Available online at www.elixirpublishers.com (Elixir International Journal)

Statistics

Elixir Statistics 114 (2018) 49492-49496

ISSN: 2229-712X

Using Transformation of Variable for Treating Autocorrelation in Simple Regression (First Difference & Cochrane-Occutt methods)

AbuElgasim Abbas A bow Mohammed[1,2] and Khalda Osman Abd Elghafar Mohammed[2]

[1]Department of Statistics &Econometrics, Faculty of Economics and Political Science, Omdurman Islamic University, Khartoum, Sudan.

[2]Department of MIS-Stat. and Quant. Method Unit, Faculty of Business & Economics, Qassim University, Buraidah, KSA.

ARTICLE INFO

Article history:

Received: 28 November 2017;

Received in revised form:

25 December 2017;

Accepted: 6 January 2018;

Keywords

Auto-correlation,
Simple regression,
First difference,
Cochrane-Orcutt,
Malaria,
Khartoum.

ABSTRACT

The aim of the study is to examine it is sometimes necessary to use transformation of variable for treating autocorrelation of regression or time series model. Different there are many methods have been used for this purpose. This paper uses the first difference method and the Cochrane-Orcutt method in autocorrelation of malaria data in Khartoum State, treated in out-patient clinics. The data collected from the Federal Ministry of Health, National Health Information Centre, and annual reports during (2001 – 2015). By calculating the estimates for each method and errors, the paper showed that the method of differences is better in the treatment of auto-correlation, and there are limits of negatives error, because reducing these errors from least square method.

1. Introduction

Regression and time series are useful studies of phenomena that need to know their direction in the future, that consider a fit model suitable data after treating autocorrelation, there are many methods used for this purpose.

One objective of this paper is treating autocorrelation by using transformation variables. Therefore, the study uses first difference and Cochrane-Orcutt methods in malaria data in Khartoum State, treated in out-patient clinics. The data is collected from the Federal Ministry of Health, National Health Information Centre, annual reports during (2001 – 2015). The paper attempt to know the best method for treating autocorrelation by comparing estimates of these methods although estimates error.

2. Problem Statement

Where t is time, y is dependent variable, X is independent variable must be fixed , as opposed to observation taken randomly on the time scale e_t is random disturbances with mean zero and variance σ^2 in the least square analysis , the usual regression model :

$$y_t = \beta_0 + \sum \beta_i x_i + \epsilon_i \quad \ldots\ldots\ldots (1) \qquad t = 1, 2, \ldots, n$$

If ny, s are successive observation in time ,the experimenter frequently wishes to investigate the nature of the response curve over time .In this case might set x_{it}= t , or might use the method of harmonic analysis to search for periodicities in y in the other case , the assumed model might involve lagged values of y as predictor as follow

$$y_t = \beta_0 + \sum \beta_i y_{t-1} + e_t \quad \ldots\ldots\ldots (2)$$

This is an autoregressive model. Finally, one could use combined regression model with lagged y's, present ax's lagged x's and time as predictor. The method of least square is applicable for autoregressive model provided n is large.

One of the major difficulties with least square method with time series is strong possibility that the ϵ_i , s are not independent [1].

Pointed out that it is correlation of the ε,s and not of y,s which is to be avoided . It is possible that if the x's and y, s are both correlation in time the errors will be relatively uncorrelated inconsiderable amount of research been devoted to the problem.

3. Objective of paper

The aim of the paper that to compare between the Cochrane-Orcutt method and the first difference method so as to choose what is the best method for treating autocorrelation.

4. Hypothesis of paper

The paper assumes that the Cochrane-Orcutt method gives the same results as the first difference method for malaria data.

5. Literature Review

Stanley Jere and Edwin Mayo (2016) used Box-Jenkins modeling procedure for determine an ARIMA model and forecasting. They used data of malaria case from Ministry of Health (Kabwe District) –Zambia for the period (2009-2013) for age 1 to 5 years .The results showed that an appropriate model is simply an ARIMA (1, 0,0) and the forecasted malaria for January and February, 2014 are 220 and 265, respectively [2].

Sadia AbdEkareem (2012) using BOX and Jenkins Method (Identification, Diagnostics Checking of model, forecasting) to find the best forecasting model to the number of patient with malignant tumors in An bar province by using the monthly data for the period (2006-2010).The result of data analysis show that the proper and suitable model is integrated Auto regressive model of order (2) ARIMA (2, 1, 0) [3].

Tele:

E-mail address: khalda999@hotmail.com

6. Methodology

The paper was based on the theoretical method of dealing with the method of first differences and the Cochrane-Orcutt method in simple regression and time series and uses the Durbin Watson test .it also uses applied side to the malaria data in Khartoum State- Sudan, treated in out-patient clinics. Taken from the Federal Ministry of Health, National Health Information Centre annual reports during (2001 – 2015) .The paper used E-Views program and SPSS Program as analytical tools of analysis data.

Theoretical Formulation

7. Simple Regression

The paper described statistical techniques for analyzing the relationships among variables

Assume that the first independent variable X, where is the other dependent variable Y is subject to some random variation. This is known as regression analysis. The paper introduced the linear model under the assumption that its error terms could be represented by uncorrelated random variables with mean of zero and constant variance .However the paper want to significance tests for hypothesized values of its parameters α ,β and estimates of the uncertainty in predicted values of the dependent variable for anew value of predictor [4]. If the nature of data is time series, then an analytical method that will work for large data plots with much dispersion in their values have been obtained sequentially in time $\boldsymbol{t}$ where $\boldsymbol{t = 1, 2, \dots n}$ and prefer to express the independent variable, n is number of observation .In this case it's very important to determine a fit mathematical model for our data and estimate the intercept parameters. The $\boldsymbol{e_t}$ term is random variables or variants with some kind of probability distribution should make some assumption about its properties.

8. Auto-correlation

Autocorrelation, also known as serial correlation occurs when residuals from adjacent measurements in a time series are not independent of one another. Is the correlation of a signal with a delayed copy of itself as a function of delay? Informally, it is the similarity between observations as a function of the time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain.

Unit root processes, trend stationary processes, autoregressive processes, and moving average processes are specific forms of processes with autocorrelation.

In statistics, the autocorrelation of a random process in the Pearson correlation between values of the process at different times, as a function of the two times or of the time lag [5].

9. The Cochrane-Orcutt Method

Cochrane–Orcutt estimation is a procedure in econometrics, which adjusts a linear model for serial correlation in error term [6].

Consider the model

$$\boldsymbol{y_t = \alpha + X_t\beta + \varepsilon_t} \quad \text{.....(3)}$$

Where $\mathbf{y_t}$ is the value of the dependant variable of interest at time t, $\boldsymbol{\beta}$ is a column vector of coefficients to be estimated, $\mathbf{X_t}$ is a row vector of explanatory variables at time t, and $\boldsymbol{\varepsilon_t}$ is the error term at time t.

If it is found via the Durban-Watson statistic that the error term is serially correlated

Over time, then standard statistical inference as normally applied to regressions is invalid because standard errors are estimated with bias.

To avoid this problem, the residuals must be modeled. If the process generating the residuals is found to be a stationary first-order autoregressive structure [7].

$$\boldsymbol{\varepsilon_t = \alpha + \rho\varepsilon_{t-1} + e_t} \quad \text{.....(4)} \qquad \boldsymbol{|\rho| < 1}$$

With the errors $\mathbf{e_t}$ being white noise, then Cochrane-Orcutt procedure can be used to transform the model by taking a quasi- difference:

$$\boldsymbol{y_t - \rho y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + e_t} \quad \text{.....(5)}$$

In this specification the error terms are white noise, so statistical inference is valid. Then the sum of squared residuals is minimized with respect to$(\boldsymbol{\alpha, \beta})$ **conditional on** ,

This method estimates **ρ** several times until the estimates stop changing, or converge. The Cochrane-Orcutt method is often called an iterative process because it repeats certain steps over and over [8]. The steps are:

1. Estimate the regression using ordinary least squares.
2. Use the error term observations from step 1 to estimate $\boldsymbol{e_t = Pe_{t-1} = u_t}$getting an estimate for **ρ**.
3. Use the estimate for **ρ** along with the data for the dependent and independent variables to estimate the generalized difference equation,
4. using the error term observations from step 3, go back to step 2 and estimate **ρ** again. Repeat this process until the estimate of **ρ** stays about the same; then estimate the generalized difference equation one last time.

10. First Difference Procedure:

First differences is the simplest transformation procedure as it implicitly assumes r = 1. This assumption is often approximately justified because Estimates of r are often close to 1.

The relationship of SSE with r is often "flat" for values of r near the optimum, so the estimate of r does not need to be exact

The first differences transformation is thus

$$\boldsymbol{y_t = y_t - y_{t-1}} \quad \text{.....(6)}$$

$$x_t = x_t - x_{t-1} \dots\dots\dots\dots (7)$$

The first differences procedure involves two regressions with the transformed data:

1. a first regression without a constant term to estimate the regression coefficients (since the first differences transformation "wipes out" the constant term)
2. a second regression with a constant term to recalculate the D-W D statistic only (because the D-W formula requires a constant in the model) [9].

11. Durbin-Watson statistics:

In statistics, the Durbin–Watson statistic is a test statistic Used to detect the presence of autocorrelation (a relationship between values separated from each other by a given time lag) in the residuals (prediction errors) from a regression analysis. It is named after James Durbin and Jeoffrey Watson. The small sample distribution of this ratio was derived by Neumann [10].

Durbin and Watson (1950, 1951) applied this statistic to the residuals from least square regression, and developed bounds tests for the null hypothesis that the errors are serially uncorrelated against the alternative that they follow a first order autoregressive process.

Later, John Denis Sargan and Bhargave developed several von Neumann–Durbin–Watson type test statistics for the null hypothesis that the errors on a regression model follow a process with a unit root against the alternative hypothesis that the errors follow a stationary first order auto regression [11]. Note that the distribution of this test statistic does not depend on the estimated regression coefficients and the variance of the errors, [12].

If e_t is the residual associated with the observation at time t, then the test statistic is:

$$\boldsymbol{d} = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})2}{\sum_{t=1}^{T} e_t^2} \quad \ldots\ldots\ldots (8)$$

Where T is the number of observations. Note that if one has a lengthy sample, then this can be linearly mapped to the Pearson correlation of the time-series data with its lags Since d is approximately equal to $2(1 - r)$, where r is the sample autocorrelation of the residuals [13]. d = 2 indicates no autocorrelation. The value of d always lies between 0 and 4. If the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if Durbin–Watson is less than 1.0, there may be cause for alarm. Small values of d indicate successive error terms are, on average, close in value to one another, or positively correlated. If d > 2, successive error terms are, on average, much different in value from one another, i.e., negatively correlated. In regressions, this can imply an underestimation of the level of statistical significance.

To test for positive autocorrelation at significance α, the test statistic d is compared to lower and upper critical values (dL,α and dU,α):

- If d < dL,α, there is statistical evidence that the error terms are positively autocorrelated.
- If d > dU,α, there is no statistical evidence that the error terms are positively auto correlated.
- If dL,α < d < dU,α, the test is inconclusive.

Positive serial correlation is serial correlation in which a positive error for one observation increases the chances of a positive error for another observation.

To test for negative autocorrelation at significance α, the test statistic (4 − d) is compared to lower and upper critical values (dL,α and dU,α):

- If (4 − d) < dL,α, there is statistical evidence that the error terms are negatively auto correlated.
- If (4 − d) > dU,α, there is no statistical evidence that the error terms are negatively auto correlated.
- If dL,α < (4 − d) < dU,α, the test is inconclusive.

Negative serial correlation implies that a positive error for one observation increases the chance of a negative error for another observation and a negative error for one observation increases the chances of a positive error for another.

The critical values, dL,α and dU,α, vary by level of significance (α), the number of observations, and the number of predictors in the regression equation. Their derivation is complex—statisticians typically obtain them from the appendices of statistical texts.

The following assumptions should be satisfied:

1. The regression model includes a constant
2. Autocorrelation is assumed to be of first-order only
3. The equation does not include a lagged dependent variable as an explanatory variable

Steps

1: Estimate the model by OLS and obtain the residuals
2: Calculate the DW statistic
3: Construct the table with the calculated DW statistic and the dU, dL, 4-dU and 4-dL critical values.
4: Conclude

Application

Malaria data had been collected from the year (2001-2015), see table (1) below

Table (1). Malaria data.

Year	Number of patient
2001	879795
2002	857782
2003	738985
2004	661159
2005	527609
2006	496664
2007	288256
2008	273479
2009	241657
2010	155011
2011	153675
2012	146579
2013	199256
2014	258927
2015	263658

Source: Annual Health Statistical Reports

Malaria treated cases in out-patient clinics had been considered as dependent variable while time for treatment as independent variable. for analyzing the data, firstly estimate the mode of basic data in table below:

Table (2).Model Summary.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.896a	.803	.788	119836.146	.285

The Durbin Watson had been employed to check the autocorrelation, the result above shows the value (0.285) indicating existing of autocorrelation in the model. Durbin Watson critical value found that $\boldsymbol{DL = 1.08, DU = 1.36, for\ K = 1}$, so the calculated value less than lower limit then the result $\boldsymbol{\rho > 0}$ effect on the estimate and forecast .

Table (3) and table (4) show sum of square of ANOVA and coefficients below:

Table (3). ANOVA.

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	762825853882.575	1	762825853882.575	53.119	.000[b]
	Residual	186689124387.159	13	14360701875.935		
	Total	949514978269.733	14			

Table (4). Coefficients.

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	416166.133	30941.560		13.450	.000
	Time	52195.575	7161.580	.896	7.288	.000

Below is the simple regression model with one independent variable

$$\hat{\boldsymbol{y}} = \boldsymbol{416166.133 + 52195x} \quad \ldots\ldots\ldots (9)$$

Table (5). Model Summary.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.387a	.150	.079	59601.623	.150	2.112	1	12	.172	2.080

a. Predictors: (Constant), time

b. Dependent Variable: difference

Table (7). Coefficients.

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	118832.738	37173.976	-.387-	3.197	.008	-.387-	-.387-	-.387-
	Time	-5742.541-	3951.549		-1.453-	.172			

Dependent Variable: difference

Therefore $|\rho| < 1$

BY using the first difference method for analytical data we found the results below:

Table(6). ANOVA.

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	7.502E9	1	7.502E9	2.112	.172a
	Residual	4.263E10	12	3.552E9		
	Total	5.013E10	13			

a. Predictors: (Constant), time

b. Dependent Variable: difference

Table (8). Residuals Statistics.

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	32694.63	107347.66	70021.14	24022.771	14
Residual	-85334.656-	129773.047	.000	57263.387	14
Std. Predicted Value	-1.554-	1.554	.000	1.000	14
Std. Residual	-1.432-	2.177	.000	.961	14

a. Dependent Variable: difference

Table (9). Linear regression estimate of malaria cases (original model).

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	416166.1	30941.56	13.45007	0.0000
TIMEX	52195.58	7161.580	7.288277	0.0000
R-squared	0.803385	Mean dependent var		416166.1
Adjusted R-squared	0.788261	S.D. dependent var		260427.5
S.E. of regression	119836.1	Akaike info criterion		26.34920
Sum squared resid	1.87E+11	Schwarz criterion		26.44361
Log likelihood	-195.6190	Hannan-Quinn criter.		26.34820
F-statistic	53.11898	Durbin-Watson stat		0.284568
Prob(F-statistic)	0.000006			

It is observed that both intercept and slope are highly significant. The F-stat and its P- value suggesting that over all models is highly significant. The Durbin-Watson statistic is less than DL which indicated that disturbance error is an auto correlated.

WE applied on Cochrane-Orcutt method according the tables below

Table (10). Coefficient and Statistic.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	76876.20	93558.41	0.821692	0.4273
Resid(-1)	0.879918	0.144090	6.106712	0.0001
R-squared	0.717612	Mean dependent var		-3.88E-11
Adjusted R-squared	0.670548	S.D. dependent var		115477.0
S.E. of regression	66281.42	Akaike info criterion		25.31729
Sum squared resid	5.27E+10	Schwarz criterion		25.45890
Log likelihood	-186.8797	Hannan-Quinn criter.		25.31578
F-statistic	15.24738	Durbin-Watson stat		1.129549
Prob(F-statistic)	0.000507			
Inverted AR Roots	.88			

So the coefficient $\mathbf{p = 0.8799}$ will be multiplied by $\mathbf{y_{t-1}}$ so the dependent variable become $\mathbf{y_t - py_{t-1}}$ and $\mathbf{x_t - px_{t-1}}$ as independent variable then the coefficient of the model in the table

Table (11). Coefficients of Model.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-36275.55	24833.68	-1.460740	0.1698
DELTAXt	134357.8	23487.70	5.720348	0.0001
R-squared	0.731678	Mean dependent var		90014.06
Adjusted R-squared	0.709318	S.D. dependent var		78915.97
S.E. of regression	42547.51	Akaike info criterion		24.28619
Sum squared resid	2.17E+10	Schwarz criterion		24.37749
Log likelihood	-168.0034	Hannan-Quinn criter.		24.27774
F-statistic	32.72238	Durbin-Watson stat		2.274826
Prob(F-statistic)	0.000096			

The estimate model becomes:

$$\hat{y} = -36275.55 + 134357.8x \quad \ldots\ldots\ldots\ldots (10)$$

It is observed from Durbin Watson statistic 2.275 is closer to 2.0 and greater than Du so now the problem of autocorrelation is solved.

Table (12). Fundamental results of regression.

Method	b_1	s_{b_1}	r	MSE
Cochrane-Orcutt	134357.8	23487.70	0.86	2.17E10
First difference	5742.541	3951.549	1	3.552E9

12. Conclusion

In comparing the two methods of transformation it can conclude that:

1-Estimation of $\mathbf{B_1}$ for the method of first difference is less than the Cochrane-Orcutt method

2-The standard deviations of b_1 for the method of the first difference are lower than the Cochrane-Orcutt

3-The first difference method gives less estimates of the error term $\mathbf{e_t}$

The first difference method can be effective in removing auto correlation

References:

[1]R. L. Anderson, (1954). The problem of autocorrelation regression analysis, Journal of the American Statistical Association.

[2]Jere, Moyo; Stanley, Edwin. (2016). Modelling Epidemiological Data Using Box-Jenkins Procedure .copy right by authors and scientific Research publishing lic.http:\\ Creative Common. Org/ licenses / 4.o/

[3] Sadia S. AbdEkareem. Tomah, (2012).using analysis of time series to forecast number of patients malignant tumors "Al- unbar university journal of economic. 8(4). 371-393

[4]Dorland F. Morrison, (1983). Applied Linear Statistical Methods , prentice-hall, Inc., Engel wood Cliffs, New Jersey. p6-7

[5]Bisgaard, S. and M. Kulahci (2005b). "The effect of autocorrelation on statistical process control procedures,"

[6]D. Cochrane and G. H. Orcutt, "Application of Least-Squares Regressions to Relationships Containing Auto correlated Error Terms," Journal of the American Statistical Association, Vol. 44, 1949, pp. 32–61

[7]Wooldridge, Jeffrey M. (2013). Introductory Econometrics: A Modern Approach (Fifth international ed.). Mason, OH: South-Western. pp. 409–415

[8]D. Cochrane and G. H. Orcutt, "Application of Least-Squares Regressions to Relationships Containing Auto

correlated Error Terms," Journal of the American Statistical Association, Vol. 44, 1949, pp. 32–61
[9]Wooldridge, Jeffrey M. (2001). Econometric Analysis of Cross Section and Panel Data. p. 279–291
[10] Neumann, John Von (1941). "Distribution of the ration of the mean square successive difference to the variance". The Annals of Mathematical Statistics Vol. 12, No. 4, pp. 367-395
[11]Sargan, J.D.; Bhargava, Alok (1983), "Testing Residuals from Least Squares Regression for being generated by the Gaussian Random Walk". Econometrica. 51(1): 153 – 174. JSTOR 1912252
[12]Chatterjee, Samprit; Simonoff, Jeffrey (2013). Handbook of Regression Analysis. John Wiley & Sons. ISBN 1118532813.
[13]N.C Gujarati, Damodar , Porter, Dawn.Basic Econometrics (5th ed.) (2009).Boston:MCGraw-Hill Irwin ISBN 978-0-07-337577-9