EX: Find a multiple regression (i.e., hyperplane) fit of tensile strength versus composition for the following data for the hardness of steel, (adapted from [1]):

TABLE I
FERRITIC STAINLESS STEELS

Product #		Composition %						
	Tensile Strength (MPa)	C	Mn	Si	Cr	P	S	Other
405	415	0.08	1.0	1.0	11.5-14.5	0.04	0.03	0.10-0.30 Al
409	415	0.08	1.0	1.0	10.5-11.75	0.045	0.045	0.75 Ti
429	450	0.12	1.0	1.0	14.0-16.0	0.04	0.03	-
430	450	0.12	1.0	1.0	16.0-18.0	0.04	0.03	-
434	530	0.12	1.0	1.0	16.0-18.0	0.04	0.03	0.75-1.25 Mo
436	530							
442	515	0.20	1.0	1.0	18.0-23.0	0.04	0.03	-
444	415							
446	515	0.20	1.5	1.0	23.0-27.0	0.04	0.03	0.25 N

SOL'N: The hyperplane is defined as follows:

There are various difficulties with the data and the vague problem statement that we wish to resolve before attempting to find the multiple regression fit. First, we observe that Mn, Si, P, and S are constant (except for the peculiar product 409). Since product 409 has one-of-a-kind composition with respect to Titanium, we ignore it. To obtain a meaningful fit, we need to either hold variables constant or have enough data to cover the possible combinations of variations of variables. Here, we will hold Mn %, Si %, P %, and S % constant, and we will model the effects of changes in C % and Cr %.

Second, we eliminate products with peculiar characteristics in the "Other" column (with the exception of product 405). Thus, we now eliminate products 434 and 446. We retain product 405 for no better reason than that we will need it in order to have enough data for the multiple regression, and intuition suggests that the small concentration of Aluminum makes little difference in the tensile strength. (Such questionable assumptions are often made when dealing with real data, and it is imperative that they be recognized, specified, and then verified. Here, we proceed without verification.)

Third, we eliminate products 436 and 444, owing to their lack of values for composition.

Fourth, we must decide how to deal with the Cr values that are given as ranges. An arbitrary decision is made to us the highest value in the range. This leaves us with only the four data points in Talbe II. The values that enter into the multiple regression are shown in bold. This means our solution is applicable only when Mn, Si, P, and S have the fixed values in the table.

TABLE II
FERRITIC STAINLESS STEELS SELECTED DATA

Product #		Composition %						
	Tensile Strength (MPa)	C	Mn	Si	Cr	P	S	Other
405	**415**	**0.08**	1.0	1.0	**14.5**	0.04	0.03	0.10-0.30 Al
429	**450**	**0.12**	1.0	1.0	**16.0**	0.04	0.03	-
430	**450**	**0.12**	1.0	1.0	**16.0**	0.04	0.03	-
442	**515**	**0.20**	1.0	1.0	**23.0**	0.04	0.03	-
	y	**x1**			**x2**			

The hyperplane for the regression fit is defined as follows:

$$y = (b, a_1, a_2) \circ (1, x_1, x_2) = b + a_1 x_1 + a_2 x_2$$

We write matrix terms for the data and hyperplane fit:

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & x_{13} & x_{23} \\ 1 & x_{14} & x_{24} \end{bmatrix} = \begin{bmatrix} 1 & 0.08 & 14.5 \\ 1 & 0.12 & 16.0 \\ 1 & 0.12 & 18.0 \\ 1 & 0.20 & 23.0 \end{bmatrix} \qquad (b, \vec{a}) = \begin{bmatrix} b \\ a_1 \\ a_2 \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} 415 \\ 450 \\ 450 \\ 515 \end{bmatrix}$$

The matrix equation for the data and hyperplane fit are as follows:

$$X \bullet (b, \vec{a}) = \vec{y}$$

A pseudoinverse of X yields the least-squares (i.e., regression) solution:

$$(b, \vec{a}) = X^{+} \vec{y}$$

where

$$X^{+} = (X^T X)^{-1} X^T$$

NOT'N: $X^T \equiv X$ transpose

Using Matlab™, we obtain the following result:

$$X^{+} = \begin{bmatrix} 0.25 & 5.125 & -3.0 & -1.375 \\ -13.75 & 38.125 & -30.0 & 5.625 \\ 0.1 & -0.550 & 0.4 & 0.050 \end{bmatrix}$$

and

$$(b,\vec{a}) = X^{+}\vec{y} = \begin{bmatrix} 351.875 \\ 846.875 \\ -0.250 \end{bmatrix}$$

Plugging the (x_1, x_2) coordinates into the hyper-plane equation yields the following estimated y values:

$$\hat{\vec{y}} = X \bullet (b,\vec{a})$$

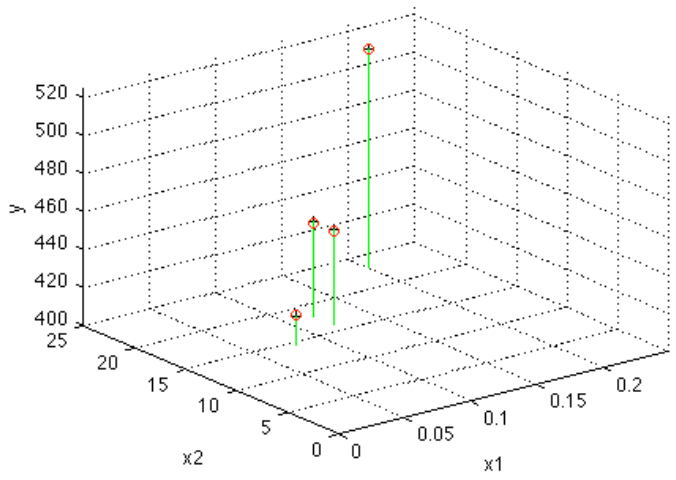
or

$$[\vec{x}_1,\vec{x}_2,\hat{\vec{y}}] = \begin{bmatrix} 0.08 & 14.5 & 416.0 \\ 0.12 & 16.0 & 449.5 \\ 0.12 & 18.0 & 449.0 \\ 0.20 & 23.0 & 515.5 \end{bmatrix}$$

The errors are found by subtracting estimated y values from actual y values:

$$\vec{e} = \begin{bmatrix} 415 \\ 450 \\ 450 \\ 515 \end{bmatrix} - \begin{bmatrix} 416.0 \\ 449.5 \\ 449.0 \\ 515.5 \end{bmatrix} = \begin{bmatrix} -1.0 \\ 0.5 \\ 1.0 \\ -0.5 \end{bmatrix}$$

In the plot below, the estimates are shown as circles:



NOTE: In this problem, we might do just as well by using linear interpolation with the x coordinates on a line connecting the first and last points and collapsing the middle two points into a single point with $y = 450$.

NOTE: More data would yield a more credible fit. With so few data points, extrapolating from the above fit would be hazardous.

REF: [1] Vladimir B. Ginzburg, *Steel-Rolling Technology, Theory and Practice,* New York, NY: Marcel Dekker, 1989, (Tables 3.13 and 3.15).