STABILITY OF LINEAR REGRESSION MODELS

S. Mustafa[1], K. Riaz[2], Q. Perveen[3]
[1,2]Department of Mathematics and Statistics, PMAS, Arid Agriculture University Rawalpindi
1. saimamustafa28@gmail.com, 2. bestamongbetter@yahoo.com
[3]Department of Education, PMAS, Arid Agriculture University Rawalpindi
3. qaisarach@yahoo.com

ABSTRACT: *Linear model is an integral part in every field and stability of these models relates to the various parameters involved in the model. The objective of this research is to investigate the stability of linear models. The estimated values of the parameters of linear models have been obtained through ordinary least squares (OLS) method and their stability is analyzed by Recursive test. We have investigated the stability of linear models based on the structural changes when the number of observations is small. The point where structural change occurs is a key point. So when prior knowledge, where the change occur is not available then recursive least square and recursive residual test have been used.*

Keywords: Linear models, Ordinary least square Method, Recursive Test

INTRODUCTION:

The roots of general linear model have the origin of mathematical thought. The general linear model is the result of emergence of theory of algebraic invariant in 1800. The theory of algebraic invariants developed more ideas about linear models in 19th century. The innovation of linear model has been done for economic relationships. It has also been developed for better understanding of science and technology. In statistics, generalized linear regression is generalizing form of linear models. In generalized linear regression, the linear model to be related to the response variable via link function and magnitude of the variance of each measurement to be a function of its predicted value. These linear models were introduced by John Nelder and Robert Wedderburn, [1].

The linear regression model should be linear in parameter is $Y = \beta_0 + \beta_1 X + \varepsilon_i$, where β_0, β_1 are unknown estimated parameters and ε_i is the error term. Here error term should be normally distributed with mean zero and variance σ^2. We generally used the method of least square to estimate the parameter of linear models. The method of ordinary least square is widely used for regression analysis because it is much simpler than maximum likelihood parameters which are estimated by ordinary least square (OLS) gives Best Linear Unbiased Estimators (BLUE). This method is fairly simple and attractive as compared to other different econometric techniques. A system or model would be stable when it remains close to its actual or original value. Recursive least squares (RELS) test is applied to check data is stable or not. Andrews [2] investigated that stability is an estimator's property and defined the effect of single observation in the sample on the realized value of estimator. He determined and compared stability exponents for a wide variety of estimators and econometric models and also find its dependence on maximal moments of estimator's influence curve. Lindsey [3], focused on generalized linear models which are reasonably well known with the exception of logistic, log-linear, and some survival models. At the same time, he analyzed the generalized linear modeling methodology is used in powerful methods, involving wider classes of distributions, non-linear regression, censoring and dependence among responses, are required.

MATERIAL AND METHODS

Ordinary least square (OLS) technique has been applied to get estimated parameters. To check stability of linear regression model Recursive least square test has been applied. In linear model stability of model depend on its parameters. The basic ideas behind Recursive least square (RELS) is illustrated by an example. Suppose the data is available during the year of 1977 to 1990. Firstly data for 1977 to 1979 has been used, estimated the model and obtained the parameters ξ_0 and ξ_1 then same data has been used for 1977 to 1980 and re-estimated the model. By adding a data point on X and Y, the process has been done until the entire sample is exhausted. Every time regression gives a new estimate of ξ_0 and ξ_1 and the estimated values of these parameters have been plotted against each iteration. The model is considered to be stable if the change in the estimated value is small otherwise not. One of the main goals of study is to check the structural change of our data. This structural change may be due to the different external forces for example policy changes, stock exchange stock prices and various different causes. To check weather related linear data is stable or not, recursive test is applied. The point where structural change occurs is a key point. It is very sensitive to the selection of time where parameter of model changes. So when prior knowledge, where the change occur is not available then recursive least square and recursive residual test is used.

DATA ANALYSIS

The data set for checking stability of linear models is taken from Introduction to the Practice of Statistics by Moore and McCabe [4]. The population consists of the number of manatees killed by boats in Florida and powerboat registrations (in thousands) in the years 1977 to 1990.

Y= number of manatees killed by boats.

X= powerboat registrations (in thousands).

Another data set is on the telephone cable manufacturer for the period 1968 to 1983 is tabulated in Gujarati [5, p.290]. The relationship of annual sales of telephone cable with gross national product, housing starts, and unemployment rate and customer line gains has been checked.

Y= annual sales in million paired feet.

X_2= gross national product (GNP), $, billion.

X_3= housing start, thousands of unit, %
X_4= unemployment rate, %
X_6= customer line gains, %

PRESENTATION AND ANALYSIS OF DATA

Most common assumptions for OLS regression analysis have been checked. For this purpose several tests have been performed. The normality of our data is check by using histogram for response variable (number of manatees killed by boats). In fig.1 standardize residual of response variable (Manatees Killed) are normally distributed. Normality assumption is also checked for multiple linear model by using histogram for regressand variable (annual sales of telephone cable) in fig.2, where standardize residual of predictor variable (annual sales of telephone cable) are normally distributed.

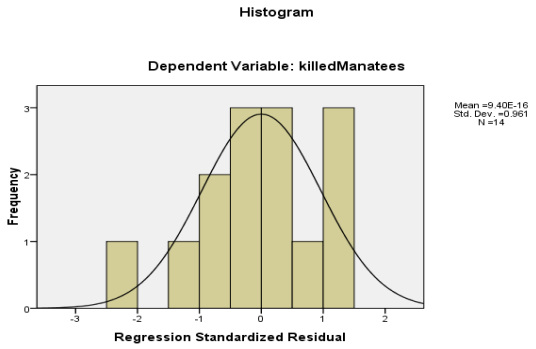


Fig. 1

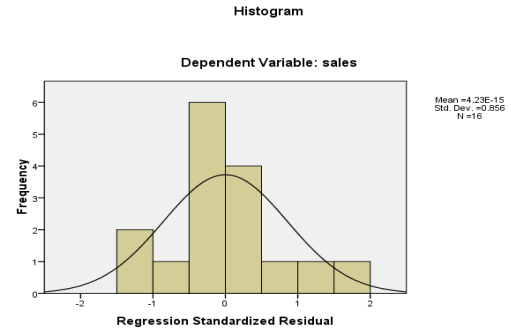


Fig 2

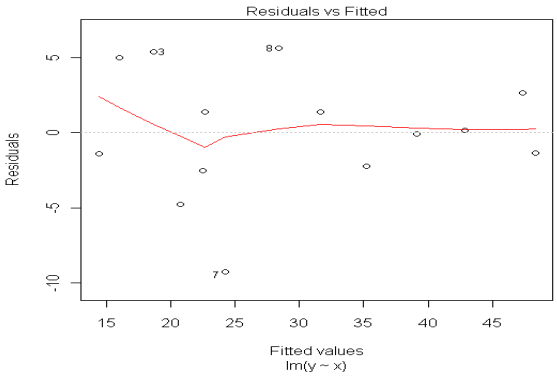


Fig 3 Plot of residual vs fitted for killed manatees

This study is intended as a time series data to check the relationship between powerboats and Manatees Killed and then checked stability of linear model. Same procedure is done for multiple linear models where relationship between gross national product, housing starts, and unemployment rate and customer line gains annual and sales of telephone cable has been checked.

Table 1: OLS Regression Summary for the number of manatees killed by boats

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-41.4304	7.4122	-5.589	0.000118 ***
Powerboat registrations	0.1249	0.13	9.675	<5.11e-07***

On 12 degrees of freedom, Residual standard error: 4.276
Multiple R-squared: 0.8864, Adjusted R-squared: 0.8769
F-statistic: 93.61 on 1 and 12 DF, p-value: < 5.109e-07.

The data analyses of the particular study consist of an Ordinary Least Squares (OLS) regression technique, diagnostic tests of several types have also been performed to test the data before doing the OLS regression. Plots of residual against fitted value show a megaphone pattern, which helps to conclude that the homogeneity assumption is fulfilled. Histogram and normal Q-Q plot are used to test the normality and linearity.

The results of the OLS regression have been summarized in Table 1 for the number of manatees killed by boats with one independent variable. Here R-square is 0.8864 which interprets that only 89% of the variation is described by the independent variable. It tells that there is a strong relationship between numbers of killed or injured manatees and power boat registration.

In each point of linear model gives an estimate. The value of ξ_1 (=0.1249) which is a slope of linear model show that within sample range of boats between 447 and 719 per boat, as no of boat registration increases on the average number of manatees killed by boats increases amount to about 0.1249.

Table 2: OLS Regression Summary for the annual sales of telephone cable

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	6031.9195	2250.9072	2.680	0.021417 *
Gross National Product(GNP)	5.0524	1.3637	3.705	0.003472 **
Housing Start	2.3092	0.4879	4.733	0.000616 ***
Unemployment Rate	-824.3777	168.2115	-4.901	0.000471 ***
Customer Line Gains	-864.4400	233.2971	-3.705	0.003469 **

Residual standard error: 598.6 on 11 degrees of freedom
Multiple R-squared: 0.8226,, Adjusted R-squared: 0.7581
F-statistic: 12.75 on 4 and 11 DF, p-value: < 0.0004084.

The results of the OLS regression for the annual sales of telephone cable have been summarized in Table 2 with four explanatory variables. Here R-square is 0.8226 which interprets that only 82% of the variation is described by the independent variable. The intercept value of about 6032 interpreted meaningfully because when variable like gross national product (GNP), housing start, unemployment rate and customer line gains, have practically zero values, annual sales of telephone cable will be very high, which make a particular sense.

The p-values of two of the variables (housing start, unemployment rate) are less than 0.001 alpha levels with positive coefficients of housing start and negative coefficient unemployment rate. So it indicates that housing start is positively related to the annual sales of telephone cable and unemployment is negatively related to annual sales of telephone cable as unemployment decreases annual sale increases. The p-values of the other two variables (gross national product (GNP), customer line gains) are less than 0.01 alpha levels. It conclude that gross national product (GNP) is positively related to annual sales of telephone cable ($b=5.0524$, $p<0.01$) and customer line gains is negatively related to annual sales of telephone cable ($b=-864.44$, $p<0.01$).

Stability test for linear model of number of manatees killed by boats

Recursive least square test has been used for checking parameter stability of linear model for manatees killed. In fig.3 shows recursive coefficients estimates where estimated values of parameters for response variable (manatees killed) and predictor variable (boats) are plot against each iteration. C (1), C (2) are just notations which is used in e views by default. These are estimated parameters used in fig. 4. They represent the notation of ξ_0 and ξ_1 used in linear regression model .

Plots of recursive coefficient test for intercept term of killed manatees

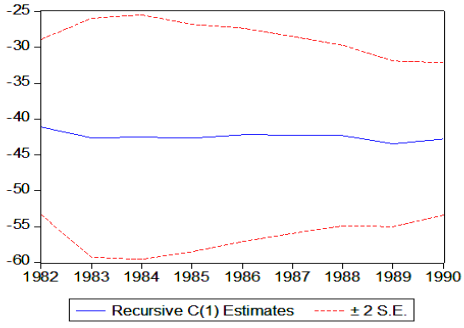


Fig 4

C (1) is intercept term obtain by regressing the data firstly for 1977 to 1982. Then again same data has been used for 1977 to 1983 and re-estimate the killed manatees model. Same procedure has been done until entire sample space has been exhausted. Every time when regression has been applied on response variable the value of C (1) is obtained with its standard error which is shown in figure 4. As C(1) do not show a dramatic change so it show the data is stable at every point.

Plots of recursive coefficient test for slope of killed manatees

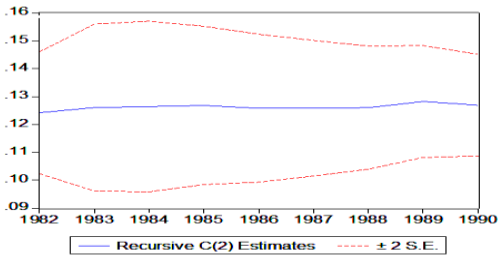
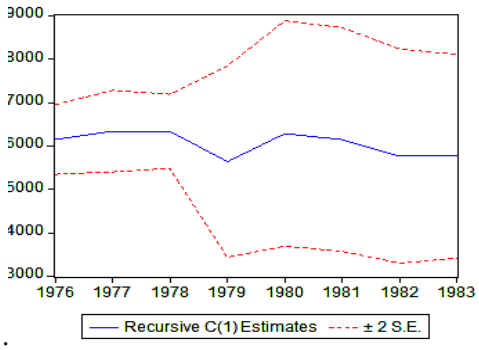


Fig 5

In fig.5 shows C (2) slope of linear model i.e. (manatees killed) is estimated at each iteration. Also two standard error bands are shown around the estimated coefficients C(2) i.e. (slope of linear model). figure for estimated coefficient C (2) do not shows significant variation or dramatic jumps which show data is stable each and every point. It means relation between number of killed manatees and boat is stable during the year 1977 to 1990.

Stability test for linear model of annual sales

Recursive least square test has been also applied for checking parameter stability for annual sales of telephone cable. In fig. 6 to fig. 9 shows recursive coefficients estimates where estimated values of parameters for regressand variable i.e. annual sales of telephone cable and explanatory variables i.e. gross national product (GNP), housing start, unemployment rate, and customer line gains are plot against each iteration. C (1), C (2) , C (3), C(4), C(6) are also notations which is used in e views by default. These are estimated parameters used in fig. 5 to fig. 9. They represent the notation of ξ_0 ,ξ_2,ξ_3,ξ_4 and ξ_6 used in multiple linear models. C (1) is intercept term obtain by regressing the data firstly for 1968 to 1976. Then again same data has been used for 1968 to 1977 and re-estimate the annual sales income model. In fig. 6 Recursive coefficient C (1) has been shown, which is intercepts of model represents stability of data. The graph shows the stability analysis during the year 1976-1983. During the interval 1976-1978 there is a slight change in stability, while the year 1978-1980 a significant change occurs, after 1980-1983 the data is looking more stable than before



.

Fig 6

Same procedure has been done until entire sample space has been exhausted. Similarly C (2), C (3), C(4), C(6) are coefficient of gross national product, housing start, unemployment rate, and customer line gains of linear model for annual sales of telephone cable, is estimated at each iteration for next year. Also two standard error band are shown around the estimated coefficients C(1) i.e. (intercept term for response variable) and C (2), C (3), C(4), C(6) i.e. (slopes of linear model). In fig. 7, C (2) represents recursive coefficient of GNP of model shows stability up to 1978. The graph shows the stability during the year 1976-1983. During the interval 1976-1978 there is a slight change in stability, while the year 1978-1980 a significant change occurs, after 1980-1983 the data is looking more stable than before. Every time when regression has been applied on response variable the value of C (1), C (2), C (3), C(4), C(6) is obtained with its standard error which is shown in figure 7.

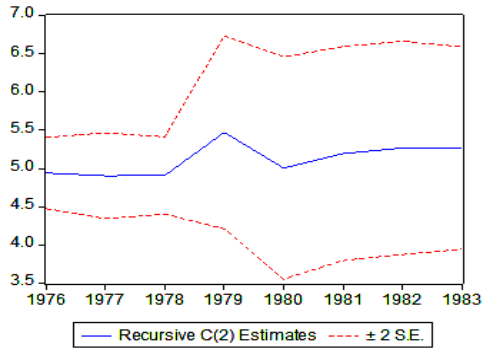


Fig 7

In fig.8, C (3) represents recursive coefficient of housing start of model. The graph shows the stability during the year 1976-1983. During the interval 1976-1978 there is a slight change in stability, while the year 1979-1980 a significant change occurs, after 1980-1983 the data is looking more stable than before.

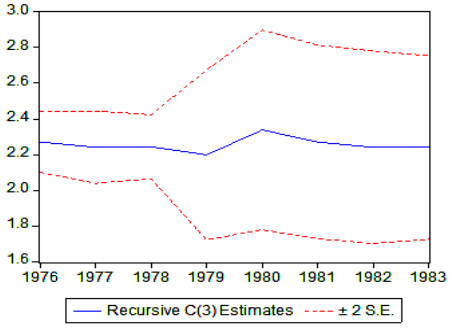


Fig 8

In fig. 9, C(4) represents recursive coefficient of unemployment rate of model. The graph shows the stability during the year 1976-1983. During the interval 1976-1983 there are very slight changes in stability which mean unemployment rate remains stable.

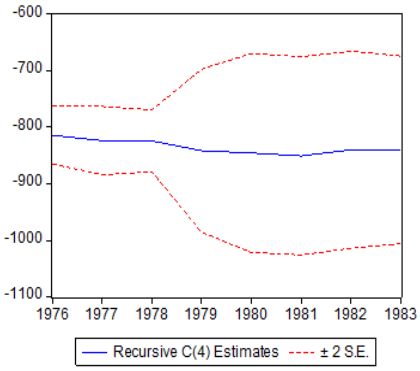


Fig 9

In fig.10, C(6) represents recursive coefficient of customer line gains of model .The graph shows that during the year 1978-1980 the dramatic jump is observed at 1979. After 1980 the behavior is less stable as compared to the behavior before 1978.

It is concluded that all recursive coefficient show drastic behavior during the year 1978 to 1980 except C (4) which exhibits the stable behavior during the given years.

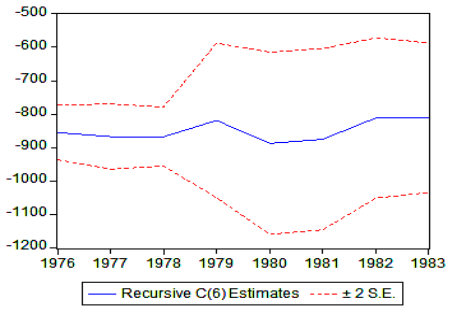


Fig 10

CONCLUSION

We have concluded that the stability of linear models based on the structural changes when the number of observations is small. The point where structural change in time series data occurs is a basic point. Recursive least square and recursive residual has provided better information when prior knowledge, regarding change occurs is not given.

REFERENCES

1 Nelder, J. A. & Wedderburn, R. W. M. Generalized linear models. J. Royal Sta. Soc., Ser. A., 135: 370-384, 1972.

2 Andrews, D. W. K. Stability comparisons of estimators. J. Econometrica., 54(1): 1207-1235, 1986.

3 Lindsey, J. K.. A review of some extension to generalized linear models.J. stat. med., 18: 2223-2236, 1999.

4 David S. Moore, George P. McCabe, Introduction to the Practice of Statistics,4th Edition, 2003.

5 Gujarati N. Damodar, Basic Econometrics, 4th Edition, 2004.