Regression Statistics

Background

Engineers often deal with fitting data to equations. For example, in the course of an experiment, the correlation between two observables may appear linear so the data is fit to a straight line. Engineers also use correlations developed by others in the course for their work. For example, the temperature dependence of the vapor pressure is often fit to the Antoine Equation. This aids in the dissemination of the knowledge as it is easier to create a table of compound-specific parameters than to list all of the experimental data for every compound of interest.

Whenever data is fit to an equation, a certain amount of error is present. One reason is that the data are almost never exactly predicted by the model. Another is that the data may just not follow the model. Thus, using correlations introduces error into subsequent calculations and affects analyses and conclusions. This document outlines how to use statistics to describe the errors associated with regression.

Fitting Data to Straight Lines

The Model

Many phenomena encountered by engineers may be described by straight lines. Expressed mathematically, fitting data to a straight line is described by

$$Y_i = b_0 + b_1 X_i + e_i \tag{1}$$

where b_0 is the intercept, $b_1$is the slope, X_i is the i-th value of the independent variable, Y_i is the i-th value of the dependent variable corresponding to X_i and e_i is the residual or error between what the model predicts and what the data show. Mathematically $e_i = Y_i - \hat{Y}_i$ where Y_i is the value of the dependent variable corresponding the X_i. The *sum squared error* (SS_E) is an important quantity when fitting equations as is defined as

$$SS_E = \sum_{i=1}^{n} e_i^2 \tag{2}$$

When you fit data to a line in Excel using the solver, you minimize SS_E by changing the slope and intercept.) Another important quantity is the *mean square error* (MS_E) because it is an estimate of the variance of the fit ($\hat{\sigma}^2$). It is defined as

$$MS_E = \hat{\sigma}^2 = \frac{SS_E}{n-2} \tag{3}$$

The predicted value for the ith dependent variable, $\hat{Y}_i$, at X_i is given by

$$\hat{Y}_i = b_0 + b_1 X_i \tag{4}$$

Confidence Intervals on the Slope and Intercept

Consider a set of n points of (X, Y) data. When the data are fit to a straight line, the resulting slope and intercept are statistically only *estimates* of the *true* slope and intercept. One way to state the uncertainty of these parameters is using *confidence intervals*. For *significance level* α, the $(1-\alpha)100\%$ confidence interval on the intercept is

$$b_0 \pm S_{b_0} t_{n-2,1-\frac{\alpha}{2}} \tag{5}$$

where $t_{n-2,1-\frac{\alpha}{2}}$ is the value of the Student's T distribution for n-2 degrees of freedom and at significance level $\frac{\alpha}{2}$ and S_{b_0} is the standard error of the intercept defined by

$$S_{b_0} = \left(\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X_i - \bar{X})^2} \right)^{0.5} \hat{\sigma} \tag{6}$$

Here, $\bar{X}$ denotes the average of the X data. The corresponding confidence interval on the slope is

$$b_1 \pm S_{b_1} t_{n-2,1-\frac{\alpha}{2}} \tag{7}$$

where S_{b_1} is the standard error of the slope defined by

$$S_{b_1} = \left(\frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)^{0.5} \hat{\sigma} \tag{8}$$

The R^2 Statistic

A common statistic used to determine the overall goodness of the fit is the R^2 statistic. It is defined as

$$\begin{aligned} R^2 &= \frac{SS\ due\ to\ regression}{(Total\ SS, corrected\ for\ the\ mean\ \bar{Y})} \\ &= \frac{SS_R}{SS_T} \\ &= \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \end{aligned} \tag{9}$$

Statistically speaking, R^2 gives the "proportion of total variation about the mean $\bar{Y}$ explained by the regression." Said another way, R is the correlation between Y and $\hat{Y}$ and is termed the "Pearson's Correlation Coefficient" or the "multiple regression coefficient.

Confidence Interval of a Predicted Y Value

Once a line is fit to obtain a slope and intercept, new Y values can be predicted for any X. The $(1-\alpha)100\%$ confidence interval for this predicted Y value, is

$$\hat{Y}_0 \pm S_{\hat{Y}} t_{n-2,1-\frac{\alpha}{2}} \tag{10}$$

where $\hat{Y}_0$ is the predicted value of the dependent variable for a given value of the independent variable X_0 The standard error of this predicted Y value, $S_{\hat{Y}}$, is given by

$$S_{\hat{Y}} = \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)^{0.5} \hat{\sigma} \tag{11}$$

Notice that $S_{\hat{Y}}$ is dependent upon a particular value of X. As you move further away from the mean of the X data, the confidence interval become wider.

Expected Range of Collected Data

If the experiment is repeated several times, the data collected will come from the same underlying distribution as the original experiments. We can therefore predict, with a certain confidence, where future data will lie. The $(1-\alpha)100\%$ range in which a future observation Y_0 at the value of X_0 will be found is given by

$$Y_0 = \hat{Y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^{0.5} \hat{\sigma} \tag{12}$$

Notice that this range will never approach zero no matter how many data points are taken. The smallest it can become is $\hat{\sigma}$.

Inverse Prediction

Inverse prediction refers to finding a value for X corresponding to a given Y. This problem occurs often in engineering. For example, the temperature of saturated steam is usually determined by measuring the saturated pressure calculating the corresponding temperature from the Antoine Equation. If the governing equation was determined by regression of data, there is error associated with calculated X value. Given Y_0, the predicted value for X, termed $\hat{X}_0$, is given by

$$\hat{X}_0 = \frac{Y_0 - b_0}{b_1} \tag{13}$$

The $(1-\alpha)100\%$ confidence interval for $\hat{X}_0$ is given by

$$\hat{X}_0 + \frac{b_1(Y_0 - \bar{Y})}{\lambda} \pm \frac{t_{n-2,1-\frac{\alpha}{2}}}{\lambda} \hat{\sigma} \left(\frac{(Y_0 - \bar{Y})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \lambda\left(\frac{n+1}{n}\right)\right) \tag{14}$$

where

$$\lambda = b_0^2 - t_{n-2,1-\frac{\alpha}{2}}^2 S_{b_1}^2 \tag{15}$$

Matrices and Fitting to Any Linear Model

Matrices offer a useful notation for describing regression. Suppose we have the following data that we wish to fit to a straight-line model.

X	Y
21	186
24	214
32	288
47	425
50	455
59	539
68	622
74	675
62	562
50	453
41	370
30	274

The following matrices may be defined.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 21 \\ 1 & 24 \\ \vdots & \vdots \\ 1 & 41 \\ 1 & 30 \end{bmatrix} \quad \boldsymbol{Y} = \begin{bmatrix} 186 \\ 214 \\ \vdots \\ 370 \\ 274 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \quad \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n-1} \\ e_n \end{bmatrix} \tag{16}$$

Here, $\boldsymbol{X}$ is a matrix of the independent variable in its functional form for each term of the model, $\boldsymbol{Y}$ is the vector of "Y" data corresponding to each "X" value, $\boldsymbol{b}$ is the vector of parameters in the model, and $\boldsymbol{e}$ is a vector of the errors.

If we wanted to fit the data to a second-order polynomial, namely

$$Y_i = b_0 + b_1 X_i + b_2 X_i^2 + e_i \tag{17}$$

similar matrix notation could be used. The only changes come in the $\boldsymbol{X}$ and $\boldsymbol{b}$ matrices which become

$$\boldsymbol{X} = \begin{bmatrix} 1 & 21 & 441 \\ 1 & 24 & 576 \\ \vdots & \vdots & \vdots \\ 1 & 41 & 1681 \\ 1 & 30 & 900 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \tag{18}$$

Notice that $\boldsymbol{X}$ now has an extra column which is the X data squared and $\boldsymbol{b}$ has an extra row. A similar procedure could be followed to add any number of terms. The analysis below gives the least squares solution for any *linear* regression.

In matrix notation, the model for any linear equations is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{19}$$

Solving for $\boldsymbol{b}$ requires taking derivatives and the derivation is found in many textbooks on statistics. The solution is

$$\boldsymbol{b} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T Y} \tag{20}$$

where $\boldsymbol{X^T}$ is the transpose of $\boldsymbol{X}$. The *sum squared error*, written in matrix notation, is

$$SS_E = \boldsymbol{Y^T Y} - \boldsymbol{b^T X^T Y} \tag{21}$$

The *mean squared error* is given by

$$MS_E = \hat{\sigma}^2 = \frac{SS_E}{n-p} \tag{22}$$

where p is the number of fitting parameters (the number of rows in $\boldsymbol{b}$).

Confidence Intervals for the Parameters

The $(1-\alpha)100\%$ confidence intervals for each of the parameters can be calculated from

$$b_i \pm S_{b_i} t_{n-p,1-\frac{\alpha}{2}} \tag{23}$$

where S_{b_i} is the standard error of b_i which is calculated as the square root of the i-th diagonal term of the matrix $\left(\boldsymbol{X^T X}\right)^{-1} \hat{\sigma}^2$

Confidence Interval (Region) of a Predicted Y Value

The $(1-\alpha)100\%$ confidence interval for this predicted Y value, is

$$\hat{\hat{Y}}_0 \pm S_{\hat{Y}} t_{n-p,1-\frac{\alpha}{2}} \tag{24}$$

where $\hat{Y}_0$ is the predicted value of the dependent variable for a given value of the independent variable X_0. The standard error of this predicted Y value, $S_{\hat{Y}}$, is given by

$$S_{\hat{Y}} = \left(\boldsymbol{X_0^T}\left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X_0}\right)^{0.5} \hat{\sigma} \tag{25}$$

The R^2 Statistic

The R^2 statistic for a fit to any linear model is given by Equation 8. In matrix notation, this equation can be written

$$R^2 = \frac{\boldsymbol{b^T X^T Y} - n\bar{Y}^2}{\boldsymbol{Y^T Y} - n\bar{Y}^2} \tag{26}$$

This statistic can be deceptive for higher-order models. For example, if a certain data set had 10 points, a 10th order polynomial would exactly fit each point and the R^2 would be 1. Though the fit is perfect, it is probably not useful for predicting trends as it would have many loops. For this reason an *adjusted* R^2 value, which gives a more reliable indication on the predictive power of the correlation, is calculated by

$$R_a^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-p)} \tag{27}$$