Lecture 25-27 - Linear Regression - Theory

(A) Theory
(B) Problems
(C) MATLAB Applications

Text: Chapter 11, Walpole, Myers, and Myers

25.1 What is Linear Regression?

Linear Regression is a technique used to translate a collection of data into a model of best fit. Linear regression provides a means to say whether a model is even capable of describing the data.

25.2 An example

For example, consider Einstein's relation for the diffusivity of a substance. Einstein said that the true, long-time diffusivity of a substance, D, is proportional to the ratio of the mean square displacement (MSD), $\left\langle \Delta x(\tau)^2 \right\rangle$, over the observation time, τ. Einstein's relation can be expressed as

$$D = \lim_{\tau \to \infty} \frac{\left\langle \Delta x(\tau)^2 \right\rangle}{6\tau} \tag{25.1}$$

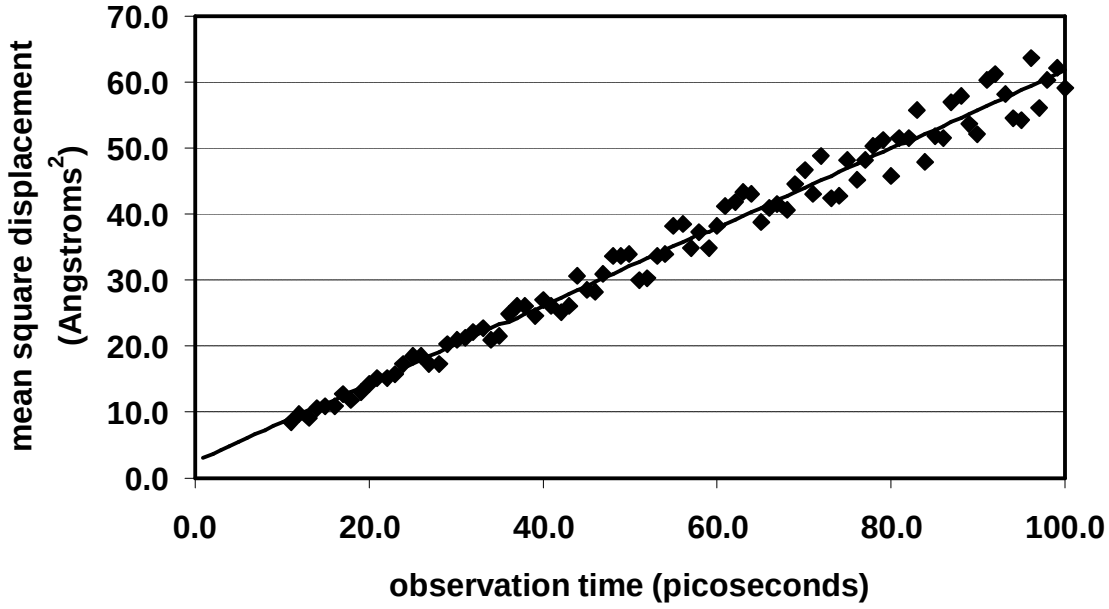
which can be rearranged as

$$\lim_{\tau \to \infty} \left\langle \Delta x(\tau)^2 \right\rangle = 6D\tau \tag{25.2}$$

Since the short-time behavior does not obey this model, the form of Einstein's relation with which data is analyzed is

$$\left\langle \Delta x(\tau)^2 \right\rangle = 6D\tau + D_o \tag{25.3}$$

Once can perform molecular dynamics computer simulations of Nuclear Magnetic Resonance (NMR) experiments to obtain data for the MSD, $\left\langle \Delta x(\tau)^2 \right\rangle$, as a function of the observation time, τ. The data for such an experiment looks like this:



One then wants to calculate what the diffusivity is from this data. Linear regression provides a means to do that. The data points are the diamonds and the solid line is the "best fit model" obtained by linear regression. From equation (3) we see that the slope of this best fit line is $6D$ and the y-intercept is D_o. So we can obtain our diffusivity by dividing the slope of the best fit line by 6. Linear regression will give us the slope and the y-intercept of the best fit model.

25.3 Terminology

In the above example, the mean square displacement is called the "dependent variable" or the "response" and is usually denoted by y.

The observation time is an "independent regressor variable", usually denoted by x.

The slope, $6D$, and y-intercept, D_o, of the best-fit model are usually referred to as b and a respectively.

The equation for the best-fit model is then

$$\hat{y} = bx + a \qquad (25.4)$$

where $\hat{y}$ is the best-fit prediction of y at a value x.

25.4 Derivation of Least Square Regression

The data points do not fall exactly on the best-fit line. Each data point has some error, e, from the best-fit line, so that for the i^{th} point

$$y_i = bx_i + a + e_i \tag{25.5}$$

The best-fit model will minimize the error. In particular we wish to minimize the Sum of the Square of Errors, SSE, defined as

$$SSE \equiv \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - bx_i - a)^2 \tag{25.6}$$

In order to minimize the function SSE with respect to b and a, we take the partial differential of SSE with respect to b and a, and set them equal to zero, and then solve for b and a.

$$\frac{\partial SSE}{\partial a} = -2\sum_{i=1}^{n} (y_i - bx_i - a) = 0 \tag{25.7}$$

$$\frac{\partial SSE}{\partial b} = -2\sum_{i=1}^{n} (y_i - bx_i - a)x_i = 0 \tag{25.8}$$

Equations 25.7 and 25.8 represent a system of two linear equations and two unknowns. Solving for b and a yields

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{25.9}$$

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x} \tag{25.10}$$

where $\bar{x}$ and $\bar{y}$ are the average values of the set of x and y respectively.

25.5 The Variance of the Regression Coefficients a and b

From page 366, of Walpole, Myers and Myers.

$$\sigma_b^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{25.11}$$

$$\sigma_a^2 = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2 \tag{25.12}$$

where σ^2 is the model error variance. An unbiased estimate of σ^2 is s^2 where

$$\sigma^2 \approx s^2 = \frac{SSE}{n-2} \tag{25.13}$$

25.6 Obtaining a "measure of fit", ANOVA

The measure of fit, MOF, provides a way to determine if the best-fit model is a good model. A common definition of the MOF is

$$MOF = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{25.14}$$

where the sum of the squares of the errors, SSE, was defined in equation (25.6), and where the Sum of the Squares of the Regression, SSR, is

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{25.15}$$

and the Sum of the Squares of the Total variance, SST, is

$$SST = SSR + SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{25.16}$$

A value of MOF of 1.0 means the model fits the data perfectly. The farther the value of the MOF is below 1.0, the worse the fit of the model.

A more rigorous method for evaluating the goodness of the fit is to use the f-distribution, (Walpole, Myers, & Myers, p. 232-233). We can define a variable, f, by

$$f = \frac{SSR}{s^2} \tag{25.17}$$

When

$$f > f_\gamma(1, n-2) \tag{25.18}$$

there is a significant amount of variation in the response of the dependent variable, y, accounted for by the model. $f_\gamma(1, n-2)$ is the critical value for the F-distribution as located in Appendices, Table A.6, pp. 687-690, WMM. n is the number of data points. γ is the level of confidence.

25.7 An example employing the measure of fit

Let's return to the data in example 25.2. There we decided to model the data after Einstein's relation. However, that is not the only such theory for the relationship between the mean square displacement, MSD, and the observation time. One-dimensional hard-rod theory (1DHRT) says that the MSD is proportional to the *square root* of the observation time.

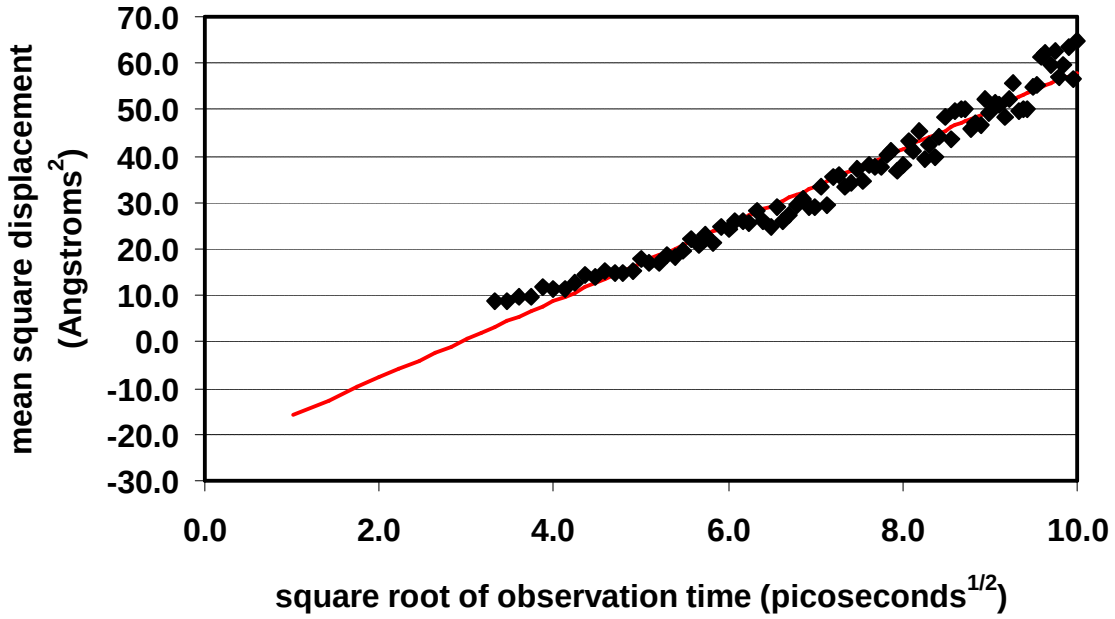
$$\alpha = \lim_{\tau \to \infty} \frac{\left\langle \Delta x(\tau)^2 \right\rangle}{2\sqrt{\tau}} \tag{25.19}$$

where α is a mobility. This equation can be rearranged into a form analogous to that of Einstein's relation (equation (25.3)),

$$\left\langle \Delta x(\tau)^2 \right\rangle = 6\alpha\sqrt{\tau} + \alpha_o \tag{25.20}$$

where α is a mobility. Now, given a set of data, it is necessary to determine if the MSD displacement better fits the model of Einstein or 1DHRT.

First of all, one must see that the independent regressor variable x is $\sqrt{\tau}$ for 1DHRT. (x was τ for Einstein's relation. Plotting the same data in example 25.2 versus $\sqrt{\tau}$, and fitting to 1DHRT in equation (25.18), we have:



Does Einstein's model or 1DHRT give a better fit to the data? Performing a linear least squares regression on the data using both the Einstein relation (25.3) and 1DHRT (25.18), we find the following results:

parameter	**equation**	**Einstein's Relation**	**1-D Hard Rod Theory**
b	25.9	0.596 Å^2/ps	8.17 Å^2/ps$^{1/2}$
a	25.10	1.96 Å^2	-23.8 Å^2
σ_b^2	25.11	6.28E-05 Å^4/ps^2	0.0248 Å^4/ps
σ_a^2	25.12	0.236 Å^4	1.37 Å^2
s^2	25.13	3.81 Å^4	7.82 Å^4
SSE	25.6	335.4 Å^4	688.3 Å^4
SSR	25.15	21621.6 Å^4	21268.7 Å^4
SST	25.16	21957.0 Å^4	21957.0 Å^4
MOF	25.14	0.985	0.969
$n-2$		88	88
f	25.17	5672.4	2719.2
$f_{.01}(1,88)$	Table A.6	7	7

The MOF is better (closer to 1.0) for Einstein's relation. The value of f both cases satisfies equation (25.18). However the Einstein's relation case is much higher, indicating a better fit.

25.8 Multiple Linear Regression

In many instances, the dependent variable, y, is a function of several independent regressor variables, $\{x\}$. This relationship can be expresses in the following form where the i subscript indicates a particular data point and the j subscript refers to the j^{th} regressor variable. m is the number of independent regressor variables (as specified by the model).

$$y_i = b_0 + \sum_{j=1}^{m} b_j x_{j,i} + e_i \tag{25.21}$$

The model is then

$$\hat{y}_i = b_0 + \sum_{j=1}^{m} b_j x_{j,i} \tag{25.22}$$

The method for finding the best-fit parameters, $\{b\}$, for this system is exactly analogous to what we did for the single-variable linear regression. We define, the Sum of the Squares of the Error, SSE, exactly as we did before in equation (25.6)

$$SSE \equiv \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{m} b_j x_{j,i} \right)^2 \tag{25.23}$$

We take the partial derivatives of SSE with respect to each of the parameters in $\{b\}$ and set them equal to zero. (Remember there are n data points and m regressor variables. The i index counts n and the j index counts m.) This gives $m+1$ independent, linear equations of the form:

$$b_0 \sum_{i=1}^{n} x_{0,i} x_{j,i} + b_1 \sum_{i=1}^{n} x_{1,i} x_{j,i} + b_2 \sum_{i=1}^{n} x_{2,i} x_{j,i} + \ldots + b_m \sum_{i=1}^{n} x_{m,i} x_{j,i} = \sum_{i=1}^{n} y_i x_{j,i} \tag{25.24}$$

for j = 1 to m and where $x_{0,i} = 1$ for all i. This set of equations is linear in $\{b\}$. It can be solved using any standard technique for the solution of linear equations. The equation in matrix form looks like:

$$\underline{\underline{A}}\,\underline{b} = \underline{g} \tag{25.25}$$

where

$$A_{j,k} = \sum_{i=1}^{n} x_{k-1,i} x_{j,i} \tag{25.26}$$

$$b_j = b_j \tag{25.27}$$

and

$$g_j = \sum_{i=1}^{n} y_i x_{j,i} \tag{25.28}$$

To determine the variances of the parameters in multiple linear regression, we use analogous equations as for the single-variable case. We have our SSE. The variances are

$$\sigma_{b_j}^2 = A_{j,j}^{-1} \sigma^2 \tag{25.29}$$

where $A_{j,j}^{-1}$ is the j,jth element of the inverse of $\underline{\underline{A}}$ (which is not the inverse of the j-jth element of $\underline{\underline{A}}$). The covariances are

$$\sigma_{b_j,b_k}^2 = A_{j,k}^{-1} \sigma^2 \tag{25.30}$$

where, as in the single-variable case, an unbiased estimate of σ^2 is given by s^2, which is defined as

$$\sigma^2 \approx s^2 = \frac{SSE}{n-m-2} \tag{25.31}$$

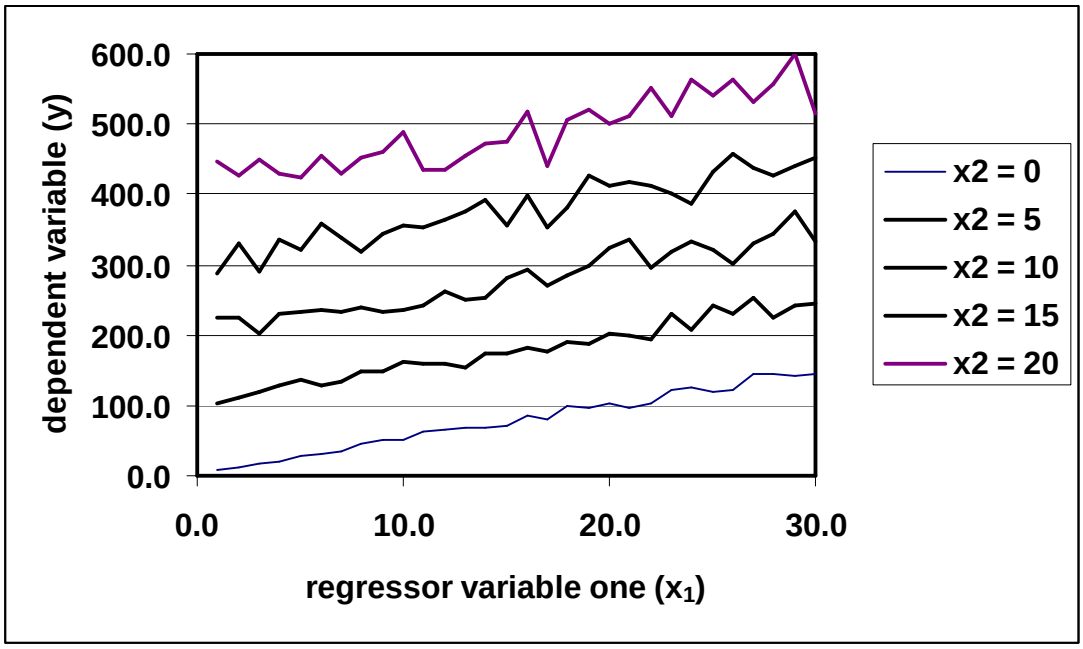
The measure of fit for the multiple regression case is defined in exactly the same way as the single-variable regression case (equation (25.14)).

25.9 An example of multiple linear regression

We are given data, plotted in the graph below, which shows that y has some functional dependence on two variables, x_1 and x_2. Using a model of the form:

$$\hat{y}_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} \tag{25.32}$$

find the parameters of the model and state whether the model is a good one for the data.



Applying the multiple linear regression techniques, we find:

$b_0 = 0.56$ $\sigma_{b_0} = 3.20$ $MOF = 0.989$

$b_1 = 4.90$ and $\sigma_{b_1} = 0.15$ and $f = 6845.4$

$b_2 = 20.40$ $\sigma_{b_2} = 0.18$ $n - m - 2 = 147$

What conclusions can we make? We have good estimates of b_1 and b_2 because the standard deviations, σ_{b_1} and σ_{b_2} are much smaller than the values of the parameters. However, the value standard deviation for b_0 is much larger than the actual value. What this can be interpreted to say is that the value of b_0 is 0.56 +/- 3.20, which means it is between -2.64and 3.76. Therefore, very little importance should be attached to the figure of 0.56.

In terms of the fit of the model, we get a very nice MOF, indicating that the model is right. We would just require more data points to pin down the b_0 more precisely. Do the parameters make physical sense? Yes. In the plot, y rises with x_1 and it rises with x_2. That is what the positive values of b_1 and b_2 tell us it should do.

I generated the data in the plot randomized around the following model:

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} \qquad (25.33)$$

where $b_0 = 2.0,\ b_1 = 5.0,\ b_2 = 20.0$. You can see how well the data was fit.

25.10 Polynomial Regression

Data can be fit to a polynomial in precisely the same way that it can be fit to multiple variables. In multivariate regression, we fit to a model like:

$$\hat{y}_i = b_0 + \sum_{j=1}^{m} b_j x_{j,i} \tag{25.22}$$

If we make the substitution that

$$x_{j,i} = \left(x_{1i}\right)^j \tag{25.34}$$

then we have a polynomial model of the form:

$$\hat{y}_i = b_0 + \sum_{j=1}^{m} b_j x_i^{\,j} \tag{25.35}$$

where m is now the order of the polynomial. This model is solved in precisely the same manner as the multivariate regression. In matrix notation, the problem becomes:

$$\underline{\underline{A}}\,\underline{b} = \underline{g} \tag{25.36}$$

where

$$A_{j,k} = \sum_{i=1}^{n} x_i^{k-1} x_i^j = \sum_{i=1}^{n} x_i^{j+k-1} \tag{25.37}$$

and

$$g_j = \sum_{i=1}^{n} y_i x_i^j \tag{25.38}$$

25.11 An example of Polynomial Regression

We are given the data in the following plot (solid diamonds) and asked to determine the functional relationship between y and x. These suggested models are supplied:

a linear model:

$$y_i = b_0 + b_1 x_i \tag{25.39}$$

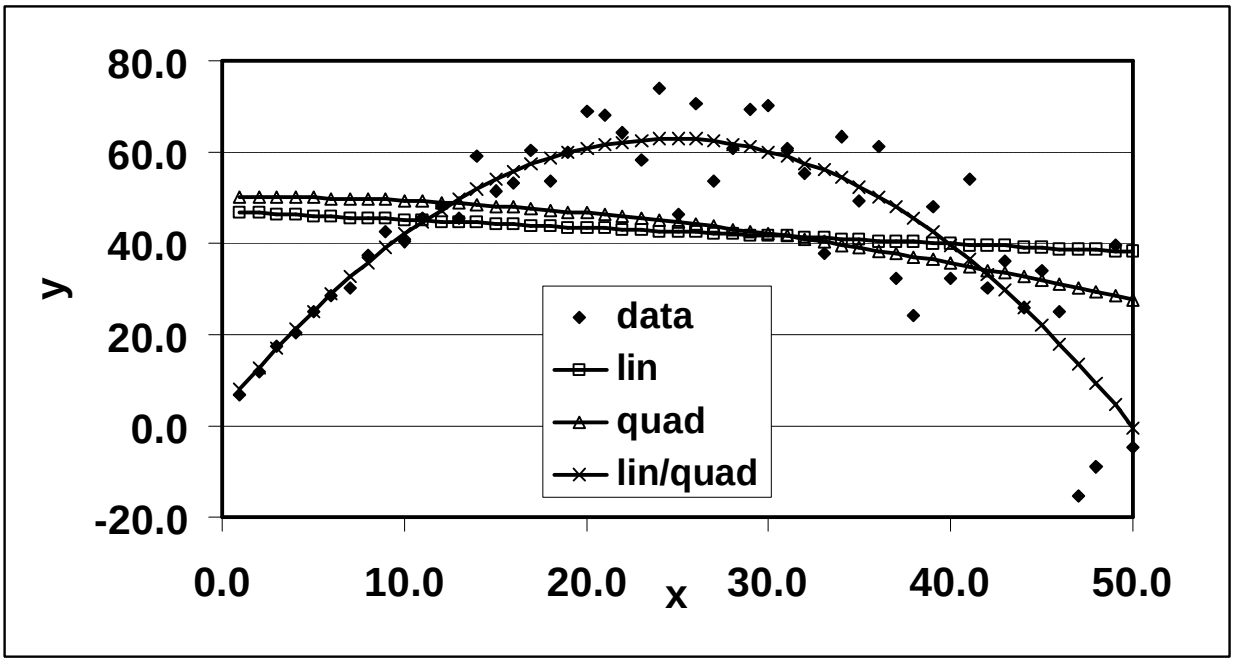
a quadratic model:

$$y_i = b_0 + b_2 x_i^2 \tag{25.40}$$

a linear and quadratic model:

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 \tag{25.41}$$

Determine which of these models bests fits the data and find the parameters, $\{b\}$, for the model of best fit.



Calculating the regression for each of these models we find:

parameter	linear model	quadratic model	linear and quadratic model
b_0	46.95161	50.31831	3.564038
b_1	-0.17569	NA	4.83057
b_2	NA	-0.00914	-0.09816
σ_{b_0}	6.143767	4.345695	4.68266
σ_{b_1}	0.209684	NA	0.423575
σ_{b_2}	NA	0.003792	0.008052
MOF	0.014415	0.107968	0.76321
$n-2$	48	48	47
f	0.702031	5.809732	75.74414

The measure of fit clearly indicated that the linear and quadratic model best fits the data. The parameter set for the linear and quadratic model is

$$b_0 = 3.56$$
$$b_1 = 4.83$$
$$b_2 = -0.098$$

Actually, I generated this data using randomization around a linear and quadratic model with the parameter set:

$$b_0 = 2$$
$$b_1 = 5.0$$
$$b_2 = -0.1$$

25.12 Regression in MATLAB

On the website, there is a code called regress.m. This code will perform:

- single-variable linear regression
- multivariate linear regression
- single-variable polynomial regression

in a straight-forward manner. It is also capable of performing with multivariate polynomial regression with a modicum of ingenuity. It is not capable of performing non-linear regression.

The description for how to use the file can be obtained by opening MATLAB, moving to the directory where you have downloaded the regress.m file, and typing
`help regress`
This yields:

```
regress(type,k,n,bo,'fname')
  This script will perform a least squares regression on a set of input data

  type = 1 for multivariate linear regression
  type = 2 for single variable polynomial regression
  For type = 1, k = number of variables in linear regression
  For type = 2, k = order of polynomial regression
  n = the number of data points in the input file
  b0 = 0, force the zero-order parameter to be zero, otherwise include it
  fname is the input filename
  For this program, the file name must be 'file.anything.dat' !!!
  (You need single quotes around the filename.)
  This program creates an output file 'regress.out'.

  Author:  David Keffer Date: October 22, 1998
```

So, for example, if you entered, at the MATLAB command line interface,

```
regress(1,2,100,1,'file.problem1.dat')
```

this would perform a single-variable quadratic fit of the form

$y = m_0 + m_1 x + m_2 x^2$ on 100 data points in the file file.problem1.dat. The format of the data in file.problem1.dat should have 2 columns, the first with y values and the second with x-values, for example:

```
15.10065279  8.1
20.4980324   1.3
25.56136963  19.4   etc…
```

Or, if you entered, at the MATLAB command line interface,

```
regress(1,2,200,0,'file.problem2.dat')
```

this would perform a multivariate linear fit of the form

$y = m_1 x_1 + m_2 x_2$ on 200 data points in the file file.problem2.dat. The format of the data in file.problem2.dat should have 3 columns, the first with y values, the second with x_1-values, and the third with x_2-values, for example:

```
-83.62305844   18   10
-86.66736601   20   10
29.37498574    2    20   etc…
```

Note, the input argument bo is just a logical key. If the input argument bo is zero, then the constant model parameter is defined to be zero. Otherwise, the constant model parameter is a variable and will be calculated to best fit the data.

25.13 Adding Confidence Intervals to Linear Regression Fits

(Taken from "Applied Statistics and Probability for Engineers", Montgomery, D.C., Runger, G.C., 2nd Ed., Wiley, 1999, page 545 ff.)
(Added to notes April 2, 1999)

Confidence Interval on Slope and Intercept

Under the assumption that the observations are normally and independently distributed, a $100(1-2\alpha)$ confidence interval on the slope, b, is given by

$$P(\hat{b} - t_{\alpha,n-2}\sqrt{\sigma_b^2} < b < \hat{b} + t_{\alpha,n-2}\sqrt{\sigma_b^2}) = 1-2\alpha \tag{25.42}$$

and a $100(1-2\alpha)$ confidence interval on the intercept, a, is given by

$$P(\hat{a} - t_{\alpha,n-2}\sqrt{\sigma_a^2} < a < \hat{a} + t_{\alpha,n-2}\sqrt{\sigma_a^2}) = 1-2\alpha \tag{25.43}$$

where

$$\sigma_b^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{25.11}$$

$$\sigma_a^2 = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2 \tag{25.12}$$

where σ^2 is the model error variance. An unbiased estimate of σ^2 is s^2 where

$$\sigma^2 \approx s^2 = \frac{SSE}{n-2} \tag{25.13}$$

Confidence Interval at any point

Under the assumption that the observations are normally and independently distributed, a $100(1-2\alpha)$ confidence interval on the slope, b, is given by

$$P(\hat{y}(x_o) - t_{\alpha,n-2}\sqrt{\sigma^2_{y(x_o)}} < y(x_o) < \hat{y}(x_o) + t_{\alpha,n-2}\sqrt{\sigma^2_{y(x_o)}}) = 1-2\alpha \qquad (25.44)$$

where

$$\hat{y}(x_o) = bx_o + a \qquad (25.4)$$

and where

$$\sigma^2_{y(x_o)} = \left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}\right]\sigma^2 \qquad (25.45)$$

where σ^2 is the model error variance. An unbiased estimate of σ^2 is s^2 where

$$\sigma^2 \approx s^2 = \frac{SSE}{n-2} \qquad (25.13)$$

This allows you to evaluate the confidence interval at every value of x, giving upper and lower confidence limits that are functions of x.