RESEARCH HUB – International Multidisciplinary Research Journal (RHIMRJ)
Research Paper
Available online at: www.rhimrj.com

Clustering on The Basis of Regression Equations Under Heteroscedastic Errors

Jagabandhu Saha
Department of Economics and Politics,
Visva-Bharati University, India

Abstract: ***The Chow test is not robust under heteroscedasticity. The presence of heteroscedasticity will affect level of significance as well as power of the test, especially when the sizes of the samples are small. The present paper not only resolves the problem of heteroscedasticity in the error terms, but also extends the existing method of comparing two linear regressions to one where it is possible not only to compare the equality between the sets of coefficients in the two linear regressions, but also, in case they are not equal, to provide detailed information about the inequality of the sets as well as to accomplish all these for not just only two linear regressions but for the two linear regressions of all possible pairs of linear regressions out of any number of given linear regressions, and then to use the results of all these comparisons in order to form clusters among the regressions on the basis of some principle stated therein. The procedure is then illustrated through comparison of the decadal growth rates of the population of India, using NSSO data.***

1. Introduction

Testing the equality between sets of coefficients in two linear regressions by Chow test (Chow 1960) [1] is well known. Chow however assumed homoscedasticity of the regression errors. It is already demonstrated in the literature that the Chow test is not robust under heteroscedasticity (Toyoda 1974 [2], Schmidt and Sickles 1977 [3], Ali and Silver 1985 [4] and Tansel 1987 [5]). The presence of heteroscedasticity will affect level of significance as well as power of the test. This means that if there is heteroscedasticity in the errors, but we perform Chow test assuming homoscedasticity then the result may be different from the actual especially when the sizes of the samples are small.

In the Chow test, if the null hypothesis of equality between the sets of coefficients is not rejected then there is no problem (as in the examples in his paper). But if rejected, then, naturally, one is probed to the questions: a). at which component/s the sets differ, and b). for each of those components, between the two coefficients of the two regressions concerned, which one is larger/smaller. Chow test does not provide answer to any of these questions. This problem can be resolved with some modifications of the model (Saha and Pal 2014) [6]. Saha and Pal introduced the concept of "component wise complete comparison" (CCC)[1] in order to overcome this problem. The test procedure for CCC between every two successive regressions out of any number of given successive regressions was developed. If heteroscedasticity is present then the problem of CCC aggravates and needs further modifications. The earlier paper [6] was then extended in order to incorporate heteroscedasticity (Saha 2018) [7]. The test procedure for CCC between every two successive regressions out of any number of given successive regressions, when the errors are heteroscedastic, was developed. The present paper extends the last one [7] in order to develop a test procedure for CCC between not only every two successive regressions out of any number of given successive regressions, but between the two linear regressions of *all possible pairs of regressions out of any number of given regressions*, and then to use the

[1] By complete comparison between any two parameters a and b we mean to decide whether a < b or a = b or a > b. By component wise complete comparison (CCC) between two vectors of parameters of the same size ($a_1 a_2 \ldots a_m$) and ($b_1 b_2 \ldots b_m$) we mean complete comparison between (a_1 and b_1), (a_2 and b_2), ... and (a_m and b_m). By CCC between/of/for two regressions with same no. of parameters we mean CCC between the two vectors of parameters of these regressions. In the paper by Saha and Pal, CCC is done between every two successive regressions out of any number of given successive regressions with same no. of parameters.

  ISSN: 2349-7637 (Online)

results of all these comparisons in order to form clusters among the regressions on the basis of this simple principle: *all those regressions which satisfy the condition that the vectors of coefficients of any two of these regressions do not differ from each other significantly will form a cluster.* The rest of the paper can be outlined as follows. In Section 2, we put the problem in the formal terms, problem of finding test procedure for CCC between the two regressions of all possible pairs of regressions out of several given regressions, when the regression errors are heteroscedastic. Section 3 is devoted for the methodology for solving this problem and then for discussion about the process of clustering. In Section 4 we consider a numerical example in order to illustrate the methodology and the process of clustering while in Section 5 we put our conclusions.

2. The Model: We consider the problem of finding test procedure for CCC between the two regressions of all possible pairs of regressions out of m given regressions as follows:

$$y^{(1)} = a_1^{(1)} + a_2^{(1)} x_2^{(1)} + a_3^{(1)} x_3^{(1)} + \ldots\ldots + a_k^{(1)} x_k^{(1)} + u^{(1)},$$

$$y^{(2)} = a_1^{(2)} + a_2^{(2)} x_2^{(2)} + a_3^{(2)} x_3^{(2)} + \ldots\ldots + a_k^{(2)} x_k^{(2)} + u^{(2)},$$

……

$$y^{(m)} = a_1^{(m)} + a_2^{(m)} x_2^{(m)} + a_3^{(m)} x_3^{(m)} + \ldots\ldots + a_k^{(m)} x_k^{(m)} + u^{(m)}, \quad \ldots\ldots\ldots\ldots\ldots (1)$$

where, the superscripts denote the individual regressions, $n_1, n_2, \ldots, n_m$ are the nos. of observations for these regressions and the errors are heteroscedastic in the sense as formulated by the assumptions given under:

$$\text{i). } E(u^{(i)}) = 0_{nx1}, \quad \forall\, i = 1, 2, \ldots\ldots\ldots, m, \quad I$$

$$\text{ii). } E((u^{(i)})(u^{(i)})') = \sigma_i^2 I_{nxn}, \quad \forall\, i = 1, 2, \ldots\ldots\ldots, m, \quad I \quad \ldots\ldots\ldots\ldots\ldots(2)$$

$$\text{iii). } E((u^{(i)})(u^{(j)})') = 0_{nxn}, \quad \forall\, i \neq j = 1, 2, \ldots\ldots\ldots, m, \quad I$$

$$\text{where, } n_i = n \ , \ \forall\, i = 1, 2, \ldots\ldots, m,$$

(i.e., the sample sizes for the different regressions are the same, say, n).

It is admitted that a particular type of heteroscedasticity of the errors in the regressions in (1) has been adopted --- this is in the light of the model adopted in the Zellner's (1962) [8] SURE Estimation Procedure, and the solution here is, also, similar to that of Zellner's.

For the problem stated we proceed as follows. Firstly, for CCC between first and second regressions in (1), first and third regressions, …, first and m-th regressions, one requires to decide whether the differentials:

$$c_j^{12} = a_j^2 - a_j^1 < 0 \text{ or } = 0 \text{ or } > 0, \text{ for all } j = 1, 2, \ldots, k,$$

$$c_j^{13} = a_j^3 - a_j^1 < 0 \text{ or } = 0 \text{ or } > 0, \text{ for all } j = 1, 2, \ldots, k,$$

……

$$c_j^{1m} = a_j^m - a_j^1 < 0 \text{ or } = 0 \text{ or } > 0, \text{ for all } j = 1, 2, \ldots, k.$$

Similarly, for second and third regressions, second and fourth regressions, …, second and m-th regressions, one requires to decide whether the differentials:

$$c_j^{23} = a_j^3 - a_j^2 < 0 \text{ or } = 0 \text{ or } > 0, \text{ for all } j = 1, 2, \ldots, k,$$

$$c_j^{24} = a_j^4 - a_j^2 < 0 \text{ or } = 0 \text{ or } > 0, \text{ for all } j = 1, 2, \ldots, k,$$

……

$$c_j^{2m} = a_j^m - a_j^2 < 0 \text{ or } = 0 \text{ or } > 0, \text{ for all } j = 1, 2, \ldots, k.$$

ISSN: 2349-7637 (Online)

……

Lastly, for (m-1)-th and m-th regressions, one requires to decide whether the differentials:

$c_j^{m-1,m} = a_j^m - a_j^{m-1} < 0$ or $= 0$ or > 0, for all $j = 1, 2, \ldots, k$.

A moment's reflection shows that the desired CCC, i.e., CCC for all pairs of regressions out of the m regressions in (1), will be over when all the decisions enlisted just above, in (m-1) groups, so to say, are completed. Our task is to device tests for all these decisions.

3. The Methodology: As indicated at the end of the preceding Section, Methodology consists of (m-1) steps as ahead.

1). To build up tests in order to decide upon the differentials $c_j^{12}, c_j^{13}, \ldots, c_j^{1m}$, for all $j = 1, 2, \ldots, k$:

Firstly, we combine the m regressions in (1) into a single regression equation model as follows.

$$\begin{pmatrix} y_1^{(1)} \\ \vdots \\ y_n^{(1)} \\ y_1^{(2)} \\ \vdots \\ y_n^{(2)} \\ y_1^{(m)} \\ \vdots \\ y_n^{(m)} \end{pmatrix} = \begin{pmatrix} 1 & x_{21}^{(1)} & \cdots & x_{k1}^{(1)} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{2,n}^{(1)} & \cdots & x_{k,n}^{(1)} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & x_{21}^{(2)} & \cdots & x_{k1}^{(2)} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & x_{2,n}^{(2)} & \cdots & x_{k,n}^{(2)} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & x_{21}^{(m)} & \cdots & x_{k1}^{(m)} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & x_{2,n}^{(m)} & \cdots & x_{k,n}^{(m)} \end{pmatrix} \begin{pmatrix} a_1^{(1)} \\ \vdots \\ a_k^{(1)} \\ a_1^{(2)} \\ \vdots \\ a_k^{(2)} \\ a_1^{(m)} \\ \vdots \\ a_k^{(m)} \end{pmatrix} + \begin{pmatrix} u_1^{(1)} \\ \vdots \\ u_n^{(1)} \\ u_1^{(2)} \\ \vdots \\ u_n^{(2)} \\ u_1^{(m)} \\ \vdots \\ u_n^{(m)} \end{pmatrix} \quad \ldots(3)$$

The solution for the single equation model is same as that of finding solution separately for each equation in (1). The benefit of writing a single equation model is that the error terms in (1) are now constituents of the one error vector of this single equation model and hence we can now conceive the heteroscedasticity of these error terms easily, heteroscedasticity in the sense as formulated by the model (2). In addition to introducing heteroscedasticity of the error terms we also want to decide upon the differentials $c_j^{12}, c_j^{13}, \ldots, c_j^{1m}$, for all $j = 1, 2, \ldots, k$. For that we slightly change the model further as follows.

$$\begin{pmatrix} y_1^{(1)} \\ \vdots \\ y_n^{(1)} \\ y_1^{(2)} \\ \vdots \\ y_n^{(2)} \\ y_1^{(m)} \\ \vdots \\ y_n^{(m)} \end{pmatrix} = \begin{pmatrix} 1 & x_{21}^{(1)} & \cdots & x_{k1}^{(1)} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{2,n}^{(1)} & \cdots & x_{k,n}^{(1)} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 1 & x_{21}^{(2)} & \cdots & x_{k1}^{(2)} & 1 & x_{21}^{(2)} & \cdots & x_{k1}^{(2)} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{2,n}^{(2)} & \cdots & x_{k,n}^{(2)} & 1 & x_{2,n}^{(2)} & \cdots & x_{k,n}^{(2)} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{21}^{(m)} & \cdots & x_{k1}^{(m)} & 0 & 0 & \cdots & 0 & \cdots & 1 & x_{21}^{(m)} & \cdots & x_{k1}^{(m)} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{2,n}^{(m)} & \cdots & x_{k,n}^{(m)} & 0 & 0 & \cdots & 0 & \cdots & 1 & x_{2,n}^{(m)} & \cdots & x_{k,n}^{(m)} \end{pmatrix} \begin{pmatrix} a_1^{(1)} \\ \vdots \\ a_k^{(1)} \\ c_1^{12} \\ \vdots \\ c_k^{12} \\ c_1^{1m} \\ \vdots \\ c_k^{1m} \end{pmatrix} + \begin{pmatrix} u_1^{(1)} \\ \vdots \\ u_n^{(1)} \\ u_1^{(2)} \\ \vdots \\ u_n^{(2)} \\ u_1^{(m)} \\ \vdots \\ u_n^{(m)} \end{pmatrix} \quad \ldots(4)$$

Let us, for convenience, rewrite (4) as:

$$Y = X c + U, \qquad \ldots\ldots (5)$$

where, Y_{Nx1} = the Y-vector in (4), X_{NxK} = the X-matrix in (4), c_{Kx1} = the coefficient-vector in (4) and U_{Nx1} = the disturbance-vector in (4), N = nm and K = km.

We can now run regression with (5), estimate c and perform tests in order to decide upon the differentials c_j^{12}, c_j^{13}, ..., c_j^{1m}, for all j = 1, 2, …, k. But model (5) is a Generalised Least Squares Model (GLSM) [9]. The estimation procedure will depend on the variance- covariance matrix of the regression error, which is given as:

$$(D(U))_{NxN} = V, \text{ say, } = \begin{pmatrix} \sigma_1^2 I_{nxn} & 0_{nxn} & 0_{nxn} \\ 0_{nxn} & \sigma_2^2 I_{nxn} & 0_{nxn} \\ & \ddots & \\ 0_{nxn} & 0_{nxn} & \sigma_m^2 I_{nxn} \end{pmatrix},$$

or, $$V = \sum_{mxm} \otimes I_{nxn} \text{ , where,} \qquad \ldots\ldots (6)$$

$$\sum_{mxm} = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ & \ddots & \\ 0 & 0 & \sigma_m^2 \end{pmatrix}.$$

The GLS estimator of c based on (5) is given as:

$$c^* = (X'V^{-1}X)^{-1}X'V^{-1}Y, \qquad \ldots\ldots (7)$$

and its dispersion matrix is:

$$D(c^*) = (X'V^{-1}X)^{-1}. \qquad \ldots\ldots (8)$$

But due to (6), V is unknown as $\sum_{mxm}$ is so. So it is not possible to use (7) and (8) in practice, particularly for the testing purposes we are aimed at. Henceforth let us proceed following Zellner [8].

Firstly, we need to estimate V. For that we need to estimate $\sum_{mxm}$ and that is done as follows. The steps are:

i). Apply OLS separately to each of the regressions in (1); let the residual vector for the i-th regression be denoted as e^i, for all i = 1, 2, …, m,

ii). Estimate σ_i^2 as : $s_i^2 = (e^{i\prime}e^i) / (n-k)$, for all i = 1, 2, …, m.

Then, estimated $\sum_{mxm,}$ say, S_{mxm}, is:

$$S_{mxm} = \begin{pmatrix} s_1^2 & 0 & 0 \\ 0 & s_2^2 & 0 \\ & \ddots & \\ 0 & 0 & s_m^2 \end{pmatrix}.$$

Then, V is estimated as:

$$\hat{V} = S_{mxm} \otimes I_{nxn}. \qquad \ldots\ldots (9)$$

Now we replace V in (7) by $\hat{V}$ as given by (9) and form the estimator:

$$c^{**} = (X'(\hat{V})^{-1}X)^{-1}X'(\hat{V})^{-1}Y. \qquad \ldots\ldots (10)$$

Then it follows that $(n^{1/2})$ $(c^{**} - c)$ has asymptotic normal distribution and the dispersion matrix of c** is:

$$D(c^{**}) = (X'(\hat{V})^{-1}X)^{-1} + {}_0(n^{-1}),$$

where ${}_0(n^{-1})$ denotes terms of high order of smallness than n^{-1}.

So, for large value of n, c** is normally distributed. Also, evidently, for large n, ${}_0(n^{-1})$ is negligible and then,

$$D(c^{**}) \simeq (X'(\hat{V})^{-1}X)^{-1}.$$

ISSN: 2349-7637 (Online)

So, for large n, we have:

$$c^{**} \sim N_K(c_{Kx1} , (X'(\hat{V})^{-1}X)^{-1}). \quad \text{..... (11)}$$

Now, the tests that we require are obvious, provided that n is sufficiently large which we assume for the rest of the paper. Representing,

$$c^{**} = (c^{**}_1 \; c^{**}_2 \; c^{**}_3 \; \ldots\ldots \; c^{**}_K)' ,$$

$$c = (c_1 \quad c_2 \quad c_3 \quad \ldots\ldots \; c_K)' , \text{ and}$$

$$(X'(\hat{V})^{-1}X)^{-1} = (a_{ij})_{KxK},$$

we have: $(c^{**}_i - c_i) / (a_{ii})^{1/2} \sim N(0 , 1)$, for all $i = 1, 2, 3, \ldots\ldots , K$.

Hence for the null hypothesis: $H_0 : c_i = 0$,

the test statistic is: $T = c^{**}_i / (a_{ii})^{1/2}$, and ….. (12)

$T \sim N(0 , 1)$, under H_0 , for all $i = 1, 2, 3, \ldots\ldots , K$.

This completes tests for CCC for (m-1) pairs of regressions: first and second, first and third, ..., first and m-th. Needless to say that each of the tests here is a normal test.

2). To build up tests in order to decide upon the differentials $c_j^{23}, c_j^{24}, \ldots, c_j^{2m}$, for all j = 1, 2, ..., k:

The entire procedure here is exactly similar as in the step 1). but with the last (m-1) regressions in (1) instead of all the m regressions in (1).

This completes tests for CCC for (m-1) pairs of regressions: second and third, second and fourth, ..., second and m-th.

……

m-1). Lastly, to build up tests in order to decide upon the differentials $c_j^{m-1,m}$, for all j = 1, 2, …, k.

Here also the entire procedure is exactly similar as in the step 1). but with only the last two Regressions in (1) ((m-1)th and mth).

This completes tests for CCC for one and the last pair of regressions: (m-1)th and mth.
Thus, tests for CCC for all possible pairs of regressions out of the m given regressions in (1) is complete.

Having done this one can partition the set of the given regressions into, say, clusters, *every cluster consisting of those regressions which satisfy the condition that the vectors of coefficients of any two of these regressions do not differ from each other significantly*. In order to sort out the clusters easily we first present the outcomes of all the above comparisons neatly by introducing a matrix, let it be called *Indicator Matrix* (IM), as:

$$\mathbf{i}_{mxm} = (r_{ij}) , \quad \text{..... (13)}$$

where, r_{ij} indicates whether the ith and the jth regressions in (1) differ from each other significantly or not and is defined as follows:

$r_{ij} = 0$, when $i = j$,

0, when $i \neq j$ and the two regressions concerned do not differ from each another significantly,

1, when $i \neq j$ and the two regressions concerned differ from each another significantly,

ISSN: 2349-7637 (Online)

where, i, j = 1, 2, ..., m[2].

It is this IM which will clearly show the clusters: number of clusters as well as the regressions in each of these clusters. Needless to say that IM is a symmetric matrix of order mxm with all the diagonal elements as zeros and it will be a null matrix (mxm) iff all the m regressions coincide.

4. Illustration: In the context of rate of growth of population in India, we consider three regression equations (m = 3) as follows. With state level population of India, we first define the following four variables:

X_1 = size of the population in a state of India in 1981,

X_2 = size of the population in a state of India in 1991,

X_3 = size of the population in a state of India in 2001,

X_4 = size of the population in a state of India in 2011,

the sources of these data being Census of India (1981, 1991, 2001, 2011) [10].

Let us now define variables Y_1, Y_2, Y_3 as follows:

$Y_1 = X_2 - X_1$ (i.e., growth/increase of population during: 1981 to 1991)

$Y_2 = X_3 - X_2$ (i.e., growth/increase of population during: 1991 to 2001)

$Y_3 = X_4 - X_3$ (i.e., growth/increase of population during: 2001 to 2011).

We now consider three regressions as follows:

$$\left.\begin{aligned} Y_1 &= \beta_1 X_1 + U_1 \\ Y_2 &= \beta_2 X_2 + U_2 \\ Y_3 &= \beta_3 X_3 + U_3 \end{aligned}\right] \quad \ldots (14)$$

β_1, β_2 and β_3 are nothing but the rates of growth of population over the decades: 1981 to 1991, 1991 to 2001 and 2001 to 2011 respectively (to be referred as first decade, second decade and so on).

Now, for CCC of all possible pairs out of the three regressions in (14), following the last Section we have here two steps ((m-1) steps with m = 3) as follows.

1). First consider tests for CCC of two pairs of regressions: first and second, first and third.

Now, following the last Section, we apply OLS separately to each of the above three regressions in (14). (It may be noted that each of these regressions is a regression without an intercept term.) (The no. of observations for each regression here is n = 32(no. of States in India); so, we have here: n = 32, k = 1, m = 3.)

The Residual vectors of these three regressions are first obtained. Then we get the sum of squares for the residual vectors and hence the estimates of σ_i^2's (s_i^2's) using the formula as given in the previous section. We then use the following steps to get the value of c** in (10) and D(c**) in (11), the value of c** representing as: first component gives the estimate of the growth rate in the first decade, and the second and the third components give respectively the estimates of changes in the growth rates over first decade to second decade and over first decade to third one.

1. Construct the matrix S_{3x3} and compute $(S_{3x3})^{-1}$.

2. Compute $(\hat{V})^{-1}_{96x96} = (S_{3x3})^{-1} \otimes I_{32x32}$.

3. Compute c** as: $c^{**} = (X'(\hat{V})^{-1} X)^{-1} X'(\hat{V})^{-1} Y$ and its dispersion matrix, D(c**), as:

[2] Of course, IM provides limited knowledge that for every two regressions, whether they coincide or not, nothing more.

$D(c^{**}) = (X'(\hat{V})^{-1} X)^{-1}$.

The estimate of growth rate in the first decade is 2.408 and the estimates of the changes concerned are respectively -2.270 and -2.227 with the corresponding T-values, given by (12), as 132.029, -86.159 and -83.461 respectively. Compared with table-values, it may be concluded that the second T-value indicates that there is a decline in growth rate as one moves from the first decade to the second decade and the third T-value also indicates in the similar way, i.e., there is a decline in growth rate as one moves from the first decade to the third one[3]. Hence it may be concluded that the first regression differs significantly from both of second and third regressions.

2). Next and lastly we consider test for CCC of one pair of regressions: second and third.

We consider now the last two regressions in (14) and proceed exactly in the similar way as in the step 1). with these two regressions only. The results come out as follows.

The estimate of growth rate in the second decade is 0.138 and the estimate of the change concerned, i.e., the estimate of the change in growth rate as one moves from second decade to the third one, is 0.043 with the corresponding T-values, given by (12), as 6.952 and 1.362 respectively. Compared with table-value, it may be concluded that the second T-value indicates that the change is insignificant. Hence it may be concluded that the second regression does not differ from the third one significantly!

This evidently completes tests for CCC for all possible pairs of regressions out of the three given regressions in (14).

Now, in order to get the clusters in the present context, following the discussions above, we first construct the IM defined by (13) which comes out to be as follows:

$$\mathbf{i}_{3x3} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (15)$$

Hence, representing the decades concerned as D_1, D_2, D_3 respectively, we get here two clusters as follows:

$C_1 = \{D_1\}$ and

$C_2 = \{D_2, D_3\}$.

So, it may be concluded that in respect of the relation of size of population to growth of population, the behaviour of the first decade do not match with that of the other decades while the other decades behave in the same manner.

5. Conclusions: The above test procedure enables one to perform CCC for the two regressions in each and every pair of regressions possible out of several given regressions.

Needless to say, this generalises the Chow test in two directions. This, further, has the following important implications.

Suppose the regressions are now arranged successively in a definite order for investigating existence of structural change. Then, obviously, if this procedure is carried out, CCC between every two successive regressions out

[3] All tests throughout this Illustration is carried out at 5% level of significance.

ISSN: 2349-7637 (Online)

of the given regressions, arranged now successively, is automatically done and hence detailed informations regarding all structural changes are obtained, if there is any such at all [4].

Again, once this procedure is performed, i.e., CCC between the two regressions of all possible pairs of regressions out of the several given regressions are done, one can partition the set of the given regressions into, say, clusters, *every cluster consisting of those regressions which satisfy the condition that, as already stated, the vectors of coefficients of any two of these regressions do not differ from each other significantly.* An example has already been illustrated.

It is to be noted that number of clusters need not always be two; it may be more than two as well as be one when and only when all the regressions coincide, i.e., the Indicator Matrix becomes a null matrix of an appropriate order.

It may be noted that the clustering introduced here is quite different from that which is obtained by using Mahalanobis D^2 Statistic [4] or using k-means Method [5] --- clustering by means of each of the later approaches is based on a single variable/a vector of variables while that by means of our procedure is based entirely on *relationship of variables*.

It is a hunch that the procedure introduced here can be extended for the purpose of clustering on the basis of not only one relationship of variables but more than one relationship together!

REFERENCES

1. Chow, Gregory C. (1960): Tests of Equality between Sets of Coefficients in Two Linear Regressions, Econometrica, Vol. 28, No. 3, pp. 591-605.
2. Toyoda, Toshihisa (1974): Use of the Chow Test under Heteroscedasticity, Econometrica, Vo. 42, No. 3, May, pp. 601-608.
3. Schmidt, P and Sickles, R (1977): Some Further Evidence on the Use of the Chow Test under Heteroscedasticity, Econometrica, Vol 45, pp. 1293-1298.
4. Ali, Mukhtar M and Silver J. Lew (1985): Tests for Equality between Sets of Coefficients in Two Linear Regressions under Heteroscedasticity, J. of the Am. Stast. Assoc., Vol. 80, No. 391, pp. 730-734.
5. Tansel, Aysit (1987): Testing for Structural Change under Heteroscedasticity: A Note and an Application, Commun. Fac Sci., Univ. Ank., Ser. A, V. 36, No. 2, pp. 55-67.
6. Saha, J. & Pal, M (2014): A Modified Chow Test Approach Towards Testing Differences in The Engel Elasticities, Asian-African Journal of Economics and Econometrics, Vol.14, No.1, 2014; 57-67.
7. Saha, J (2018): Test for Structural Change under Heteroscedastic Errors: The Case of Successive Regressions, Research Hub – International Multidisciplinary Research Journal, Volume-5, Issue-01, January 2018.
8. Zellner, Arnolds (1962): An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias, Journal of the American Statistical Association, Vol.57, 1962, pp.348-368.
9. Johnston J. (1984): Econometric Methods, Third Edn., McGraw-Hill, New York.
10. Census of India (1981, 1991, 2001, 2011): Government of India.
11. Mahalanobis P.C. (1936): On the generalised distance in statistics, *Proceedings of the National Institute of Sciences of India 2 (1): 49–55*. Retrieved 2012-05-03.
12. R in Action, Second Edition, Manning publishing house.

[4] It is to be noted that if one requires only CCC between every two successive regressions out of given several successive regressions and nothing more, it is not necessary to carry out the procedure described here; for that purpose it is sufficient to work out only the procedure laid for that purpose in the paper by Saha and Paul (2014) referred earlier [6].