

Multivariate Regression: A Very Powerful Forecasting Method

Vasilopoulos, Ph.D.
ST. JOHN'S UNIVERSITY
The Peter J. Tobin College of Business
CIS/DS department

ABSTRACT

Regression Analysis is at the center of almost every Forecasting technique, yet few people are comfortable with the Regression methodology. We hope to greatly improve the level of comfort with this article. Here we briefly discuss the theory behind the methodology and then outline a step-by-step procedure, which will allow almost everyone to construct a Regression Forecasting function for both the linear and Multivariate case. The Linear Regression is shown to be a special case of the multivariate problem. Also discussed, in addition to model formation and estimation, is model testing (to establish statistical significance of factors) and the Procedure by which the final regression equation is obtained from the estimated equation. The Final Regression Equation is retained and used as the forecasting equation. A hand solution is derived for a relatively small sample problem, and this solution is compared to the MINITAB-derived solution to establish confidence in the statistical tool, which then can be used exclusively for larger problems.

Key Words: Multivariate Regression, Matrix Algebra, Linear Regression, Estimated Equation, Final Equation

INTRODUCTION AND MODEL ESTIMATION FOR THE MULTIVARIATE PROBLEM

Multivariate Regression analysis, in which an equation is derived that connects the value of one dependent variable (Y) to the values of p independent variables $X_1, X_2, \dots, X_p$, starts with a given multivariate data set and uses the Least Squares method to assign the **best** possible values to the unknown multipliers found in the model we wish to estimate. The multivariate data set used to estimate the multivariate model consists of n p-tuples of values:

$$(x_{11}, x_{21}, \dots, x_{p1}, y_1), (x_{12}, x_{22}, \dots, x_{p2}, y_2), \dots, (x_{1n}, x_{2n}, \dots, x_{pn}, y_n)$$

1) Estimation of the Model

The multivariate model is given by:

$$Y = a + b_2 X_2 + b_3 X_3 + \dots + b_p X_p \quad (1)$$

$$\text{or } Y = b_1 X_1 (=1) + b_2 X_2 + b_3 X_3 + \dots + b_p X_p \quad (2)$$

Note that the first 2 terms of the multivariate model given by equation (1) are identical to the linear model

$Y = a + bX$ and in equation (2) we introduced a variable X_1 , whose value is always equal to 1 (if we wish the model to have a constant term), to make the handling of the multivariate model easier, using matrix operations. Note also that the 'a' in equation (1) is set equal to b_1 in equation (2). To estimate the Multivariate model, we use the Least Squares Methodology, which calls for the formation of the Quadratic function:

$$\begin{aligned}
 Q(b_1, b_2, \dots, b_p) &= \sum_{i=1}^n [y_{actual_i} - y_{Multi\ variable\ function_i}]^2 \\
 &= \sum_{i=1}^n [y_i - b_1 - b_2 X_{2i} - b_3 X_{3i} - \dots - b_p X_{pi}]^2
 \end{aligned} \quad (3)$$

To derive the “Normal Equations for the Multivariate model”, from which the values of: $b_1, b_2, b_3, \dots, b_p$ are derived, we take partial derivatives of the $Q(b_1, b_2, b_3, \dots, \text{and } b_p)$ function with respect to $b_1, b_2, b_3, \dots, \text{and } b_p$ respectively, and set each equal to 0; i.e. we obtain, and set equal to zero:

$$\begin{aligned}
 \frac{\partial Q}{\partial b_1} &= 0 \\
 \frac{\partial Q}{\partial b_2} &= 0 \\
 &\dots \\
 \frac{\partial Q}{\partial b_p} &= 0
 \end{aligned} \quad (4)$$

However, when attempting to solve the set of equations (4) algebraically, the results are very complicated, and it is advisable to state the resulting “Normal equations” in a matrix form, by which they are stated as:

$$(X^t X)b = (X^t Y) \quad (5)$$

where: X = Matrix formed from the values of the p independent variables (X has n rows and p columns, or X is a $n \times p$ matrix),

X^t = Transposed matrix X (X^t has p rows and n columns, or X^t is a $p \times n$ matrix)

Y = Column Vector (or $n \times 1$ matrix) of the given Y values

b = Column Vector (or $p \times 1$ matrix) of the unknown multipliers $b_1, b_2, b_3, \dots, b_p$

The Multivariate data set, from which the matrices: X , X^t , Y , and b are defined, has the structure shown below:

$$\begin{array}{cccccc}
 X_1 & X_2 & X_3 & \dots & X_p & Y \\
 X_{11} & X_{21} & X_{31} & \dots & X_{p1} & Y_1 \\
 X_{12} & X_{22} & X_{32} & \dots & X_{p2} & Y_2 \\
 X_{13} & X_{23} & X_{33} & \dots & X_{p3} & Y_3 \\
 \dots & \dots & \dots & \dots & \dots & \dots \\
 X_{1n} & X_{2n} & X_{3n} & \dots & X_{pn} & Y_n
 \end{array} \quad (6)$$

The values under variable X_1 (i.e. $X_{11}, X_{12}, X_{13}, \dots, X_{1n}$) can each be set equal to 1 to make sure the multivariate equation has a constant term. Then, from equation (6) we define the matrices X , X^t , Y , and b , and form the Matrix Products $X^t X$ and $X^t Y$ needed in equation (5). We obtain:

$$X = \begin{pmatrix} X_1 & X_2 & X_3 & \dots & X_p \\ 1 & X_{21} & X_{31} & \dots & X_{p1} \\ 1 & X_{22} & X_{32} & \dots & X_{p2} \\ 1 & X_{23} & X_{33} & \dots & X_{p3} \end{pmatrix} \quad (7)$$

$$X^t = \begin{array}{ccccc} \cdots & & & & \\ 1 & X_{2n} & X_{3n} & \cdots & X_{pn} \\ & & & & \\ 1 & & 1 & & 1 & \cdots & 1 \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2n} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3n} \\ & & & & \\ \cdots & & & & \\ X_{p1} & X_{p2} & X_{p3} & \cdots & X_{pn} \end{array} \quad (8)$$

where X^t = The Transposed of matrix X, has the rows and columns interchanged such that if X is an $n \times p$ matrix, X^t is a $p \times n$ matrix.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}; \text{ i.e. } Y \text{ is a column vector or } n \times 1 \text{ matrix} \quad (9)$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}; \text{ i.e. } b \text{ is a column vector or } p \times 1 \text{ matrix} \quad (10)$$

The matrix products appearing in equation (5) are all defined and have the dimensionalities:

$$\begin{aligned} X^t X & \text{ is } p \times p \text{ matrix} \\ X^t Y & \text{ is } p \times 1 \text{ matrix} \end{aligned} \quad \begin{aligned} (X^t X)b & \text{ is } p \times 1 \text{ matrix} \end{aligned} \quad (11)$$

$$\begin{aligned} & y_1 + y_2 + y_3 + \dots + y_n \\ & y_1 x_{21} + y_2 x_{22} + y_3 x_{23} + \dots + y_n x_{2n} \\ \text{where } X^t Y = & y_1 x_{31} + y_2 x_{32} + y_3 x_{33} + \dots + y_n x_{3n} \\ & \dots\dots\dots \\ & y_1 x_{m1} + y_2 x_{m2} + y_3 x_{m3} + \dots + y_n x_{mn} \end{aligned} \quad (12)$$

and

$$X^t X = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ X_{21} & X_{22} & X_{23} & \dots & X_{2n} \\ X_{31} & X_{32} & X_{33} & \dots & X_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ X_{p1} & X_{p2} & X_{p3} & \dots & X_{pn} \end{bmatrix} \begin{bmatrix} 1 & X_{21} & X_{31} & \dots & X_{p1} \\ 1 & X_{22} & X_{32} & \dots & X_{p2} \\ 1 & X_{23} & X_{33} & \dots & X_{p3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{2n} & X_{3n} & \dots & X_{pn} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{i=1}^n X_{2i} & \sum_{i=1}^n X_{3i} & \dots & \sum_{i=1}^n X_{pi} \\ \sum_{i=1}^n X_{2i} & \sum_{i=1}^n X_{2i}^2 & \sum_{i=1}^n X_{2i}X_{3i} & \dots & \sum_{i=1}^n X_{2i}X_{pi} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^n X_{pi} & \sum_{i=1}^n X_{pi} & \dots & \dots & \sum_{i=1}^n X_{pi}^2 \end{bmatrix} \quad (13)$$

The matrix solution to equation (5) is given by:

$$\begin{matrix} b_1 \\ b_2 \\ b = b_3 = (X^t X)^{-1} (X^t Y) \\ \dots \\ b_p \end{matrix} \quad (14)$$

where $(X^t X)^{-1}$ is the Inverse of Matrix $X^t X$ (see equation (13) above) which can be found using either the Gauss-Elimination method or the Adjoint Matrix method.

Note: If the $X^t X$ matrix is Diagonal, i.e. it has non-zero elements only along the main diagonal, finding the Inverse matrix is trivial.

$$\text{For example, if } X^t X = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix}, \quad (X^t X)^{-1} = \begin{bmatrix} \frac{1}{d_1} & 0 & 0 \\ 0 & \frac{1}{d_2} & 0 \\ 0 & 0 & \frac{1}{d_3} \end{bmatrix}. \quad (15)$$

To complete the estimation of the multivariate model we need to first find the variances $V(b_1), V(b_2), V(b_3), \dots, V(b_p)$ from which then we can obtain: $(b_1) = \sqrt{V(b_1)}, \dots, (b_p) = \sqrt{V(b_p)}$.

The variance of the b vector ($b = \begin{matrix} b_1 \\ b_2 \\ \dots \\ b_p \end{matrix}$) is given by:

$$V(b) = (X^t X)^{-1} \wedge^2 \quad (16)$$

$$\text{where } \wedge^2 = \frac{Y^t Y - b^t X^t Y}{n - p} = \frac{Q^*}{n - p}, \quad (17)$$

and $Y^t Y = \sum_{i=1}^n y_i^2 = y_1^2 + y_2^2 + \dots + y_n^2$, $X^t Y$ was derived in equation (12) and b^t is the transposed of vector b, or $b^t = (b_1 \ b_2 \ \dots \ b_p)$.

After equation (17) is substituted into equation (16) and the multiplication of the matrix

$(X'X)^{-1}$ by $\hat{\beta}$ takes place, $V(\hat{\beta})$ assumes the form:

$$V(\hat{\beta}) = \begin{bmatrix} V(b_1) & & & \\ \text{Co variance Terms} & V(b_2) & & \\ & V(b_3) & & \\ & & \dots & \\ & & & V(b_p) \end{bmatrix}$$

Therefore, the variances $V(b_1), V(b_2), \dots, V(b_p)$ are the values along the main diagonal of the $V(\hat{\beta})$ matrix, while the off-the-main-diagonal terms are Covariance terms.

Note: At this point, we have, for the given data set:

$$b_1, b_2, b_3, \dots, b_p \text{ and } \sigma(b_1), \sigma(b_2), \sigma(b_3), \dots, \sigma(b_p).$$

II) MODEL TESTING & EXAMPLE

Now that our model of interest has been estimated, we need to test for the significance of the terms found in the estimated model. This is very important because the results of this testing will determine the final equation which will be retained and used for Forecasting purposes.

Testing THE MULTIVARIATE MODEL $\hat{y} = b_1X_1 + b_2X_2 + \dots + b_pX_p$

Testing of this model consists of the following 3 steps:

A) To test for the significance of each factor separately

The values of $b_1, b_2, b_3, \dots, b_p$ are obtained from equation (14) and the values of $\sigma(b_1), \sigma(b_2), \sigma(b_3), \dots, \sigma(b_p)$ from equations (16) and (17). Then, we test for the significance of each factor separately by either:

1) Testing the hypotheses: $H_0: \beta_i = 0$ vs. $H_1: \beta_i \neq 0$ by calculating $Z_i^* = \frac{b_i}{\sigma(b_i)}$ or $t_i^* = \frac{b_i}{\sigma(b_i)}$, for $1 \leq i \leq p$. Then, $H_0: \beta_i = 0$ is rejected if $Z_i^* > Z_{\alpha/2}$ (or if $Z_i^* < -Z_{\alpha/2}$), when $n \geq 30$ or if $t_i^* > t_{n-p}(\alpha/2)$ (or if $t_i^* < -t_{n-p}(\alpha/2)$), if $n < 30$

or 2) By constructing the confidence intervals

$$P[b_i - Z_{\alpha/2} \sigma(b_i) \leq \beta_i \leq b_i + Z_{\alpha/2} \sigma(b_i)] = 1 - \alpha, \text{ if } n \geq 30$$

or

$$P[b_i - t_{n-p}(\alpha/2) \sigma(b_i) \leq \beta_i \leq b_i + t_{n-p}(\alpha/2) \sigma(b_i)] = 1 - \alpha, \text{ if } n < 30.$$

If the value $\beta_i = 0$ is outside of these Confidence intervals, $H_0: \beta_i = 0$ is rejected.

B) To test for the Significance of the entire Regression (including the constant)

The hypotheses being tested are:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0 \text{ vs. } H_1: \text{The } \beta_i \text{ are not all equal to 0}$$

or H_0 : The entire regression (including the constant) is not significant

vs.

H_1 : The entire regression (including the constant) is significant.

It is carried out by calculating:

$$F_{Total}^* = \frac{RSS / DOF}{ESS / DOF} = \frac{b^t X^t Y / p}{(Y^t Y - b^t X^t Y) / n - p} \text{ and comparing to } F_{n-p}^p(\alpha).$$

If $F_{Total}^* > F_{n-p}^p(\alpha)$, H_0 is rejected and we conclude that the entire regression (including the constant) is significant (to the calculation of the Y value).

C) To test for the Significance of the entire Regression (excluding the constant)

The hypotheses being tested are:

$H_0: \beta_2 = \beta_3 = \dots = \beta_p = 0$ vs. $H_1: \beta_2, \beta_3, \dots, \beta_p$ are not all equal to 0

or H_0 : The entire regression (excluding the constant) is not significant

vs.

H_1 : The entire regression (excluding the constant) is significant.

It is carried out by calculating:

$$F_{Total-1}^* = \frac{(RSS - SS_a) / (p - 1)}{ESS / (n - p)} = \frac{(b^t X^t Y - SS_a) / (p - 1)}{(Y^t Y - b^t X^t Y) / (n - p)}, \text{ where } SS_a = \frac{(\sum Y_i)^2}{n}$$

and comparing it to $F_{n-p-1}^p(\alpha)$. If $F_{Total-1}^* > F_{n-p-1}^p(\alpha)$, H_0 is rejected and we conclude that the entire regression (excluding the constant) is significant.

D) Determination of the Final equation

Any variable X_i , for which the hypothesis: $H_0: \beta_i = 0$ (vs. $H_1: \beta_i \neq 0$) is not rejected, is to be dropped from the regression equation. The remaining terms are used to form the "Final Equation" which is then retained and used for Prediction/Forecasting purposes.

E) A Multivariate Example

The sales manager of a certain firm believes that Sales Ability depends on a salesman's Verbal Reasoning Ability and Vocational Interest. He is interested in constructing a regression equation to use in future hiring, to predict a candidate's success as a salesman, i.e. he wants to derive the regression equation.

Y = Sales Ability

$= b_1 X_1 (=1) + b_2 X_2 (= \text{Verbal Reasoning}) + b_3 X_3 (= \text{Vocational Interest})$

To verify this belief, 10 salesmen are selected at random from his staff and given 2 tests: One for Verbal Reasoning Ability, the other for Vocational Interest. The results are shown below:

Salesman	1	2	3	4	5	6	7	8	9	10
Y=Average sales in a month	1	1	1	2	2	4	3	5	6	6
X_2 =Verbal Reasoning Ability	1	1	2	2	3	3	4	4	5	5
X_3 =Vocational Interest	2	1	1	3	2	4	3	5	4	6

1) Find the estimated regression equation for this data set.

2) Find TSS, RSS_b , $Q^* = ESS$, and SS_a

$$\text{Also, } X'Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 3 & 3 & 4 & 4 & 5 & 5 \\ 2 & 1 & 1 & 3 & 2 & 4 & 3 & 5 & 4 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 4 \\ 3 \\ 5 \\ 6 \\ 6 \end{pmatrix} = \begin{pmatrix} 31 \\ 118 \\ 124 \end{pmatrix}$$

$$\text{Then: } b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = (X^T X)^{-1} (X^T Y)$$

$$= \frac{1}{1740} \begin{array}{ccccc} 989 & 189 & 80 & 31 & 1563 \\ 189 & 249 & 180 & 118 & 1203 \\ 80 & 180 & 200 & 124 & 1080 \end{array} = \frac{1}{1740}$$

$$\text{and } b_1 = \frac{1563}{1740} = 0.8982, b_2 = \frac{1203}{1740} = 0.6913, b_3 = \frac{1080}{1740} = 0.6207,$$

Therefore, the regression equation is:

$$1) \hat{y} = b_1 X_1 (=1) + b_2 X_2 + b_3 X_3 = -0.8982 + 0.6913 X_2 + 0.6207 X_3$$

$$2) \text{ TSS} = Y'Y = \sum_{i=1}^{10} y_i^2 = y_1^2 + y_2^2 + \dots + y_{10}^2 = 133.00$$

$$RSS_b = b' X' Y = \begin{bmatrix} 0.8982 & 0.6913 & 0.6207 \end{bmatrix} \begin{matrix} 31 \\ 118 \\ 124 \end{matrix} = 130.7047$$

$$ESS = Q^* = \frac{Y^t Y}{n} - b^t X^t Y = 133.0000 - 130.7047 = 2.2953$$

$$SS_a = \frac{\sum_{i=1}^{10} y_i^2}{n} = \frac{(y_1 + y_2 + \dots + y_{10})^2}{10} = \frac{(31)^2}{10} = \frac{961}{10} = 96.1$$

3) From equation (17) $\hat{\rho}^2 = \frac{ESS}{n - p} = \frac{Q^*}{n - p} = \frac{2.2953}{10 - 3} = \frac{2.2953}{7} = 0.3279$

Then: $V(b_1) = \frac{989}{1740} \cdot s^2 = \frac{989}{1740} (0.3279) = 0.1864$, and $(b_1) = 0.4316$

$$V(b_2) = \frac{249}{1740} \cdot \frac{1}{2} = \frac{249}{1740} (0.3279) = 0.0469, \text{ and } (b_2) = 0.2163$$

$$V(b_3) = \frac{200}{1740} \cdot 0.3279 = 0.0377, \text{ and } (b_3) = 0.1942$$

4) Test: $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$, $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$,
 $H_0: \beta_3 = 0$ vs $H_1: \beta_3 \neq 0$,

$$\begin{aligned} \text{Calculate: } t_1^* &= \frac{b_1}{(b_1)} = \frac{0.8982}{0.4316} = -2.081 \\ t_2^* &= \frac{b_2}{(b_2)} = \frac{0.6913}{0.2163} = 3.192 \\ t_3^* &= \frac{b_3}{(b_3)} = \frac{0.6207}{0.1942} = 3.196 \end{aligned} \quad \left\{ \begin{array}{l} \text{and compare} \\ \text{each against:} \\ t_{n-p}(\alpha/2) = \\ t_{10-3(0.05/2)} = \\ t_{7(0.025)} = 2.365 \end{array} \right.$$

Since $-2.365 < t_1^* = -2.081 < 2.365$, do not reject $H_0: \beta_1 = 0$; Therefore, $\beta_1 = 0$

$t_2^* = 3.192 > 2.365$, reject $H_0: \beta_2 = 0$; Therefore, $\beta_2 \neq 0$

$t_3^* = 3.196 > 2.365$, reject $H_0: \beta_3 = 0$; Therefore, $\beta_3 \neq 0$

5) Since $n = 10 < 30$ the sampling distributions of b_1, b_2, b_3 are $t_{n-p} = t_7$. Therefore, the Confidence Intervals are to be obtained from:

$$P[b_i - t_{7(\alpha/2)} (b_i) \leq \beta_i \leq b_i + t_{7(\alpha/2)} (b_i)] = 1 - \alpha$$

and they are (when the appropriate values are substituted in):

a) $P[-1.919 \leq \beta_1 \leq 0.123] = 0.95$; since $\beta_1 = 0$ falls inside this Confidence Interval, we do not reject $H_0: \beta_1 = 0$; Therefore, $\beta_1 = 0$

b) $P[0.180 \leq \beta_2 \leq 1.202] = 0.95$; since $\beta_2 = 0$ falls outside this Confidence Interval we reject $H_0: \beta_2 = 0$; Therefore, $\beta_2 \neq 0$

c) $P[0.161 < \beta_3 < 1.080] = 0.95$; since $\beta_3 = 0$ falls outside this Confidence Interval we reject $H_0: \beta_3 = 0$; Therefore, $\beta_3 \neq 0$

6) We calculate:

$$F_{Total}^* = \frac{(b'X'Y)/p}{(Y'Y - b'X'Y)/n-p} = \frac{130.7047/3}{(2.2953)/7} = \frac{43.5682}{0.3279} = 132.87$$

and compare to: $F_{n-p}^p(\alpha) = F_7^3(\alpha) = \begin{matrix} 4.35, & \text{if } \alpha = 0.05 \\ 8.45, & \text{if } \alpha = 0.01 \end{matrix}$

Since $F_{Total}^* > F_7^3(\alpha)$, for both α values, H_0 : (The entire equation including the constant is not significant) is rejected, and we conclude that the entire equation is significant.

7) We calculate:

$$\begin{aligned} F_{Total-1}^* &= \frac{(b'X'Y - SS_a)/p-1}{(Y'Y - b'X'Y)/n-p} = \frac{(130.7047 - 96.1)/2}{(2.2953)/7} = \frac{\frac{34.6047}{2}}{\frac{2.2953}{7}} \\ &= \frac{17.30235}{0.3279} = 52.77 \end{aligned}$$

and compare to: $F_{n-p-1}^p(\alpha) = F_7^2(\alpha) = \begin{matrix} 4.74, & \text{if } \alpha = 0.05 \\ 9.55, & \text{if } \alpha = 0.01 \end{matrix}$

Since $F_{Total-1}^* > F_7^2(\alpha)$, for both α values, we reject H_0 (H_0 : The entire equation (excluding the constant) is not significant) and conclude that the entire regression equation, excluding the constant, is significant.

8) Because $H_0: \beta_1 = 0$ is not rejected but $H_0: \beta_2 = 0$ and $H_0: \beta_3 = 0$ are rejected, the final regression equation is:

$$\hat{y} = b_2 X_2 + b_3 X_3 = 0.6914 X_2 + 0.6207 X_3$$

which should be retained and used to predict the sales ability of future candidates.

F) MINITAB SOLUTION to the Problem

We enter the given data and issue the regression command as shown:

MTB > Set C1

DATA > 1 1 1 2 2 4 3 5 6 6

DATA > end

MTB > Set C2

DATA > 1 1 2 2 3 3 4 4 5 5

DATA > end

MTB > Set C3

DATA > 2 1 1 3 2 4 3 5 4 6

DATA > end

MTB > print C1 C2 C3

Data Display

Row	C1	C2	C3
1	1	1	2
2	1	1	1
3	1	2	1
4	2	2	3
5	2	3	2
6	4	3	4
7	3	4	3
8	5	4	5
9	6	5	4
10	6	5	6

MTB > REGRESS c1 2 c2 c3;

SUBC > CONSTANT;

SUBC > BRIEF 2.

Regression Analysis: C1 versus C2, C3

The regression equation is

$$C1 = -0.898 + 0.691 C2 + 0.621 C3$$

Predictor	Coef	SE Coef	T	P
Constant	-0.8983	0.432	-2.08	0.076
C2	0.6914	0.2168	3.19	0.015
C3	0.6207	0.1943	3.2	0.015

S = 0.5730 R-Sq = 93.8% R-Sq(adj) = 92.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	34.602	17.301	52.69	0
Residual Error	7	2.298	0.328		
Total	9	36.9			

Source	DF	Seq SS
C2	1	31.25
C3	1	3.352

Unusual Observations

Obs	C2	C1	Fit	SE Fit	Residual	St Resid
9	5	6	5.041	0.359	0.959	2.15R

R denotes an observation with a large standardized residual

When we compare the MINITAB and hand solutions, they are identical, within the rounding of fractional numbers. From the p values calculated by MINITAB we see, once again, that $H_0: \beta_1 = 0$ is not rejected because $p = 0.076 > \alpha = 0.05$ but $H_0: \beta_2 = 0$ and $H_0: \beta_3 = 0$ are rejected because $p = 0.015 < \alpha = 0.05$. The F value of 52.69, shown in the MINITAB Analysis of Variance is almost identical to the hand value $F_{Total}^* = 52.77$ obtained to test the hypothesis H_0 : The entire regression (excluding the constant) is not significant. Also, the values of $\hat{\beta}^2$ are almost identical.

SUMMARY OF MULTIVARIATE PROCEDURE PROCEDURE FOR SOLVING MULTIVARIATE/BIVARIATE PROBLEMS

A) Model Estimation

Given a Multivariate (or bivariate) data Set:

- 1) Identify the matrices: Y, X, b, Z which form the matrix equation $Y = Xb + Z$
- 2) Calculate X^tX and X^tY
- 3) Calculate The Inverse of Matrix $(X^tY) = (X^tX)^{-1}$
 - a. If X^tX is a diagonal Matrix, $(X^tX)^{-1}$ is easy to find
 - b. If X^tX is not Diagonal, Finding $(X^tX)^{-1}$ is more difficult
 - i. Use the Gauss Elimination Method
 - ii. Use the Adjoint Matrix Method

$$4) \text{ Calculate: } b = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_p \end{pmatrix} = (X^tX)^{-1} X^tY$$

5) Calculate: Y^tY , b^tX^tY , $Q^* = Y^tY - b^tX^tY$, $SSa = (\sum Y_i)^2 / n$

6) Calculate: $\sigma^2 = Q^* / n - p$

7) Calculate: $V(b) = (X^tX)^{-1} \sigma^2 \leftarrow$ The Variances $V(b_1)$, $V(b_2)$, ..., $V(b_p)$ are along the Main Diagonal

Note1: The other terms of the $V(b)$ matrix are covariances

Note2: At this point we have: $b_1, b_2, \dots, b_p$ and : $\sigma(b_1), \sigma(b_2), \dots, \sigma(b_p)$

Also available are all the sums of squares

B) Model Testing**a) To Test for the Significance of Each Factor Separately**

- 1) From the knowledge of b_i and $\sigma(b_i)$
Use Either Z (if $n \geq 30$) or t_{n-p} (if $n < 30$) to test :
 $H_0: \beta_i=0$ vs. $H_1: \beta_i \neq 0$

- 2) From the knowledge of b_i and $\sigma(b_i)$

Use either Z or t_{n-p} to construct confidence intervals

$$P[b_i - Z_{\alpha/2} \sigma(b_i) \leq \beta_i \leq b_i + Z_{\alpha/2} \sigma(b_i)] = 1 - \alpha$$

or

$$P[b_i - t_{n-p}(\alpha/2) \sigma(b_i) \leq \beta_i \leq b_i + t_{n-p}(\alpha/2) \sigma(b_i)] = 1 - \alpha$$

b) To Test for the Significance of the Entire Equation (Including the Constant)

- 1) Construct ANOVA with SSa
- 2) Test the Hypothesis : $H_0: \beta_1 = \beta_2 = \beta_3 = \dots \beta_p = 0$ vs. $H_1: \text{The } \beta_i \text{ are not all } = 0$
 - I. Calculate $F_1^* = [(b^t X^t Y)/p] / [Q^*/n-p]$
 - II. Compare F_1^* to : $F_{n-p}^p(\alpha)$
 - III. Reject H_0 if : $F_1^* > F_{n-p}^p(\alpha)$

c) To Test For the Significance of the Entire Equation (excluding the Constant)

- a. Construct ANOVA without SSa
- b. Test the hypothesis: $H_0: \beta_2 = \beta_3 = \dots \beta_p = 0$ vs. $H_1: \text{The } \beta_i \text{ are not All } = 0$
 - i. Calculate $F_2^* = \frac{(b^t X^t Y - SSa)}{p-1} \cdot \frac{Q^*}{n-p}$
 - ii. Compare F_2^* to : $F_{n-p}^{p-1}(\alpha)$
 - iii. Reject H_0 If : $F_2^* > F_{n-p}^{p-1}(\alpha)$

DETERMINE THE FINAL EQUATION

Any Variable, X_i , for which the Hypothesis: $H_0: \beta_i=0$ vs. $H_1: \beta_i \neq 0$

is not rejected, is to be dropped from the regression equation. The remaining terms are used to form the "Final Equation" which is then retained and used for Prediction/Forecasting purposes.

THE LINEAR REGRESSION MODEL AS A SPECIAL CASE OF THE MULTIVARIATE MODEL

For the Linear model, the value of one dependent variable (Y) depends on the values of one independent variable (X). In this case, the given data set is Bivariate, and given by:

X	x_1	x_2	x_3	...	x_n
Y	y_1	y_2	y_3	...	y_n

(18)

From which we can obtain:

$$\sum_{i=1}^n X_i = X_1 + X_2 + X_3 + \dots + X_n \quad (19)$$

$$\sum_{i=1}^n X_i^2 = X_1^2 + X_2^2 + X_3^2 + \dots + X_n^2 \quad (20)$$

$$\sum_{i=1}^n Y_i = Y_1 + Y_2 + Y_3 + \dots + Y_n \quad (21)$$

$$\sum_{i=1}^n Y_i^2 = Y_1^2 + Y_2^2 + Y_3^2 + \dots + Y_n^2 \quad (22)$$

$$\sum_{i=1}^n X_i Y_i = X_1 Y_1 + X_2 Y_2 + X_3 Y_3 + \cdots + X_n Y_n \quad (23)$$

$$n = \text{number of ordered pairs } (X_i, Y_i) \quad (24)$$

The ideal linear model will be represented by:

$$y = \alpha + \beta x \quad (25)$$

while the model of observations (estimated model) will be written as:

$$y_i = a + bx_i + z_i \quad (26)$$

Then, applying the matrix methodology, summarized above, we need to define the matrices: Y , β , b , and X where:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}; \beta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}; b = \begin{bmatrix} a \\ b \end{bmatrix}; X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad (27)$$

These vectors/matrices are obtained when the given data set is substituted in the model of observations (equations 26 above) which becomes the “algebraic system of equations”:

[illegible]

which can be rewritten in the matrix form as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \quad (29)$$

or:

$$Y = Xb + Z \quad (30)$$

To obtain the Normal equations for the Linear model ($X^t X b = X^t Y$), we need to first obtain $X^t Y$ and $X^t X$.

Since $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ is an $n \times 2$ matrix, X^t is $2 \times n$, and given by: $X^t = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix}$ and

X^tX becomes
a 2x2 matrix:

$$X^t X = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \quad (31)$$

and

$$X^tY = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_n \\ x_1y_1 & x_2y_2 & x_3y_3 & \cdots & x_ny_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_iy_i \end{bmatrix} \quad (32)$$

Then, the normal equations $X^t X b = X^t Y$ become:

$$\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}, \quad (33)$$

(which are identical to the algebraic normal equations) and their solution is:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \quad (34)$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \quad (35)$$

Note: The inverse matrix in (34) is obtained by using either the “Gauss Elimination Method” or the “Adjoint Matrix Method.”

Then, from equation (35) we obtain:

$$a = \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \quad (36)$$

$$b = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \quad (37)$$

$$V(a) = \left[\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \right] \sigma^2 = \frac{1}{n} \left[\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \sigma^2 \quad (38)$$

(Note: $V(a) = \frac{\sigma^2}{n}$ if $\bar{x} = 0$)

$$V(b) = \left[\frac{n}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \right] \sigma^2 = \left[\frac{n}{n \sum_{i=1}^n (x_i - \bar{x})^2} \right] \sigma^2 = \left(\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \sigma^2 \quad (39)$$

$$\text{because } n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 = n \sum_{i=1}^n (x_i - \bar{x})^2 \quad (40)$$

$$\text{Also, Covariance}(a, b) = \left[\frac{-\sum_{i=1}^n x_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} \right] \sigma^2 \quad (41)$$

$$Q^* = Y^t Y - b^t (X^t Y)$$

$$= (y_1, y_2, y_3, \dots, y_n) \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} - (a \ b) \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \quad (42)$$

$$= \sum_{i=1}^n y_i^2 - a \left(\sum_{i=1}^n y_i \right) - b \left(\sum_{i=1}^n x_i y_i \right) = ESS \quad (43)$$

$$\text{and } \sigma^2 = \frac{Q^*}{n-2} \quad (44)$$

Note: The expressions for the normal equations of the Linear model, $V(a)$, $V(b)$, $\text{Cov}(a, b)$, $Q^* = ESS$, and σ^2 are identical to those obtained directly for the Linear model.

CONCLUSIONS

1. Regression Analysis, whether Linear, Non-Linear, or Multivariate, is extremely important as a Forecasting Technique.
2. Linear Regression is relatively easy to perform using purely algebraic methods.
3. But, Linear Regression can also be considered as a special case of the more general Multivariate Regression Model, which can be analyzed efficiently by using matrix

methods.

4. A step-by-step procedure on how to solve the multivariate regression problem is included in this paper.
5. The Application of the Method (which consists of: Model Estimation, Model Testing, and the derivation of the FINAL Regression equation which is retained and used for forecasting purposes) requires an elementary knowledge of Matrix Algebra, including the calculation of the Inverse Matrix.
6. Statistical tools, such as the MINITAB, can also be used to solve the Multivariate problem by computer, and then compare the hand and MINITAB results.
7. Using MINITAB to also solve the sample problem, produces solutions which are identical to the hand solutions.
8. The MINITAB output not only estimates the model, but also generates “p-values” for all the important model parameters, which allows the testing of their significance.
9. The p-value, called the “Observed level of significance,” in contrast to the a-priori α value, has the following relationship to α :
 - a) If $p > \alpha$, do not Reject H_0 .
 - b) If $p < \alpha$, Reject H_0 .
10. The MINITAB output also provides values for R^2 (coefficient for Multiple Determination) and R^2 adjusted which tells us “how well” the model fits the given data

References

- Berenson, Marc, L.; Levine, David, M.; Krehbiel, Timothy, C.; “Basic Business
Black, Ken; Business Statistics, Wiley 2004
Canavos, George, C.; Applied Probability and Statistical Methods; Little, Brown; 1984
Childress, Robert, L.; Gorsky, Robin, D.; Witt, Richard, M.; Mathematics for Managerial Decisions, Prentice Hall, 1989
Carlson, William, L.; Thorne, Betty; “Applied Statistical Methods”, Prentice Hall, 1997
Childress, Robert, L.; Gorsky, Robin, D.; Witt, Richard, M.; Mathematics for Managerial
Chou, Ya-lun; “Statistical Analysis for Business and Economics”; Elsevier, 1992
Draper, Norman; Smith, Harry; Applied Regression Analysis John Wiley & Sons, 1966
Johnson, J. Econometric Methods, McGraw-Hill, 1963
Pindyck, Robert S; Rubinfeld, Daniel L.; Econometric Models and Economic Forecasts, 2nd Edition McGraw-Hill, 1981
Vasilopoulos, A. “Business Statistics – A Logical Approach. Theory, Models, Procedures, and Applications Including Computer (MINITAB) Solutions”, Boston, MA; Pearson Custom Publishing, 2007
Vasilopoulos, A. & Lu, F. Victor. “Quantitative Methods for Business with Computer Applications”, Boston, MA; Pearson Custom Publishing, 2006
Vasilopoulos, A. “Regression Analysis Revisited”, Review of Business, St. John’s University, Jamaica, NY; 2005