Data table or interpolation

Héctor G. Riveros
Instituto de Física UNAM, Ciudad Universitaria, CP01000, México.

E-mail: riveros@fisica.unam.mx

ISSN 1870-9095

(Received 20 February 2020, accepted 15 September 2020)

Abstract

The results of an experiment can be presented as a data table or as an equation that represents them. In the case of adjusting a polynomial, Excel allows us to change its degree and calculates the R^2 of the adjusted equation. It is considered that if it is 1 the equation goes through all the experimental points. By adjusting water vapor pressure data, it is found that the equations do not go through all the points (even with $R^2 = 1$), which is verified by calculating the differences for each point. In those cases, the best fit is the linear interpolation between consecutive points. The equation adjusted by Excel or Origin requires checking if it corresponds to the minimum in the sum of squared differences, it is possible that it can be reduced by changing the coefficients values.

Keywords: equation adjustment, interpolation, uncertainties.

Resumen

Los resultados de un experimento se pueden presentar como una tabla de datos o como una ecuación que los represente. En el caso de ajustar un polinomio, Excel nos permite cambiar su grado y calcula el R2 de la ecuación ajustada. Se considera que si es 1 la ecuación pasa por todos los puntos experimentales. Al ajustar los datos de presión de vapor de agua, se encuentra que las ecuaciones no pasan por todos los puntos (incluso con R2 = 1), lo cual se verifica calculando las diferencias para cada punto. En esos casos, el mejor ajuste es la interpolación lineal entre puntos consecutivos. La ecuación ajustada por Excel u Origin requiere verificar si corresponde al mínimo en la suma de diferencias al cuadrado, es posible que se pueda reducir cambiando los valores de los coeficientes.

Palabras clave: ajuste de ecuaciones, interpolación, incertidumbres.

I. INTRODUCTION

Experiments are performed to verify the predictions of a theory or to find an empirical relationship between variables. It is said that there is an agreement between theory and experiment if the calculated data is within the uncertainty of the measured data. If there is no agreement, it is necessary to review the theory and/or review the experiment. If an empirical relationship is sought, an equation is sought that goes through the experimental data with its uncertainty. The one that is simpler is chosen, if we do not have a theory that suggests the equation. It is at the discretion of the researcher if he presents his results in a table or in an equation, or in both presentations.

There are books that include the adjustment of equations [1] and articles that solve relevant details. We can mention the titles of some articles: True lines [2], Systematic errors and graphic extrapolation [3], Measurement of systematic errors with curve fit [4], Can students draw better fit lines? [5], The art of adjusting models to experimental results [6], Comparison of different approaches in the extraction of a parameter in a linear adjustment [6], and Analysis of data and graphs in an introductory physics laboratory: spreadsheet versus statistics suite [8] Some mention the R^2 but do not mention the need to verify the goodness of the fit, plotting the residuals. We have Excel and Origin that adjust different curves and calculate the parameters that give the minimum of the sum of the squared residues. Peterlin [8] has:

$$R^2 = 1 - [\sum (Yi - fi)^2] / \sum (Yi - \langle Y \rangle)^2 , \quad (1)$$

where fi is the calculated Y value and <Y> is its average. If $R^2 = 1$, it implies that all the residuals are zero, that is, it passes through all the data points. The best fit is the one with R^2 closer to 1.

On the Internet we have the blog of Minitab [9] that says:

"The adjusted line graph shows that this data follows a good adjusted function and the R square is 98.5%, which sounds great. However, look more closely to see how the regression line systematically over and under-predicts the data (bias) at different points along the curve. You can also see patterns on the Residual versus Fits chart, instead of the

randomness you want to see. This indicates a bad adjustment and serves as a reminder of why you should always check the waste charts."

Frost [10] says: "At first glance, R-square seems an easy to understand statistic that indicates how well a regression model fits a set of data. However, he doesn't tell us the whole story".

II. ADJUSTING THE WATER VAPOR PRESSURE

The properties of the materials and their variation with temperature are usually presented in the form of tables. Water vapor pressure is one of these properties and we can find its value in internet [11]. All digits in a table are expected to be significant, the uncertainty implied in the value 4.58447 varies from 4.58448 to 4.58446.

But if we need the water vapor pressure at other temperatures, we need to interpolate or find the adjusted equation that represents them. If we graph the data with Excel, we can adjust different types of equations and calculate the R^2 that indicates how good the fit is. If the R^2 value is 1, the calculated and measured values are equal, and the curve passes through all the experimental points. Figure one shows 3 Excel polynomial fits that seem to all go through the experimental points.

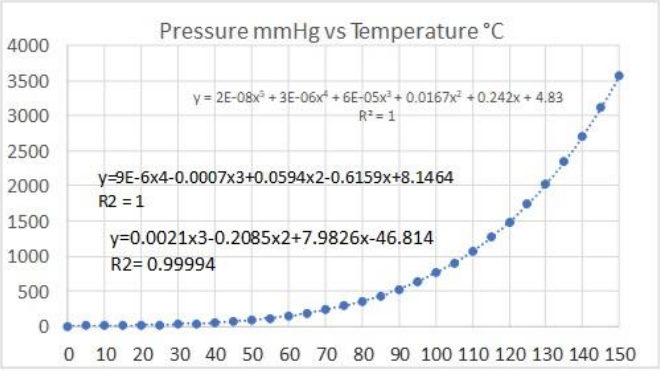

FIGURE 1. Water vapor pressure versus Temperature and 3 fitted equations.

III. HOW DO WE FIND THE BEST FIT?

The procedure consists in comparing the different adjusted equations and see which one looks more like the original data. But the graphs show that they all go through the experimental points, so it is necessary to measure the differences with the original data.

The linear adjustment and the second-degree polynomial don´t fit the data. Polynomials of 3,4- and 5-degree pass through the points plotted in Figure 1.

To be able to appreciate which one is the best, it is necessary to calculate the difference of the experimental vapor pressures and the calculated values. Figure 2 shows the result.

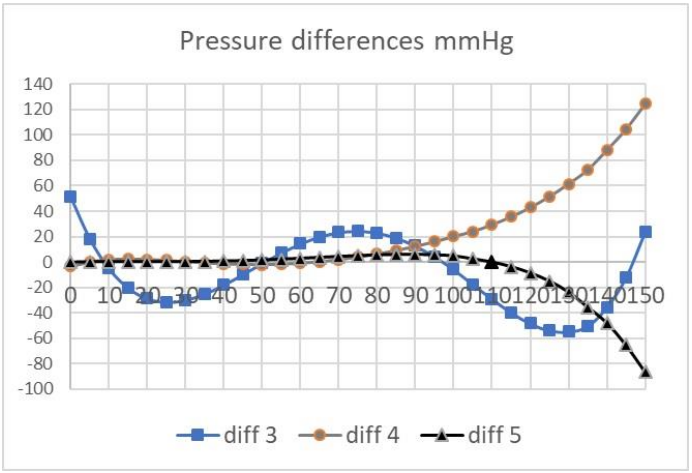

FIGURE 2. Difference between the measured vapor pressure and calculated values. Polynomials of 3, 4 and 5 degrees against temperature in °C.

The best fit is the polynomial of degree 5. Configuration 3 has 51 mmHg as the pressure difference in the first data. Note that configurations 4 and 5 have $R^2 = 1$, but their data does not pass through the experimental points. The differences graph allows us to distinguish between random errors and systematic deviations of the points. The best fit is calculated as the one with the smallest sum of the squares of the differences.

The sum of the squares of the differences is 25571 for polynomial 3, 51478 for polynomial 4 and 16378 for polynomial 5. Using a polynomial of greater degree improve the fit of the data. Excel rounds R^2 to 1 if it has more than four nines (0.99996).

The equations adjusted by Excel or Origin requires checking if it corresponds to the minimum in the sum of squared differences, it is possible that it can be reduced by changing the coefficients values of the fitted equations. Reviewing the calculations, it was found that polynomial 3 was well adjusted, but for polynomials 4 and 5 the fit could be improved. A well-made fit is found at the minimum of the sum of the squared differences. We find that by changing the coefficients you can find lower values for the sum of the squares. Table I mentions these values.

The first three rows show the values of the Excel equations and the lines 4, 5 and 6 show the optimized values of the equations. But polynomial 3 sum is 25571 and optimized is 19962. For polynomial 4 the sum was 51478 and optimized is 319. For polynomial 5 sum is 16378 and optimized is 1687. Now the best fit is polynomial 4. The figure 3 shows the differences for the new settings.

TABLE I. Values of the adjusted coefficients by changing their values. The sum of the squared differences is different from zero, as it would be if the measured and calculated values were equal.

x^5	x^4	x^3	x^2	x	constant	Sum Diff^2
Excel		0.0021	-0.2085	7.9826	-46.814	25571
	9.00E-06	-0.0007	0.0594	-0.6159	8.146	51478
2.00E-08	3.00E-06	6.00E-05	0.0167	0.242	4.830	16378
Excel		0.00209	-0.2085	7.9808	-46.800	19962
Optimized	9.23E-06	-0.0007	0.0593	-0.6289	8.170	319
1.90E-08	3.00E-06	0.000064	0.0168	0.254	5.600	1687

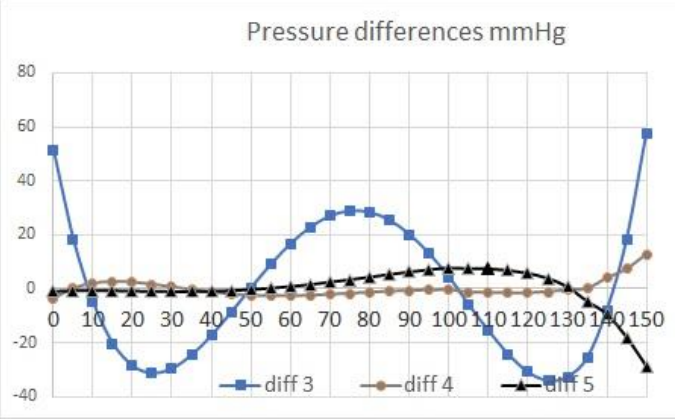

FIGURE 3. Vapor pressure difference between the measured and calculated optimized values. Polynomials of 3, 4 and 5 degrees against temperature in °C. The best fit is polynomial 4, but interpolation it is more precise.

Using Origin, we obtain Table II:

TABLE II. Data given by Origin for the three polynomials. Trying to optimize the polynomials, the changes are very small, and it is not worth trying to improve them.

x^5	x^4	x^3	x^2	x	constant	SumDiff^2
Origin		0.00209	-0.20851	7.98258	-46.81378	19962
	9.36E-06	-7.18E-04	0.05939	-0.61593	8.14645	84.24
1.56E-08	3.49E-06	5.83E-05	1.67E-02	2.42E-01	4.83E+00	2.25

Figure 4 shows the pressure differences.

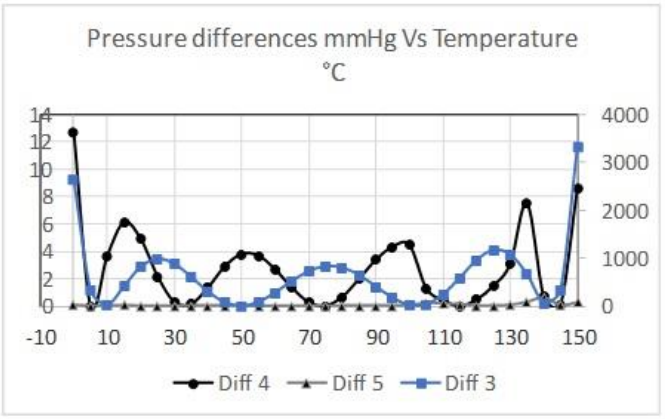

FIGURE 4. Pressure difference between measured and calculated values, in mmHg. The scale on the right corresponds to polynomial 3.

Polynomial 5 is better than the table, it allows to calculate intermediate values. Figure 5 shows the data provided by Origin. They are quite complete, but they forget to cut digits to the values with their uncertainty. Uncertainty is used to cut the measured digits and is given with one or two significant figures. When adjusting Origin values, it is found that the latest digits do not contribute to the sum of squares. Sigma Plot data is practically the same as Origin.

Equation	y = Intercept + B1*x^1 + B2*x^2 + B3*x^3 + B4*x^4 + B5*x^5
Plot	Presión
Weight	No Weighting
Intercept	4.83003 ± 0.24989
B1	0.242 ± 0.03559
B2	0.01673 ± 0.00153
B3	5.83219E-5 ± 2.6382E-5
B4	3.49434E-6 ± 1.95121E-7
B5	1.56389E-8 ± 5.17611E-10
Residual Sum of Squares	2.24565
R-Square (COD)	1
Adj. R-Square	1

FIGURE 5. Origin data for grade 5 polynomial.

IV. CAN WE TRUST THE R^2 COEFFICIENT?

It is assumed that $R^2 = 1$ implies that the adjusted curve passes through all the experimental points. The difference should be zero for all values. We note that this does not happen, and that, for Excel, polynomials 3 and 4 with $R^2 = 1$, do not go through the experimental points. We cannot rely on the square coefficient, we need to calculate the differences of each adjustment. If the differences are too large for our application, we will have to interpolate between each pair of points. If this error is too much, the best fit would be the linear interpolation between each pair of points, which is equivalent to adjusting N-1 straight lines, if N is the number of points.

V. CLAUSIUS CLAPEYRON EQUATION

The Clausius-Clapeyron equation relates the vapor pressure of a substance with its heat of evaporation L and its absolute temperature T, which can be integrated assuming constant heat of evaporation L and an ideal gas water vapor.

$$\ln(p2/p1) = -(LPm/R)*(1/T2-1/T1), \quad (2)$$

where Pm is the molecular weight and R the gas constant. Figure 6 shows the difference of measured and calculated pressures, and the square of the differences as a function of temperature

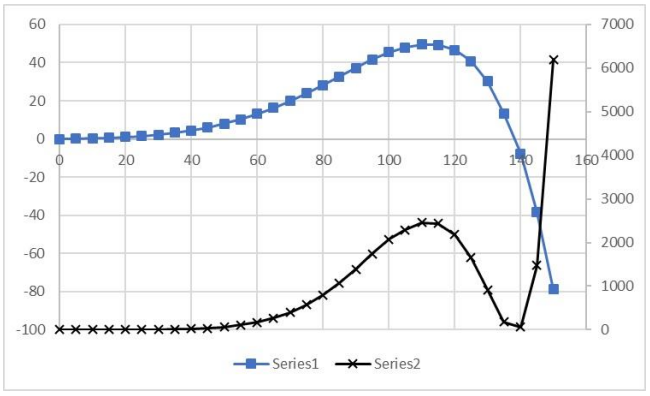

FIGURE 6. Difference in measured and calculated pressures, and the square of the differences as a function of temperature. The pressure difference is shown on the left vertical axis.

The LPm/R parameter was optimized to obtain the minimum sum of squares, such as 5147° K. But that the curve does not pass through the experimental points means that the assumption of the constancy of the evaporation heat L. is not valid. Figure 6 with LPm/H = 5235 shows that the approximation is good up to 80° C. The value of L corresponds to the initial temperature of the interval, that is T1

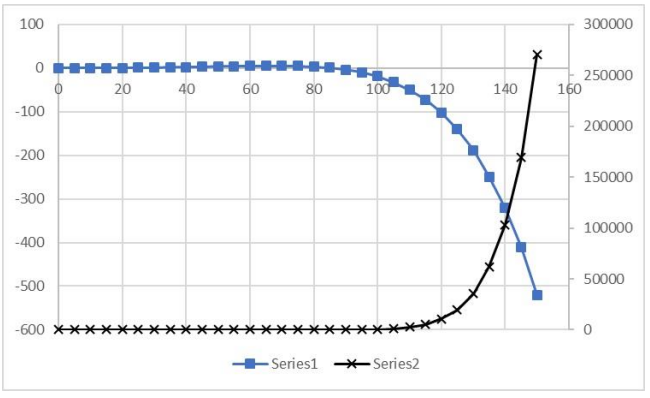

FIGURE 7. Difference in measured and calculated pressures, and the square of the differences as a function of temperature. The pressure difference is shown on the left vertical axis.

By dividing the data into 5 groups, with initial temperatures of 0, 35, 70, 100 and 130° C, the LPm/R constants can be estimated for each section. and get a very good fit for the calculated pressures.

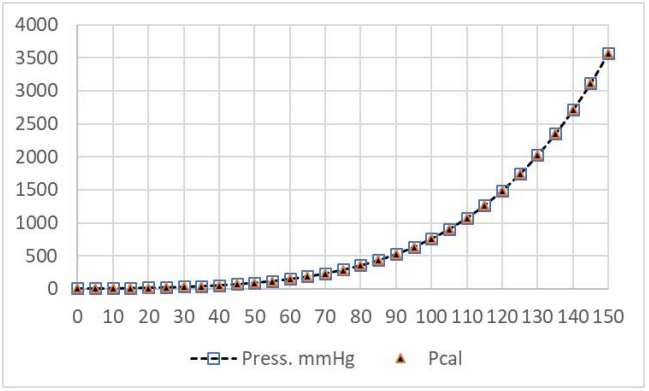

FIGURE 8. Here the graph shows that the calculated pressures and measured are equal.

Looking at the differences between the calculated and measured pressures, figure 9 is obtained.

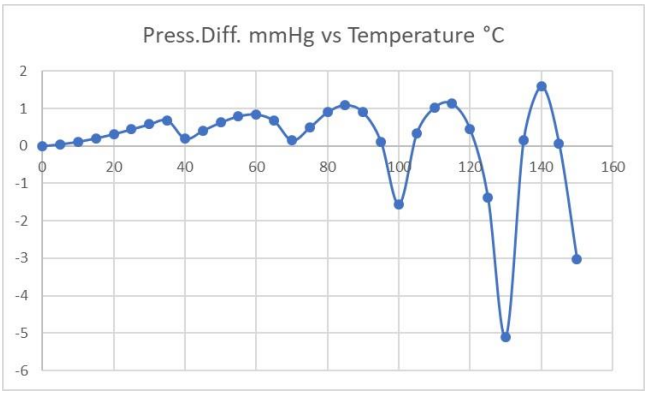

FIGURE 9. Pressure differences measured and calculated for the 5 temperature ranges.

The table III shows the values of the initial temperature of the interval and the constants L and LPm/R used in the adjustment.

TABLE III. Values of the initial temperature of the interval and the constants L and LPm/R.

T °C	0	35	70	100	130
L cal/gr	585.1	570.9	556.4	543.7	534.3
LPm/R	5300	5172	5040	4925	4840

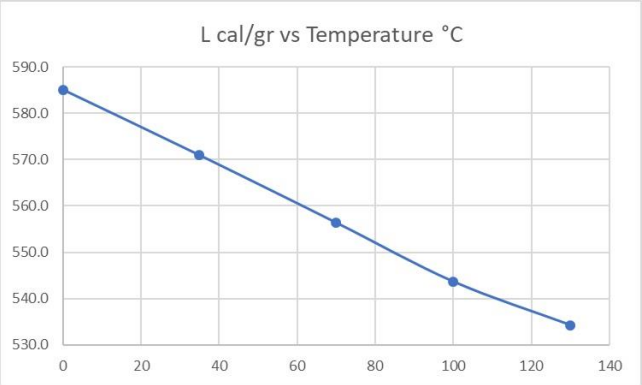

FIGURE 9. Latent heat of evaporation, in cal/gr, for each initial temperature range.

VI. CONCLUSIONS

For now, the interesting result is that, when adjusting an equation to a data set, knowing that the largest R^2 does not necessarily coincide with the best fit, it is necessary to verify it by calculating the difference between the calculated value and the measured value for the entire measured interval. I used to choose the simplest equation that would go through the experimental points, now I prefer to measure the differences to find the best fit and to improve the settings of Excel or another database. Using a higher degree equation does not guarantee that the fit is better. This graph helps distinguish between systematic or random errors. To replace a data table with an equation we must be sure that the residuals of the equation are zero, ensuring that it passes through all the experimental points.

REFERENCES

[1] Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A., *Graphical Methods for Data Analysis* (Wadsworth & Brooks/Cole, Pacific Grove, CA, 1983).
[2] Mersereau, A., Metz J., *True lines*, The Physics Teacher **36**, 174 (1998).
[3] Blikensderfer, R., *Systematic errors and graphical extrapolation*, The Physics Teacher **23**, 545 (1985).
[4] Ruprigth, M. E., *Measuring Systematic Error with Curve Fits*, The Physics Teacher **49**, 54 (2011).
[5] Zetie, K. P., *Can students draw lines of best fit?* Phys. Educ. **51**, 065017 (2016).
[6] Sebastião, P. J., *The art of model fitting to experimental results*, European Journal of Physics **35**, 015017 (2014).
[7] Vasilyeva, D., Giannotti, M., Goehl, J. F., *Comparison of different approaches in extraction of a parameter in a linear fit*, Eur. J. Phys. **36**, 045011 (2015).
[8] Peterlin, P., *Data analysis and graphing in an introductory physics laboratory: spreadsheet versus statistics suite*, Eur. J. Phys. **31**, 919 (2010).
[9] Minitab Blog, *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?* (2013). https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
[10] Frost, J., *How to Interpret R-squared in Regression Analysis* (2020). https://statisticsbyjim.com/regression/interpret-r-squared-regression/
[11] *Presión de vapor de agua líquida y hielo a varias temperaturas* (2020). http://vaxasoftware.com/doc_edu/qui/pvh2o.pdf
.