The Economic Journal of Nepal
Vol. II, No. 2. April 1979 EIC-TU

Assumptions in Regression Analysis: When Violated.

– Govinda P. Koirala★

ASSUMPTIONS IN REGRESSION ANALYSIS

Very often we turn to problems with multivariate distribution, a distribution that contains two or more variables. In other words, two or more than two observations are made on a single trial. The central problem of regression analysis involving a multivariate distribution is to determine the true relationship among the variables. But the problem is that we do not know the true relationship and have to depend upon the sample observations to estimate the true relationship. The estimation may also turn out to be wrong, because, there is always a chance to deviate the estimated relation away from the true relation in almost all the social sciences. Even if we perform trials under essentially similar conditions, there is always a chance that the outcomes of two trials are different and hence the formulated relations among the variables will deviate from the true relations. If we have sufficient theoritical background about the true relationship we can be very close to it and can predict about one of the variable (dependent variable) with the knowledge of other variable (independent variable) and this prediction can be very near to the true value. But what is a true value ? Can we always find it ? Of course, not. We cannot observe true value. The error may be measurement error or any other error due to chance. And no one can claim that he will observe the true value. Can you say what your true weight is ? Is it 65 kg ? 65.42 kg ? or 65.4205321 kg. ? or what ? Can you say how much time you work in a day ? Is it 7 hours ? 6.56 hours ? 6.63245 hours ? 7.2153 hours ? By this arguments, I don't mean that you can never observe accurate value. Cann't you answer how many children do you have ?

★Mr. Koirala is a member, Statistics Instruction Committee, T. U. Kirtipur Multiple Compus

Cann't you say how much money do you earn in a day ? Of course, yes. In such cases also, there may be some chance that you can forget some of your children (may be you have some illegal children), you may misreport about your exact income when you are asked (you may have black money and you do not want to disclose it, or to show your false prestige you may give false information). Bht in the long run when some information is collected from many individuals in the aggregate, misreporting error may cancel themselveꞩ some being upward biased error (positive error) and some being downward biased error (negative error). In regression analysis this type of assumption is thought to be valid. That is, the assumption that the expected value of the error term is zero, is valid.

Suppose the true relation among the variables X, Y, and Z is given by linear model,

$$Y= \propto +\beta X+\delta Z \quad .. . (1)$$

Then the 1st observation will be

$$Y_i= \propto +\beta X_i+\delta Z_i+U_i \quad ... ..(2)$$

Here U_i (error term or disturbance term) is included because of the reason explained above. If there were no error term, three different sets of obeservations would be sufficient to find out $\propto$, β and δ. But the problem is not only that U_i's are unknown but also that the distribution of U_i is unknown; how the value of U_i changes with the change in observations in the variables X, Y, and Z. But since X's and Z's are the independent variables, they are generally taken as the fixed set of values and no error term for these variables, where as Y is the dependent variable which depends on the values of X and Z. Error will be there corresponding to the observed values of Y.[1] We try to find out the values of $\propto$, β and δ such that error will be minimum. The purpose will be solved by ordinary least square (OLS) method. We can never find out the exact values of $\propto$, β and δ. We get only their estimated $\hat{\propto}$ $\hat{\beta}$ and $\hat{\delta}$. These estimates will tend to be true estimates only if the following assumptions about the error term are valid. The assumptions are:[2]

1. U_i's are normally distributed with mean zero and constant variance $\sigma^2 u$.
2. U_i and independent variables are not correlated.

[1] Of course this may also be unrealistic assumption that X and Z variables will be measured correctly and only Y variable cannot be measured correctly. But this is generally not questioned and is accepted that X and Z are generated as fix set of values

[2] Damodar; Gujrati Basic Econometrics M. Grew Hill book company 1978; p. 167.

3. There is no multicolinearity among the independent variables.
4. There is no autocorrelation, i.e. $E(U_iU_j)=O;\ i\neq j$.

Linearity Assumption : When Violated

Above mentioned assumptions are the major assumptions when anyone tries to estimate the values of α, β and δ. These assumptions about the error term are made after assuming that the variables X, Y, and Z are related by tbe relation (1). But how do we know that the true relation is of the type (1) ? Cann't it be of any other type ? What is that, that can tell us the true relation can be very much approximated by the relation of the type (1) ? Most people may think that R^2, the coefficient of determination is the best measure of goodness of fit. Higher the value of R^2, better will be the fit. This may be true but not in all cases. For the model to be fit, the value of R^2 should be high enough. But it should be noted that "even though $R2$ is used as a measure of the proportion of variation in the dependent variable that is explained by the regression equation it should not always be interpreted as a determinant of goodness of fit to the casual relation".[3] The high value of $R2$ may also indicate that there is the presence of multicolinearity in the explanatory variables (i.e. independent variables).

In most of the cases the linearity assumption of the model is violated. But even then we use linear model. If there is no sound theoritical basis to assume linearity, why do we use linear model ? The simplest answer to this problem is that the linear model is very simple to deal with. And it is also to be noted that for a certain region (a small interval of observations), but not outside the region, there is not much deviation in the linear and non–linear models as can be seen in the following model.

Suppose we are trying to find out the income elasticity of excise tax in Nepalese economy with respect to the GDP originating from manufacturing and cottage industries of Nepal.

[3] Rao and Miller; Applied Econometrics, Prentice-Hall of India Pvt. Ltd,. New Delhi 1972 p. 14.

Table-I

(Rupees in millon)

Year	Excise tax revenue[A]	Total tax revenue[A]	GDP origination from manufacturing and cottage industries[B]
	(Et)	(Tt)	(Gt)
1964/65	13.88	150.84	448
1965/66	20.06	177.02	577
1966/67	19.96	225.78	533
1967/68	21.48	283.86	625
1968/69	28.04	368.26	748
1969/70	38.12	411.29	787
1970/71	56.57	395.62	818
1971/72	63.60	466.80	996
1972/73	67.80	521.10	971
1973/74	77 40	642.40	1282
1974/75	119.70	843.70	1618
1975/76	132.00	1101.80	1698
1976/77	166.10	1333.60	1797
1977/78			

Source : A. Budget speeches, Ministry of Finance / Nepal.

B. National Planning Commission Secreteriat; and Central Bureau of Statistics / Nepal.

An effort has been made firstly to use the linear model to find the income elasticity of excise tax. The linear model is[4]–

$$E_t = \propto + \beta G_t + U_t \quad \dots \quad ..(3)$$

Here β may be expressed as the amount of change in E_t for a unit change in G_t, since the income elasticity of excise tax for any particular year t may be defined as:

$$\mathcal{E}_t = \frac{\%\text{ Change in } E_t}{\%\text{ Change in } G_t}$$

$$= \frac{\text{Change in } E_t}{E_t} \Big/ \frac{\text{Change in } G_t}{G_t}$$

$$= \frac{\text{Change in } E_t}{\text{Change in } G_t} \times \frac{G_t}{E_t}$$

$$= \beta \times \frac{G_t}{E_t}$$

Estimating (3) by ordinary least square method, we get

$$E_t = -- 39.1518 + 0{\cdot}1034 \; G_t \qquad R^2 = 0{\cdot}9618$$
$$\qquad\qquad\overset{@}{(4{\cdot}9074)} \quad \overset{@}{(14{\cdot}0450)}$$

Here since t–values are significant and R^2 is high, the model is fit for the data. Yearly income elasticity of excise tax may be presented as below.

Table– II

Year	Elasticity	Year	Elasticity
1964/65	3.3246	1971/72	1.6131
1965/66	2.9628	1972/73	1.4752
1966/67	2.7505	1973/74	1.7061
1967/68	2.9971	1974/75	1,3923
1968/69	2.7477	1975/76	1.3250
1969/70	2.1265	1967/77	1.1144
1970/71	1.4894		

[4] The general convention is that the suffix 't' is used for time series data and suffix 'i' for cross sectional data.

@ The values given inside paranthesis are t–values

But it is to be noted that in this model (3) we have assumed that the change in excise tax is constant in amount for the same change in income (Gt) for all the level of Gt. For every 10 million increase in income (Gt) excise tax (Et) increases by 1.03 million. This may not be as realistic as to suppose that excise tax (Et) increases proportionately with income (Gt). If this is the truth, the relation between excise tax and income (proxied by GDP originating from manufacturing and cottage industries) may be approximated by,

$$E_t = a\, G_t^{\,b} V_t \quad \ldots \;..\; (4)$$

This relation is non-linear relation and when it is graphed, the curve so obtained will be a non-linear curve. Changing the model (4) into log-linear form we have

$$\text{Log } E_t = \text{Log } a + b \text{ Log } G_t + \text{Log } V_t$$

$$\text{or, } Y_t = A + B X_t + U_t \quad \ldots \ldots (5)$$

$$\text{where } Y_t = \text{Log } E_t$$

$$A = \text{Log } a$$

$$B = b$$

$$X_t = \text{Log } G_t$$

$$U_t = \text{Log } V_t$$

The non-linear model (4) can be changed into linear model (5) by simply taking the logarithm values of Et and Gt. And we can estimate A and B as $\hat{A}$ and $\hat{B}$ by OLS technique. Estimating (5) by applying OLS technique to the logarithms of Et and Gt we get,

$$\text{Log } E_t = --2.7313 + 1.4922 \text{ Log } G_t \qquad R^2 = 0.9886$$

$$(3.2030)^{@} (5.1805)^{@}$$

Here also t–values are significant and R is high. The model is fit for the data. Comparing the two models we cannot say which model is the best fit. If we want the model to discuss with constant elasticity the second model can be studied. Also if we are interested in model with high R^2 second model is preferred. But in case if we are interested with the changing elasticities to study first model is to be taken. To estimate the values of excise tax for different values of income the reliability of the assumption about the elasticity shows the path which model is to be taken. For a small range, however, the two models yield approximately the same estimates.

Table-III

Estimated values of excise tax (Et)

Rear	Value of Gt	Linear estimate of Et	Non-linear estimate of Et.
		Et=—39.1515+0.1030Gt	Et=0.00.1857Gt 1.4922
1964/65	448	6.9941	16.78
1965/66	577	20.2815	24.48
1966/67	533	15.7494	21.75
1967/68	625	25.2257	27.58
1968/69	748	37.8951	36.07
1969/70	787	41.9122	38.91
1970/71	818	45.1053	41.22
1971/72	996	63.4400	55.30
1972/73	971	60.8649	53.23
1973/74	1282	92.8989	80.67
1974/75	1618	127.5581	114.00
1975/76	1698	135.7484	122.60
1976/77	1797	145.9457	133.50

It is generally very easy to deal with the linear model. But we cannot assume always the linear model, particularly when we have to predict the dependent variable outside certain region. In such cases when the true model is non-linear, prediction from linear model will be very much unrealistic. But if we can somehow theorize the nature of the relationship or by drawing the scattered diagram (if you have only two variables) the nature of non-linearity can be known. We can convert (in most of the cases but not always) the non-linear form into linear form (by taking the logarithms of values, or reciprocals of the values etc). But to accept any non-linear form there should be theoritical validity of the relationship.

Homoscedasticity Assumption : When Violated

In estimating the parameters of the model (1) some assumptions about the error term are made. One of the assumptions was that of the constant variance of the error term which is called the homoscedasticity assumption. However, this may not be very realistic assumption especially in cross section or micro-economic analysis. It may happen that the errors are mutually uncorrelated (that is assumption (4) may be valied) but at the same time have different variance. For example, “low income families have less variaton in consumption than for high income families. The variance of savings among high income families may be larger than the variance of savings among low income families”.[5] Thus we can say that the problem of heteroscedasticity arises when variance changes with independent variable.

But what happens to the estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\varrho}$ if we try to estimate them by OLS technique in the presence of heteroscedasticity in the error term? Will they be unbiased estimates ? Will they be efficient estimates ? In other words, will they be still best linear unbiased estimates ?

When the variances of the error terms vary, i. e. the errors are not homoscedastic, then the least square coefficients are still unbiased but are not efficient estimators. These estimators are also asymptotically inefficient and test of significance and confidence limits do not apply.

To be sure one has to test for homoscedasticity. In general, test for homoscedastity is made for the cross sectional data. Sometimes time series date are also presented for the homoscedasticity test. In estimating the relation between the excise tax and income from industries (income proxied by GDP originating from manufacturing and cottage industries), we can first detect whether or not heteroscedasticity of error term is present. Here test of homoscedasticity for small samples (Quandt–Goldfeld method) is tried. In this method observations are first ordered according to the size of that variable (X's) which is suspected to have different variance for different values of this variable. Some (c) of the middle observations are dropped. Two separate groups are then observed and separate regression lines by ordinary least square technique to the first (n–c)/2 observations and to the last (n–c)/2 are obtained provided of course that (n–c)/2 exceeds K, the number of parameters.to be estimated. Under the homoscedastic asumption the ratio of residual sum of squares $\left(\frac{S_2}{S_1}\right)$, 1 denoting that from smaller X^i values and 2 that from larger X^i values will be distributed as F—distribution with (n–c–2k)/2, (n–c–2k)/2 degrees of freedom.[6]

[5] Parasar Singh; Econometrics and Mathematical Economics, S. Chand & Company Ltd. Ramnagar New Delhi 110055; 19

[6] J. Johnston, Econometric Methods (2nd Ed.) McGraw-Hill Kogakusha Ltd. Tokyo, 1972 p. 219

The Quandt--Goldfeld test can be applied to the excise tax example used previously. The data are divided into two groups; the first including those with income values less than Rs 800 million and the second including higher income values than Rs 900 million. One middle observation is dropped. The output associated with the two separate re gression equations is as follows:

1· Low income group

$$E_t = \underset{(1.92)}{--14.5222} + \underset{(5.12)}{0.0615}\, G_t \qquad R^2=0.8868$$

Error sum of squares, $S_1 = 40.1929$

II. High income group

$$E_t = \underset{(1.96)}{--47.5776} + \underset{(6.44)}{0.1091}\, G_t \qquad R^2=0.9127$$

Error sum of squares, $S_2 = 745.4490$

The F–statistic used to test the homoscedasticity assumption is $S_2/S_1 = 18.58$. Under the homoscedasticity assumption this will be distributed as F with (4,4) degrees of freedom. Examination of the table of the F–distribution shows that the critical value of F at the 5 percent level of significance is 6.39. We conclude that we can reject the null hypothesis in favour of the alternative hypothesis of heteroscedasticity.

Instead of the actual observations we can transform them into logarithms and test for the homoscedasticity, the two separate regression equations are as follows.

I. Low income group

$$\text{Log } E_t = \underset{(4.94)}{--3.0028} + \underset{(7.17)}{1.5638}\, \text{Log } G_t \qquad R^2=0.9278$$

Error sum of squares, $S_1 = 0.0080$

II. Highinconme group

$$\text{Log } E_t = \underset{(3.99)}{--2.4037} + \underset{(7.29)}{1.4034}\, \text{Log } G \qquad R^2=0.9301$$

Error sum of squares, $S_2 = 0.01035$

The F–statistic to test the homoscedasticity assumption is $\frac{S_2}{S_1} = 1.29$ and hence we conclude in favour of the hypothesis of homoscedasticity. Thus by transforming the values of excise tax (Et) and income Gt into the logarithmic form we come to the conclusion that the problem of heteroscedasticity has been reduced. This technique cannot be advisable always unless we have enough theoritical arguement to use the logarithms of the variables.

Independency Assumption : When Violated

Another assumption in estimating the parameters of the model (1) is that the errors are independent with the values of the variables under study. "When error (e) is correlated with X (independent variable) the OLS estimator is no longer consistent. To be concrete, suppose e and X are positively correlated. Then positive values of e tend to be associated with positive values of the deviation $X - \bar{X}$. Consequently, the OLS fit will have too large a slope. This bias in $\hat{\beta}$ will persist even with a very large sample".[7]

The independency assumption is very commonly violated in the cases when we try to estimate the paramaters from a single equation where the truth is the system of simultaneous equation. Consider the Keynesian model, where consumption is related with income by

$$C_t = \alpha + \beta Y_t + U_t \quad \dots \quad \dots \quad \dots \quad (6)$$

The model (6) is a very common linear model. Suppose we are trying to estimate the marginal propensity to consume (β). Simple OLS technique can be applied to estimate β. But because we have here ignored the fact there is another identity to be true[8] i.e.

$$Y_t = C_t + I_t \quad \dots \quad \dots \quad \dots \quad (7)$$

I_t being the investment, the estimate β will be not only bias but also inconsistent. It can be very easily shown that U_t and Y_t are correlated i.e. $E\ (Y_t U_t) \neq 0$. To get the consistent estimate $\hat{\beta}$ any of the instrumental variable technique, indirect least square technique, two

[7] Wonnacott and Wonnacott; Econometrics, John Wiley & Sons Inc. 1970 pp. 152-153.

[8] In this case there is no disturbance term because relation (7) is not an equation but an identity.

stage least square technique may be applied.[9] All these techniques will yie'l the identical results.

Multicollinearity Assumption: When Violated

The multicollinearity assumption is associated in the multiple regression model and the assumption is that there exists no exact linear relationship between the independent variables in the model. If such linear relationship does exist it is said that the independent variables are perfectly collinear or that perfect collinearity exists. With perfect collinearity we cannot calculate the least square parameter estimates.

Consider the following example[10]

$$S_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + U_t \quad \ldots\ldots (8)$$

Where S is sales revenue, X_1 is the number of left shoes sold, X_2 is the number of right shoes sold, and X_3 and X_4 are other products. The revenue comes from selling both the right and the left shoes, therefore each has legal claim in explaining the movements in sales revenue, but then some of the parameters in equation (8) have no meaningful interpretation. The parameter $\beta_{1,}$ for example, is the partial derivative of S with respect to left shoes keeping all the variables including right shoes constant. That is, β_1 is the change in sales revenue for unit change in the number of left shoes sold (X1) when all the other variables are kept constant. Such a situation is never observed because shoes are always sold in pairs. Also note that the estimates of β_1 and β_2 can never be obtained and even if they were somehow obtained, they could not be interpreted.

Whatever the source of the fixed relation the problem can be changed by redefining the variables in such a way as to make the parameters subject to interpretation. For example, in the above case instead of specifying the independent variables as left aud right shoes separately, a new variable defined as "a pair of shoes" may be used.

[9] For detail see Johnston: Econometric Methods, McGraw Hill Kogakusha Ltd. Tokyo.

[10] Rao and Miller; op. cit; foot note 3; p. 47.

In multiple regression, when a regression equation has several parameters to be estimated, the ordinary least square technique for $\hat{\beta}$ is usually expressed in matrix notation as

$$\hat{\beta} = (X' \ X)^{-1} X' \ Y \quad .. \quad .. \ (9)$$

When a linear relation exists between independent variables then the matrix (X/X) is singular and has no inverse. But since in most practical situations exact linetarity among the X variables may not be seen due to measurement error or due to chance factor, then $(X/X)^{-1}$ may be calculated and hence may by found out. Generally in the presence of multicollinearity in the variables F–test may be significant but the t–values are individually insignificant. "Collinearity is otter suspected when R^2 is high and when zero order correlations (i. e. simple correlation efficients) are also high but none or very few of the partial regression coefficients are individually statistically significant on the basis of the conventional t–test. Although high zero order correlations may suggest collinearity, it is not necessary that they be high to have collinearity in any specific case".[11] If there are only two explanatory variables the presence of high simple correlation coefficient will guide to the presence of multicollinearity. But in the models involving more than two explanatory variables the simple correlation will not provide the infallible guide to the presence of nulticollinearity. As a result it is suggested that one should look not only on the simple correlation but also at the partial correlation coefficients. If R^2 is high but the partial correlations are low, multicollinearity is a possibility. Although a study of the partial correlation may be useful, there is no guarantee that they will provide unfalling guide to multicollinearity; for, it may happen that both R^2 and all the partial correlations are sufficiently high.

Because multicollinearity is dependent directly upon the sample of observations, little can be done to resolve it unless more information abouc the process in question is available. When multicollinearity is suspected, the easiest way to tell whether multicollinearity is causing problems is to examine the standard errors of the coefficients.[11] If several coefficients have high standard errors, and the dropping of one or more variables from equation lowers the standard errors of the remaining variables, multicollinearity will usually be the source of the problem.[12] When multicollinearity is present, the exclusion of the variables from the regression equation does not decrease the explanation of the dependent variable. The entire influence of the excluded

[11] Damodar; Gujrati, op. cit. foot note 2

[12] Pindyck RS and Rubin Feld; DL; Econometric models and Economic forecasts, McGraw Hill Kogakusha Ltd. 1976

variable will be captured by the included variable (after detecting that these two excluded and included variables are collinear) and the other coefficients will be unaffected. If we are interested in explaining the movement of the dependent vriable or in predicting the values of dependent variable, then it makes no difference whether the variable (excluded) is in the regression or not. The higher the value of R^2 the better will be the prediction provided that the collinearity among the variables will also continue in future. When our objective is to estimate the coefficients of the other explanatory variables then again exclusion of variable (excluded) will not damage the estimate. "The problem of which variables are to be included in a regression equation is a major problem in applied econometrics. Rules are helpful, but they cannot make decisions for the applied econometrician.[13]

"In passing note that multicollinearity, as we have just discussed it, excludes only linear relationship among the X variables. It does not rule out non-linear relationship among them. For example, consider the following regression model:

$$Y_i = \beta_o + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + U_i \quad \dots \dots (10)$$

where, say, Y = total cost of production and X output. The variables X_i^2 (output squared) and Xi^3 (output cubed) are obviously related to Xi but the relationship is non–linear therefore mndels such as (10) do not violate the assumption of no multicollinearity. As a matter of fact, to depict the U—shaped average and marginal cost curves of economic theory model (10) is very appropriate."[14]

Autocorrelation Assumption : When Violated

The assumption that errors corresponding to different observations are uncorrelated often breaks down in time-series studies, but can be a problem in some cross section work as well. When the error from different (usually adjacent) time periods (or cross section observations) are correlated we say that the error term is autocorrelated or serially correlated. Autocorrelation occurs in time-series, when the errors associated with observations in a given time period carry over into future time periods.

[13] Rao and Miller op. cit. foot note 3. p. 52.

[14] Damodar, Gujrati, op. cit. foot note 2. p. 173.

Just as, when homoscedasticity assumption is violated, the estimates remain unbiased but inefficient, similarly when autocorrelation assumption is violated the estimates will still be unbiased but they will be inefficient estimaters and tests of significance will not be valid. "Thus as in the case of heteroscedasticity the t–tests of the significance of coefficients and F–tests of the significance of the entire regression will in general be invalid. It is therefore important to test for autocorrelation and if found present to correct for-it."[15] The conventional Durbin-Watson 'd'–statistic may be calculated to test the presence of autocorrelation.

The most important type of autocorrelation generally assumed is the first–order–linear autocorrelation scheme in which case the error at any time period t is obtained to be correlated with the error at the previous time t–1. Thus, for a simple model.

$$Y_t = \alpha + \beta X_t + U_t \quad \ldots\ldots\ldots \quad (11)$$

$$U_t = \rho U_{t-1} + V_t \text{ for all t and } |\rho| < 1 \ \ldots\ldots\ldots (12)$$

Here ρ is an unknown parameter and Vt is a residual stochastic disturbance term, which is assumed to satisfy the assumption of the basic linear regression model, including absence of serial correlation. Then by lagging the equation one period and subtracting from the original equation. We obtain.

$$Y_t - \rho Y_{t-1} = \alpha(1-P) + \beta(X_t - \rho X_{t-1}) + U_t - \rho U_{t-1}$$

$$\text{or, } Y_t^* = \alpha^* + \beta X_t^* + V_t \quad \ldots \quad (13)$$

$$\text{where, } Y_t^* = Y_t - \rho Y_{t-1}$$

$$X_t^* = X_t - \rho X_{t-1}$$

$$\alpha^* = \alpha(1-\rho)$$

We can perform the regression on these transformed variables. But this can be done only when ρ is known. However, ρ is unknown. An obvious way of estimating ρ is to use the residuals from the equation estimated by OLS. The suggested method of estimating ρ is by[16]

[15] Intriligator, Michall D; Econometric Models, Techniques and Applications, Prentice Hall, Inc. Eaglewood cliefs. 1978

[16] M. J. C., Surrey, An Introduction to Econometries, Clarendon Press Oxford, 1974. p. 57.

$$\hat{\rho} = \frac{\Sigma\rho_t \; \rho_{t-1}}{\Sigma\rho_t^{\;2}} \quad \ldots\ldots \quad (14)$$

$$\text{where,} \quad \rho_t = Y_t - \hat{Y}_t$$

It is to be noted that $\hat{\rho}$ is a biased estimator of ρ .17 One should be very careful in using $\hat{\rho}$ for ρ if $\hat{\rho}$ is very different from ρ. The OLS estimates from the transformed variables may even less efficient than those yielded by OLS estimation from the untransformed variables.

Even if there is no serial correlation in the error terms one might suspect of this. The use of Durbin–Watson test statistic may lead to this conclusion, that there is a serial autocorrelation in the disturbance term when calculated a simple linear regression, say, of total tax revenue of Nepal (T) on the time in years (N) from 1964/65 to 1977/78 [data used here are taken from table I.]

The linear model is

$$Tt = 25.0874 + 82.2280 \; Nt \qquad R^2 = 0.9042$$

Durbin–Watson Statistic, d=0.4093.

Here, since $d < d_l$ (tabulated value of DW Statistic), we have to reject the null hypothesis of no autocorrelation (i.e. $H_0 : \rho = 0$) in favour or of positive first order autocorrelation.

But here the presence of autocorrelation has appeared, not because there is actually its presence but because we have approximated the true model by linear where the true model is somewhere near to the non-linear one. The simple non-linear regression of total tax revenue of Nepal (T) on the time in years (N) from 1964/65 to 1977/78 is obtained to be

$$Tt = (163.2)\,(1.172).^{Nt} \qquad R^2 = 0.9836.$$

DW Statistic d=1.6008.

In this case, $d > d_u$ and we accept the null hypothesis that there is no autocorrlation.

"If first order serial correlation is diagnosed using the Durbin–Watson test, there are two possible treatments, each involving a change in the specification of the original model. One

[17] M. J. C. Surrey, ibid p. 58

treatment is to incl de other explanatory variables. To some extent the stochastic disturbance term may be representing the actions of these variables, which, if explicitly accounted for, would reduce the serial correlation. The other treatment, which is used, if such additional variables are not readily available involves not the original model but the transformed model (Equation 13). To do so, however, requires an estimate of the unknown parameter."[18]

Conclusion

Unlike scientists in their pure field of knowledge regarding the experiments economists generally have little control over their data. More often than not, economists depend on secondary data, that is data collected by someone else, such as governments and other organizations. As long as this is understood the theory can only be legitimately applied in parctie. But the theory can only be the guide, many difficulties will arise and one has to solve then and there. To quote Rao and Miller once again "no statistical tool or economic guide is a good substitute for theory. Guidelines indicate where to look in case of trouble but not necessarily how te solve the problem"[19]

[18] M. D. Intriligatcr; op. cit. foot note 13.

[19] Rao and Miller; op. cit. foot note 3.