MODELING CROSS SECTION DATA CONTAINS EQUALITY CONSTRAINTS WITH APPLICATION

نوزاد محمد احمد*

0- ABSTRUCT:

sometimes we are agree to identify a suitable regression model , for a such kind of data . But the behaviors of data , restrict us to use some procedure to fitting those models , specially if data consists some undesirable behaviors , like some constraints among explanatory variables , and nonlinearity .

So the main problems in fitting models , is multicolinearity and also serial auto – correlations among the serial generated residuals When model was fitted , Which they makes the fitted model insignificant .

In general the one of mor applicable models for alinear and independent cross-section data , is multiple linear regression , (ordinary least square estimation , OLS) , to estimate models parameters .

The OLS method is not appropriate for data contains multicolinearity that arise with respect to the constraints , exists on the data .

There are several approach of estimation to treat this condition , in this survey , the researcher used , (restricted least square method , RLS) , to fit model with (equality constraints) in the data under consideration.

0-1 Sample survey :

Three random variables taken , as a cross – section data , specified as follow :

Y : Production of several kinds of clothes , measured with (New Iraqi Dinar ,ID).

X1: The labour cost \ production unit , named by (hours) , with (ID).

مدرس / جامعة السليمانية / كلية الادارة والاقتصاد / قسم الاحصاء*

مقبول النشر بتاريخ 2005/12/1

X2: The capital cost \ production unit , named by (capital), with (ID).
Data, taken monthly for each variable , begin from (01 -01 – 2003) to (31-12 – 2004) ,with (24)observations.

0-2 SURVEY ASSUMPTION :

Since the data are not linearly work then the (OLS) method is not an appro-perate to fit the model , so the researcher used the mathematical transformations (Natural Logarithm) , to achieve linearity for variables .
Rather than non –linearity the existence of multicolinearity , made the researcher use (RLS) , in addition to (OLS) , and comparing these two models after fitting them , by the efficiency of the parameters in each method .

0-3 THE OBJECTIVE:

The fist goal is to show that , for data under covsideration the RLS , estimation procedure is more efficient , comparing with ,OLS estimation , for data contains equality constraints .
The second , is to fit a mor applicable model to define the behaviors of these random variables . and the last is to use the predicted model to evaluate changes in production due to any small changes in explanatory variables , having minimum mean square of residuals generated with the best fitted model .

Table (1) :

Data shown in table (1),illustrate the natural logarithm for datas .
(logarithm transformation to achieve linearity)

	In (Product)	In (labour cost/hour)x1	In (capital) x2
1	2.411818	-3.43594	5.58024
2	2.61387	-2.79132	9.60583
3	1.05284	-2.48216	8.5655
4	0.44446	-2.44854	8.91801
5	3.27799	-4.87069	7.09506
6	2.46055	-3.36865	10.37051
7	2.68695	-3.6367	10.15868
8	0.99783	-1.89842	10.66812
9	3.35017	-1.95606	9.93309
10	2.91034	-1.61485	10.21852
11	2.98611	-4.16603	5.39022
12	2.56777	-3.65951	5.59157
13	3.27298	-4.51688	5.64457
14	0.059349	-2.12369	4.17045
15	1.34256	-2.69044	4.54897
16	2.99052	-4.04532	6.73953
17	3.42933	-4.64912	5.89393
18	2.2482	-3.49611	3.57198
19	1.04602	-2.54413	4.81501
20	1.83873	-3.161	5.03331
21	2.94671	-4.16581	5.98897
22	2.89316	-4.06482	5.67651
23	2.93152	-3.89741	5.20147
24	3.56874	-4.25487	4.56451

Data reference : Sulaimani manufacturing clothes factory . (2003 – 2004).

1- (OLS) procedure estimation :

let the production model is as follow :

Pi = (Hi) ^ B1 * (Ci) ^ B2 * ei i : 1,2,3,4……,24 …………… (1-a)

P: Cost production / unit .
H: Labour cost / unit .
C: Capital cost / unit .
Ei: Residuals .

Ln (pi) = B1 * Ln (Hi) + B2 * Ln (ci) + Ln (ei) ………………….(1-b)
Let (2-a) reduce to:
Yi = B1 * X_{i1} +B2 * X_{i2} + ai ……………………………………….(1 – c)
Ln (pi) = Yi , Ln (Hi) = X_{i1} , Ln(ei) = ai
ai : Normal (0 , σ^2_a) , n= 24 obs .

The sums and cross products , Which evaluated from table (1) , are :

$$\underset{\sim}{F} = (x' * X) = \begin{bmatrix} n & \sum X1 & \sum X2 \\ \sum X1 & \sum X1^2 & \sum X1 * X2 \\ \sum X2 & \sum X2 * X1 & \sum X2^2 \end{bmatrix}$$

$$\underset{\sim}{F} = \begin{bmatrix} 24 & -79.9383 & 163.9457 \\ -79.9388 & 286.8228 & -524.2456 \\ 163.9457 & -524.2456 & 1242 \end{bmatrix}$$

LET $\underset{\sim}{f}$ **= inverse (F)**

$$x' * Y = \begin{bmatrix} \sum Y \\ \sum X1 * Y \\ \sum X2 * Y \end{bmatrix} = \underset{-}{g}$$

$$\underset{-}{g} = \begin{bmatrix} 56.32.84 \\ -201.8363 \\ 384.8646 \end{bmatrix}$$

If $\underline{b}_{(ols)}$ is to be a vector of ordinary least square estimators , then by OLS estimation method :

$$b_{(ols)} = \underset{\sim}{f} * \underset{-}{g} = \begin{bmatrix} -1.5447 \\ -0.8540 \\ 0.1533 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \quad \text{.........(1-e-1)}$$

Then the suggested model with (OLS) method is as follow :

Yi = -1.5447 - 0.8540 * Xi1 + 0.1533 * Xi2 ....(1-e-2)

To test the hypothesis :

H_0 : suggested model , (1-e-2) , is not significant .

Verses H_1 : suggested model, (1-e-2) , is significan.

ANOVA

Variation sources	D.F	Sum squares	Mean sum squares	Fc
Regression sum square	2	12.1488	6.07440	11.64
Residuals sum square	21	10.9556	0.52169 = S^2a	----------
Total (corrected)	23	23.1044	------------------	-----------

Comparing Fc = 11.64 , with ,F(tabular), d.f = (2,21), and level of significant α = 0.05
Indicated that the model (1-e-2)is weakly significant , such that the correlation of Determination (R^2)which is given by :

$$R^2 = \frac{\text{Reg.ss}}{\text{Total.ss}} * 100 = 53\ \% \quad \text{................................(1- f)}$$

In other hand 47% of the variations are explained by residuals , this statistically said that the model is inadequate . In addition to inexistent of serial autocorrelations among serial residuals *(Dorbin Watson statistic

greater than 1.4) ,there exist a high linear relation ship among the explanatory variables ,(see the variance – covariance , and correlation matrix for (OLS) , estimators , that indicated strongly the existence of multico – linearity .

$$v \underset{\sim}{\hat{-}} c\, \mathbf{b}_{(ols)} = S^2 a(ols) * \underset{\sim}{f} = 0.52169 * \underset{\sim}{f}$$

$$v \underset{\sim}{\hat{-}} c\, \mathbf{b}_{(ols)} = \begin{bmatrix} 0.8691 & 0.1424 & -0.0546 \\ 0.1424 & 0.0313 & -0.0056 \\ -0.0546 & -0.0056 & 0.0053 \end{bmatrix} \qquad \text{.............(1-g1)}$$

$$\underset{(ols)}{\text{correlation matrix}} = \begin{bmatrix} 1.0000 & -0.8095 & 0.6417 \\ -0.8095 & 1.0000 & 0.6417 \\ 0.6417 & -0.1373 & 1.0000 \end{bmatrix} \qquad \text{........(1-g2)}$$

2-0 : Restricted least square estimation . (RLS) :

From the behaviors of datas , and the nature of the production process , in this factory there are some linear constraints in it .

Let the matrix ($\underset{|}{R}$) , represented the matrix of constraints ,formulated due to the following hypothesis:

$$\begin{aligned} B0 &= G \\ B1 + B2 &= 1 \\ 3B1 - B2 &= 0 \end{aligned} \qquad \text{..(2-a-1)}$$

(*) See graph (4) : the estimated autocorrelation coefficients shows that , the serial residuals appears , they are randomly distributed , because they are inter the normality boundaries , such that the random residuals , distributed normally with zero mean , and variance (1/n).

Practically , if ρ (k) is an autocorrelation function , and (k) is lag times (k:1,2,3..,..,..,max lag) then the upper and lower boundaries of randomness is given by :

$$-1.96 * (1/n \wedge 0.5) < \hat{\rho}(k) < 1.96 * (1/ n \wedge 0.5)$$

for all k : 1,2,3,4,…,…,…,..k (k : maximum lag time)

it is very necessary to say that formulating these constraints depends on the behaviors of data fist , and second , the experience of researcher .(*)

From the constraints , (2-a-1).

$$R=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 3 & -1 \end{bmatrix} \quad \text{……………………(2-a-2)}$$

Let , (r) represented the right hand side for the set of constraints ,(2-a-1) :

$$r=\begin{bmatrix} G \\ 1 \\ 0 \end{bmatrix}$$ **, usually , G, is replaced by , bo (ols)**

such that :

$$r = R * b\ (ols) \quad \text{……………………………..(2-b)}$$

the (RLS) , estimators are given by :

$$b_{(RLS)} = b_{(oLS)} + f * R * \text{inverse } [R*f*R] * (r - R*b_{(ols)})$$
……………………. (2-c)

$b_{(RLS)}$: the vector of estimators, with restricted least square method . (**)
The variance covariance matrix for (RLS), estimaters is given by :

$$V - C\ b_{(RLS)} = \sigma^2{}_a * \left[f - f * R * \text{inverse } \{R * F * R\} * R*f \right] \quad \text{…………..(2-d)}$$

This matrix can be estimated as:

$$V - C\ b_{(RLS)} = S^2_a\ (RLS) * \left[f - f * R * \text{inverse } \{R*f*R\}*R*f \right] \text{………(2-e)}$$

(*) The researcher used more than one constraint matrix in order to reach to the optimum (R). Please see the conclusions .
(**) The (RLS) estimators are unbiased , and minimum variances , comparing with (OLS) estimators , for more mathematical details see , references,(1,6).

Such that :

$$S^2_{a\,(RLS)} = \frac{Y * Y - b_{(RLS)} * X * Y}{n - m - 1}$$ **m=2 , (no. of explanatory variables)……...(2-f)**

To estimate the efficiency of the (RLS) parameters , with respect to , (OLS) 's , use the following formula .
eff (RLS) J denoted the efficiency of (RLS) , parameter (j), j : 0,1,2

$$eff(RLS)\ j = \frac{var\ bj_{(RLS)}J}{var\ bj_{(ols)}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2\text{-}g)$$

Or it can be calculated from :
$$eff(RLS)J = I_{K+1} - R * inverse\ (R * f * R) * R * f \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2\text{-}h)$$

$$f = inverse(F) = \begin{bmatrix} 1.6669 & 0.2731 & -0.1048 \\ 0.2731 & 0.0600 & -0.0107 \\ -0.1048 & -0.0107 & 0.0101 \end{bmatrix}$$

Using eq (2-c), we can estimate $b_{(RLS)}$ vector of estimators as follow :

$$b_{(RLS)} = \begin{bmatrix} -1.5447 \\ 0.2500 \\ 0.7500 \end{bmatrix} \quad , b_{(RLS)} * (X * Y) = 151.01789 \text{ (including } nY^2)$$

Using eq (2-f) , to estimate sample variance is :

$S^2_{a\,(RLS)} = 0.19664$
Also using eq (2-e) to calculate estimated var – cov matrix of (RLS) estimaters as :

$$V - C\ (RLS) = \begin{bmatrix} 0.3278 & 0.0537 & -0.0206 \\ 0.0535 & 0.0118 & -0.0021 \\ -0.0206 & -0.0021 & 0.0020 \end{bmatrix} \qquad \ldots\ldots\ldots\ldots(2\text{-}i)$$

2-1 Comparison :

From the two estimated (V – C) matrix in both , (OLS) , and (RLS) , the following table can be displayed :

Table (2)

$b_{(ols)}$ = -1.5447 -0.8540 0.1533	$B_{(RLS)}$ = -1.5447 0.2500 0.7500
Estimated variance $b0_{(OLS)}$ = 0.8691	Estimated variance $b0_{(RLS)}$ = 0.3278
Estimated variance $b1_{(OLS)}$ = 0.0313	Estimated variance $b1_{(RLS)}$ = 0.0118
Estimated variance $b2_{(OLS)}$ = 0.0053	Estimated variance $b2_{(RLS)}$ = 0.0020

For all estimators the variance (RLS) estimators , is smaller than (OLS), moreover the covariance (bi , bj) , i = j also smaller than (OLS) . These , all indicates that model fitted by (RLS), for this data is more suitable than (OLS) estimation method .

To calculate the efficiency of the (RLS) ,estimators , use eq (2-g):

$$\begin{aligned} &\text{Eff } b0\,(RLS) = 0.377 < 1 \\ &\text{Eff } b1(RLS) = 0.376 < 1 \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2-j) \\ &\text{Eff } b1(RLS) = 0.378 < 1 \end{aligned}$$

These efficiency coefficients showed that , the model fitted with (RLS) , is more applicable for illustrating the behaviors of the production process , having raw materials (labor , and – capital) . so the best regression model for this process is :

Ln (cost product / unit) = -1.5447 – 0.2500 Ln (labor cost / unit) + 0.15533 Ln (capital cost / unit) ………………………(2-k)

2 -2 Analysis of variance for model (2 –k):

the test of this model come from testing the following hypotheses:

H0 : the model (2-k) is not significant .
Verses H1 : the model (2-k) is significant .

Variations sources	D.F	Sum squares	Mean sum squares	FC
Regression sum square	2	18.9748	9.4874	48.257
Residuals sum square	21	4.1294	0.1966= S^2a(RLS)	-------
Total (corrected)	23	23.1044	---------------------	-------

The coefficient of determination (R= 82.12%).
Comparing Fc > Ft = 3.47 , (2,21)d.f, and level of significant = 0.05 , indicated that the model under test , is significant . the value of Durbin Watson statistic = (1.6) , greater than (1.4), is a good evidence for non – existing a serial autocorrelation in residuals having generated with model (2-k) . and finally in addition to these tests, the covariances between (bi , bj), (i= j), , in (RLS) , is small as it is in (OLS) , that is also a good evident that (RLS) estimates made the model excluded multicolinearity problem .

3-1 conclusions :

during the analysis of several sides in this study , the researcher concluded some important results that illustrated as follow .

1/ Parito test , of normality arise , that all variables under consideration , are normally distributed , because more than 70% of frequencies falling into normality curve .

2/ the ordinary least square estimation method , and fitting regression model , gave only 53% of total variations explained by model (1-e-2) , but restricted estimation method fitted the model (2-k) , that explained , 87% of these variations .

3/ Restricted least square estimation method (RLS) , treated and removed the multico-linearity , among explanatory variables . this can be seen clearly by increasing coeff- cient of determination to (87%) , and weakly covariance among estimaters , see(2-i), moreover the efficiency for estimators , in the model (2-k), where are all (less than 1) (see table 2).

4/ the residuals generated by applying model (1-e-2) , having no serial autocorrelation and appears weakly randomness due to the normality range of autocorre- lation , given by the following 95% confidance interval:

$$- 1.96 * (1/n \wedge 0.5) < \quad < 1.96 * (1/n \wedge 0.5)$$

k : Lags time .

n : number of observations(24) . box – pierce statistic value (*) (Q1)=18.944$_{(24\ D.F)}$

But the randomness of autocorrelation for residuals for model (2-k) , is more stronger than the model (1-e-2) , because of having no pattern , and smaller coefficient values , with same previous normality range . (see graphs 4,5) , and compare).

BOX – Pierce statistic value (Q2) = 9.6504 $_{(24\ D.F)}$

5/ To arrive to an appropriate and optimum results , and an adequate model , the researcher tried with more than one case for the matrix of restrictions , as follow:

CASE (1):

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$

implies sum square regression = 4.5037 (including nY^2)

CASE (2):

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 2 & -1 \end{bmatrix}$$

implies sum square regression = 102.30 (including nY^2)

CASE (3):

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 3.5 & -1 \end{bmatrix}$$

implies sum square regression = 173.02 (including nY^2)

The conditions (1,2)didn't reach the optimal R , and case (3) , made the sum square of Regression , exceed the total sum square (correct), Which is not allowed then we concluded that slop of variable (X1), doesn't exceed three times absolute slop of variable (X2).

(*) Box – Pierce statistic given by :
Q = (n) * $\sum$ $_{(k)}$. k: 1,2,3,…..,……k (maximum lag)
Calculate (Q) for each models , and compare them , with (x^2) tabular value with , (n) d.f and (=0.05) , level of significant . to make decision about non – significant of residuals estimated autocorrelation coefficients , to test the hypothesis :
H0 : =0 verses H1:

6/ The model (2-k) , is suitable to use for predicting the response (production) , corresponding to any changes occurs in explanatory variables (labor , and capital).

3-2 Recommendations :

1/ From the conclusions in section (3-1) , when try to fit these models , with these kinds of datas , one must be very carefully treat it , and take care of the relations between variables , specially when one deals with these non – linear relationships . also it is very important to note the strategy of the companies about their main goals , if these goals is a tools for economic development , and growth economy , or not . because strategies are effectible directly to specify the constraints matrix , which is very useful to be near of the process nature , to make the fitted models proper .

2/ Model (2-k) , can be used as a process to control the amount of production , if the company or factory has information about market , and national demand's volume on his productions .

References:

1/ Wannacatt , T.H. , & Wannacatt R . J . (1990) , Introductory Statistics For Business & Economics . New York : Wiley .

2/ Ramanathan , R. (1992) . Introduction to Econometrics with Applications ., 2nd ed. Fort Worth , The Dryden Press .

3/ Ramanathan , R. (1993). Statistical Method In Econometric . SanDiego : Academic Press .

4/ Eengel , R . F .(1995) . Selected Readings . Advanced Texts In Econometrics . Oxford , And New York : Oxford University Press .

5/ Spyros M. Steven , C . W . , & Rob , J . H .(1998) . Forecasting : Methods and Applications . 3 rd , ed . Jhon Wiley & Sons , New York .

6/ Professor , Dr , Amory , H . K. & Bassim , S, M, (2002) .Advance Econometrics , (Theory and Application) . Al – Taif Press , Baghdad .(Arabic).