Obtaining Uncertainty Measures on Slope and Intercept of a Least Squares Fit with Excel's LINEST

Faith A. Morrison
Professor of Chemical Engineering
Michigan Technological University, Houghton, MI 39931

17 July 2014

Most of us are familiar with the Excel graphing feature that puts a trendline on a graph. For example, some experimental data of temperature versus time are shown in Figure 1. The trendline was inserted as follows: Right click on data on chart, Add trendline, Linear, Display Equation on chart, Display R-squared value on chart. The trendline function, however, does not give us the value of the variances that are associated with the slope and intercept of the linear fit. If we wish to report the slope within a chosen confidence interval (95% confidence interval, for example), we need the values of the variance of the slope, s_m^2. Excel has a function that provides this statistical measure; it is called LINEST. In this handout, we give the basics of using LINEST.

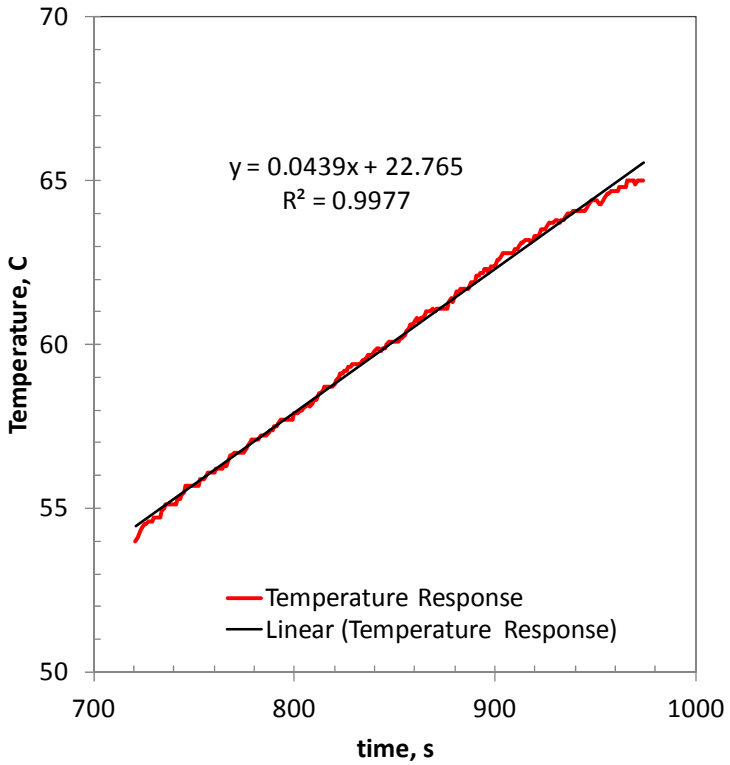

Figure 1: Temperature read from a thermocouple as a function of time. The trendline feature of Excel has been used to fit a line to the data; the equation for the line and the coefficient of determination R^2 values are shown on the graph.

Excel Array Function LINEST

The MS Excel function LINEST carries out an ordinary least squares calculation. For the data shown in Figure 1, we apply LINEST as follows (instructions are for a PC):

1. Select a blank range of five rows by two columns (10 cells total) to store the output of the function; we choose B1:C5 as shown shaded in Figure 2.
2. Click on *Formulas* and then "Insert Function."
3. In the *Insert Function* window, choose category "Statistical" and function "LINEST;" then click on OK.
4. Select the y and x data ranges; for "Const," enter TRUE (TRUE=calculate an intercept rather than having zero be the intercept) and for "Stats," also choose TRUE (TRUE=list the error estimates); click on OK.
5. Specify that LINEST is an array function by selecting the formula in the entry field and pressing CTRL-SHIFT-ENTER (Note: the Analysis ToolPak-VBA must be activated before this step; often this is already the case in later editions of Excel, but for Excel 2007 you may need to do this manually). The ten selected output cells will populate with statistics associated with the fit as labeled in Figures 2 and 3 and discussed below.

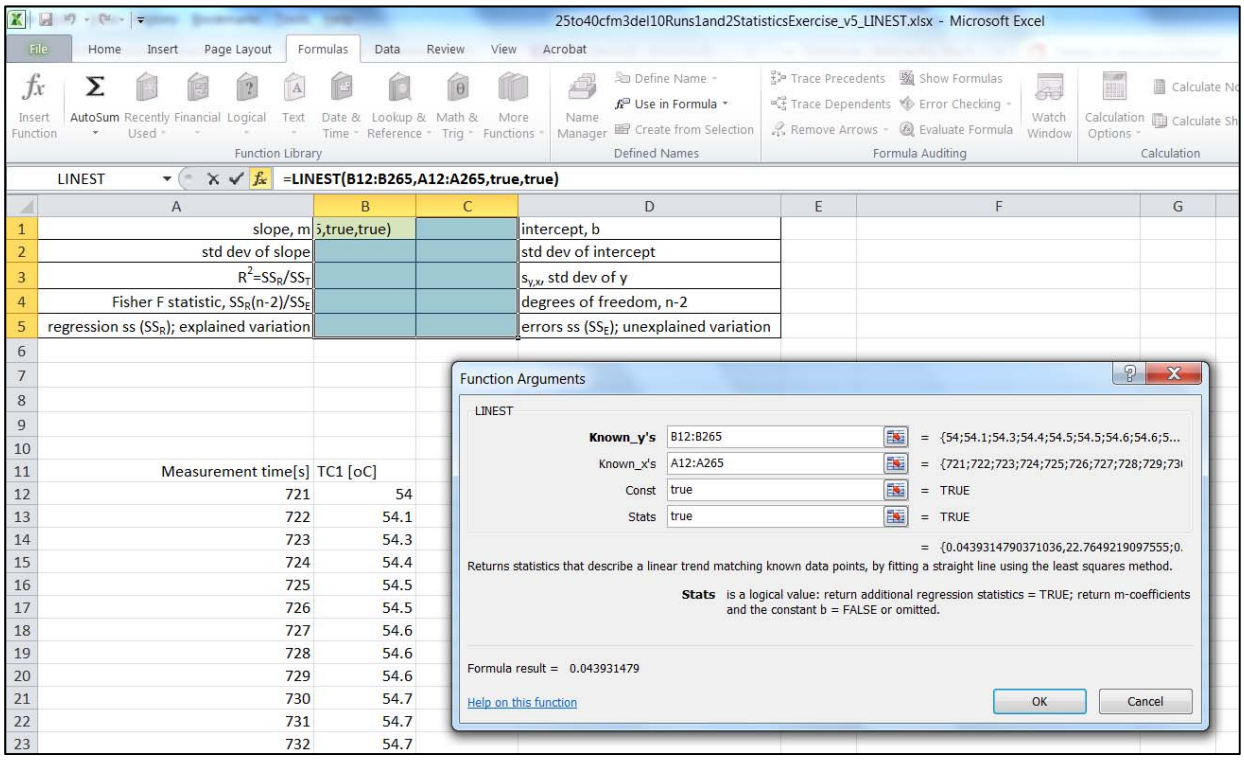

Figure 2: Following the instructions in the text, we fill in the function arguments of LINEST as shown. After clicking on OK there is one final important step: highlight the function call "=LINEST(B12:B265,A12:A265,true,true)"and press CTRL-SHIFT-ENTER.

A	B	C	D
slope, m	0.043931479	22.76492191	intercept, b
std dev of slope	0.000131455	0.111823909	std dev of intercept
$R^2=SS_R/SS_T$	0.997748762	0.153614504	$s_{y,x}$, std dev of y
Fisher F statistic, $SS_R(n-2)/SS_E$	111686.4212	252	degrees of freedom, n-2
regression ss (SS_R); explained variation	2635.510932	5.946548808	errors ss (SS_E); unexplained variation

Figure 3: After specifying that LINEST is an array function, the ten cells B1:C5 populate with the statistics shown. The labels are not provided by Excel; symbols are defined in the text.

Meaning of LINEST Results

LINEST performs an ordinary least squares calculation (Wikipedia, 2014b). The least squares process of solving for the slope and intercept for the best fit line is to calculate the sum of squared errors between the line and the data and then minimize that value. In ordinary least squares it is assumed that there are no errors in the x-values. For other assumptions of this analysis, see Appendix A. See the literature (Montgomery and Runger, 2014; McCuen, 1985) for detailed derivations; we give a brief discussion here.

The values (x_i, y_i) are a set of n data pairs to which we wish to fit a line; $\bar{y} \equiv (\sum_{i=1}^{n} y_i)/n$ is the mean value of the y_i values, and the linear model we are fitting is $\hat{y}(x) = \hat{m}x + \hat{b}$. To explain the values returned by Excel, we begin by defining three sums of squares: SS_{yy}, SS_E, and SS_R.

Total sum of squares
$$SS_T = SS_{yy} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (1)$$

Error Sum of Squares
$$SS_E \equiv \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (2)$$

Regression Sum of Squares
$$SS_R \equiv SS_T - SS_E \quad (3)$$

SS_{yy} is the sum squared error between the data and the mean of the data $\bar{y}$; SS_E is the sum of squared error between the data and the line $\hat{y} = \hat{m}x + \hat{b}$; SS_R is the difference between these two and represents the portion of the total sum of squares that can be explained by the linear model. In ordinary least squares we minimize SS_E. Two more useful sums of squares that appear in the least-squares formulas and LINEST results are

$$SS_{xx} \equiv \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad (4)$$

$$SS_{xy} \equiv \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \tag{5}$$

where $\bar{x} \equiv (\sum_{i=1}^{n} x_i)/n$ is the mean value of the x_i values. SS_{xx} is directly calculable with the Excel function DEVSQ(xrange) and SS_{yy} is available with the Excel function DEVSQ(yrange).[1] More is said about the various sums of squares below.

We seek to fit the n data points (x_i, y_i) to the linear model given here:

$$\hat{y} = \hat{m}x + \hat{b} \tag{6}$$

The ten statistics calculated by LINEST are (note that the order used below differs from the order of Excel's ten LINEST cells):

1. $\hat{m}$, Least Squares Estimator of the Slope – the slope of the ordinary least squares best-fit line; also available with the Excel function SLOPE(yrange,xrange).

$$\hat{m} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{SS_{xy}}{SS_{xx}} \tag{7}$$

   The two calculation formulas given in equation 7 may be shown to be equivalent by straightforward algebra.

2. $\hat{b}$, Least Squares Estimator of the Intercept – the intercept of the ordinary least squares best-fit line; also available with the Excel function INTERCEPT(yrange,xrange).

$$\hat{b} = \frac{(\sum_{i=1}^{n} x_i)^2(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i y_i)(\sum_{i=1}^{n} x_i)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \bar{y} - \hat{m}\bar{x} \tag{8}$$

   The two calculation formulas given in equation 8 may be shown to be equivalent by straightforward algebra.

3. $n - p$, Least Squares Degrees of Freedom. There are n data points, and $p = 2$ regression parameters. Before performing the least squares calculation we have n degrees of freedom. We use two degrees of freedom in calculating the slope and intercept, leaving $n - 2$ degrees of freedom in subsequent calculations.

4. $s_{y,x}$, Standard deviation of $y(x)$ (square root of the variance $s_{y,x}^2$ of $y(x)$; also available with the Excel function STEYX(yrange,xrange):

$$s_{y,x}^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^{n} (y_i - \hat{y})^2 = \frac{SS_E}{n-2} \tag{9}$$

[1] SS_{xy} is not directly available in Excel but may be calculated as follows: SUMPRODUCT(xrange,yrange)-COUNT(xrange)*AVERAGE(xrange)*AVERAGE(yrange).

The variance is defined as the error sum of squares divided by the degrees of freedom. This quantity is used in constructing confidence intervals and prediction intervals (error bars) for values of y; see the discussion in the section *Predictions with the Model*.

5. s_m, Standard Deviation of Slope, $\hat{m}$ (square root of s_m^2, the variance of $\hat{m}$).

$$s_m^2 = \frac{s_{y,x}^2}{SS_{xx}} \tag{10}$$

where $s_{y,x}^2$ is the variance of $y(x)$ (see equation 9). The formulas for s_m^2 in equation 10 and for s_b^2 in equation 14 (below) may be derived from a propagation of error calculation based on equations 7 and 8 (see Appendix B).

To construct confidence intervals around the calculated $\hat{m}$ and $\hat{b}$, we use the t-distribution and $n-2$ degrees of freedom (Montgomery and Runger, 2011; p 421). Note that for degrees of freedom greater than or equal to 6, $t_{0.025,n-2\geq 6} \approx 2$ (to one significant figure on error).

$$\text{95\% Confidence Interval on slope: } \hat{m} \pm t_{0.025,n-2} s_m \tag{12}$$

$$\cong \hat{m} \pm 2s_m \ \ (n-2 \geq 6) \tag{13}$$

The value of $t_{\alpha/2,n-2}$ may be obtained from Excel with the call $-\text{T.INV}(\alpha/2, n-2)$, where for 95% confidence $\alpha = 0.05$.[2]

6. s_b, Standard Deviation of Intercept $\hat{b}$ (square root of s_b^2, the variance of $\hat{b}$). Confidence intervals on $\hat{b}$ are constructed with s_b and the t-distribution with $n-2$ degrees of freedom.

$$s_b^2 = \frac{s_{y,x}^2 \sum_{i=1}^{n} x_i^2}{nSS_{xx}} = s_{y,x}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right) \tag{14}$$

$$\text{95\% Confidence Interval on intercept: } \hat{b} \pm t_{0.025,n-2} s_b \tag{15}$$

$$\cong \hat{b} \pm 2s_b \ \ (n-2 \geq 6) \tag{16}$$

The two versions of s_b^2 in equation 14 may be shown to be equivalent by straightforward algebraic manipulations.

7. 7a) Total sum of squares $SS_T = SS_{yy}$ (not given by LINEST but easily calculated from the LINEST results by summing two quantities that are given, SS_E and SS_R). SS_T is the total sum of the squares of the difference between the data y_i and the average $\bar{y}$; it is a measure of the total error made when assuming the y-data are constant and equal to the mean (equation 1 repeated below):

$$SS_T = SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{17}$$

[2]The correct call is to the function "T.INT," which in Excel returns a left-tailed inverse of the Student's t distribution (a negative number). The function "TINT" (without the period) also works, but this returns the two-tailed inverse of the Student's t distribution, which is positive and twice the value of the T.INT result.

7b) Residual sum of squares of errors SS_E—sum of squares of difference between the data and the linear model $\hat{y}_i$; a measure of the error made in assuming the y-data are characterized by the linear model. This is the leftover SS_T variation that is unexplained by the model. When $SS_E \to 0$, all the total error SS_T is explained by the linear model, and we can conclude that the linear model is a very good fit (equation 2 repeated below).

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{18}$$

8. Regression sum of squares SS_R—the portion of SS_T that is explained by the model. SS_R is defined as (equation 3 repeated below):

$$SS_R = SS_T - SS_E \tag{19}$$

See the coefficient of determination discussion (R^2, equation 20) and the Fisher F Statistic discussion (equation 21) for more uses of SS_R.

9. R^2 Coefficient of Determination—fraction of the variability of the y_i accounted for by the linear model:

$$R^2 = \frac{\text{explained error}}{\text{total error}} = \frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T} \tag{20}$$

When the model is a very good fit, there is little deviation between the data and the model; then, $SS_E \to 0$ and $R^2 = 1$. Note, however, that if the model is a horizontal line, the model is $\hat{y} = \bar{y}$, and SS_T is equal to SS_E, and R^2 is zero. The coefficient of determination is a measure of goodness of fit except when the data are nearly constant.

10. Fisher F Statistic—used in a test of the regression to see if using two parameters (slope and intercept) is justified over using one parameter ($\hat{y} = \bar{y}$; i.e. zero slope and $\bar{y} =$ intercept). The F statistic for a regression is calculated as the ratio of two quantities, the variance explained by the model to variance unexplained by the model (McCuen, 1985; p191; Wikipedia, 2014a):

$$F = \frac{\text{"lack of fit" sum of squares}/\nu_1}{\text{"pure error" sum of squares}/\nu_2} = \frac{SS_R/\nu_1}{SS_E/\nu_2} = \frac{(SS_T - SS_E)}{s^2_{y,x}} \tag{21}$$

where $\nu_1 = 1$ and $\nu_2 = n - 2$ are the degrees of freedom for each of the sources of variation shown. This ratio is the computed value of a random variable having an $F(\nu_1, \nu_2)$ distribution (another common population distribution; compare to the normal or Student's t distributions) with degrees of freedom $\nu_1 = 1$ and $\nu_2 = n - 2$. If $F > F_{crit}$, then using the linear model $\left(\hat{y} = \hat{m}x + \hat{b}\right)$ is justified (at the $(1 - \alpha)$% confidence level) over using the model $\hat{y} = \bar{y}$. F_{crit} corresponds to the cumulative distribution function of the $F(\nu_1, \nu_2)$ distribution with α equal to the desired confidence level and degrees of freedom ν_1 and ν_2. See the literature for more on the Fisher F statistic (Wikipedia, 2014c; McCuen, 1985).

Predictions with the Model $\hat{y} = mx + b$

In equations 13 and 16, we gave the 95% confidence intervals for the two model parameters $\hat{m}$ and $\hat{b}$. These confidence ranges are appropriate to use in error propagation calculations when the model parameters $\hat{m}$ and $\hat{b}$ are used directly in subsequent calculations.

When the model equation is used to estimate values of y at a chosen value of x, different error limits are appropriate. The two most common cases are discussed here.

1. Estimate the best value of y at a chosen value of x. The best value of y at any point will be the mean of all possible observed values of y at that point. Let the chosen value of x be x_p, and the best estimate of y at that point be y_p, which is given by

$$y_p = \hat{m}x_p + \hat{b} \tag{22}$$

The variance for y_p is calculated from equation 22 and an error propagation calculation (see Appendix B), with the complication that the slope and the intercept are not independent quantities, and thus there is a nonzero covariance between $\hat{m}$ and $\hat{b}$. The result for the variance of the mean value of y at x_p is

Variance of the mean value of y at x_p

$$s_{y,x}^2\left(\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}\right) \tag{23}$$

(Recall that SS_{xx} may be calculated with the Excel call DEVSQ(xrange).) The appropriate confidence interval for the mean value of y at x_p is obtained from the standard deviation of the mean value along with the t-distribution with $n - 2$ degrees of freedom (Montgomery and Runger, 2011; p 422):

Confidence interval for the mean value of y at x_p

$$\left(\hat{m}x_p + \hat{b}\right) \pm t_{\alpha/2,n-2}s_{y,x}\sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}} \tag{24}$$

Equation 24 is the appropriate interval to use for error bars on y-values obtained from the least squares best fit (Figures 3 and 4). Note that the error bars are narrowest near the center point of the regression $(\bar{x}, \bar{y})$ and fan out towards the ends. This reflects the fact that uncertainty in the slope makes values at the ends of the x-range less certain than points near the center.

2. Estimate the predicted *new* value of y at a chosen value of x. The n experimental values $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (x_n, y_n)$ were used in the calculation of the least squares results. If an additional point (x_p, y_p) is now taken, its expected value is just the mean value of y at the desired x, $y_p = \hat{m}x_p + \hat{b}$. Thus, the expected values of both the mean value of y at x_p and of a *new* value of y at x_p are the same.

The *uncertainties* in these two quantities at x_p are not the same, however. The mean value of y at x_p can be quite well known: as n increases, both terms in equation 23 go to zero, making the variance go to zero as well. Taking more data makes us more certain of the mean response at x_p.

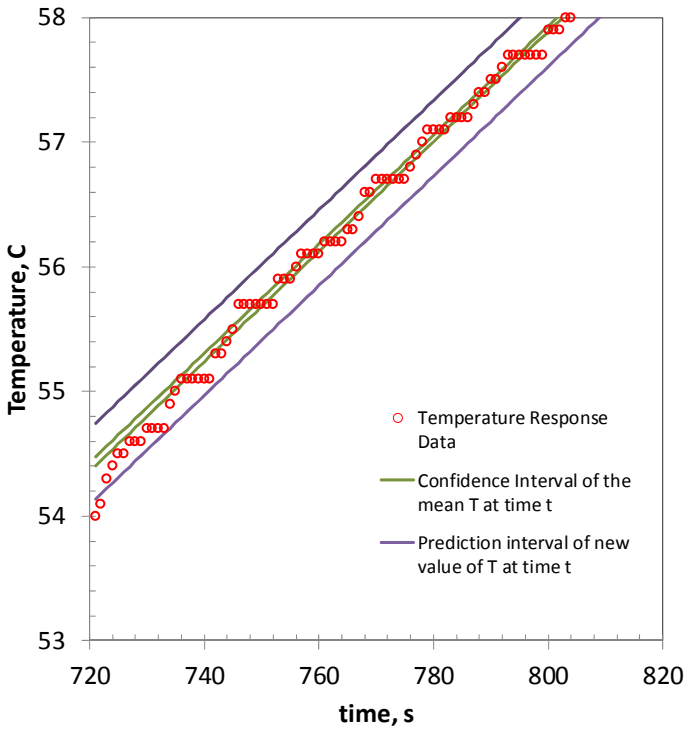

Figure 3: A portion of the data from Figure 1 is shown along with (inner pair of green lines), the 95% confidence interval on the mean values of y (temperature) at each value of x (time). The outer pair of lines (purple) reflect the 95% prediction interval on new values of y at each value of x.

A new value of y at x_p will be subjected to the random variations of the process being studied, and thus will always be more uncertain than the mean response. The correct calculation of the variance of a predicted new value of y at x_p is (Montgomery and Runger, 2011; p 423)

Variance of the predicted new value of y at x_p

$$s_{y,x}^2\left(1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{SS_{xx}}\right) \tag{25}$$

Note that as n increases, the variance does not go to zero, but rather goes to $s_{y,x}^2$. The prediction interval for the new value of y at x_p is given by the t-distribution with $n-2$ degrees of freedom; this interval is appropriate for showing the expected uncertainty in this prediction (Montgomery and Runger, 2011).

Prediction interval for the new value of y at x_p

$$\left(\hat{m}x_p+\hat{b}\right)\pm t_{\alpha/2,n-2}s_{y,x}\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{SS_{xx}}} \tag{26}$$

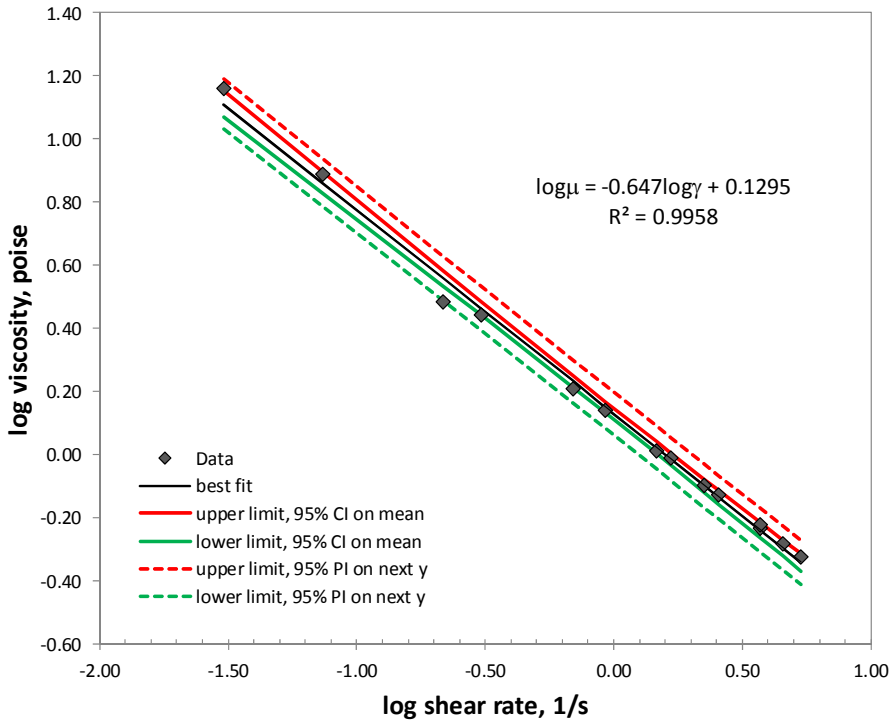

Figure 4: Confidence intervals of the mean indicate how well we know the average value of y at every x; prediction intervals for the next value of y indicate the expected span of the data. Usually, as is true here, most of the data will fall in the 95% prediction interval of the next value of y. Data shown are for viscosity of star branched polylstyrene at 379°C, MW=333 kg/mol; from G. Kraus and J. T. Gruver, *J. Polym. Sci. A*, **3**, 105 (1965).

As n goes to infinity, the terms under the square root go to one and the t-distribution goes to the normal distribution. Note that the error bars for new values of y, like those on the mean values of y, are narrowest near the center point of the regression $(\bar{x}, \bar{y})$ and fan out towards the ends. The prediction intervals for a new value of y are wider than the confidence intervals on the mean values of y (Figures 3 and 4).

References

P. R. Bevington (1969) *Data Reduction and Error Analysis for the Physical Sciences* (McGraw Hill, New York)

W. A. Fuller (1987) *Measurement Error Models* (Wiley, New York).

R. H. McCuen (1985) *Statistical Methods for Engineers* (Prentice Hall, Englewood Cliffs, NJ).

D. C. Montgomery and G. C. Runger (2011) *Applied Statistics and Probability for Engineers*, 5th edition (Wiley, New York)

B. C. Reed (1989) “Linear least-squares fits with errors in both coordinates,” *Am. J. Phys*. **57**(7), 642-646; also, erratum *Am. J. Phys***. 58**(2) 189 (1990).

Wikipedia (2014a) “Lack-of-fit sum of squares,” Wikipedia, the Free Encyclopedia, en.wikipedia.org/wiki/Lack-of-fit_sum_of_squares, accessed 14 July 2014.

Wikipedia (2014b) “Ordinary Least Squares,” Wikipedia, the Free Encyclopedia, en.wikipedia.org/wiki/Ordinary_least_squares, accessed 14 July 2014.

Wikipedia (2014c) “F-Test,” Wikipedia, the Free Encyclopedia, en.wikipedia.org/wiki/F-test, accessed 15 July 2014.

Appendix A: Assumptions of Ordinary Least Squares

In developing the equations for linear regression the following assumptions are made (Montgomery and Runger, 2011; Wikipedia, 2014b):

1. Weak exogeneity. This assumption is that there are no errors in the x values, only in the y values. In Figure A-1 we show the difference between assuming y-errors only versus allowing errors in both x and y. Methods exist for the case where both errors are accounted for (Reed, 1989; Fuller, 1987)
2. Linearity. The mean of the response variable y is a linear combination of the parameters (slope and intercept) and the predictor variable (x).
3. Constant variance $s_{y,x}^2$ (homoscedasticity). Constant variance implies that different response variables (y_i) have the same variance in their errors, regardless of the values of the predictor variable (x_i).
4. Independence of errors. The errors of the response variables are uncorrelated with each other, that is, the errors in each of the values of y_i are not correlated (no covariance terms among the y_i)
5. Lack of multicollinearity in the predictors (We cannot fit to x and 2x as two different predictor variables in our model; there would be a redundancy if we did so.)

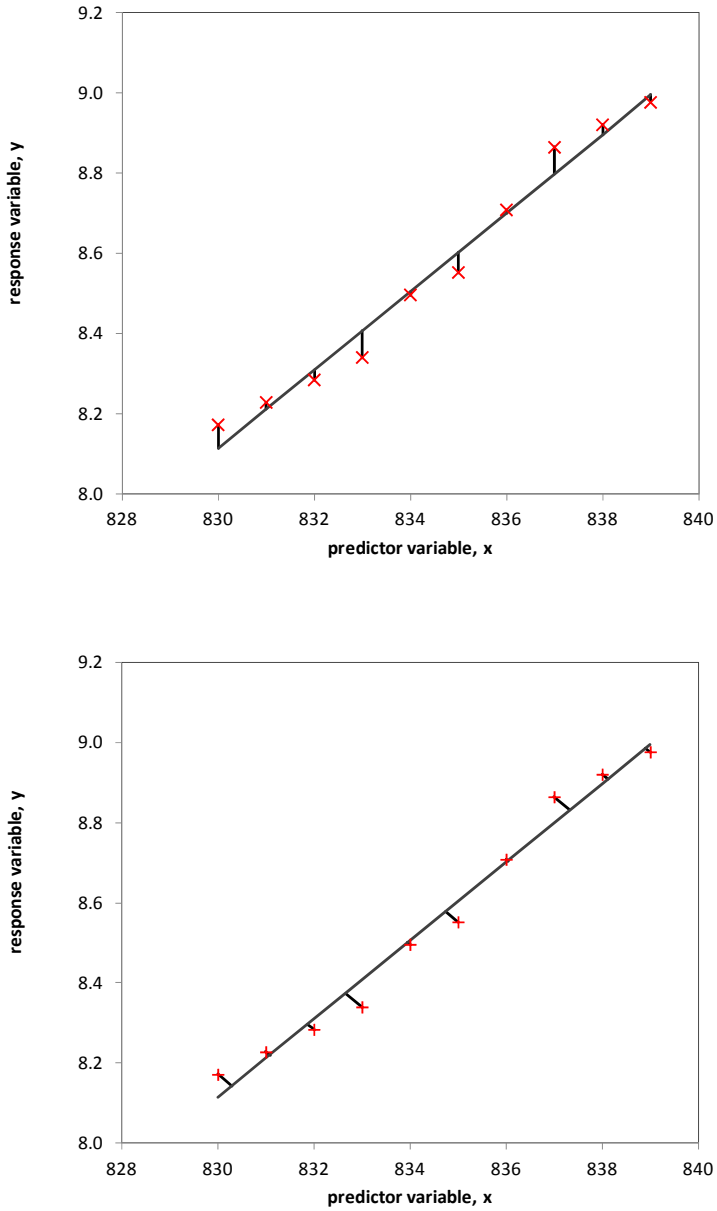

Figure A-1: The ordinary least squares algorithm assumes that there are no errors in the x-values of the data; the residuals of the response variable y are calculated as $y_i - \hat{y}_i$, the vertical distance between the point and the line (top). The shortest distance between the point and the line is the perpendicular distance as shown in the plot on the bottom; for this case, errors in x are assumed to be present as well (Reed, 1989; Fuller, 1987).

Appendix B: Variance of a Function of Random Variables

For a linear function y of p variables x_i,

Linear function f

$$y = f(x_1, x_2, \ldots x_p) = c_1x_1 + c_2x_2 + c_3x_3 \ldots c_px_p = \sum_{i=1}^{p} c_ix_i \tag{B-1}$$

the variance of the function is given by (Montgomery and Runger, 2011; p182):

Variance of linear function f

$$\sigma_f^2 = \sum_{i=1}^{p} c_i^2\sigma_i^2 + 2\sum_{i<j}^{p}\sum_{j=2}^{p} c_ic_j\mathrm{Cov}(x_i, x_j) \tag{B-2}$$

where $\mathrm{Cov}(x_i, x_j)$ is the covariance of the variables x_i and x_j. If the variables are all mutually independent, the covariance is zero. In error propagation usage of equation B-2, the covariance terms may be nonzero if there are systematic errors causing the individual variables x_i to be correlated.

For a general, nonlinear function $y = f(x_1, x_2, \ldots x_p)$, the equation for the variance of the function is

Nonlinear function f

$$y = f(x_1, x_2, \ldots x_p) \tag{B-3}$$

Variance of nonlinear function f

$$\sigma_f^2 = \sum_{i=1}^{p}\left(\frac{\partial f}{\partial x_i}\right)^2\sigma_{x_i}^2 + \sum_{i<j}^{p}\sum_{j=2}^{p} 2\left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_j}\right)\mathrm{Cov}(x_i, x_j) \tag{B-4}$$

We can obtain equation B-2 by straightforward evaluation of equation B-4 with equation B-1. We can obtain equations 10 and 14 from evaluation of equation B-4 with equations 7 and 8 (the variables are the y_i; the x_i are assumed constant). Also, we can calculate equation 23 and 25 from equations B-4 and equation 22; note, however, that since $\hat{m}$ and $\hat{b}$ are both calculated from $y_1, y_2, \ldots y_n$, they are correlated, and $\mathrm{Cov}(\hat{m}, \hat{b})$ is not zero. It may be shown that the covariance of slope and intercept is given by (Montomery and Runger, 2011):

Covariance of least squares slope and intercept

$$\mathrm{Cov}(\hat{m}, \hat{b}) = -\frac{s_{x,y}^2\bar{x}}{SS_{xx}} \tag{B-5}$$