Relative Efficiency of Linear Probability Model on Paired Multivariate Data

Aldwin M. Teves [1], Adelfa C. Diola [2]

[1] Professor , Institute of Arts and Sciences , Southern Leyte State University, Sogod, Philippines,
[2] Institute of Arts and Sciences , Southern Leyte State University, Sogod, Philippines.
Email: [1]tevesaldwinm@gmail.com, [2]adelfadiol06@gmail.com

Abstract

The investigation advanced the comparative efficiency of Liner Probability Model on paired multivariate data. On the basis of simulation, vector of means having common variance-covariance matrix were taken into account as data sets subjected to analyses. The linear probability model exhibited a more power tool to detect the presence of significant difference among variables compared to the multivariate paired (Hotellings'-T2) t-test. The Linear probability model is 31.33% and 44.67%., more efficient than the usual counterpart containing two and three predictor variables, respectively. It pays further, that under the regression analysis, individual variable is directly identified in relation to its significant contribution to the dependent variable. This observation is tantamount on determining that such vector of mean disparity is not significantly different from zero against the usual hypothesis. This procedure gains added advantage such that a sweeping generalization of whether vector of means are significantly or insignificantly different is avoided.

Keywords— efficiency, simulation, vector of means, variance-covariance matrix

I. INTRODUCTION

The term linear probability model (LMP) is used to denote a regression model in which the dependent variable is in a form of dichotomous expression. Here, the variable is measured in only two ways and can be designated with observations of success or failure, winner or loser, pro or anti, positive or negative, before and after. In this study, we consider only a few explanatory variables, (x's) for simplicity sake. The variable y is an indicator variable that denotes the occurrence or non-occurrence of an event. For instance, in the analysis of the determinants of unemployment, we have data on each person that shows whether or not the person is employed, and we have some explanatory variables that determine the state of employment. Here, the event under consideration is unemployment.

We define the dichotomous variable in various contention. In this context, it should be understood that under this consideration, the dependent variable has only two possible outcomes or measures. These dichotomized observations is strictly denoted by 1's and 0's. Problem with defined dichotomous dependent variable can be performed by a particular regression method known as Linear Probability Model (LPM). It has been observed in literature, that there are situations wherein the linear probability model clearly exhibits problem in the case of forecasting probability. However, there are many common situations where the linear model is just fine, and even has advantages. Besides, the focus of this study is not on the prediction of the probability of the occurrence, but rather on determining whether those define variables are significantly different relative to the tool routinely used in the paired multivariate. The T^2-Hotelling test. The linear probability model has the following advantages to wit:

i) The Linear Probability Model used in this study does not intend to forecast probability values, rather, based from the coefficients it determines the significance of the variables;
ii) The significant difference of the variable can be gleaned from the straight forward effect on the dichotomy of the dependent variable. When the coefficients is significant, it indicates that the variable has significant difference on the dichotomy;
iii) It has a direct recognition on the basis that individual effects are readily observable; unlike the t-test that when the vector of deltas is significantly different from zero, there might be a case when one of the deltas is not significantly different from zero.

This investigation determines the power of the Linear Probability Model as a tool in analyzing data in the paired multivariate case. In the multivariate case, the hypothesis focus on the differences between the before and after or the dichotomize categories of some defined variables. In the usual analysis confronted by the paired multivariate data, the multivariate t-test (T^2-Hotellings) is routinely employed. If the result of the analyses revealed that the vector of differences is significantly different from zero, there seems to be a sweeping generalization that all the variables are significantly different. In practice, it can be considered that not all or only a few of these are significantly different, hence, another question will be raised which of the variable that exhibits difference between categories. This proceeds to going back to the univariate case where, individual differences will be scrutinized or determined. This seems to be different from the usual objective of the multivariate analysis.

II. CONCEPTUAL FRAMEWORK

The basic idea sprouts in determining a statistical test in which given a significance level for testing a compound null hypothesis (H_0) against a compound alternative (H_a), the power is not less than the power of any other statistical test for testing H_o against H_a of the same significance level. The statistical test must be able to reject the false null hypothesis given the information contained in the data at hand. Such a test is considered as the most powerful test. Neyma-Pearson Lemma (1959) focused on this subject matter to determine the most powerful test.

In this study, two statistical tools are compared in terms of detecting the false null hypothesis, the paired multivariate t-test and the linear probability model. The comparison proceed from the data generated through simulation under the paired multivariate test. The data were generated through simulation, thus, the nature and properties of the data were known ahead. The number of variables have been considered in the comparison and their efficiency to detect the false null hypothesis.

III. OBJECTIVES

This study intends to determine the efficiency of Linear Probability Model (LPM) compared to the usual Hotelling's-T^2 in a paired multivariate case. The measure of efficiency is based on the number of significance (*false null hypothesis rejected*) detected by the LMP over the Hotelling's-T^2.

IV. METHODOLOGY

A. *Ordinary Least Square Method using Linear Probability Model*

Let's start by comparing the two models explicitly. If the outcome y is a dichotomy with values 1 and 0, define

$$y = \beta_o + \sum_{i=1}^{k} (\beta_i x_i) + \varepsilon_j$$

which is just the probability that y is 1, given some value of the regressor x . The linear probability model has an intuitive appeal over the related tool such as logistic model in the since that the coefficients are easily interpretable. For instance, a coefficient of 0.005 for a certain variable x_i would be interpreted as a 0.005 increase in probability per unit increase in x_i. Then the linear probability model is given to be:

$$\mu = E(y \mid x_1, x_2, ..., x_k) = \beta_o + \sum_{i=1}^{k} (\beta_i x_i) + \varepsilon_j$$

where:

β_i - is the regression coefficient at of the ith variable

x_i - predictor variables which are independent from each other

e_j - is random error in the jth observation.

Using the multiple linear model, the parameters are estimated with the following procedure

$$\hat{\beta} = (X'X)^{-1} X'Y$$

where:

$$\beta = \begin{pmatrix} \beta_o \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{pmatrix}$$, vector of unknown parameters

X =matrix of constant

Y =vector of observed dependent variable

The expression $\hat{y} = a + b_1 x_1 + ... + b_k x_k$ is tested whether y is attributed by the $x's$ and subsequently testing which particular x has a significant attribution to y. The linear model assumes that the probability of y is written as a linear functional manner of the predictor variables involved. The null hypothesis under the regression analysis is to test the contention that

$$H_o = \begin{pmatrix} \beta_o \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}$$

Such hypothesis is validated after the test of linearity using the analysis of variance on whether or not the variation of the predictor/independent/regressor variables significantly attribute to the dependent variable. The individual regression coefficients is tested using the t-test. It is in this fashion that the liner probability model exhibits appeal for it directly determines which particular variable in the coupled are significantly different or have significant contribution to probability measure.

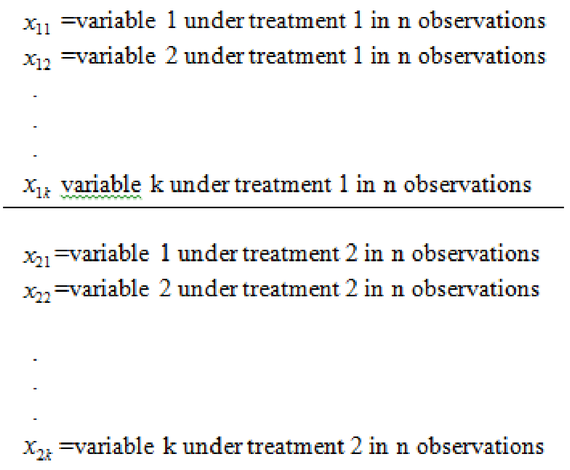
B. Multivariate Paired t-test

In carrying out any statistical analysis, it is always important to consider the assumptions for the analysis and confirm that all assumptions are satisfied.

Recall the four assumptions underlying the Hotelling's-T^2 test.

1. The data from population *i* is sampled from a population with mean vector μ_i.
2. The data from both populations have common variance-covariance matrix Σ.
3. **Independence**. The subjects from both populations are independently sampled.
4. **Normality**. Both populations are multivariate normally distributed.

Now let us consider the multivariate case. All scalar observations will be replaced by vectors of observations. Some notations are for the multivariate distinction of observations under the paired comparison procedure. It is necessary to distinguish between k responses, two treatments, and n experimental units. We label the observation accordingly in the following manner



x_{11} =variable 1 under treatment 1 in n observations
x_{12} =variable 2 under treatment 1 in n observations
.
.
.
x_{1k} variable k under treatment 1 in n observations

x_{21} =variable 1 under treatment 2 in n observations
x_{22} =variable 2 under treatment 2 in n observations
.
.
.
x_{2k} =variable k under treatment 2 in n observations

and the k paired-difference random variables becomes

$d_1 = x_{11} - x_{21}$ in n observations

$d_2 = x_{12} - x_{22}$ in n observations

.

.

.

$d_k = x_{1k} - x_{21}$ in n observations

Their corresponding expectations are as *follow*

$$E(d_i) = \sum_{j=1}^{n} (d_{ij} / n) = \delta_i = 0, \ (\forall \, i = 1,...,k)$$

Paired T²-Hotelling's test statistic is given by the expression below. Considering k-variable, we have the following contention for the null hypothesis, where the vector of the difference assume to be zero is plausible.

$$H_o : \delta = \begin{pmatrix} \delta_1 \\ \delta_2 \\ . \\ . \\ . \\ \delta_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}.$$

Here, δ_i, Type equation here., correspond to the theoretical difference in an ith variable between the before and after set of coupled observations. The expected difference between the before and after scenario is zero (0).

$$T^2 = n(\bar{d} - \delta)' S^{-1} (\bar{d} - \delta)$$

The null hypothesis will be rejected at alpha level when

$$T^2 > \frac{k(n-1)}{n-k} F_{(k,n-k)}$$

C. Data Generation

Data shall be generated through simulation. There will be two sets of data portraying the before and after scenario or we termed as coupled data. In each, there will be related variables being observed. These data can be analyzed with the use of the multivariate paired t-test or simply the Hotelling's- T². However, this type of data can also be analyzed using a regression analysis where the variables of the before and after scenario corresponds the to a dependent variable with measure of 0 for before and 1 for after, or to designate such measure to dichotomous observations. Moreover, these associated variables can serve as the explanatory variables whose significance need to be tested and will be the basis for the comparison between the methodologies.

The power of the tests shall be compared based on the how each test detects a false null hypothesis. The more the test detects a false null hypothesis the more powerful the test is. In this case, the efficiency is based on the most powerful tool of the two test.

Algorithm of Generating Data

1. Equal vector of means
   i) Generate multivariate data set 1 containing k-variable;
   ii) Generate multivariate data set 2 containing k-variable with slightly difference but not statistically significant variance-covariance matrix from data set 1;
   iii) Compare the decision rule on both data sets based on the null hypothesis;
   iv) Repeat the 100 times for comparison purposes of the two methodologies;
   v) Tabulate results for comparative purposes of the two methodologies.
2. Unequal vector of means
   i) Generate multivariate data set 1 containing k-variable
   ii) Generate multivariate data set 2 containing k-variable with unequal vector of means with slightly different but not significant variance-covariance matrix from the data set 1;
   iii) Compare the decision rule on both data sets based on the null hypothesis;

iv) Repeat the 100 times for comparison purposes of the two methodologies;

v) Tabulate results for comparative purposes of the two methodologies.

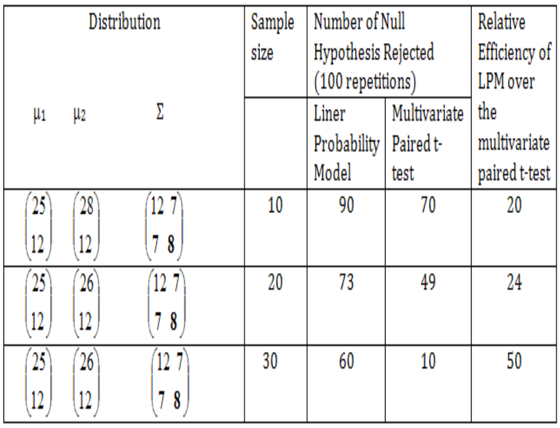
V. RESULT AND DISCUSSION

The efficiency of the Linear Probability Model (LPM) and the Paired Multivariate t-test were compared based multivariate data containing two variables and three variables. In each data description, the vector of means were adjusted as well as the sample sizes. This allows modification from one set of runs to the other set.

The null hypothesis that the difference in vector of means is zero were recorded after 100 runs of simulation. Results were tabulated below.

A. Case of two Dependent Variables

Table 1. Summary result of the LPM and multivariate paired t-test with the hypothesis that the differences of vector of means is zero under the distribution containing two variables .



Distribution			Sample size	Number of Null Hypothesis Rejected (100 repetitions)		Relative Efficiency of LPM over the multivariate paired t-test
μ1	μ2	Σ		Liner Probability Model	Multivariate Paired t-test	
$\begin{pmatrix}25\\12\end{pmatrix}$	$\begin{pmatrix}28\\12\end{pmatrix}$	$\begin{pmatrix}12 & 7\\7 & 8\end{pmatrix}$	10	90	70	20
$\begin{pmatrix}25\\12\end{pmatrix}$	$\begin{pmatrix}26\\12\end{pmatrix}$	$\begin{pmatrix}12 & 7\\7 & 8\end{pmatrix}$	20	73	49	24
$\begin{pmatrix}25\\12\end{pmatrix}$	$\begin{pmatrix}26\\12\end{pmatrix}$	$\begin{pmatrix}12 & 7\\7 & 8\end{pmatrix}$	30	60	10	50

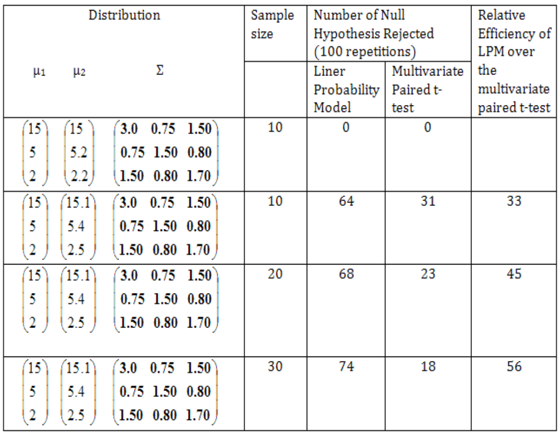
For data set containing two variables with small size of ten, the LPM rejected a false null hypothesis by 90 out of 100 repetition. While the multivariate paired t-test rejected false null hypothesis by 70 out of 100 repetitions. As the sample size has been increased to 20, both tests tend to reduce their power to reject false null hypothesis. The LPM rejected 73 out of 100 repetition while 49 for the multivariate paired t-test. Eventually, as the size has been increase to 30 the same trend has been observed. There were 60 false null hypothesis rejected by the LPM while 10 for the multivariate paired- t-test.

The LPM is sensitive to reject differences for small than the multivariate paired t-test, however, as the sample size has been increased such test lessen its sensitivity to reject the null hypothesis. The same is through with the multivariate paired t-test, however, the LPM is still more powerful than its counterpart.

It is note worthy to observe that as the sample sizes have been increased, the LPM tends to be more efficient in rejecting the false null hypothesis compared to the multivariate paired t-test. The LPM is 20%, 24% and 50% efficient for sample size of 10, 20, and 30, respectively. Generally, it is 31.33 % efficient compared its competing multivariate counterpart.

B. Case of three Dependent Variables

Table 2. Summary result of the LPM and multivariate paired t-test with the hypothesis that the differences of vector of means is zero under the distribution containing three variables .



Distribution			Sample size	Number of Null Hypothesis Rejected (100 repetitions)		Relative Efficiency of LPM over the multivariate paired t-test
μ1	μ2	Σ		Liner Probability Model	Multivariate Paired t-test	
$\begin{pmatrix}15\\5\\2\end{pmatrix}$	$\begin{pmatrix}15\\5.2\\2.2\end{pmatrix}$	$\begin{pmatrix}3.0 & 0.75 & 1.50\\0.75 & 1.50 & 0.80\\1.50 & 0.80 & 1.70\end{pmatrix}$	10	0	0	
$\begin{pmatrix}15\\5\\2\end{pmatrix}$	$\begin{pmatrix}15.1\\5.4\\2.5\end{pmatrix}$	$\begin{pmatrix}3.0 & 0.75 & 1.50\\0.75 & 1.50 & 0.80\\1.50 & 0.80 & 1.70\end{pmatrix}$	10	64	31	33
$\begin{pmatrix}15\\5\\2\end{pmatrix}$	$\begin{pmatrix}15.1\\5.4\\2.5\end{pmatrix}$	$\begin{pmatrix}3.0 & 0.75 & 1.50\\0.75 & 1.50 & 0.80\\1.50 & 0.80 & 1.70\end{pmatrix}$	20	68	23	45
$\begin{pmatrix}15\\5\\2\end{pmatrix}$	$\begin{pmatrix}15.1\\5.4\\2.5\end{pmatrix}$	$\begin{pmatrix}3.0 & 0.75 & 1.50\\0.75 & 1.50 & 0.80\\1.50 & 0.80 & 1.70\end{pmatrix}$	30	74	18	56

For data set containing three variables with small size of ten, the LPM and the multivariate paired t-test failed to reject that the differences of the vector of means is zero with a very slight differences. Such difference is defined as the ratio of the difference of the ith variable to its standard deviation, δ_i/σ. When the slight differences among means have been widen, the false null hypothesis has been detected by both tests. However, they detected in different proportions in 100 repetitions. The LPM has rejected 64 out of 100 while the multivariate paired t-test rejected 31 out of 100 repetitions. When the sample size has been increased to 20, it was observed that the number of false null hypothesis rejected using the LPM also increase while the multivariate paired t-test decreases. They rejected 68 and 23 out of 100 repetitions, respectively. With the same differences in the vector of means while the sample size has been increased to 30, the number of false null hypothesis rejected using the LPM continually increased while the multivariate paired t-test declines. The LPM rejected 74 out of 100 repetition while the multivariate paired t-test rejected 18 out of 100 repetition.

As the number of variables has increased from two to three, the LPM tends to detect false null hypothesis compared to its multivariate counterpart. Similarly, it was further noted that its efficiency tend to increase as the sample size gets larger and larger. It has a relative efficiency of 33%, 45% and 56%. With similar slight differences in the vector of means, the LPM detects more false null hypothesis than the multivariate paired t-test. Thus, it is more efficient to detect as the sample size gets larger. Generally, it has a relative efficiency of 46.67% across sample sizes.

VI. CONCLUSION AND RECOMMENDATION

The LPM exhibits efficiency in detecting false null hypothesis under two and three explanatory variables. Its efficiency improves as the sample size tends to increase compared to the multivariate paired t-test. Moreover, the relative efficiency of LPM has increased from two explanatory variables up to three explanatory variables across sample sizes accordingly, based from the generated simulated data.

It is recommended to check the comparison of these two methodologies for larger number of explanatory variables. It is also recommended that the comparison of means from individual variable-wise of $[(\mu_{1j}-\mu_{2j})/\sigma_j] < 1$, $[(\mu_{1j}-\mu_{2j})/\sigma_j] = 1$, and $[(\mu_{1j}-\mu_{2j})/\sigma_j] > 1$ be the bases of differences.

REFERENCES

1. Anderson, T. W. (1984, 2nd Ed). An Introduction to multivariate statistical analysis, N.Y.: Wiley
2. Bock, R. D. (1975). Multivariate statistical methods in behavioral research, N.Y.: McGraw Hill.
3. Carroll, J. D., Green, P. E. & Chaturvedi, A. (1997, 2nd Ed.). Mathematical tools for applied multivariate analysis. N.Y.: Academic Press
4. Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. N. Y.: Wiley.
5. Flury, B. (1997). A first course in multivariate statistics. N.Y.: Springer
6. Gifi, A. (1990, 2nd Ed.). Nonlinear Multivariate analysis. Chichester: Wiley
7. Gnanadesikan, R. (1997, 2nd Ed.). Methods for statistical data analysis of multivariate observations, N.Y.: Wiley.
8. Jonhson, R. A., and Wichern, D.W. (1998). Applied multivariate statistical analysis. Prentice-Hall International Incorporated. Simon & Schuster, Upper Saddle River, New Jersey 07458. ISBN: 0-13-080084-5.
9. Kendall, M. G. (1980, 2nd Ed.). Multivariate analysis. London: Griffin
10. Madala,G.S. (2001, 3rd). Introduction to econometrics. John Wiley & Sons Limited. Baffins Lane, Chichester,West Sussex PO19 IUD, England. ISBN: 9971-51-383-8.
11. Santos-Pereira, C.M. and Pires, A.M. (2002). Detection of outliers in multivariate data: a method based on clustering and robust estimators. Technical University of Lisbon Portugal.
12. Simon, M.K. (2006). Probability Distributions Involving Gaussian Random Variables. A Handbook for Engineers, Scientists and Mathematicians. Springer.
13. Scheaffer, R.L. and Young, L.J. (2010, 3rd Ed). Introduction to Probability and Its Application. Brooks/Cole CENGAGE Learning. International Edition.

14. Snedecor, George.W. and William G. Cochran (1980 7th Edition). Statistical Methods 1980. The Iowa State University Press, USA.
15. Staudte, R.G. and Simon J. Sheather (1990). Robust Estimation and Testing. A Wiley- Interscience Publication. John Wiley & Sons, Incorporated.
16. Teves, Aldwin M. (2017). Test of Homogeneity of based on geometric mean of variances. Volume 3 Issue 2, pp. 306 - 316Date of Publication: 06th September, 2017DOI. https://dx.doi.org/10.20319/pijss.2017.32.306316.