Journal of Computer Science and Information Technology
December 2017, Vol. 5, No. 2, pp. 18-25
ISSN: 2334-2366 (Print), 2334-2374 (Online)
Copyright © The Author(s). All Rights Reserved.
Published by American Research Institute for Policy Development
DOI: 10.15640/jcsit.v5n2a3
URL: https://doi.org/10.15640/jcsit.v5n2a3

# Multivariate Regression: A Very Powerful Forecasting Method

**A. Vasilopoulos[1] Ph.D.**

## 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.

**Keywords:** Multivariate Regression, Matrix Algebra, Linear Regression, Estimated Equation, Final Equation

## I) 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 $X_1, X_2, ...X_p$ , 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:

$$(x_{11}, x_{21}, ..., x_{p_1}, y_1), (x_{12}, x_{22}, ..., x_{p_2}, y_2), ...(x_{1n}, x_{2n,} ...x_{p_n}, y_n)$$

1) Estimation of the Model The multivariate model is given by:

$$Y = a + b_2X_2 + b_3X_3 + ... + b_pX_p \qquad (1)$$

$$\text{or} \quad Y = b_1X_1(=1) + b_2X_2 + b_3X_3 + ... + b_pX_p \qquad (2)$$

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:

[1] ST. John's University, The Peter J. Tobin College of Business, CIS/DS department, 800 Utopia Parkway, Jamaica, N.Y. 11439

$$\mathrm{Q}(b_1,b_2,\ldots,b_p)=\sum_{i=1}^{n}\left[y_{actual_i}-y_{Multivariable\,fuction_i}\right]^2$$

$$=\sum_{i=0}^{n}\left[y_i - b_1 - b_2X_2 - b_3X_3 - \ldots - b_pX_p\right]^2 \qquad (3)$$

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

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

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:

$$(X^tX)b=(X^tY) \qquad (5)$$

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

$X^t$ = Transposed matrix $X$ ($X^t$ has p rows and n columns, or $X^t$ is a p x n matrix)
$Y$ = Column Vector (or n x 1 matrix) of the given $Y$ values

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

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

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

The values under variable $X_1$ (i.e. $X_{11}, X_{12}, X_{13}, \ldots, X_{1n}$) 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 $X^tX$ and $X^tY$ needed in equation (5). We obtain:

$$X = \begin{pmatrix} X_1 & X_2 & X_3 & \cdots & X_p \\ 1 & X_{21} & X_{31} & \cdots & X_{p1} \\ 1 & X_{22} & X_{32} & \cdots & X_{p2} \\ 1 & X_{23} & X_{33} & \cdots & X_{p3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{pn} \end{pmatrix} \quad (7)$$

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

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

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

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

The matrix products appearing in equation (5) are all defined and have the dimensionalities:
$X^t X$is p x p matrix
$(X^t X)$b is p x 1 matrix (11)
$X^t Y$ is p x 1 matrix

$$\text{where } X^t Y = \begin{bmatrix} y_1 + y_2 + y_3 + \ldots + y_n \\ y_1 x_{21} + y_2 x_{22} + y_3 x_{23} + \ldots + y_n x_{2n} \\ y_1 x_{31} + y_2 x_{32} + y_3 x_{33} + \ldots + y_n x_{3n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_1 x_{p1} + y_2 x_{p2} + y_3 x_{p3} + \ldots + y_n x_{pn} \end{bmatrix} \quad (12)$$

And

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

$$= \begin{bmatrix} n & \sum_{i-1}^{n} X_{2i} & \sum_{i-1}^{n} X_{3i} & \dots & \sum_{i-1}^{n} X_{p1} \\ \sum_{i-1}^{n} X_{2i} & \sum_{i-1}^{n} X_{2i}^{2} & \sum_{i-1}^{n} X_{2i} X_{3i} & \dots & \sum_{i-1}^{n} X_{2i} X_{pi} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i-1}^{n} X_{pi} & \sum_{i-2}^{n} X_{pi} & \dots & \dots & \sum_{i-1}^{n} X_{pi}^{2} \end{bmatrix} \textbf{(13)}$$

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

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

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 Ad joint 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 matrixis trivial.

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

To complete the estimation of the multivariate model we need to first find the variances V($b_1$), V($b_2$), V($b_3$), …, V($b_p$) from which then we can obtain: $\sigma(b_1) = \sqrt{V(b_1)}, \dots, \sigma(b_p) = \sqrt{V(b_p)}$ . The variance of the b vector (b = $\begin{pmatrix} b_1 \\ b_2 \\ \dots. \\ b_p \end{pmatrix}$) is given by:

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

Where $\hat{\sigma}^2 = \frac{Y^t Y - b^t X^t Y}{n-p} = \frac{Q*}{n-p}$ , (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 $(X^t X)^{-1}$by $\hat{\sigma}^2$ takes place, V (b) assumes the form:

$$\text{V (b)} = \begin{bmatrix} V(b_1) & & & & \\ & V(b_2) & & Co\,\mathrm{var}\,iance & \\ & & V(b_3) & Terms & \\ & Co\,\mathrm{var}\,iance & & ............. & \\ & Terms & & & V(b_p) \end{bmatrix}$$

Therefore, the variances V ($b_1$), V($b_2$), …, V($b_p$) are the values along the main diagonal of the V(b) matrix, while the off-the-main-diagonal terms are Covariance terms.

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

$b_1$, $b_2$, $b_3$, …,$b_p$ and **σ ($b_1$), σ ($b_2$), σ($b_3$), …, σ ($b_p$).**

**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 $\hat{y} = b_1X_1(=1)+b_2X_2+...+b_pX_p$

Testing of this model consists of the following 3 steps:

**A) To test for the significance of each factor separately**

The values of $b_1$, $b_2$, $b_3$, …,$b_p$ are obtained from equation (14) and the values of σ ($b_1$), σ ($b_2$), σ ($b_3$), …, σ($b_p$) 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 $Z_i^* = \dfrac{b_i}{s\,(b)_i}$ or $t_i^* = \dfrac{b_i}{s\,(b_i)}$, for $1 \le i \le p$.

Then, $H_0 : \beta_i$ =0 is rejected if $Z_i^* > Z_{\alpha/2}$ (or if $Z_i^* < -Z_{\alpha/2}$), when n ≥ 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

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

or

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

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:

$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_p = 0$ vs. $H_1$: The $\beta_i$ 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:

$H_0 : \beta_2 = \beta_3 = ... = \beta_p = 0$ vs. $H_1 : \beta_2, \beta_3, ..., \beta_p$ 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:

$$F^*_{Total-\beta_1} = \frac{(RSS - SS_a)/p-1}{ESS/n-p} = \frac{(b^t X^t Y - SS_a)/p-1}{(Y^t Y - b^t X^t Y)/n-p}$$

and comparing it to $F^{p-1}_{n-p}(\alpha)$. If $F^*_{Total-\beta_1} > F^{p-1}_{n-p}(\alpha)$, $H_0$ 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.

**III) 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) Calculate: $b = \begin{pmatrix} b_1 \\ b_2 \\ .. \\ .. \\ b_p \end{pmatrix} = (X^tX)^{-1} X^tY$

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: $V(b) = (X^tX)^{-1}\sigma^2$ ← 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: $b_1$, $b_{2,...,}b_p$ and : $\sigma(b_1)$, $\sigma(b_2)$,… $\sigma(b_p)$
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 $t_{n\text{-}p}$ ( if $n < 30$) to test :
   Ho: $\beta_i = 0$ vs. $H_1 = \beta_i \neq 0$
2) From the knowledge of $b_i$ and $\sigma(b_i)$
   Use either Z or $t_{n\text{-}p}$ to construct confidence intervals
   $P[b_i - Z_{\alpha/2}\, \sigma(b_i) \leq \beta_i \leq b_i + Z_{\alpha/2}\, \sigma(b_1)] = 1\text{-}\alpha$
   or
   $P[b_i - t_{n\text{-}p}\, (\alpha/2)\sigma(b_i) \leq \beta_i \leq b_i + t_{n\text{-}p}(\alpha/2)\sigma(b_1)] = 1\text{-}\ \alpha$

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

1) Construct ANOVA with SSa
2) Test the Hypothesis : Ho : $\beta_1 = \beta_2 = \beta_3 = \ldots\ \beta_p = 0$ vs. $H_1$ : The $\beta_i$ are not all=0

I. Calculate $F_1^* = [(b^t\, X^tY)/p]\ /\ [\ Q^*/n\text{-}p]$

II. Compare $F_1^*$ to : $\boldsymbol{F}^{\,p}_{n-p}\ (\alpha)$

III. Reject Ho if : $F_1^* > \boldsymbol{F}^{\,p}_{n-p}\ (\alpha)$

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

a. Construct ANOVA without SSa
b. Test the hypothesis: Ho: $\beta_2 = \beta_3 = \ldots\ \beta_p = 0$ vs. $H_1$ : The $\beta_i$ are not All=0

i. Calculate $F_2^* = \underline{(b^t\, X^tY\text{- SSa})/\ p\text{-}1}$
$Q^*/\ n\ \text{-}p$

*ii.* Compare $F_2^*$ to *:* $\boldsymbol{F}^{\,p-1}_{n-p}$ *(a)*

*iii. Reject Ho If :* $F_2^* > \boldsymbol{F}^{\,p-1}_{n-p}$ *(a)*

### IV) Determine the Final Equation

Any Variable, $X_i$, for which the Hypothesis: Ho: $\beta_i = 0$ 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.

### (V) 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 $\alpha$ value, has the following relationship to $\alpha$:

a) If $p > \alpha$, do not Reject Ho.
b) If $p < \alpha$, Reject Ho.

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