MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 11, NUMBER 3 SUMMER 1982

MULTIVARIATE NONPARAMETRIC ANALYSIS OF VARIANCE THROUGH MULTIPLE REGRESSION - THE TWO GROUP CASE

Bradley E. Huitema
University of Western Michigan

Abstract

The computation of the multivariate nonparametric analysis of variance requires matrix manipulations that are not familiar to many researchers. It is shown that the multivariate test statistic for the two group case can easily be computed with the aid of a conventional multiple linear regression computer program.

Presented at the annual AERA meeting March 19, 1982, New York City.

Introduction

In the randomized two group univariate analysis of variance case, situations arise where the nonparametric Mann-Whitney test is recommended in place of the parametric ANOVA F or t test or the corresponding regression analog. The choice between these parametric and nonparametric alternatives should generally be based on the nature of the population distributions and the adequacy of the measurement of the response variable. In the case that the population distributions approximate normality and the response measures are known to be

Presented at AERA 1982, MLR Special Interest Group
Not refereed by editorial staff

carefully obtained, the parametric procedures are generally chosen. This is because the relative efficiency (both asymptotic and small sample) of the nonparametric test relative to the parametric test is about .95. That is, if we compute the ratio of the sample sizes associated with the parametric and nonparametric tests having the same power and probability of Type I error, we find that fewer subjects are required for t or F than for the Mann-Whitney. Alternatively, when the sample size is constant, the power of the parametric test is greater. Many data analyzers appear to discount the usefulness of nonparametric alternatives for this reason and because the F test is said to be "robust" or insensitive to departures from distribution assumptions. It turns out, however, that a good case can be made for employing nonparametric statistics in certain situations.

If the population distributions are clearly nonnormal (e.g., exponential, rectangular, two-tailed exponential or long-tailed Cauchy) the parametric test is reasonably robust (using the typical textbook definition of robustness) but this does not mean that the inferences concerning the population means based on the sample means are equally good under all types of nonnormal distributions. The point here is that there is a difference between the effects of different types of nonnormality on a test criterion (such as F) and the effects on inferences made about paramaters. The former has to do with the concept of "criterion robustness" whereas the latter issue is that of "inference robustness". The reader is referred to Box and Tiao (1973) as the basic source on this distinction. The issue here is that the sample arithmetic means associated with a conventional parametric ANOVA may be inappropriate as estimates of the corresponding population means with certain types of nonnormality. The next point has to do with relative efficiency under nonnormality.

It was pointed out earlier that parametric t or F is generally preferable to the Mann-Whitney when normality is present because the relative efficiency of the latter is about .95. But what happens to the relative efficiency or power when the population distributions are clearly not normal?

If the deviation from normality is one of the long-tailed distributions, the Mann-Whitney test is far more efficient. For example, the asymptotic relative efficiency of the Mann-Whitney when the populations are two-tailed exponential is 150%. If the population distributions are Cauchy the asymptotic relative efficiency of the Mann-Whitney is ∞ (infinity) and the efficiency of t or F is zero.

The practical data analyzer should not conclude that there is no use for nonparametric tests such as the Mann-Whitney just because he/she does not encounter extreme nonnormality. There is a second reason why one should consider the use of nonparametrics.

It is not unusual, especially in large studies that involve many variables, to encounter "outliers" or scores that are extreme relative to others in the distribution. Sometimes these extreme scores can be attributed to instrumentation failures or clerical errors. In these situations it makes sense to eliminate the obviously invalid scores from the analysis. But it is frequently the case that we don't know whether an extreme observation is the result of invalid measurement or not. When this happens it is not clear whether the observation should be discarded or left in the sample. A reasonable strategy in this situation is to transform the data in such a way that the extreme score(s) has less influence in the estimation of parameters than when raw data are employed. The ranking transformation, which is a part of the computation of the Mann-Whitney test, is a simple and effective way of decreasing the in-

fluence of outliers. Since the chance of encountering an outlier increases with the number of variables analyzed, it is argued here that nonparametric procedures should be given serious consideration in large exploratory studies.

Purpose of Nonparametric Multivariate Analysis of Variance

When multiple dependent variables are employed in a two-group study it is frequently suggested that a multivariate analysis of variance or the mathematically equivalent Hotelling T^2 be computed. These approaches are employed rather than (or in addition to) univariate tests on each dependent variable for two reasons. First, the univariate approach ignors possibly useful information concerning the covariances among the various response measures. Second, the multivariate methods control the probability of Type I error for the whole family of response measures. That is, the probability of making one or more Type I errors in the whole collection of dependent variable tests is equal to or less than the alpha level selected for the analysis. When studies containing multiple dependent variables are analyzed using univariate tests the probability of making a Type I error is greater than the nominal alpha associated with each test. Hence the multivariate approach involves running an overall test that simultaneously considers all dependent variables at once.

In the case of the two-group multivariate nonparametric analysis of variance, the null hypothesis is written as follows:

$$H_o: \underline{\nu}_1 = \underline{\nu}_2 \text{ or } \begin{bmatrix} \nu_{11} \\ \nu_{21} \\ \cdot \\ \cdot \\ \cdot \\ \nu_{p1} \end{bmatrix} = \begin{bmatrix} \nu_{12} \\ \nu_{22} \\ \cdot \\ \cdot \\ \cdot \\ \nu_{p2} \end{bmatrix}$$

where ν_{ij} is the location parameter associated with the i^{th} dependent variable and the j^{th} population and

$\underline{\nu}_1$ and $\underline{\nu}_2$ are the vectors of the location parameters associated populations 1 and 2.

This is the hypothesis that the two populations are identical with respect to the p response measures. If this overall hypothesis is rejected there are several procedures that are appropriate for the identification of the dependent variable(s) responsible for the overall test conclusion. A simple approach is to run a Mann-Whitney test on each dependent variable. Issues associated with employing tests subsequent to the overall multivariate test are beyond the scope of the present paper.

The nonparametric multivariate techniques are virtually unused at the present time because they have been developed recently and the basic references (e.g., Puri and Sen, 1971) have been written primarily for mathematical statisticians rather than research workers. The purpose of this paper is to describe a simple procedure for computing the two group nonparametric multivariate analysis of variance with the aid of the output of a conventional multiple linear regression computer program.

Conventional Computation

The Puri and Sen nonparametric multivariate ANOVA procedure involves the computation of the test statistic $(N - 1)tr\underline{BT}^{-1}$

$\underline{B}$ is the between or among group sum of products of ranks matrix and

$\underline{T}^{-1}$ is the inverse of the total sum of products of ranks matrix.

This test statistic* is evaluated as a chi square with $p(J - 1)$ degrees of freedom where p is the number of dependent variables and J is the number of groups.

*While Puri and Sen (1971) have shown that their test statistic $Ntr\underline{BT}^{-1}$ is asymptotically distributed as chi square, the small sample properties are

Regression Procedure

The multiple regression solution requires the following steps:

1. Construct a data matrix that contains a dummy variable to identify subjects in the two groups (column 1), all other columns contain the ranks associated with the p dependent variables included in the design.
2. Regress the group membership dummy variable on the ranks of the dependent variable scores to obtain the multiple rank correlation coefficient R_s
3. Square R_s
4. Multiply N-1 times R_s^2 to obtain the test statistic. That is, $(N-1)R_s^2 = \chi^2$.

It can be seen from a comparison of the conventional and regression approaches that the test statistics are $(N-1)tr\underline{BT}^{-1}$ and $(N-1)R_s^2$ respectively. It follows that,

$$tr\underline{BT}^{-1} = R_s^2.$$

A proof is presented in the Appendix.

---

not known (Puri, 1974). I have chosen to define the test statistic as $(N-1)tr\underline{BT}^{-1}$ because (a) this statistic is also asymptotically distributed as chi square with p degrees of freedom under the null hypothesis of identical populations and (b) this statistic reduces (exactly) to the Kruskal-Wallis chi square statistic in the case of one dependent variable. Since the small sample properties of the Kruskal-Wallis statistic have been found to differ little from the asymptotic results, it would be suprising if the small sample properties of the multivariate generalization suggested here differ from the theoretical results. There will be almost no difference in the results obtained using these two formulas with respectable sample sizes.

Table 1 Example Raw and Ranked Data from a Two Group Design with Three Dependent Variables

Raw Scores

Group I			Group II		
y_1	y_2	y_3	y_1	y_2	y_3
3	10	12	21	56	11
17	17	7	27	57	10
20	51	5	35	62	6
70	53	0	38	63	1

Ranked Scores

Group I			Group II		
y_1 ranks	y_2 ranks	y_3 ranks	y_1 ranks	y_2 ranks	y_3 ranks
1	1	8	4	5	7
2	2	5	5	6	6
3	3	3	6	7	4
8	4	1	7	8	2

Computational Example

The computation of the multivariate test statistic for the data contained in Table 1 is summarized below for the conventional and regression solutions.

Conventional Solution

$$\underline{B} = \begin{bmatrix} 8.00 & 16.00 & 2.00 \\ 16.00 & 32.00 & 4.00 \\ 2.00 & 4.00 & 0.50 \end{bmatrix}$$

$$\underline{T}^{-1} = \begin{bmatrix} .12877 & -.07241 & .06746 \\ -.07241 & .06857 & -.02732 \\ .06746 & -.02732 & .06319 \end{bmatrix}$$

$$\underline{BT}^{-1} = \begin{bmatrix} .00649 & .46319 & .22886 \\ .01298 & .92638 & .45772 \\ .00162 & .11579 & .05722 \end{bmatrix} \quad \text{and}$$

$$\text{tr } \underline{BT}^{-1} = .00649 + .92638 + .05722 = .99009.$$

The test statistic is $(N - 1)\underline{trBT}^{-1} = (7).99009 = 6.93$. Since the critical value of chi square based on $p(J - 1) = 3(1) = 3$ degrees of freedom is 7.81 for alpha = .05, the overall multivariate null hypothesis is retained.

Regression Solution

Step 1 Construct the data matrix as shown below.

(1)	(2)	(3)	(4)
Group Membership	Y_1	Y_2	Y_3
Dummy Variable	Ranks	Ranks	Ranks
1	1	1	8
1	2	2	5
1	3	3	3
1	8	4	1
0	4	5	7
0	5	6	6
0	6	7	4
0	7	8	2

It can be seen that all subjects in the first group have been assigned the dummy score of one and all subjects in the second group have been assigned

the dummy score of zero.

Step 2 Regress the group membership dummy variable (column 1) on the ranks of the dependent variable scores (columns 2, 3, and 4). The resulting multiple correlation coefficient (actually the multiple rank correlation coefficient R_s) is .99503.

Step 3 Square R_s. R_s^2 is .99009.

Step 4 Multiply R_s^2 by N-1. (8-1) .99009 = 6.93 = χ^2. Notice that this is the same value obtained with the conventional computation procedure.

Since the obtained chi square does not exceed the critical value of 7.81 the following hypothesis is retained:

$$H_o: \begin{bmatrix} \nu_{11} \\ \nu_{21} \\ \nu_{31} \end{bmatrix} = \begin{bmatrix} \nu_{12} \\ \nu_{22} \\ \nu_{32} \end{bmatrix}$$

There is insufficient data to conclude that the population distributions are not identical. Since the overall hypothesis is not rejected there is no justification for additional tests on the individual dependent variables.

In conclusion, the nonparametric multivariate analysis of variance is a useful method for dealing with long tailed population distributions, possible outliers, and increased probability of Type I error associated with multiple response measures. It is easily computed with the aid of any multiple regression computer program.

Epilog

There is an alternative to the multivariate nonparametric analysis of variance for handling the problem of increased Type I error that is simple, effective and easily understood. This approach is described elsewhere (Huitema, forthcoming).

SUMMARY

There are two situations in which nonparametric procedures such as the Mann-Whitney test should be considered as useful alternatives to the parametric analogs: (1) when the population distributions are of certain nonnormal forms and (2) when the data contain unknown outliers. If responses are obtained on multiple dependent variables both of these problems are more likely to occur than in the univariate case.

An additional problem associated with the multivariate case is an increase in the probability of Type I error; that is, as the number of dependent variables is increased the probability of making a Type I error increases. One method of controlling Type I error is to employ the Puri-Sen nonparametric multivariate analysis of variance. It appears that the Puri-Sen method has virtually never been used. This is so because (a) the original papers presenting this procedure were written for mathematical statisticians (and are inscrutable for the typical research worker), (b) there are no secondary sources that describe the procedure, and (c) there are no widely distributed computer programs available to carry out the analysis.

The Puri-Sen test statistic can easily be computed for the two-group case by regressing a group membership dummy variable on the rank-transformed dependent variables and multiplying the resulting R^2 by N-1.

REFERENCES

Huitema, B. E. Bonferroni Statistics: Quick and Dirty Methods of Simultaneous Statistical Reference. Forthcoming.

Puri, M. L. Personal communication. February 13, 1974.

Puri, M. L. & Sen, P. K. On a class of multivariate multisample rank-order tests, Sankhyā, 1966, 28A, 353-376.

Puri, M. L. & Sen, P. K. Nonparametric Methods in Multivariate Analysis, New York: Wiley, 1971.