MULTIPLE LINEAR REGRESSION APPROACH TO ANALYSIS OF VARIANCE

BASSAM YOUNIS IBRAHIM[1]

ABSTRACT :

This paper deals with analysis of variance (ANOVA) as a special case of multiple linear regression analysis approach when categorical independent variables take values "0" or "1" are used. Numerical example is used to show that the two approaches lead to identical results, but the regression analysis convergence to the results without any additional step such as Duncan test as in ANOVA technique.

الملخص:

تهدف هذه الدراسة إلى استخدام تحليل التباين كحالة خاصة من حالات تحليل الانحدار الخطي المتعدد عندما تكون قيم جميع المتغيرات المستقلة وهمية (صفر أو واحد). حيث تم اعتماد مثال عددي يوضح توصل الطريقتين إلى نفس النتائج، إلا أن تحليل الانحدار الخطي المتعدد يعطي نتائج مباشرة من خلال اختبار معاملاته، في حين يتوجب استخدام طرق إضافية أخرى مثل اختبار دانكن للبحث عن معنوية المتغيرات في حالة استخدام أسلوب تحليل التباين.

Regression and ANOVA Models :

The regression and the analysis of variance models are treated as two separate and unrelated topics. ANOVA can be treated as a special case of a multiple linear regression model. In this paper I present the relationship between the two models and show how the ANOVA technique can be developed through regression approach.

Consider the two models, multiple linear regression model[1]:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \qquad \ldots\ldots\ldots(1)$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and $E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$.

And one-way analysis of variance model[2]:

[1] Dept. of Applied Statistics, College of Science, SUST.

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad i=1,2,...,k \; ; \; j=1,2,...,n \qquad (2)$$

where $\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)$ and $E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$.

Traditionally, the two models are presented as a method for treating different practical problems, the regression model being a mean of arriving at a procedure for predicting some response as a function of one or more quantitative independent variables, and the analysis of variance model for arriving at significance tests on multiple population means. Any mathematical model which is linear in the parameters such as (2), can be considered as a special case of (1). We can use conventional matrix notation to describe how each observation is expressed as a function of the parameters for the two models.

In the analysis of the sample data set presented in the next section we coded the level variable so that level I were represented by the value "1" and level j by the value "0". Such codes for categorical variables are called "dummy" variables since the actual values used have no intrinsic meaning; the numerical values which are coded are representatives for the categories. If "k" represents the number of categories of a particular variable, then the number of variables, which must be generated, and dummy coded to fully represent the original categories variable is k-1.

For the regression model:

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon} \qquad .....(3)$$

or more explicitly

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{21} & \cdots & X_{k1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where $\underline{Y}$ is a vector of responses in the experiment, the $\underline{X}$ matrix has already been described, and the $\underline{\beta}$ is a vector of parameters appearing in the model, and $\underline{\varepsilon}$ is an error vector. The

least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ of the parameters $\beta_0, \beta_1, \cdots, \beta_k$ are obtained by solving the normal equations:

$$\underline{A}\underline{\hat{\beta}} = \underline{g} \qquad \cdots\cdots(4)$$

where $\underline{A} = \underline{X'X}$ is a non-singular matrix and $\underline{g} = \underline{X'Y}$. Thus the estimates are given by

$$\underline{\hat{\beta}} = \underline{A}^{-1}\underline{g} \qquad \cdots\cdots(5)$$

To test the hypothesis $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$, i=1,2,....k we use the statistic :

$$t_i = \frac{\hat{\beta}_i}{S.E(\hat{\beta}_i)} \sim t_{(n-2)\alpha/2}$$

Hence the fitted model is obtained.

For the ANOVA model, in matrix form :

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n} \\ \cdots \\ Y_{2n} \\ Y_{2n} \\ \vdots \\ Y_{2n} \\ \cdots \\ \vdots \\ \vdots \\ \cdots \\ Y_{k1} \\ Y_{k2} \\ \vdots \\ Y_{kn} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ & & \cdots & & \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ & & \cdots & & \\ \vdots & \vdots & \vdots & & \vdots \\ & & \cdots & & \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \vdots \\ \tau_k \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n} \\ \cdots \\ \varepsilon_{2n} \\ \varepsilon_{2n} \\ \vdots \\ \varepsilon_{2n} \\ \cdots \\ \vdots \\ \vdots \\ \cdots \\ \varepsilon_{k1} \\ \varepsilon_{k2} \\ \vdots \\ \varepsilon_{kn} \end{bmatrix}$$

$kn \times 1 \quad kn \times (k+1) \quad (k+1) \times 1 \quad kn \times 1$

Again each observation is expressed as a function of the parameters. Here the very important matrix $\underline{X}$ (the matrix of experimental conditions), consists of zeroes and ones. Applying the least squares approach to the one-way ANOVA model, the normal equations are given by

$$\underset{(k+1)\times(k+1)}{\begin{bmatrix} nk & n & n & \cdots & n \\ n & n & 0 & \cdots & 0 \\ n & 0 & n & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ n & 0 & 0 & \cdots & n \end{bmatrix}} \underset{(k+1)\times}{\begin{bmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \\ \vdots \\ \hat{\tau}_k \end{bmatrix}} = \underset{(k+1)\times}{\begin{bmatrix} T_{..} \\ T_{1.} \\ T_{2.} \\ \vdots \\ T_{k.} \end{bmatrix}}$$

the significance tests are performed on the population means $\mu_i = \mu + \tau_i$ and in formulating the test procedure, the linear constraint $\sum_{i=1}^{k} \tau_i = 0$ is applied. Thus, the τ_i's takes on the role of deviations (plus or minus) of the treatments or population means then becomes equivalent to testing that τ_i's are all zero.

The estimating equations can be solved to yield:

$$\hat{\mu} = \frac{T_{..}}{nk} = \bar{Y}_{..} \qquad \ldots\ldots\ldots(6)$$

$$\hat{\tau}_i = \frac{T_{i.}}{n} - \frac{T_{..}}{nk} = \bar{Y}_{i.} - \bar{Y}_{..} \quad i=1,2,\ldots,k \qquad \ldots\ldots\ldots(7)$$

To approach the hypothesis-testing problem on the one-way ANOVA model from the multiple regression procedure, we compute the regression sum of squares for the estimators $\hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_k$. These estimators take on the same role as the coefficients $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ in the multiple linear regression model. We would then compute the regression sum of squares :

$$R(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k) = SSR$$

$$= \hat{\beta}_0 g_0 + \hat{\beta}_1 g_1 + \cdots + \hat{\beta}_k g_k - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}\right)^2}{nk}$$

$$= \hat{\beta}_1 g_1 + \hat{\beta}_2 g_2 \cdots + \hat{\beta}_k g_k$$

$$= \hat{\tau}_1 g_1 + \hat{\tau}_2 g_2 \cdots + \hat{\tau}_k g_k$$

The R.H.S. of the estimating equations give

$$g_i = T_{i.} \quad \forall\ i = 1,2,\ldots,k$$

Hence

$$R(\hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_k) = \sum_{i=1}^{k}\left(\frac{T_{i.}}{n} - \frac{T_{..}}{nk}\right)T_{i.}$$

$$= \sum_{i=1}^{k}\frac{T_{i.}^2}{n} - \frac{T_{..}^2}{nk}$$

$$= SSB$$

The hypothesis $H_0 : \tau_1 = \tau_2 = \cdots \tau_k$ is tested by forming the ratio

$$F = \frac{R(\hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_k)/(k-1)}{SSE/k(n-1)} = \frac{SSB/(k-1)}{S^2} \quad \ldots\ldots\ldots(8)$$

The significance F at (k-1) and k(n-1) d.f's is needed to look for the required treatments and in order to do that, we must use one of the methods such as Tukey, Dunnett, Schiffe, or Duncan methods, Steel & Torrie[3].

NUMERICAL EXAMPLE:

The following example given by Neter & Wasserman[2] is used. The Kenton Food Company wishes to test four different package designs for a new breakfast cereal. Ten stores, with approximately equal sales volumes, were selected as the experimental units. Other relevant conditions beside package design, such as price, amount and location of shelf space, and special promotional efforts, were kept the same for all stores in the experiment. Sales, in number of cases, were observed for the study period and the results are recorded in table (1).

Table(1): Number of cases soled by stores for each package design

	1	2	3	4
Stores	12	14	19	24
	18	12	17	30
		13	21	

Regression Approach:

Since the independent variables has four classes (factor levels), we use three categorical (indicator) variables in the regression model. Let us define them as follows:

$$X_1 = \begin{cases} 1 \text{ if observation from factor 1} \\ 0 \text{ otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 \text{ if observation from factor 2} \\ 0 \text{ otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 \text{ if observation from factor 3} \\ 0 \text{ otherwise} \end{cases}$$

Table (2) gives the transformed data of table (1).

Table(2): Data using regression approach

Y	X_1	X_2	X_3
12	1	0	0
18	1	0	0
14	0	1	0
12	0	1	0
13	0	0	0
19	0	0	1
17	0	0	1
21	0	0	1
24	0	0	0
30	0	0	0

The regression model is :

$$Y_i = 27 - 12X_{1i} - 14X_{2i} - 8X_{3i}$$

Table (3) indicates the significance of the three variables.

Table(3): ANOVA table for the regression model

S.O.V	D.F	S.S	M.S	F
Regression	3	258	86	11.2*
Error	6	46	7.67	
Total	9	304		

*: means significant at 5%

To test $H_0 : \beta_i = 0$

versus $H_1 : \beta_i \neq 0$ i=1,2,3

we found t_1=4.33 , t_2=5.53 , t_3=3.17 which are all significant at 5%. Therefore the fitted model is :

$$Y_i = 27 - 12X_{1i} - 14X_{2i} - 8X_{3i}$$

ANOVA Approach :

The ANOVA table for the data in table (1) is given in table (4).

Table(4): ANOVA table for the data in table (1)

S.O.V	D.F	S.S	M.S	F
Between Designs	3	258	86	11.2*
Error	6	46	7.67	
Total	9	304		

*: means significant at 5%

We conclude that the factor level means are not equal, or that the four different package designs do not lead to the same mean sales volume. Therefore there is a relation between package designs and sales volume. Now, to search about designs which are effective on sales volume we use Duncan's multiple range test as follows :

Interchange Y by X_0, then :

$$S_{\bar{X}_i} = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{7.67}{10}} = 0.8758 \quad , i=0,1,2,3$$

The values of S.S.R according to the number of means are compared, number of d.f's for error 6 and at 5% are: Snedecore & Cochran[4]

S.S.R= 3.46 , 3.58 , 3.64

and since:

$L.S.R=S.S.R \times S_{\bar{D}}$

therefore:

L.S.R=3.0303 , 3.1354 , 3.1879

Table(5): Duncan Test for the data in table (1)

Variables	$\bar{X}$	L.S.R	$\bar{X}-\bar{X}_1$	$\bar{X}-\bar{X}_2$	$\bar{X}-\bar{X}_3$
X_0	18	3.0303	17.8*	17.7*	17.7*
X_2	0.3	3.1354	0.1	0.0	
X_3	0.3	3.1879	0.1		

*: means significant at 5%.

From table (5), we conclude that there is a significance difference between Y and each of the three independent variables.

CONCLUSIONS:

- The ANOVA and regression analysis leads to identical results, and the ANOVA approach can be treated as a special case of multiple linear regression model when the independent variables are all categorical.
- The regression analysis gives the results easier and faster than the ANOVA, since the model depends on testing of the coefficients of the variables, whereas the multiple regression procedure used after the ANOVA procedure specially when the null hypothesis is rejected.

REFERENCES:

1- Myers, R.H. (1986); "Classical and Modern Regression with Applications", PWS Publishers.

2- Neter, J. & W. Wasserman (1974); "Applied Linear Statistical Models", Home Wood, III .

3- Steel, G.D. & J.W. Torrie (1981); "Principles and Procedures of Statistics", McGraw-Hill book Company.

4- Snedecore, G.W. & W.G. Cochran (1967); "Statistical Methods", 6th ed , Iowa State Univ., Press Ames, Iowa.

5- Walpole, R.E. & R.H. Myers (1972); "Probability and Statistics for Engineers and Scientists", Macmillan Company, New York.