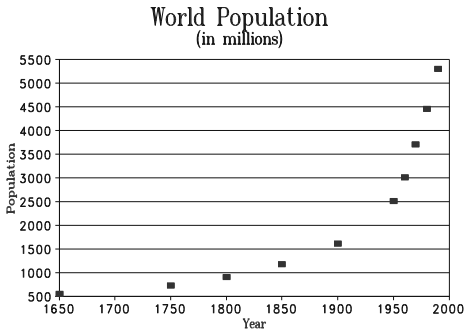
Modelling with the TI-85

Suppose you have a data set consisting of ordered pairs, $\{(x_i,y_i),\ 1 \le i \le n\}$, and you want to create a mathematical model of this data. Most applications consist of determining the linear function that best approximates the data. I want to look at polynomial approximations; how is the best polynomial approximation obtained and how good is it.

The data used to motivate this model was that of world population between 1650 and 1990 as shown in the adjacent figure. I have used polynomial approximations of data for years as examples of applications of matrix theory to real world problems.



Consider the model

$$y = \sum_{j=0}^{k-1} a_j x^j$$

where the a_j are constants to be determined. If the data set given above is substituted into this model we obtain a system of n equations in k unknowns where n > k. Write this system as

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{k-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{k-1} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 1 & x_n & x_n^2 & \dots & x_n^{k-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_{k-1} \end{bmatrix} .$$

Denote the n x k matrix of x-values by X and write this system as

$$\mathbf{y} = X\mathbf{a}$$

where **y** is the n-vector of y-values and **a** is the k-vector of coefficients to be determined. To "solve" such a system premultiply both sides by the transpose of X to obtain the normal equations:

Normal Equations: $X^T\mathbf{y} = (X^TX)\mathbf{a}$.

The matrix X^TX is a symmetric k x k matrix and it is known that this matrix is invertible if the x_j are distinct. The solution is given by

$$\mathbf{a} = (X^TX)^{-1}X^T\mathbf{y}$$

All of the above is "old hat" to anyone who has looked for a

best model of data in the least-squares sense. The TI-85 does an excellent job of obtaining these coefficients for all cases with k = 2, 3, 4, and 5. However, the correlation coefficient is given only in the case k = 2. Recall from elementary statistics that the correlation coefficient, r, satisfies

$$r^2 = 1 - \frac{SSE}{SSy}$$

where

$$SSE = \mathbf{y}^T\{I - X(X^TX)^{-1}X^T\}\mathbf{y}$$

and

$$SSy = \mathbf{y}^T\mathbf{y} - (1/n)(\mathbf{u}^T\mathbf{y})^2.$$

Here, I is the k x k identity and **u** is the n-vector consisting of all 1s.

When this modelling approach was applied to the population data exhibited above, both **a** and r^2 coincided with those given by the TI-85 in the case k = 2. The vector **a** coincided with the TI-85 in the case k = 3 but began to differ slightly with the case k = 4. However, in the case k = 5, r^2 was computed to be greater than 1; recall that $0 \leq r^2 \leq 1$.

Built in functions in the TI-85 can be used to compute r^2 using

$$SSE = [\ \text{norm}(\mathbf{y} - X\mathbf{a})\ ]^2$$

and

$$SSy = [\ \text{norm}(\ \mathbf{y} - (1/n)\ \det(\mathbf{u}^T\mathbf{y})*\mathbf{u}\ )\ ]^2.$$

The models for the population (P) data given for the years (Y) in the above scatter plot are given at the end of this paper.

This resolved the problem but it does not totally close the book. The following possibly related questions should be resolved:

1. Why do the models begin to differ when k is larger than 3?
2. Why is the matrix X^TX ill-conditioned; that is, unstable for relatively small values of k?

Data Comparisons on world population (P) figures between 1650 and 1990 (Y) using the above model and STAT on the TI-85 are given below. The subscript M indicates the matrix approach and the T subscript indicates the TI-85's norm approach to computing r^2. STAT is used to indicate the model obtained using the TI-85 statistics mode.

Linear:

Model $P = -21011 + 12449Y$,
$r_M^2 = 0.73016$, $r_T^2 = 0.73016$

STAT $P = -21011 + 12449Y$,
$r^2 = 0.73015$

Quadratic:

Model $P = 199614 - 229171Y + 65877Y^2$
$r_M^2 = 0.914655$, $r_T^2 = 0.914655$

STAT $P = 199614 - 229171Y + 65877Y^2$

Cubic:

Model $P = -2650495 + 4491394Y - 2533999Y^2 + 476164Y^3$
$r_M^2 = 0.97628$, $r_T^2 = 0.97631$

STAT $P = -2650572 + 4491521Y - 2534069Y^2 + 476176Y^3$

Quartic:

Model $P = 37057172 - 82850984Y + 69398149Y^2 - 25812443Y^3 + 3597374Y^4$
$r_M^2 = 1.05265$ (oops!), $r_T^2 = 0.99649$

STAT $P = 35500679 - 79427908Y + 66579633Y^2 - 24782608Y^3 + 3456477Y^4$