

Curve Fitting: Linear and Nonlinear Least Squares (Physics 1210 Notes, Appendix D)

1. PREFACE

Appendix C detailed the major components that comprise an effective graph and also discussed the functional relationships which produce straight lines on linear, semi-log or log-log graphs. This Appendix demonstrates the use of regression analysis to obtain a “best fit” functional expression for set of experimental data.

2. LINEAR REGRESSION

Engineers utilize two types of formulas to mathematically describe the relationship between a dependent variable and its independent variables:

1. Theoretical relationships
2. Empirical relationships.

In order to show the difference between these two relationships, let's consider the following example:

You are given the job to determine an accurate force versus velocity functional relationship for a certain 0.5 kg object as it undergoes a **vertical** free fall through a given fluid. Table D1 presents a typical acceleration versus velocity experimental data set and the corresponding force values which was computed from Newton's second law.

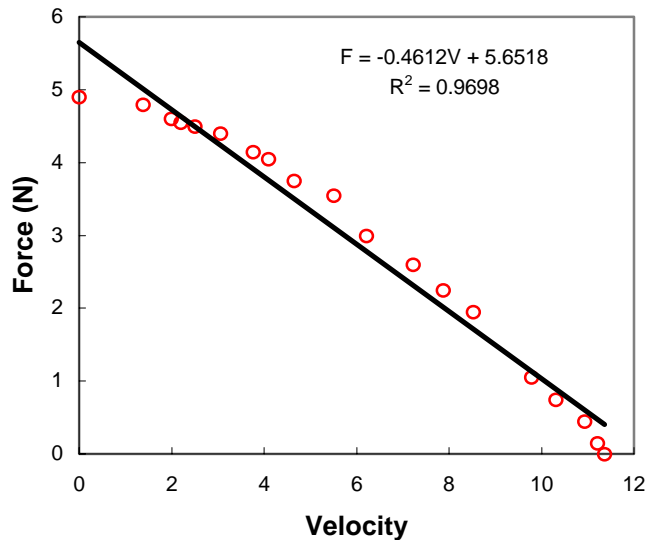
Table D.1 Dynamics of Free Fall Experiment

Velocity (m/s)	Accel. (m/s ²)	Force (N)
0	9.8	4.9
1.38	9.6	4.8
1.99	9.2	4.6
2.2	9.1	4.6
2.51	9.0	4.5
3.06	8.8	4.4
3.77	8.3	4.2
4.09	8.1	4.1
4.65	7.5	3.8
5.51	7.1	3.6
6.21	6.0	3.0
7.22	5.2	2.6
7.88	4.5	2.3
8.53	3.9	2.0
9.79	2.1	1.1
10.31	1.5	0.8
10.93	0.9	0.5
11.21	0.3	0.2
11.37	0.0	0.0

where:

- F = net force on the object (N)
- $mass$ = the object's mass = 0.5 kg
- V = velocity (m/s)
- dV/dt = acceleration (m/s²)

$$F = mass \frac{dV}{dt}. \tag{D.1}$$



This tabular data between the net force and its corresponding velocity is also graphed in Figure D1. If there is an available theoretical or an

Figure D1. Linear Curve Fit of Force vs. Velocity Data

accepted empirical relationship that supposedly described this phenomena, you should use it to see how well it matches the experimental results. One must however always remember that valid experimental results reflect reality whereas theoretical or empirical correlations are usually very restricted in their range of applicability.

In cases where the physics of the phenomena or the system's properties are not well understood, the form of satisfactory empirical relationship must be deduced from the experimental data alone. In this case, a simple linear relationship between the force and velocity is the simplest produces a somewhat reasonable fit. That is

$$F = mV + b \quad (D.2)$$

where

m = slope of the resulting straight line (Ns/m)

b = y intercept of the straight line (N) or Y_0 .

The best least square linear fit to the above data set can be easily obtained by superimposing a “trendline” as shown in Figure D1. The Excel procedure to affix a trendline and its corresponding equation that has any arbitrary Y_0 intercept and the coefficient of determination, r^2 , to a graph is described in the ES 1060 text [1]. A mathematical relationship for r^2 is given by Equation D.6 and its physical significance is described in the ES 1060 text. A value of 1 indicates a perfect match between the data and the corresponding predicted values while a value of 0 implies that the trendline is no better than just using the mean value as the predicted value for all points. The ES1060 text states that “values of r^2 above 0.8 or so are considered good”. By this criterion, the linear trendline depicted in Figure D1 should be an excellent fit to the data since it has a very high r^2 of 0.9698. Despite this, it is fairly obvious from Figure D1 that the force versus velocity relationship is not linear.

3. Estimated Errors

Knowledge of the estimated error in m and Y_0 is vital to permit propagation of error calculations whenever m and Y_0 are used in subsequent calculations. The above trendline method does not produce this information but, for linear relationship only, Excel provide a convenient way to calculate the least squares estimators, error in the least squares estimators, and the coefficient of determination.

- (1) Click on **Tools** in the main menu bar
- (2) Click on **Data Analysis** in the pull down menu
 - If **Data Analysis** is not an option, then
 - (a) Click on **Add-Ins** in the pull down menu
 - (b) Click on **Analysis ToolPak** in the Add-Ins dialog box (the check box must be checked)
 - (c) Click **OK**
 - (d) Click on **Data Analysis**
- (3) Click on **Regression** in the Data Analysis dialog box
- (4) Click on **OK**

The above procedure loads the data analysis tools add-in and launches the regression analysis tool. The regression analysis dialog box should now be visible on the screen. This dialog box contain four regions: **Input**, **Output**, **Residuals**, and **Normal Probability**. The **Residuals** and **Normal Probability** regions should not be changed unless you have an understanding of advanced statistics. In the Input region, you must provide the appropriate cell references for the dependent variable **Y** (in this case, the force) and the independent variable **X** (velocity). In addition, you need to specify where to put the results of the regression analysis, and this information is conveyed in the **Output** region.

The statistical relationship should be forced to comply with any known boundary conditions of the functional relationship. The most common boundary condition is that the intercept value (Y_0) must be zero. This boundary condition is easily handled by checking the box labeled **Constant is Zero** in the **Input** region of the regression analysis dialog box. The output results of a linear regression analysis of the data in Table D.1 are presented in Table D.2.

Table D.2. Regression statistics summary table from Excel on the data in Table D.1.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.9848057
R Square	0.9698422
Adjusted R Square	0.9680682
Standard Error	0.3080752
Observations	19

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	51.887577	51.887577	546.70112	2.2971E-14
Residual	17	1.6134754	0.0949103		
Total	18	53.501053			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	5.6517517	0.1366055	41.372802	1.661E-18	5.36353892	5.9399645	5.3635389	5.9399645
X Variable 1	-0.4611782	0.0197239	-23.381641	2.297E-14	-0.5027922	-0.4195643	-0.5027922	-0.4195643

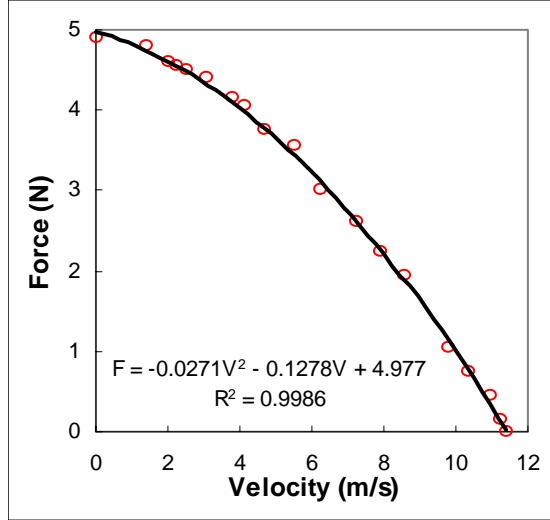
There are five items of interest in the regression statistics summary table given in Table D.2:

- (1) R square = r^2 = **0.9698**
- (2) coefficient of the intercept = b (N) = **5.652**N
- (3) standard error of the intercept = Δb (N) = **0.1366** N
- (4) coefficient of X_1 = m (N/m) = **-0.4612** N/m
- (5) standard error of X_1 = Δm (N/m) = **0.01972** N/m.

Note that the above b and m coefficients agree with the trendline value given in Figure D1 and that their relative standard errors are both fairly small. The small relative standard coefficient errors and the high r^2 value would normally imply an “excellent fit It should be noted that Excel used the default number of significant digits and that it is the students' responsibility to determine the appropriate number of significant digits.

4. Nonlinear Trendline

Besides linear trendline, Excel has the capability of fitting logarithmic, polynomial of arbitrary order, power or exponential functions to data. In the case presented in Figure 1D, it appears that a quadratic relationship should



produce an excellent fit. Figure D2 substantiates this in that this quadratic trendline has a r^2 of 0.9986 as compared to a value of 0.9698 for the linear fit. Higher order polynomials may be used but any increase in r^2 that is obtained by this increased complexity is rather superficial.

5. Nonlinear Optimizing Solver

If we start over on this problem and apply some basic dynamics to this free fall problem, the summation of forces in this case must be equal to the gravitational body force ($mass \bullet g$) in the downward direction plus a drag force in the upward direction that is some unknown function of velocity. Therefore theory implies that the force versus velocity relationship must have the following general form:

$$F = mass \bullet g - Drag(V) \quad (D.3)$$

but it does not supply any information about how the drag varies with velocity. Our own personal experience indicates that the drag force increases with velocity and extensive experimental testing over the years has shown that power laws can be used frequently to correlate velocity-drag data over limited velocity ranges. If this is assumed to be the case here, then

$$F = mass \bullet g - aV^b \quad (D.4)$$

Theory and some empirical insight has therefore been combined to obtain a possible function form between velocity and force in terms of two arbitrary constants (a , b) that is based upon the physics of phenomena and not just blind curve fitting as was done in the linear and quadratic curve fit examples.

The values of a and b that give the best fit with the experimental data can be determine through the use of the Excel nonlinear optimizing solver which was also covered in ES 1060 [1]. The first requirement of using the nonlinear optimizing solver is the development of a regression function that you what to optimize in terms of minimizing or maximizing its value or obtaining a specified value. The trendlines that are presented in the previous two curve fits are based upon least square regression in which the following regression function is minimized

$$\sum_{i=1}^n (\underline{F}_i - F_i)^2 \quad (D.5)$$

where \underline{F}_i is the measured force and F_i is the corresponding predicted value in the data set that contains n values. In this case Equation D.4 would be substituted for F_i . Instead of doing this, lets maximize r^2 . That is

$$r^2 = 1 - \frac{\sum_{i=1}^n (\underline{F}_i - F_i)^2}{\sum_{i=1}^n (\overline{F} - F_i)^2} \quad (D.6)$$

where \overline{F} is the mean force of the experimental data set. Excel provides a nonlinear optimizing solver for minimizing functions such as Equation D.6. However, the problem must be prepared properly to obtain an appropriate solution. Table D.3 presents a copy of the spreadsheet (see file **Appendix D.xls** for the actual spreadsheet) that was used to

determine a & b . This table contains five columns: **column 1** is the independent variable (velocity); **column 3** is the dependent variable (force); **column 4** is the predicted dependent variable (the force calculated from Equation D.4); **column 5** is the square of the difference between columns 3 and 4; and **column 6** is the square of the difference between **column 3** and the average force which is calculated at the end of **column 3**. The **columns 5 & 6** are then summed and these values are used to calculate the r^2 value for a guess set of coefficients (a , b). For instance, the guess of (1,1) produces a very poor r^2 value of -5.88.

Appendix D.xls

$$a = 0.0852316 \text{ N/(m/s)}^b \quad g = 9.8 \quad (\text{m/s}^2)$$

$$b = 1.6632532 \quad m = 0.5 \quad (\text{kg})$$

Velocity (m/s)	Accel. (m/s ²)	Force (N)		(F _i - F _i) ² N ²	(F _{av} - F _i) ² N ²
		Measured F _i	Predicted* F _i		
0	9.8	4.9	4.9	0.00E+00	3.93
1.38	9.6	4.8	4.8	2.08E-03	3.54
1.99	9.2	4.6	4.6	1.04E-03	2.83
2.2	9.1	4.6	4.6	1.13E-03	2.66
2.51	9.0	4.5	4.5	3.75E-05	2.50
3.06	8.8	4.4	4.4	2.27E-03	2.20
3.77	8.3	4.2	4.1	6.16E-04	1.52
4.09	8.1	4.1	4.0	1.39E-03	1.28
4.65	7.5	3.8	3.8	2.67E-03	0.69
5.51	7.1	3.6	3.4	1.14E-02	0.40
6.21	6.0	3.0	3.1	1.51E-02	0.01
7.22	5.2	2.6	2.6	2.78E-04	0.10
7.88	4.5	2.3	2.3	8.30E-05	0.45
8.53	3.9	2.0	1.9	3.97E-03	0.94
9.79	2.1	1.1	1.1	3.72E-03	3.49
10.31	1.5	0.8	0.8	4.17E-04	4.70
10.93	0.9	0.5	0.3	1.02E-02	6.09
11.21	0.3	0.2	0.2	1.33E-05	7.66
11.37	0.0	0.0	0.0	1.63E-03	8.52
	Fav =	2.9	Sum =	5.80E-02	53.50
				R² = 0.998916	= 1 - SUM(F _i - F _i) ² /SUM(F _{av} - F _i) ²

* F_i = see Module Force(m,g,V,a,b)
Force(m,g,V,a,b)

Table D.3. Excel table used to perform nonlinear regression.

Excel uses an iterative approach to solve the nonlinear regression problem once it has an initial guess set to start this iterative process. In this case, the program will systematically vary a and b to determine the local gradient of r^2 and thereby determine how the (a , b) set should be varied to maximize r^2 . In order to use the solver tool, the tool must be loaded into Excel. The solver can be loaded by:

- (1) Click on **Tools** in the main menu bar
- (2) Click on **Solver** in the pull down menu
 - If **Solver** is not an option, then
 - (a) Click on **Add-Ins** in the pull down menu
 - (b) click on **Solver Add-In** in the Add-Ins dialog box (the check box must be checked)

- (c) Click **OK**
- (d) Click **Solver**

The Solver dialog box is now visible. The first menu item is the **target cell** which is r^2 in this case. The second item delineates what action is to be performed on the target cell. In this example we wish to **maximize** the target cell. The third item specifies which cells may have their values varied to accomplish the objective which in this case are cells containing the guess values of the regression parameters a and b . Note that named cells can be utilized in specifying the cell locations of the target cell and the adjustable cells. As an option, you can set numerical constraints on the adjustable cells. A little thought about the physics of this problem indicates that a and b are both positive and these constraints may be added. In some problems you may wish to change the default **Precision** and **Tolerance** values by first clicking the **Options** button. Now click **OK**, and Excel will attempt to find the optimum solution and replace the guess values of the regression parameters with the optimum values. Table D.3 indicates that combined theoretical/empirical correlation

$$F = 4.9 - 0.0852V^{1.663} \quad (\text{D.7})$$

produces a r^2 of 0.9989 which is slightly better than the quadratic but has more physical significance. Instead of basing the curve fit on r^2 , try using the least squares regression method to compute the coefficients and compare your results.

One word of caution: nonlinear functions often contain more than one solution and that a given guess set may produce a local solution (in this case, a local maximum) instead of a global solution. Highly nonlinear problems may also require a fairly accurate initial guess to obtain a global solution or any solution and you may have to resort to plots to produce an accurate initial guess.

See Nonlinear Regression.xls for another example.

References

1. Introduction to Engineering Computing, B. R. Dewy, McGraw-Hill Primus, 1994, pg. XL1-XL20. (Your ES 1060 text)
2. Physics 1210/1310 Laboratory Manual, University of Wyoming, Department of Physics and Astronomy, Kendall/Hunt Publishing, 1992, pg. 106-114.