International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 2 Issue 8 || August. 2014 || PP-47-54

On Regression Approach to ANOVA Designs

Ukwu Chukwunenye
Department of Mathematics, University of Jos P.M.B. 2084, Jos, Plateau State, Nigeria

ABSTRACT : Regression approach to one-way (one-factor) Analysis of Variance problems was investigated using indicator variables to represent the selection of data from alternative levels (treatments) of the factor. A matrix, *U* of indicator variables X_{ij} was obtained, from which $(U^tU)^{-1}U^ty$ yielded the means of the output responses

corresponding to given treatments right-off- the-bat. The key benefit is that the matrices U^tU, $(U^tU)^{-1}$, U^ty and $(U^tU)^{-1}U^ty$ have very simple structures, which could be deployed for an easy generation of a desired ANOVA table and the establishing of Confidence and Prediction intervals.

***KEYWORDS*:** *ANOVA, Indicator, Matrices, Regression, Structure*.

I. INTRODUCTION

Every ANOVA table is motivated by the need to obtain the *F* statistic, called the variance ratio as a basis for the acceptance or the rejection decision on the null Hypothesis: all population means are equal against the alternative Hypothesis: not all population means are equal; that is, at least one population mean differs from the others. A positive assessment of any feasible method for securing the *F* statistic is predicated on its reduction of the computing complexity associated with ANOVA table generation and the simplicity of its mathematical/statistical structures for easier and more enlightened exposition on the subject. See Ukwu [1], Farnum [2], and Juran [3].

II. AIMS AND OBJECTIVES

This research is aimed at enhancing ANOVA computations using a simplified matrix-oriented mathematical structure. This also translates to a better understanding and greater mathematical appreciation of the subject. Matrix-based presentations turn out quite often to be the icing on the cake, especially if the base matrices follow simple recognizable patterns.

III. METHODS AND MATERIALS

Representations of one-way (one-factor) Analysis of Variance problems will be achieved using indicator variables to represent the selection of data from alternative levels (treatments) of the factor. A matrix, *U* of indicator variables X_{ij} will be obtained, from which $(U^tU)^{-1}U^ty$ will hopefully yield the means of the output responses corresponding to given treatments right-off-the bat. The expected key benefit is that the matrices

$$U^tU,\ (U^tU)^{-1},\ U^ty \text{ and } (U^tU)^{-1}U^ty$$

should have very simple structures, which could pave the way for an easy generation of a desired ANOVA table.

4.Solution Procedure

The procedure is as follows: Suppose that there are *k* treatments of a given factor,

$$\text{let: } X_j = \begin{cases} 1, & \text{if data are taken from treatment } j \\ 0, & \text{otherwise} \end{cases} \quad (1).$$

Suppose that the j^{th} treatment sample is of size n_j so that the entire data set is of size:

$$N = \sum_{j=1}^{k} n_j \quad (2).$$

This and the definition of X_j induce the following definitions.

For $j = 1, 2, \cdots, N$,

$$X_{ij} = \begin{cases} 1, \text{ if the } i^{\text{th}} \text{ observed (output)} \\ \text{response is taken from treatment } j \\ 0, \quad \text{otherwise} \end{cases} \tag{3}$$

Let the treatments be taken in increasing serial order, with the observed responses exhausted before the next treatment is sampled; needless to say that the treatments could simply be renamed to conform to this order. Then:

$$X_{ij} = \begin{cases} 1, \text{ if } 1 + \sum_{s=1}^{j} n_{s-1} \le i \le \sum_{s=1}^{j+1} n_{s-1} \\ 0, \quad \text{otherwise} \end{cases} \tag{4}$$

where: $n_0 = 0$.

In other words, the column vector, X_j with components, X_{ij} has a contiguous block of n_j ones in its rows $1 + \sum_{s=1}^{j} n_{s-1}$ through $\sum_{s=1}^{j+1} n_{s-1}$, and zeros elsewhere in its column.The resulting matrix, *U* of indicator variables is immediate. In fact:

$$U = \begin{pmatrix} X_1 & X_2 & \cdots & X_k \\ 1 & 0 & \cdots & 0 \\ 1 & \vdots & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ 0 & \vdots & \cdots & 0 \\ 0 & 1 & & 0 \\ 0 & 0 & & 0 \\ 0 & 0 & & 0 \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{5}$$

The created indicator variables $X_1, X_2, \cdots, X_k$ are now adopted as the input variables for the regression model.

The observed response vector is:

$$y = (y_1, y_2, \cdots, y_k)^t \tag{6}$$

where:

$$y_j = \left(y_{1j}, y_{2j}, \cdots, y_{n_j j}\right)^t \tag{7}$$

Clearly: $y = (v_1, v_2, \cdots, v_N)^t$ (8)

where: $y_j = (1 + v_{N_j}, \cdots, v_{N_{j+1}})^t$, (9)

$$N_j = \sum_{s=1}^{j} n_{s-1}, \text{ and } n_0 = 0. \tag{10}$$

In (9), the subscripts of N in the first and last entries are *j* and $j+1$ respectively. y_1 is just the 1st n_1 entries in (8), y_2 the next n_2 entries, $\cdots, y_k$ the last n_k entries.

The associated Regression model is:

$$\hat{y} = a_0^* + \sum_{j=1}^{k} a_j^* \bar{X}_j \tag{11}$$

$$\text{where: } a_0^* = \bar{y} - \sum_{j=1}^{k} a_j^* \bar{X}_j \tag{12}$$

$$\text{and } \left(a_1^*, a_2^*, \cdots, a_k^*\right)^t = \left(U^t U\right)^{-1} U^t y = \left(\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_k\right)^t \tag{13}$$

To establish (13), let I_j be a column vector with n_j onesin its rows $1 + \sum_{s=1}^{j} n_{s-1}$ through $\sum_{s=1}^{j+1} n_{s-1}$ and column*j*. Then, from (5), the following is clear:

$$U = \begin{pmatrix} I_1 & 0 & \cdots & 0 \\ 0 & I_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 \\ 0 & 0 & \cdots & I_k \end{pmatrix} = \text{Diag}\left(I_1, I_2, \cdots, I_k\right) \tag{14}$$

$$U^t = \begin{pmatrix} (I_1)^t & 0 & \cdots & 0 \\ 0 & (I_2)^t & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 \\ 0 & 0 & \cdots & (I_k)^t \end{pmatrix} \tag{15}$$

$$= \text{Diag}\left((I_1)^t, (I_2)^t, \cdots, (I_k)^t\right) \tag{16}$$

U and U^t are respectively *N* by *k*, and *k* by *N* block diagonal matrices. Therefore:

$$U^t U = \begin{pmatrix} (I_1)^t I_1 & 0 & \cdots & 0 \\ 0 & (I_2)^t I_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 \\ 0 & 0 & \cdots & (I_k)^t I_k \end{pmatrix} \tag{17}$$

$$= \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 \\ 0 & 0 & \cdots & n_k \end{pmatrix} = \text{Diag}(n_1, n_2, \cdots, n_k) \tag{18}$$

Consequently:

$$\left(U^t U\right)^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{pmatrix} = \text{Diag}\left(\frac{1}{n_1}, \frac{1}{n_2}, \cdots, \frac{1}{n_k}\right) \tag{19}$$

$$U^t y = \left((I_1)^t y_1, (I_2)^t y_2, \cdots, (I_k)^t y_k\right)^t \tag{20}$$

$$U^t y = \left((I_1)^t y_1, (I_2)^t y_2, \cdots, (I_k)^t y_k\right)^t = \left(\sum_{i=1}^{n_1} y_{i1}, \sum_{i=1}^{n_2} y_{i2}, \cdots, \sum_{i=1}^{n_k} y_{ik}\right)^t \tag{21}$$

Hence :

$$\left(a_1^*, a_2^*, \cdots, a_k^*\right)^t = \left(U^t U\right)^{-1} U^t y = \left(\frac{1}{n_1}\sum_{i=1}^{n_1} y_{i1}, \frac{1}{n_2}\sum_{i=1}^{n_2} y_{i2}, \cdots, \frac{1}{n_k}\sum_{i=1}^{n_k} y_{ik}\right)^t = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_k \end{pmatrix} \tag{22}.$$

Thus, the optimal parametervector $(a_1^*, a_2^*, \cdots, a_k^*)^t$ is just the vector of the treatment means $\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_k$. Thus:

$$a_0^* = \bar{y} - \sum_{j=1}^{k} a_j^* \bar{X}_j, \tag{23}.$$

where: $\bar{y}$ is the grand mean; that is:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij} \tag{24}$$

$$N = \sum_{j=1}^{k} n_j; \ \bar{X}_j = \frac{n_j}{N}.$$

Therefore:

$$\sum_{j=1}^{k} a_j^* \bar{X}_j = \frac{1}{N}\sum_{j=1}^{k} n_j \bar{y}_j = \frac{1}{N}\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij} \tag{25}.$$

It follows from (23) that:

$$a_0^* = 0, \tag{26}$$

which makes perfect sense, having already identified the means of the treatments by $a_1^*, a_2^*, \cdots,$ and a_k^*.

The usual hypotheses on the population parameters $a_1, a_{2,\ldots}, a_k$ or $\mu_1, \mu_2, \cdots, \mu_k$ are as follows:
Null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu \tag{27}$$

versus
Alternative hypothesis:

$$H_1 : \mu_s \neq \mu_t, \text{ for some } s \neq t;\ s, t \in \{1, 2, \cdots, k\}.$$

A level of significance α is chosen and H_0 is either accepted /not rejected or rejected based on the calculated *F*-ratio (variance ratio) being less than the critical *F*-value, $F(k, n-k-1; \alpha)$ for acceptance/non-rejection and greater than $F(k, n-k-1; \alpha)$ for rejection. Obtaining this ratio requires the generation of an ANOVA table.

Next, confidence intervals are established for the population parameters $\mu_1, \cdots, \mu_k$ using the critical *t*-value, $t_{N-k-1;\, \alpha/2}$, the pooled (averaged) standard deviation:

$$s_p = \left(\frac{\sum_{j=1}^{k} (n_j - 1) s_j^2}{N-k} \right)^{\frac{1}{2}}, \tag{28}$$

and the matrix $(U^t U)^{-1}$, where:

$$s_j^2 = \frac{1}{n_{j-1}} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y}_j \right)^2 \tag{29}$$

is the variance of the j^{th} treatment.

Note: it follows from (29) that:

$$s_p^2 = MS(\text{resid}) \equiv MSE(\text{Mean square error}) \equiv s_{e_y}^2 \tag{30}.$$

The *df* of $SS(\text{resid}) \equiv SSE$ is adjusted to $N-k$, since *k* parameters and not $k+1$ need to be estimated (having dropped a_0 from contention). The standard errors of the estimators, a_j^* are given by:

$$s_{e_j} = s_p \sqrt{d_{jj}} \tag{31}$$

where, d_{jj} is the j^{th} diagonal element $(U^t U)^{-1}$.
Therefore:

$$s_{e_j} = s_p \sqrt{\frac{1}{n_j}} = \sqrt{\frac{MSE}{n_j}} \tag{32}.$$

The confidence intervals for the population parameters
$\mu_1, \mu_2, \cdots, \mu_k$ are given by:

$$\begin{aligned} CI(\mu_j) &= CI(a_j) = a_j^* \pm t_{N-k;\, \alpha/2}\ s_{e_j} \\ &= \bar{y}_j \pm t_{N-k;\, \alpha/2} \sqrt{\frac{MSE}{n_j}} \\ &= \bar{y}_j \pm t_{N-k;\, \alpha/2} \frac{s_p}{\sqrt{n_j}} \end{aligned} \tag{33}$$

Set

$$D = (U^t U)^{-1}.$$

For any selected treatment$j$, $j \in \{1, 2, \cdots, k\}$, the Prediction interval for a future mean response value is given by:

$$PI = a_j^* \pm t_{N-k;\,\alpha/2}\, s_p \sqrt{1 + \frac{1}{N} + \frac{1}{n_j}}.$$

This follows from the definition of X_j and from the fact that for any appropriate indicator vector variable X_j, $j \in \{1, 2, \cdots, k\}$:

$$X_j^t D X_j = \frac{1}{n_j} \tag{34}.$$

Special Cases

(*i*) If

$$n_1 = n_2 = \cdots = n_k = n, \text{ say, then, } U^t U \text{ and } (U^t U)^{-1}$$

reduce respectively to:

$$n \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \text{ and } \frac{1}{n} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{35}$$

scalar multiples of the identity matrix of order *k*.

(*ii*) If $k = 2$, a two-sided $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ can be established following the appropriate test of hypothesis using a *t* test statistic. The steps are as follows:

$$H_0: \mu_1 = \mu_2 \qquad (\mu_1 - \mu_2 = 0)$$

versus

$$H_1: \mu_1 \neq \mu_2 \qquad (\mu_1 - \mu_2 \neq 0).$$

The test statistic, *t*is given by:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\text{postulated value of } [\mu_1 - \mu_2])}{\text{standard error of } (\bar{y}_1 - \bar{y}_2)},$$

where: (36)

$\bar{y}_1 - \bar{y}_2 =$ estimated value $\mu_1 - \mu_2$.

The standard error of $(\bar{y}_1 - \bar{y}_2)$ is given by:

$$\sqrt{s_{e_1}^2 + s_{e_2}^2} \tag{37}$$

$$= \sqrt{s_p^2\,(d_{11} + d_{22})}$$

$$= s_p \sqrt{d_{11} + d_{22}}$$

$$= s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{38}.$$

It follows that:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \equiv t_{\text{calculated (calc)}} \qquad (39).$$

This is the observed (calculated) *t*-statistic. t_{calc} is then compared to the critical *t* value,

$$t_{\text{crit}} = t_{n_1+n_2-2;\,\alpha/2}$$

H_0 is accepted if $t_{\text{calc}} < t_{\text{crit}}$.

H_0 is rejected if $t_{\text{calc}} > t_{\text{crit}}$.

The confidence interval for $\mu_1 - \mu_2$ is established as:

$$CI(\mu_1 - \mu_2) = \bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2;\,\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (40).$$

$$\text{where: } s_p = \sqrt{MSE} = \left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)^{\frac{1}{2}}.$$

Remarks

(*i*) It is computationally prohibitive to use the *t* test for problems of size $k \geq 3$; to do so must require $\binom{k}{2}$ tests of hypotheses, corresponding to the different *t*-values. For example, if $k = 5$, then, $\binom{5}{2} = 10$ *t*-tests would need to be performed. Clearly, the impracticality or computing complexity associated with *t*- tests for problems of reasonable sizes has been established.

(*ii*) The type 1 error associated with $\binom{k}{2}$ hypotheses tests arising from the use of *t*-tests on

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

versus

$$H_1: \mu_s \neq \mu_t, \text{ for some } s,t \in \{1, 2, \cdots, k\},\ s \neq t.$$

is $1-(1-\alpha)^{\binom{k}{2}}$.

(Mutual independence of these tests is assumed here).

A cursory glance at the table below will be quite revealing:

α	k	$\binom{k}{2}$	Type 1 error
0.05	5	10	0.4013
	6	15	0.5367
	7	21	0.6594
	8	28	0.7622

Clearly, such high levels of type 1 error are unacceptable. Above remarks motivate the use of the *F*test (a unique test for a given α and given degrees of freedom) in ANOVA designs.

Every ANOVA table is motivated by the need to obtain the variance ration or *F* statistic,

$F = \dfrac{MS\left(\text{reg}\right)}{MS\left(\text{resid}\right)}$, where $MS\left(\text{reg}\right) = \dfrac{1}{k-1}\sum_{j=1}^{k}\left(\overline{y}_j - \overline{y}\right)^2$, as a basis for the acceptance or rejection decision on the null hypothesis:all population means are equal against the alternative hypothesis: the population means are not all equal; that is, at least one population mean differs from the others.

IV. CONCLUSION

This paper used vector indicator variables and linear regression method to model one-way ANOVA designs. These indicator variables served as an appropriatevehicle for the selection of data from alternative levels (treatments) of the factor. The key benefit is that certain relevant matrices have very simple structures which could pave the way for easy generation of desired ANOVA tables as well as the establishing of confidence and prediction intervals. In the sequel the paper provided a motivation for the use of variance ratios and the unsuitability of t tests due to their unacceptably high type 1 error levels with increasing treatment sizes.

REFERENCES

[1] Ukwu, C. (2014t). A technique for $n2^k$ factorial designs. *International Journal of Mathematics and Statistics Studies (IJMSS).* Vol. 2, No.2, June 2014.

[2] Farnum, N.R., "Statistical Quality Control and Improvement", Duxbury Press,Belmont, California, 1994

[3] Juran, J.M. & Gryna, F.M., "Juran's Quality Control HAND BOOK, 4th Ed.McGraw-Hill International Editions, Industrial Engineering Series, NY.1988.