LABORATORY USE OF A SPECIALLY PROGRAMMED EXCEL USER FORM FOR POLYNOMIAL REGRESSION AND FOR EVALUATING THE UNCERTAINTY OF POLYNOMIAL REGRESSION MODELS

Sheldon M. Jeter

Georgia Institute of Technology

INTRODUCTION

Regression models are widely used in engineering practice, especially in mechanical and chemical engineering and in related fields. They are used to represent data and to calibrate instruments among other applications. Standard textbooks address linear regression models well, and some also address the associated statistical uncertainties of linear models. This uncertainty of a model is the range of uncertainty about the systematic dependence of the dependent variable on the independent variable(s).

Unfortunately, none of the popular texts reviewed for this paper adequately address polynomial models and their uncertainties, probably because polynomial models seem to be common mostly in engineering applications. In contrast, polynomial models are not so common in fields such as medicine and social sciences that seem to attract more interest from professional statisticians. Nevertheless, it has been shown elsewhere (Jeter, 2003) that Error Propagation Analysis (EPA), which is already familiar to most experimental engineers, can be used to find the uncertainty of both linear and polynomial models.

While the underlying philosophy and mathematics concerning the uncertainty of polynomial regression models is not especially complicated, the practical implementation requires multiple executions of auxiliary regressions. These extra steps are quite time consuming when each step must be defined manually, and the extra manual steps are likely to induce procedural errors. To make the calculation and plotting of the results simple and easy, a special Excel utility routine called a User Form that is described in this paper was programmed.

In the balance of this paper, the statistical and mathematical background for this technique will be reviewed, the algorithm for the implementing the technique will be outlined, and a couple of representative practical examples from mechanical engineering will be presented.

BACKGROUND

A typical polynomial model is this general quadratic formulation, which relates the dependent variable *y* to a polynomial in terms of the independent variable *x*,

$$y_{\text{est}} = c + b_1 x \ + b_2 x^2 \tag{1}$$

It is relatively well known that the result of least residual squared error regression analysis can always be written in a more convenient form by using centered variables. The alternative formulation for the quadratic model above is actually

$$y_{\text{est}} = y_{\text{ave}} + b_1 \left(x - x_{\text{ave}}\right) + b_2 \left(x^2 - x_{\text{ave}}^2\right) \tag{2}$$

This centered formulation is much more convenient for uncertainty analysis. Note that the specified model had a constant term. Otherwise the data could not be centered.

The exact relationship between the independent and dependent variables will always elude experimentation because the data is always contaminated by random error. The experimental engineer and the user of experimental data need not only the experimental model but also some reasonable estimate of the range of uncertainty in the model. As reviewed below and as has been shown elsewhere, Error Propagation Analysis (EPA) can be used to find the statistical uncertainty of the model, which is the uncertainty about the systematic dependence of *y* on *x*.

In general EPA is based on the following familiar formula for calculating the combined uncertainty when multiple independent sources of uncertainty exist. In this case the sources of uncertainty are the set of experimental parameters, which are identified as the p_i set in the equation

$$U_y^2 = \left(\frac{\partial y}{\partial p_1} U_1\right)^2 + \left(\frac{\partial y}{\partial p_2} U_2\right)^2 + \cdots + \left(\frac{\partial y}{\partial p_i} U_i\right)^2 + \cdots \tag{3}$$

The contributing uncertainties are the set of U_i values that correspond to each experimental parameter p_i.

This uncertainty calculation is unusually complicated in polynomial models because the coefficients are strongly correlated when a polynomial relationship applies. For example, consider a quadratic relationship in which *y* tends to increase monotonically with *x*. In this case, a relatively small value of the linear coefficient, b_1, is likely to be correlated with a relatively large value of the quadratic coefficient, b_2, and conversely. In consequence, the familiar formula for combining sources of error must be modified to adjust for this correlation. Specifically, the correct uncertainty of the polynomial model must be expressed using conditional uncertainties. The resulting formula for the so-called Standard Uncertainty, which is analogous to the standard deviation, is as follows

$$u^2_{\text{poly-model}} = \left(\frac{\text{SEE}}{\sqrt{n}}\right)^2 + (x - x_{\text{ave}})^2 u(b_1|p_0)^2 + (x^2 - x^2_{\text{ave}})^2 u(b_2|p_0)^2 + \cdots \quad (4)$$

Here n is the number of data points. The symbol $u(b_1|p_0)$ stands for the conditional Standard Uncertainty of the i-th coefficient with all the other parameters, represented by p_0, held fixed. Note that the preceding equation begins with the well-known formula for the uncertainty of an average, which is itself easily obtained by EPA.

Evaluation of the preceding formula requires values of the conditional uncertainties. This slightly complicated auxiliary calculation requires extra regression steps. Specifically, the influence of any one coefficient must be isolated. The first step is executing a full regression analysis that computes all of the parameters - the constant and all the coefficients.

The next step is to sequentially evaluate the conditional uncertainty of each of the coefficients. This evaluation is done by conventional regression analysis after first correcting the dependent variable data for the influence of the other coefficients. Specifically, the correction is done by subtracting all the terms involving the other coefficients. The corrected independent variable corresponding to the coefficient b_i has the following general formulation,

$$y_{\text{corr},i} = y_{\text{data}} - \sum_{j \neq i} b_j (x - x_{\text{ave}})^j \quad (5)$$

Here the summation is over all indices and powers not equal to the specific i. The corrected y data are now regressed on the corresponding x^i data only. Of course, within numerical accuracy, the original constant and coefficient will be returned. In addition, all standard regression packages, including the Excel data analysis tool package, will return the Standard Error of the Coefficient. Typically, this statistic is only used for significance testing, but it is also the best available estimate of the needed conditional uncertainty of this coefficient. This process is repeated until all of the needed conditional uncertainties have been computed.

The extra steps are time-consuming and complicated; consequently, most students and practitioners avoid finding the uncertainties of such models. In the next section, a very handy Excel User Form that completely automates this task will be presented.

Occasionally regression models with the constant arbitrarily set to zero are desired. There is little readily available published guidance about the proprietary of such models, but there are valid practical reasons for adopting them. For example, many secondary instrumentation transducers can be adjusted or programmed to virtually ensure a zero output with a zero input. In such cases, a homogeneous calibration formula seems essentially mandatory. The User Form also has an option to calculate the uncertainty of such homogeneous models. When the constant term is excluded, the data cannot be

centered. Then the formula for combining uncertainties gives, in the quadratic case for example,

$$u^2_{\text{hom-model}} = 0 + x^2\, u\left(b_1 \middle| p_0\right)^2 + \left(x^2\right)^2 u\left(b_2 \middle| p_0\right)^2 \tag{6}$$

As before, p_0 stands for the other parameter(s). The zero term is included here just to emphasize that the uncertainty of the constant must be zero if it has been set to the arbitrary exact value of zero; consequently, the uncertainty of the model itself must be exactly zero at the origin. This feature emphasizes the severity of this restriction.

Hopefully this section has now adequately addressed the formulation of the Standard Uncertainty of general and homogeneous polynomial models. In practice, the 95 % uncertainty limit or Expanded Uncertainty is needed. This statistic is calculated as

$$U_{\text{A}} = k_{\text{c}}\, u_{\text{model}} \tag{7}$$

In the preceding formula, the multiplier k_{c} is the appropriate coverage factor. Assuming with good confidence that small sample statistics apply, the coverage factor for the 95 % range is computed with the classical t-distribution. Note that the number of statistical degrees of freedom will be the number of data minus the number of parameters. This Expanded Uncertainty is identified as the Uncertainty A of the model because, in compliance with modern usage (Taylor and Mohr, 1999), it is the uncertainty calculated by statistical analysis of repeated measurements. Uncertainty A was formerly and conventionally known as imprecision. Ultimately this Uncertainty A will be combined with a user supplied value for the Uncertainty B or range of possible bias according to the general formula for combining uncertainties,

$$U_{\text{C}} = \sqrt{U^2_{\text{A}} + U^2_{\text{B}}} \tag{8}$$

To complete the uncertainty analysis, the form will also compute and plot the Uncertainty A of the data by the familiar formula,

$$U_{\text{A, MODEL}} = k_{\text{c}}\, SEE \tag{9}$$

Here the *SEE* is the usual Standard Error of Estimate, which is essentially the square root of averaged squared deviation of the data from the model. The *SEE* is obviously analogous to the Sample Standard Deviation (*SSD*) for a simple sample. Recall that the *SSD* is essentially the square root of the averaged squared deviation from the mean. This uncertainty limit is useful for comparing the data with the model to inspect. For example, it can be used to scan for possible spurious outliers.

The User Form presented in this paper computes and plots the Uncertainty A for the model defined above and the simpler Uncertainty A of the data in Equation (9).

Obviously this User Form can be very useful to experimentalists and beneficial to users of data, and it will be described and outlined in the next section.

DESCRIPTION OF IMPLEMENTATION

The algorithm for this uncertainty analysis has been implemented in an Excel utility called a User Form. The dialog box for the form is shown in Figure 1. Note that the dialog box includes inputs called ComboBoxes that allow the user to identify the ranges for the dependent "Y Data" and the independent "X Data". Recall that the block of X-Data must be contiguous columns.

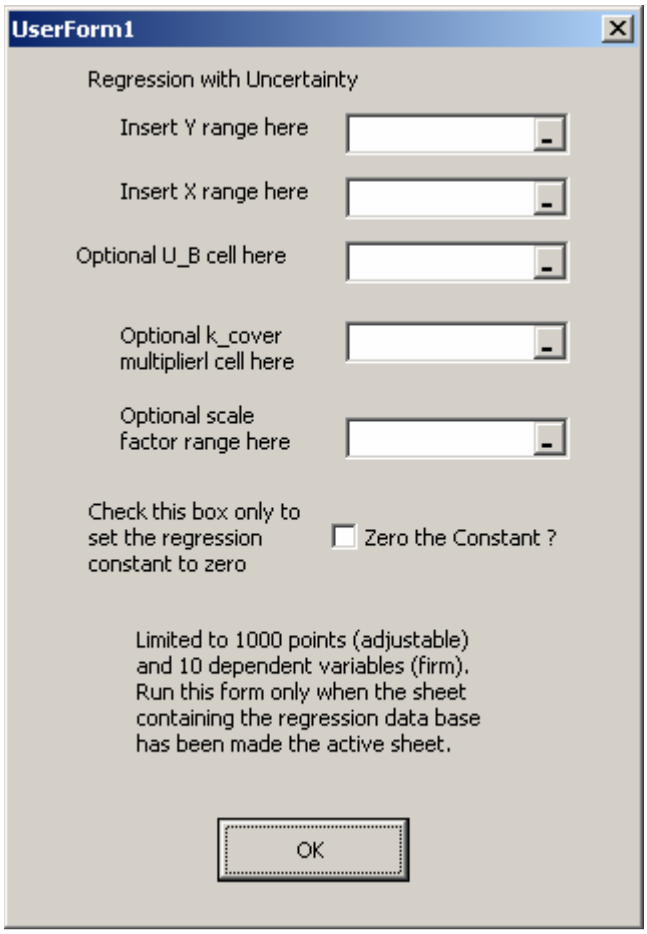

Figure 1. The Dialog Box for the User Form Called Poly_Regress

The form also has a text box to identify a cell containing a constant value for the Uncertainty B of the data. Recall that the Uncertainty B or inaccuracy is the range of possible built in bias in the data. The Uncertainty B cannot be evaluated from statistical

analysis. Instead, it requires physical analysis that usually involves Error Propagation Analysis. Ultimately an estimate of the Uncertainty B should be evaluated, but it is not required for the form to execute. Indeed, the output sheet of the form has been designed so that Uncertainty B data can be added and graphed even after the form has executed. The Uncertainty B will be combined with the statistical Uncertainty A to form the Combined Uncertainty of the model.

The form also includes another purely optional cell for a factor to arbitrarily inflate the coverage factor so that the error limits can be made large enough to be visible. This inflation is only desirable for illustration, but in the fortunate cases when the uncertainties are relatively small, it is a very desirable feature. This multiplier can also be adjusted in the output sheet after the form has executed.

The next box was added to resolve an apparent problem encountered in Excel 2003, but not in earlier versions. In some applications when the polynomials differ by several orders of magnitude, such as with absolute temperatures, numerical difficulties probably related to ill-conditioned matrices were encountered. This problem disappeared when the higher order variables (i.e., x^2, x^3) were renormalized to make them roughly the same order of magnitude as the linear terms. The User Form would not need to be aware of the scaling except for its plotting functions; however, if the user scales the independent variables, the scaling factors must be identified for the output plots to be meaningful. Most conveniently the factors should be in a row just above the block of independent variables. It is mandatory that the scale factor for x be explicitly unity, and the other scale factors are constants used as follows,

$$x^j_{\text{Scaled}} = k_{\text{S}}\, x^j \tag{10}$$

This problem appeared only in Excel 2003 and not in earlier versions. In any event, scaling to normalize the independent variables is probably always desirable in critical applications to minimize numerical errors.

The last box is a check box used to select a homogeneous model with the constant set equal to zero.

The Visual Basic for Applications (VBA) code for this User Form is very cluttered with bookkeeping and plotting functions, so it is not included in this paper. The form is readily available on the author's academic web site (Jeter, 2004). In outline, the subprogram is organized as follows:

(1) Input data and parameters are identified, a new output page is created, and some information and preliminary data are posted on the output page.

(2) The basic full regression, which identifies the regression parameters, the constant and coefficients, is performed, some summary output data are posted, and the regression model is calculated and posted on the output sheet.

(3) The corrected dependent variables, according to Equation (5) are calculated, the corresponding conditional uncertainties are computed, and the results are posted on the output sheet.

(4) Preparations are made to plot the data points and the smooth curve representing the model, the roughly parallel curves representing Uncertainty A of the data, and the roughly hour-glass shaped curves representing the Combined Uncertainty of the model. The model and the error limits are tabulated in 41 rows to ensure that smooth plots result. Actual Excel formulas, not fixed values, are tabulated for the Uncertainty A cells and the Combined Uncertainty cells so that the user can later update the coverage factor inflator and the Uncertainty B data at will.

(5) Some summary data including the trivial average Uncertainty A of the data and the important average Combined Uncertainty of the model are computed. These data are posted in cells P2 and P3 on the output sheet.

(6) Finally the smooth versions of the model, the Uncertainty A error limits for the data, the Combined Uncertainty error limits of the model, and the experimental data are plotted on a separate chart.

Note that the code uses the regression utility from the Excel Data Analysis Tool Package. This usage absolutely mandates that the VBA version of that Tool Package be installed and be identified as a so-called Excel Add-in. In addition, the form assumed that a particular and convenient default chart has been defined by the user. This chart called Typ-XY is incorporated in the Excel workbook called Default.xls that is available from the author's web site (Jeter, 2002).

EXAMPLE APPLICATIONS

The first example is processing of vapor pressure data in a typical undergraduate laboratory exercise. In this experiment, vapor pressures of the modern refrigerant R-134A are measured over a range of temperatures. The data are processed and then regressed and plotted according to the classical Clausius-Clapeyron model. In this model, the log of the vapor pressure normalized by unit pressure is regressed on a polynomial of the inverse temperature. Such models are almost universally used to represent vapor pressure data. The two models considered are the linear formulation,

$$\ln\left(\frac{P_v}{P_0}\right) = c + b\frac{1}{T} \tag{11}$$

and the quadratic formulation, which is represented by the following equation,

$$\ln(\frac{P_v}{P_0}) = C + b_1 \frac{1}{T} + b_2 \frac{1}{T^2} \tag{12}$$

An example output plotted by the User Form called Poly_Regress is shown in Figure 2. In the figure, note that the model has some slight curvature, which tends to justify the use of the quadratic model for this data. Also note the roughly parallel curves representing the Uncertainty A of the data and the roughly hour-glass shaped curves representing the Combined Uncertainty of the model. This uncertainty analysis would be very tedious if done manually, but the User Form makes this fairly challenging example very easy to execute.

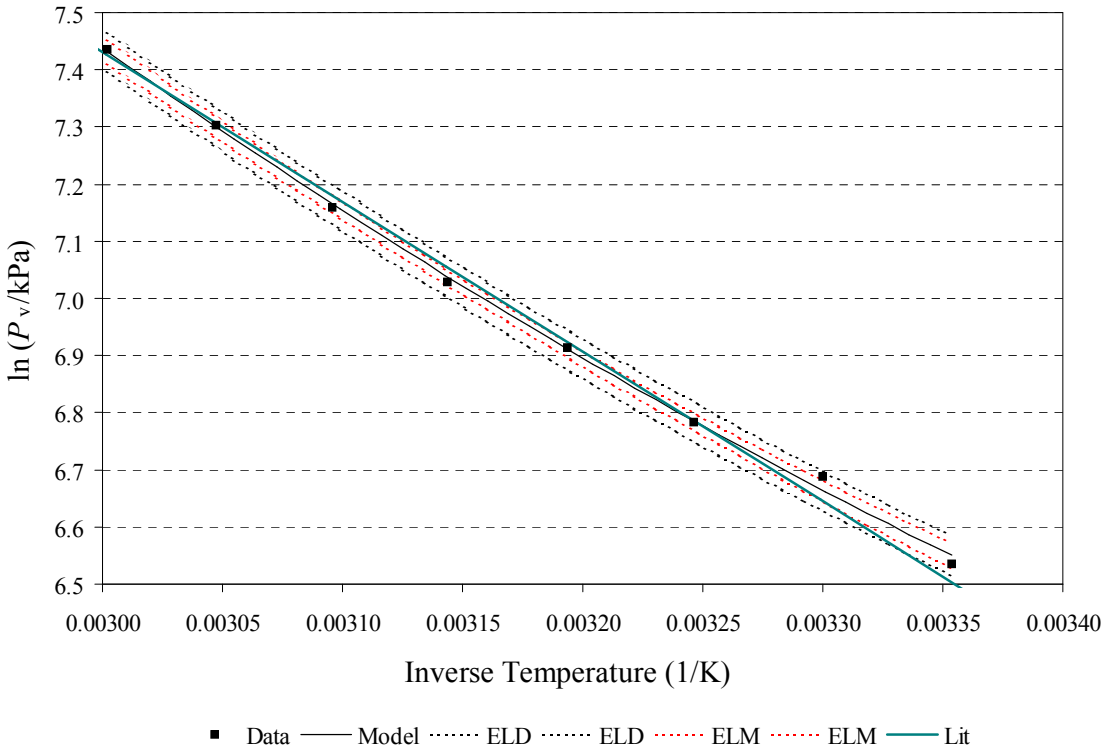

Figure 2. Plots of Vapor Pressure Data, Model, and Uncertainty Limits.

The next example involves calibration. Figure 3 shows the results of regression analysis of calibration data for a research quality constant temperature thermal anemometer.

The thermal anemometer data is processed in the form of a normalized wind speed versus a homogeneous fourth degree polynomial of a normalized bridge voltage. The calibration function recommended for this instrument is shown in the following equation,

$$Y = b_1 X + b_2 X^2 + b_3 X^3 + b_4 X^4 \tag{13}$$

In Equation (13) the normalized velocity, Y, is 10 times the ratio of the current velocity to the maximum used in the calibration

$$Y = 10 \frac{V}{V_{\max}}$$

and the normalized voltage, X, is calculated using the current voltage, the still air voltage e_0, and $e_{\max}$ the voltage at maximum wind speed, or

$$X = \left(\frac{e - e_0}{e_{\max} - e_0} \right)$$

Note that the normalization makes the use of a homogeneous function at least reasonable if not mandatory.

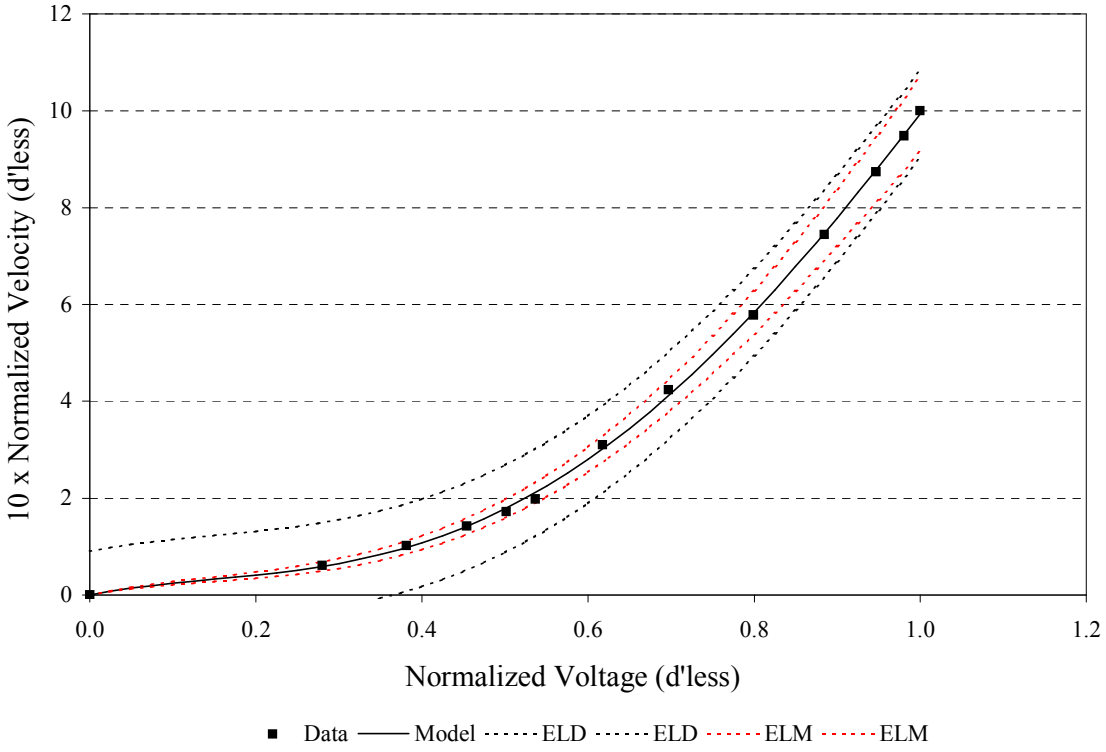

Figure 2. Plots of Thermal Anemometer Calibration

In this example the Uncertainty B of the calibration was arbitrarily set to zero to emphasize that the Uncertainty A of the model was required to be zero at the origin by the choice of a homogeneous model. Note also that the influence coefficient multiplier

was set to 5 here to exaggerate the error limits, which otherwise would be almost invisible.

CONCLUSION

An Excel User Form especially programmed for polynomial regression analysis and the accompanying uncertainty analysis has been described and presented. This form makes the extra rather complicated steps of error analysis related to polynomial models simple and easy. The underlying theory and two practical examples were also presented.

REFERENCES

Jeter, S. M., 2002, "Spreadsheet Default.xls for Setting a Convenient Default Graph", the George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, available on line at <me.gatech.edu/sheldon.jeter>.

Jeter, S. M., 2003, "Evaluating the Uncertainty of Polynomial Regression Models Using Excel", Proceedings of the 2003 ASEE Conference and Exposition, Knoxville, TN, June 2003.

Jeter, S. M., 2004, "Spreadsheet Regress_04.xls Containing the User Form Poly_Regress.frm for Regression Analysis of Polynomial Models", the George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, available on line at <me.gatech.edu/sheldon.jeter>.

Taylor, B. N. and P. J. Mohr, 1999, "The NIST Reference on Constants, Units, and Uncertainty", NIST Physics Laboratory, NIST, Gaithersbery, MD, 23 July 1999, available online at <http://physics.nist.gov/cuu/Uncertainty/index.html>.

Biography

SHELDON M. JETER is Associate Professor of Mechanical Engineering at the George W. Woodruff School of Mechanical Engineering at Georgia Tech. He has degrees from Clemson University, the University of Florida, and Georgia Tech. He has been on the academic faculty at Georgia Tech since 1979. His research interests are thermodynamics, experimental engineering, heat and mass transfer, and energy systems