DATA TABLE, EQUATION FIT OR INTERPOLATION

Héctor G. Riveros

Instituto de Física UNAM, Mexico

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 R2 of the adjusted equation. It is considered that if it is 1 the equation goes through all the experimental points. By adjusting Tungsten resistivity data, it is found that the equations do not go through all the points (even with R2 = 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, Matlab 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.

1. 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: spradsheet versus statistics suite [8] Some mention the R2 but do not mention the need to verify the goodness of the fit, plotting the residuals. We have Excel. MatLab and Origin that adjust different curves and calculate the parameters that give the minimum of the sum of the squared residues. Peterlin [8] has:

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

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"

2. ADJUSTING THE RESISTIVITY OF TUNGSTEN

The properties of the materials and their variation with temperature are usually presented in the form of tables. Tungsten resistivity is one of these properties and we can find its value in Espe [11]. All figures in a table are expected to be significant, the uncertainty implied in the value 5.48 varies from 5.475 to 5.485.

But if we need the resistivity 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 R2 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 several adjustments that seem to all go through the experimental points,

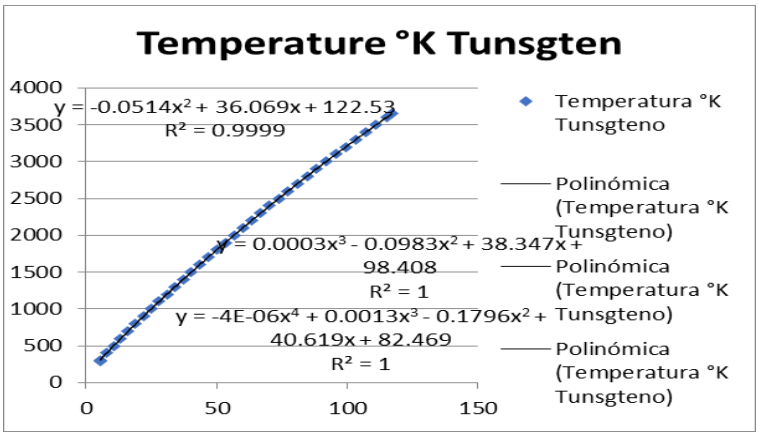

Figure 1

3. 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. Figure 1 shows the resistivity on the horizontal axis in μohm.cm.

The linear adjustment has as equation a: T = 32.552 * ρ +164.82 and R2 = 0.9974 and does not go through all the points. The configuration of polynomial 2 is T = -0.0514 * ρ2 + 36.069 * ρ +122.53 and R^2 = .9999. The polynomial configuration 3 is T = 0.0003 * ρ3 -0.0983 * ρ2 + 38.347 * ρ + 98.408 and R2 = 1. The polynomial configuration 4 is T = -4E-6 * ρ4 + 0.0013 * ρ3-0.179 * ρ2 + 4 0.619 * ρ +82.469 and R^2 = 1. All adjustments 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 temperature minus the calculated temperature and squared so that all the data are positive. Figure 2 shows the result.

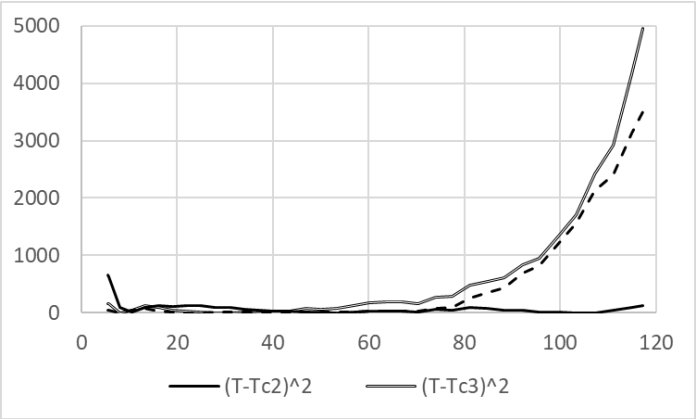

Fig. 2.- Square of the temperature difference between the measured and calculated values. Polynomials of 2, 3 and 4 degrees against resistivity ρ.

The best fit is the polynomial of degree 2. Configuration 2 has 26 ° K as the temperature difference in the first data. Note that configurations 3 and 4 have R^2 = 1, but their data does not pass through the experimental points. The sum of the squares of the differences is 3060 for polynomial2, 23223 for polynomial 3 and 8490 for polynomial 4. It is striking that using a polynomial of greater degree does not improve the fit of the data, even if it has R^2 = 1, Excel rounds to 1 if it has more than four nines (0.99999).

Reviewing the calculations, it was found that polynomial 2 was well adjusted, but for polynomials 3 and 4 the fit could be improved. A well-made fitt 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

X^4	X^3	X^2	X	cte	Sum Dif^2	R^2
		-0.0514	36.069	122.53	3060	0.99992
	0.000241	-0.095	38.168	101.5	23223	0.999395
-0.000004	0.0013	-0.1796	40.619	82.469	17081	0.999555
		-0.0514	36.0693	122.53	2433	0.99992
	0.000241	-0.0949	38.164	101.2	878	0.9999771
-4.08E-06	0.0012	-0.1684	40.2	86.4	508	0.9999868

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.

The first three lines show the values of the Excel equations and the lines 3, 4 and 6 show the optimized values of the equations. Column R^2 shows the values calculated by equation (1). The R2 values in Excel are in Figure 1. Only polynomial 2 was optimized. But polynomial 3 sum is 23223 and adjusted is 878. For polynomial 4 the sum was 17081 and optimized is 508. The figure shows the squared differences for the new settings.

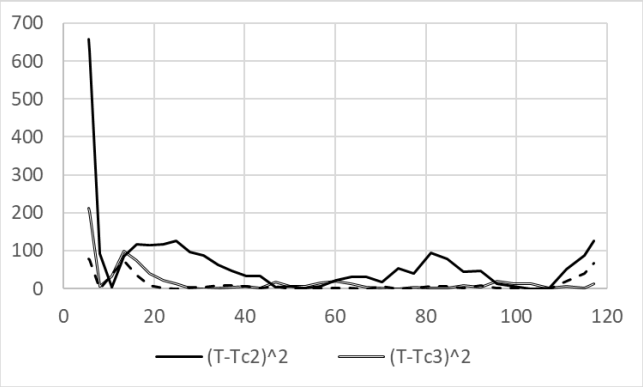

Fig. 3.- Square of the temperature difference between the measured and calculated optimized values. Polynomials of 2, 3 and 4 degrees against resistivity ρ.

Adjusting the data with Matlab, is Table II.

X^4	X^3	X^2	X	cte	Sum Dif^2	R^2
		-0.0514	36.069	122.533	3060	0.99999009
	0.0003	-0.0983	38.3474	98.4083	23267	0.99996015
-4.26E-06	0.0013	-0.1796	40.6191	82.4695	1132	0.99992027
		-0.0514	36.0693	122.53	3060	0.99999009
	0.0003	-0.102	38.33	100	1529	0.99996015
-4.33E-06	0.0013	-0.1796	40.616	82.4	380	0.99992027

Table II The first three lines are the data given by Matlab for the three polynomials. By changing the coefficients looking for the minimum in the sum of the squares of the differences, lines 4 to 6 are obtained, optimizing the adjustment.

Note that the sums went down considerably increasing the accuracy of the calculated data, except for polynomial 2 that was well adjusted. Figure 4 shows the differences calculated for each point of the Table of Espe,

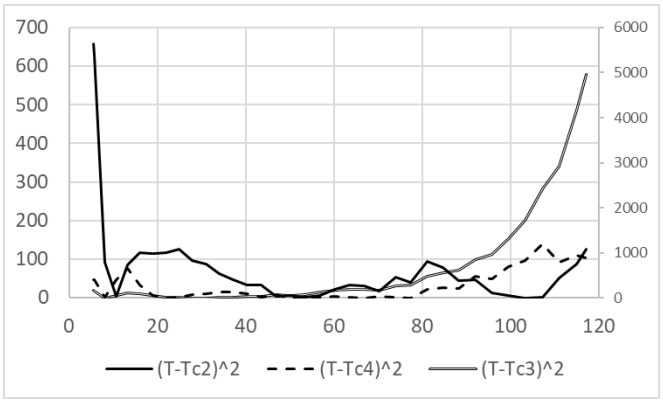

Fig. 4.- Square of the temperature difference between the measured and calculated values. Polynomials of 2, 3 and 4 degrees against resistivity ρ. The scale on the right represents (T-Tc3)^2

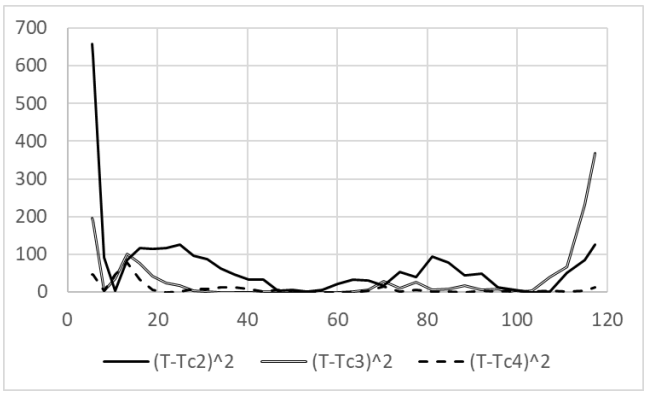

Fig. 5.- Square of the temperature difference between the measured and calculated optimized values. Polynomials of 2, 3 and 4 degrees against resistivity ρ.

From 20 to 120 the adjustments are good, and the adjustment to the fourth power is the best. Using the Origin settings, we obtain Table III:

X^4	X^3	X^2	X	cte	Suma Dif^2	R^2
		-0.0514	36.07	122.53296	3060	0.99992027
	0.00025988	-0.09832	38.35	98.40831	856	0.99997769
-0.000004269	0.00129	-0.17959	40.62	82.46947	526	0.99998627
		-0.0514	36.0693	122.53	3060	0.99992028
	2.60E-04	-0.09832	38.3474	98.4083	856	0.9999777
-0.00000424	0.00129	-0.17954	40.6193	82.5	527	0.99998626

Table III The first three lines are the data given by Origin for the three polynomials. By changing the coefficients looking for the minimum in the sum of the squares of the differences, lines 4 to 6 are obtained, optimizing the adjustment.

Origin data for polynomial 2, 3 and 4 were optimized, as you can see from Table III the Origin data was obtained. Figure 6 shows the squares of the difference.

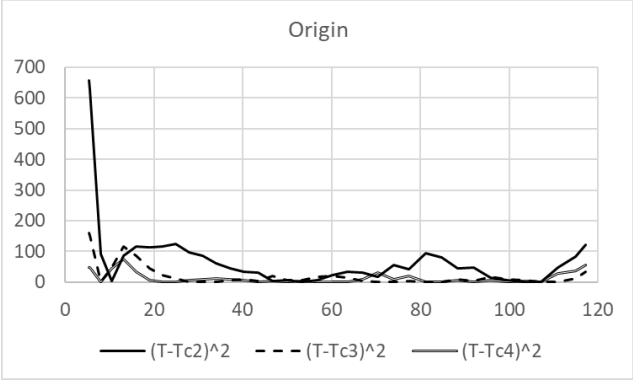

Fig. 6. The square of the difference for each measurement number. Origin presents the best settings.

Figure 6 shows the data provided by Origin. They are quite complete, but they forget to round the values with their uncertainty. The usual thing is to keep one or two figures in the uncertainty and use them to round. When adjusting Origin values, it is found that the latest figures 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
Plot	Temperatura
Weight	No Weighting
Intercept	82.46947 ± 3.53541
B1	40.61905 ± 0.40259
B2	-0.17959 ± 0.01332
B3	0.00129 ± 1.64579E-4
B4	-4.26946E-6 ± 6.75149E-7
Residual Sum of Square	373.32534
R-Square (COD)	0.99999
Adj. R-Square	0.99999

Fig. 7. Origin data for grade 4 polynomial

4. CAN WE TRUST THE R^2 COEFFICIENT?

It is assumed that R^2 equal to 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, polynomial 2 has smaller differences than polynomials 3 and 4. I expected the differences to be smaller for polynomials 3 and 4, since a more complicated equation is adjusted. Linear fit is bad. 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 adjust sections of lines or interpolate between each pair of points.

By adjusting segments of quadratic polynomials, we must obtain a better fit. Table II shows that the sum of the squared differences is reduced to 20.6 by adjusting four segments, in the temperature ranges of the first column.

Table II Adjusting segments of quadratic polynomials greatly reduces the magnitude of the differences.

	X^2	X	cte	Sum
All	-0.0514	36.069	122.53	3060.0
293-700	-0.4201	47.498	44.956	
800-2200	-0.0579	36.216	133.32	
2300-3000	-0.0368	33.648	214.72	
3100-3655	-0.0359	33.486	223.72	20.6

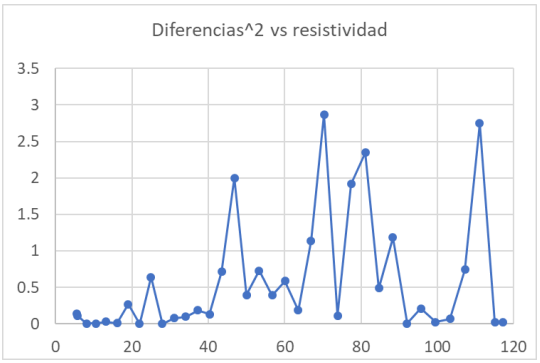

Figure 8 shows the squared differences as a function of Tungsten resistivity.

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. When interpolating between successive points, 1 ° K can be read.

5. What is the Maximum Difference in Degrees that we can Accept?

That depends on why we are adjusting the values. As a school exercise of radiation equations, or to decide the resolution of a tungsten wire to measure temperatures. Deciding between a data table and an equation depends on what we want to interpret the data for.

5. Conclusions

For now, the interesting result is that, when adjusting an equation to a data set, knowing that the largest R2 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, which does not happen for the resistivity of Tungsten.

REFERENCES

[1] Chambers J M, Cleveland W S, Kleiner B and Tukey P A 1983 Graphical Methods for Data Analysis (Pacific Grove, CA: Wadsworth & Brooks/Cole)

[2] Mersereau, A.,Metz J. 1998 True lines The Physics Teacher 36, 174 doi: 10.1119/1.879995

[3] Blikensderfer R 1985 Systematic errors and graphical extrapolation The Physics Teacher 23, 545 doi: 10.1119/1.2341909

[4] Ruprigth ME.2011 Measuring Systematic Error with Curve Fits The Physics Teacher 49, 54; doi: 10.1119/1.3527759

[5] Zetie KP.2016 Can students draw lines of best fit? Phys.Educ .51065017 https://doi.org/10.108/0031-9120/51/6/065017

[6] Sebastião PJ. 2013 The art of model fitting to experimental results European Journal of Physics, Volume 35, Number1

[7] Vasilyeva D ,Giannotti M, Goehl JF2015 Comparison of different approaches in extraction of aparameter in a linear fit Eur.J.Phys.36045011 https://doi.org/10.1088/0143-0807/36/4/045011

[8] Peterlin P 2010 Data analysis and graphing in an introductory physics laboratory:spreadsheet versus statistics suite Eur.J.Phys.31 919 https://doi.org/10.1088/0143-0807/31/4/021

[9] Minitab Blog, 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 https://statisticsbyjim.com/regression/interpret-r-squared-regression/

[11] Espe Materials of high vacuum technology Vol 1 page 46, Pergamon 1966