Multiple Linear Regression Using the Matrix Capabilities of the TI-85

Charles Bare

Chadron State College

Introduction

There are many excellent statistical software packages that will perform multiple regression analysis. Using the TI-85 to perform multiple regression should not supplant available, efficient statistical software. The intention herein is to use the graphing calculator to guide students through the processes of matrix formation and manipulation that lead to least squares regression coefficients and associated summary statistics. It is acknowledged that the following procedures could be easily programmed so that, conceivably, data input followed by a single keystroke would yield regression coefficients. However, programming the matrix manipulations would undesirably minimize student involvement in the estimation process.

Regression Models

It will be assumed that the student has studied multiple linear regression and, furthermore, the matrix approach has been sufficiently covered. The following theoretical regression models with k predictor variables will be estimated using the TI-85:

Linear: $\qquad$ $Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k + E$

Power: $\qquad$ $Y = B_0(X_1^{B}1)(X_2^{B}2) \ldots (X_k^{B}k)E$

Exponential: $Y = B_0(B_1^{X}1)(B_2^{X}2) \ldots (B_k^{X}k)E$

Data matrices **X**, **LX**, **Y**, **LY**, and **U** to be entered into the TI-85 are described below:

Predictors: $\qquad$ $\mathbf{X}_{nx(k+1)} =$

$$
\begin{bmatrix}
1 & X_{11} & X_{12} & \ldots & X_{1k} \\
1 & X_{21} & X_{22} & \ldots & X_{2k} \\
& & \cdot & \cdot & \cdot \\
1 & X_{n1} & X_{n2} & \ldots & X_{nk}
\end{bmatrix}
$$

Response: $\qquad$ $\mathbf{Y}_{nx1} = [\ Y_1,\ Y_2,\ \ldots,\ Y_n\ ]^T$

Unit vector: $\qquad$ $\mathbf{U}_{nx1} = [\ 1,\ 1,\ \ldots,\ 1\ ]^T$

LX and **LY** matrices are constructed using natural logarithmic transformations of all X_{ij} and Y_{ij} entries in the **X** and **Y** matrices, and are used to estimate power and exponential models.

Estimation

The estimated regression coefficient matrix for the linear model is $\mathbf{b} = [\ b_0,\ b_1,\ \ldots,\ b_k\ ]^T$, and is calculated using the matrix equation $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$. The formula to compute $\mathbf{b}$ can be easily entered into the TI-85 using the matrix capability. Hence,

the estimated linear model is given by

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots b_k X_k.$$

Scalars SSR, SSE, and SSTO are calculated and used to obtain the coefficient of determination R^2 (goodness-of-fit) and F statistic (statistical significance) as follows:

$$SSR = \mathbf{b}^T \mathbf{X}^T \mathbf{Y} - (1/n)(\mathbf{Y}^T \mathbf{U} \mathbf{U}^T \mathbf{Y})$$

$$SSE = \mathbf{Y}^T \mathbf{Y} - \mathbf{b}^T \mathbf{X}^T \mathbf{Y}$$

$$SSTO = \mathbf{Y}^T \mathbf{Y} - (1/n)(\mathbf{Y}^T \mathbf{U} \mathbf{U}^T \mathbf{Y})$$

$$R^2 = SSR/SSTO$$

$$F = (SSR/k)/(SSE/(n-k-1))$$

Estimated Y-values and residuals are calculated, respectively, with the matrix formulas

$$\mathbf{YEST} = \mathbf{Xb}$$

$$\mathbf{E} = \mathbf{Y} - \mathbf{YEST}$$

To estimate the power model, $ln(Y)$ is regressed on $ln(X_1)$, $ln(X_2)$, . . . , $ln(X_k)$, using **LY** and **LX** in place of **Y** and **X**, respectively, in the above matrix equations. Coefficients for the original predictor variables are $b_0 = exp(b_0')$ and $b_i = b_i'$,   i = 1, . . . , k. The estimated power regression model is given by

$$Y = b_0 (X_1{}^b 1)(X_2{}^b 2) \ldots (X_k{}^b k).$$

To estimate the exponential model, $ln(Y)$ is regressed on X_1, X_2, . . . , X_k, using **LY** in place of **Y** in the above matrix equations. Coefficients for the original predictor variables are $b_i = exp(b_i')$, i = 0, . . . , k. The estimated exponential model

is

$$Y = b_0(b_1{}^{X}1)(b_2{}^{X}2) \ . \ . \ . \ (b_k{}^{X}k).$$

Multiple Regression Example

The following problem is from <u>An Introduction to Statistical Methods and Data Analysis, 4th Ed.</u> by Lyman Ott.

The table below presents data on weight loss (pounds) of a chemical compound exposed to air for different lengths of time (hours) and different relative humidities:

Weight Loss, Y	Exposure Time, X_1	Relative Humidity, X_2
4.3	4	.20
5.5	5	.20
6.8	6	.20
8.0	7	.20
4.0	4	.30
5.2	5	.30
6.6	6	.30
7.5	7	.30
2.0	4	.40
4.0	5	.40
5.7	6	.40
6.5	7	.40

$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y}) = [.66667, 1.31667, -8.00000]^T$ and the estimated linear model is $Y = .66667 + 1.31667X_1 - 8.00000X_2$.
SSR = 31.12417, SSE = 1.34500, and SSTO = 32.46917.

R^2 = .95858 and F = 104.13291.

Power and exponential models can be estimated similarly.

Analysis of Variance

   Single-factor analysis of variance can be performed with the
TI-85 using the regression approach. For r factor levels, r-1
dummy variables can be defined as follows:

$$Z_1 = \begin{cases} 1, & \text{if the observation is from population 1} \\ -1, & \text{if the observation is from population r} \\ 0, & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1, & \text{if the observation is from population 2} \\ -1, & \text{if the observation is from population r} \\ 0, & \text{otherwise} \end{cases}$$

   •

   •

   •

$$Z_{r-1} = \begin{cases} 1, & \text{if the observation is from population r-1} \\ -1, & \text{if the observation is from population r} \\ 0, & \text{otherwise.} \end{cases}$$

Input matrices Z_{nxr} (predictors), W_{nx1} (response), and
Q_{nx1} (unit vector) yield regression coefficients from which
estimated main effects and factor-level sample means can be
calculated. The F statistic is calculated as before.

Analysis of Variance Example

The following problem is from <u>An Introduction to Statistical Methods and Data Analysis, 4th Ed.</u> by Lyman Ott.

A clinical psychologist wished to compare three methods for reducing hostility levels in university students. A certain test (HLT) was used to measure the degree of hostility. A high score on the test indicated great hostility. Eleven students obtaining high and nearly equal scores were used in the experiment. Four were selected at random from among the 11 problem cases and treated with method 1. Four of the remaining seven students were selected at random and treated with method 2. The remaining three students were treated with method 3. All treatments were continued for a one-semester period. Each student was given the HLT test at the end of the semester with the results shown below:

Method	Test Scores			
1	80	92	87	83
2	70	81	78	74
3	63	76	70	

Test at the 0.05 level the null hypothesis that the mean test scores for the different methods are equal, against the alternative hypothesis that at least one mean test score differsfrom the rest.

The coefficient vector is

$\mathbf{c}$ = $(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{W})$ = $[76.97222..., 8.52777..., -1.222...]^T$, and

the estimated regression model is

$$W = 76.97222... + 8.52777...Z_1 - 1.222...Z_2.$$

SSE = 234.416, SSR = 452.129, and F = 7.72. The test of hypothesis is significant at the 0.05 level. Factor-level sample means are given below:

Method 1   76.97222... - 8.52777... = 85.5

Method 2   76.97222... - 1.222...   = 75.75

Method 3   76.97222... - 7.30555... = 69.666...

Since the null hypothesis was rejected, factor-level sample means are necessary for multiple comparison procedures.

References

Neter, J.; Wasserman, W.; Kutner, M.H. 1985 Applied Linear
    Statistical Models 2nd Ed. Richard D. Irwin, Inc.
    Homewood IL.

Ott, R.L. 1993 An Introduction to Statistical Methods and
    Data Analysis 4th Ed. Duxbury Press. Belmont CA.