvin. In fact, if one uses the Clausius-Clapeyron equation constants to fit the water-vapor pressures above 60°C and lets DataFit set the default value of C to 1, then even after having made the appropriate conversion of the data from Kelvin into Celsius, DataFit will erroneously return a "successful" result after only one iteration that contains errors larger than the values of the parameters themselves. The Antoine equation *cannot be solved* for temperature ranges in which the denominator, (T+C), switches from negative to positive *over the range of temperatures*. If one uses the values of A and B from the Clausius-Clapeyron equation and an initial guess for C of 273.15, then the Antoine equation does converge properly to the answers below in Table 3 in the "Proper Convergence" column.

This discrepancy proved a difficult challenge for even the best students. Due to the sensitivity of the parameters for the Clausius-Clapeyron fits as to whether the data was linearized or not, and due to the slight discrepancies between the results in Table 3 and those in the literature,[4, 5] I now wonder whether the previously reported Antoine constants for other molecules may be slightly off as well.[4, 5] At a minimum, it appears the number of significant figures reported for Antoine equation constants in the literature[4, 5] is grossly overstated.

TABLE 3
Clausius-Clapeyron Constants for Vapor Pressure of Water from 0 to 60°C

Clausius-Clapeyron Constants	Linear Fit of Linearized Data	Nonlinear Fit of of Raw Data
A	9.091± 0.004	9.003 ± 0.004
B	2301± 1	2274 ± 1

TABLE 4
Antoine Curve Fitting of Vapor Pressure of Water from 0 to 60°C

Constants	250 iterations	Proper Convergence	Literature Data[4, 5]
A	6.95 ± 0.08	8.124 ± 0.002	8.10765
B	1180 ± 40	1759.8 ± 0.6	1750.286
C	186 ± 4	235.8 ± 0.1	235.000

TABLE 5
Clausius-Clapeyron Equation Parameters*

Molecule	A^L	B^L	A^N	B^N
Carbon Dioxide	7.58 ± 0.02	865 ± 4	7.58 ± 0.01	864 ± 3
Ethane	7.37 ± 0.05	837 ± 9	7.127 ± 0.008	785 ± 2
Propane	7.71 ± 0.08	1130 ± 14	7.191 ± 0.007	1128 ± 3
Isobutane	7.69 ± 0.07	1274 ± 16	7.198 ± 0.007	996 ± 2
Butane	7.61 ± 0.06	1306 ± 7	7.256 ± 0.009	1193 ± 4

*Pressures in mm Hg and temperatures in Kelvin
[L]Logarithm of pressure taken first
[N]Logarithm of pressure not taken first

ASSESSMENT

In the first class exposed to this curriculum, 17 of 20 students successfully completed both the Clausius-Clapeyron and Antoine problems. Two of the three students who failed to make a proper plot and a proper fit in DataFit attended class less than one-third of the time, and the other student, although in good attendance, turned in less than half of the homework assignments and had significant language problems. The past four years of classes have had similar results.

A similar problem, for butane vapor

pressures, has been assigned to sophomores and graduate students, using data from the NIST Chemistry WebBook.[12] All but one of 12 sampled students who came to Florida Tech from other countries for ChE graduate school sought me out for help. None of the eight students that went to Florida Tech for both bachelor's and master's degrees needed help. Ninety percent of sophomore students who took CHE 1102 as freshmen were also able to solve the butane problem successfully.

With the default guesses, DataFit failed to converge because it cannot handle the denominator changing from negative to positive, depending on temperature. When the second term exceeds A, the solution also diverges. Under some sets of initial estimates, DataFit "converges" to a flat line! When the initial estimates are reasonably close to what DataFit reports as the correct answer (A = 7.44 ± 0.04; B = 1330 ± 30; C = 294 ± 4), the solution converges to what is shown in Figure 3. *Even this is clearly incorrect*, as the low vapor pressure data is de-emphasized, because the magnitude of the error in such a small quantity is dwarfed by a small percentage error in the high vapor pressure points. This kind of error is not unique to DataFit. I have seen it in Polymath curve fits as well.

CONCLUSIONS

Of the international graduate students asked to fit vapor-pressure data for the previous problem, none had previous exposure to either Polymath or DataFit. While each of them also learned how to use Polymath in graduate school, 11 of the 12 polled said that they found DataFit easier to use. The reason that I downloaded DataFit in the first place was not because of its excellent curve-fitting capabilities, but because when I first started using it in industry in 1998, DataFit was the only program that did proper 3-D scientific plotting for less than $500. In 1999, when Florida Tech bought a site license for DataFit version 6.1, it cost only $750 for the entire campus (albeit a relatively small campus), whereas a single copy cost $100. Moreover, the site license allowed for students and faculty to use DataFit at home as long as they were doing academic work. A comprehensive set of solutions to similar problems can be found at <http://my.fit.edu/~jbrenner/dataanalysispaper1.zip>.

REFERENCES

1. Gilmore, J., DataFit, v 6.1, Oakdale Engineering, 23 Tomey Road, Oakdale, PA 15071, (724) 693-0320, sales@curvefitting.com, <http://www.oakdaleengr.com>.
2. Clough, D., "Spreadsheets Across the Curriculum," ASEE Summer School for ChE Faculty, July (2001).
3. Brenner, J.R., and E.F. Dyer, Westinghouse Savannah River Company, unpublished results, December (1997)
4. Felder, R.M., and R.W. Rousseau, *Elementary Principles of Chemical Processes*, John Wiley & Sons, 3rd Ed., New York (2000)
5. Dean, J.A., *Lange's Handbook of Chemistry*, McGraw-Hill Companies, Inc., 14th Ed., New York (1992)
6. NIST Chemistry Webbook, <http://webbook.nist.gov/chemistry> ❐

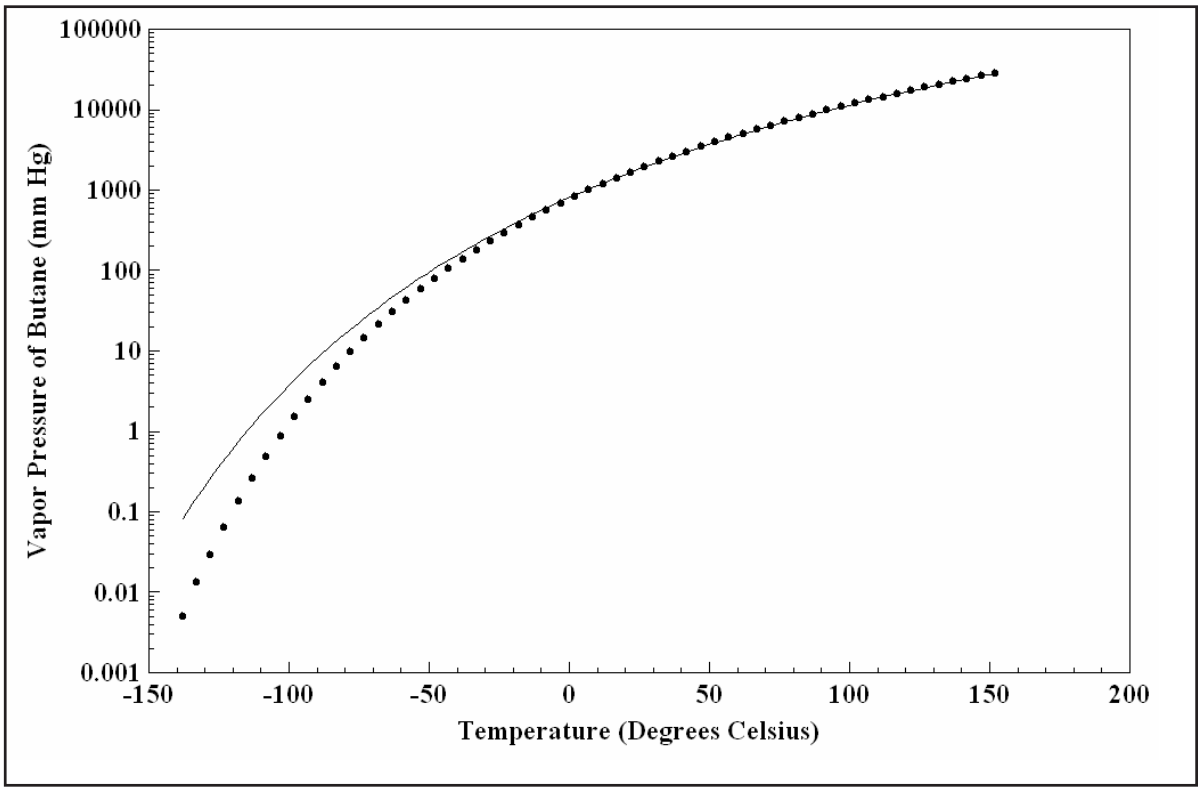

Figure 3. *Antoine fit of butane vapor pressure data clearly shows bias against low vapor pressure points.*