A Comparisons on Remedial Methods of First-order Autocorrelation Problems

Atsavin Saneechai

Department of Information Technology, Faculty of Science and Technology, Bangkok Suvarnabhumi College,

Bangkok, E-mail: atsavin555@hotmail.com

Abstract— The purpose of this study is to compare the remedial methods of first-order autocorrelation in simple linear regression analysis for the 9 levels of autocorrelation (ρ) : 0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8 and 0.9 . The three autocorrelation remedial methods are generalized differencing, Cochrane-Orcutt, and Durbin's Two-Step. The simulated sample size were set as 30 and 50 for 1000 time iterations. The Durbin-Watson test of autocorrelation was used as a criteria to specify the best methods for the remedial problem by looking at the highest percentage of data sets that showed no significant result in testing H_0: ρ=0 is considered.

The results of the study show that more autocorrelation (ρ) levels were found each time the autocorrelation levels increased, the percentages of autocorrelation remedial abilities from each method tended to decrease when the sample size was 30. All three methods showed nearly the same ability in autocorrelation of the remedial problem, but Durbin's Two-Step and Cochrane-Orcutt methods gave slightly more satisfactory results than generalized differencing. When the sample size was 50 the generalized differencing and Cochrane-Orcutt methods were best at solving the remedial problem, given autocorrelation levels.

In conclusion, the Cochrane-Orcutt method was the most suitable method for autocorrelation of remedial problems in all cases and Durbin 's Two-Step method was the best one in all sample sizes at the autocorrelation levels (ρ) 0.1-0.7 .

Keywords-First-OrderAutocorrelation;Simple Linear Regression

I. Introduction

One of the error concept patterns in the regression states that the certain number of error terms (ε_i) will not correlate with each other and being the random variable followed for normal distribution with zero average and constant in deviation [1].The health science and medical data are usually collected by means of time series. The dependent and independent variables in the regression equations is collected by the time series called "time series data". These types of data are always lacking in free error value as "ε_i" and "ε_j $(i \neq j)$". There are some correlations among time series data called "autocorrelation" or "serial correlation"[2].This study had compared the methods used to solve the problem of first order autocorrelation error within simple linear regression together with estimated their abilities in forecasting by using MSE criterion. The pattern of this research was independent variable defining by randomizing way which provided normal distribution of X_t~N(0,1) according to each sample size and defining the dependent variable pattern from the regression equation . All processes were simulated using visual studio software.

II. Research procedure

Define the regression model used in the study. The pattern of simple regression model is:

$$Y_t = b_0 + b_1 X_t + e_t \quad ; \ t = 1,2,...,n$$

Create the error term(ε_t) according to the first-order autoregressive pattern with error relation in 9 correlation levels(ρ), that is 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 (500 set for each) and use the sample size as 30 , 50 units and

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

Create the data(Y_t, X_t) with the same size as defined samples

Estimate the parameter from the data obtained by simple least square

The equation got "$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$" from simulation that has the error (ε_t) from the first-order autoregressive pattern which has the error correlation within defined correlation level. After that, the equation will be examined for the data if it has correlation or not by using the Durbin – Watson test with the confidence level $\alpha = 0.05$.

In case of data examination if the autocorrelation occur or not. Transform data in a defined way. If no autocorrelation occurs, create new data in simple regression

Remedial the autocorrelation problem by using 3 variable transformation methods: generalized differencing method Cochrane – Orcutt method and Durbin's Two-Step method

Take the post transformed variable data to the regression equation to calculate the MSE and "$e_t^{'} = \hat{Y}_t - \hat{Y}_t^{'}$"

Test the data if the autocorrelation still occurs or not by using the Durbin–Watson method with the confidence level $\alpha = 0.05$ and count the number of hypothesis accept time for each method. Then, compare the MSE value of each method.

Simulation by Visual Studio Program

A. The error property

From the definition of first-order autoregressiveof the error term "ε_t"

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

Average and deviation of "ε_t" for the first-order autoregressive

$$E(\varepsilon_t) = o$$

$$\rho^h(\varepsilon_t) = \frac{\rho^2}{1-\rho^2}$$

Suggest that "ε_t" is zero for average and contain constant deviation

Covariance between "ε_t" and "ε_{t-1}" can be replaced by "$\rho(\varepsilon_t, \varepsilon_{t-1})$" whish is

$$\rho(\varepsilon_t, \varepsilon_{t-1}) = \rho\left(\frac{\sigma^2}{1-\rho^2}\right)$$

When the correlation coefficient between "ε_t" and "ε_{t-1}" can be replaced with "$\rho(\varepsilon_t, \varepsilon_{t-1})$" that is

$$\rho(\varepsilon_t, \varepsilon_{t-1}) = \frac{\sigma(\varepsilon_t, \varepsilon_{t-1})}{\sigma(\varepsilon_t)\sigma(\varepsilon_{t-1})}$$

According to the deviation of each term is equal to "$\rho^h(\varepsilon_t) = \frac{\rho^2}{1-\rho^2}$", the coefficient then is:

$$\rho(\varepsilon_t, \varepsilon_{t-1}) = \frac{\rho\left(\frac{\sigma^2}{1-\sigma^2}\right)}{\sqrt{\frac{\sigma^2}{1-\sigma^2}}\sqrt{\frac{\sigma^2}{1-\sigma^2}}} = \sigma$$

That is autocorrelation parameter "ρ" represent the relation between the nearby error.

Covariance between "ε" with distances "S" away is

$$\rho(\varepsilon_t, \varepsilon_{t-s}) = \rho^s\left(\frac{\sigma^2}{1-\rho^2}\right),\ s \neq 0$$

And the regression coefficient between "ε_t" and "ε_{t-s}" is

$$\rho(\varepsilon_t, \varepsilon_{t-s}) = \rho^s, s \neq 0$$

Thus, when "ρ" are positive, all error will be relative. However, if the time period is more , the error relation will decrease.

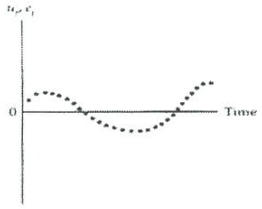

Figure 1 Demonstrated the character of autocorrelation in positive pattern

B. Solving the problem of autocorrelation with data transformation of variable

From the regression equation like:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad (1)$$

If this equation is true at the time "t", then it will also true at the time "t-1" as follow:

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1} \qquad (2)$$

Multiply both side of the equation 1 with "ρ" will gains

$$\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{t-1} + \rho\varepsilon_{t-1} \qquad (3)$$

Minus (3) From (1) will gains

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1 X_t - \rho\beta_1 X_{t-1} + (\varepsilon_t - \rho\varepsilon_{t-1})$$ From (1) $\varepsilon_t - \rho\varepsilon_{t-1} = u_t$, thus

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + u_t \qquad (4)$$

It can be written as:

$$Y_t^{'} = \beta_0^{'} + \beta_1 X_t^{'} + u_t$$

When $Y_t^{'} = (Y_t - \rho Y_{t-1})$

$X_t^{'} = (X_t - \rho X_{t-1})$

$\beta_0^{'} = \beta_1(1-\rho)$

$\beta_1^{'} = \beta_1$

Owing to "u_t" is the free random variable, thus when the independent and dependent data is transformed, the linear regression profile which has free error is obtained. Thus, the least square can be used to estimate the regression equation and the transformed variables as "X'" and "Y'". But data transformation will cause 2 problems, that is, one parameter absent and the "ρ" become unknown[3].

Because the "ρ" value is unknown, so, it needs to be estimated before data is transformed. The variable transformation require the "ρ" estimation step because the actual value is generally unknown. Defines "r" is the estimated value of "ρ". The variable transformed using "r" will be:

$$Y_t' = (Y_t - rY_{t-1})$$
$$X_t' = (X_t - rX_{t-1})$$

After data transformation, these data then is taken to create the regression equation by least square and gained regression equation is:

$$\hat{y}_t' = b_0' + b'_1 x_t'$$

If the regression equation is removed the error autocorrelation is successfully obtained, the equation can be transformed back to create the regression with former variable as:

$$\hat{Y}_t = b_0 + b_1 x_t$$

When $b_0 = \dfrac{b_0'}{1-r}$ and $b_1 = b_1'$

The standard error of regression coefficient for the former variable can be calculated from: $S_{b_0} = \dfrac{S_{b_0'}}{1-r}$

and $S_{b_1} = S_{b_1'}$

C. Autocorrelation problem remedial methods

For autocorrelation solving method when the error has the first-order autocorrelation, three solving methods are proposed in this study as follows:-

Generalized differencing method

The step to solve the error autocorrelation problem by this method is to transform the data into generalized difference equation form which is the regression equation that the equation's variable data and the error are transformed to be the difference between the variable value at present and before times. Then, the parameters are estimated using the least square which also need to estimate the "ρ" by correlation of "r" between OLS-Residual to use in data transformation (Berenson, Mark L. and David M. Levine. 1996).

The method to estimate "ρ" using correlation of "r" between OLS – Residual

From OLS – Residual, $e = y - x\hat{\beta}$, when $e = (e_1, e_2, ..., e_n)$, separate vector "e" to 2 groups as "$e_{t-1} = (e_1, e_2, ..., e_{n-1})$" and "$e = (e_1, e_2, ..., e_n)$". Then calculate the linear correlation between "e_{t-1}" and "e_t".

This time, it is estimated by $\hat{\rho} = \dfrac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$

From the equation (4)

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + u_t$$

The equation (4) called generalized difference model for new error as "$\varepsilon_t - \rho\varepsilon_{t-1}$" or "$u_t$". And bring "$\rho$" together with "$\hat{\rho}$" substitute in the generalized difference equation and rearrange in more simple form as:

$$Y_t' = \beta_0' + \beta_1 X_t' + u_t'$$

By $Y_t' = (Y_t - \hat{\rho} Y_{t-1})$

$$X_t' = (X_t - \hat{\rho} X_{t-1})$$
$$\beta_0' = \beta_1(1-\hat{\rho})$$
$$\beta_1' = \beta_1$$
$$u_t' = \varepsilon_t - \hat{\rho}\varepsilon_{t-1}$$

And the estimated parameters β_0' and β_1' in the model with least square method

Cochrane-Orcutt method

The Cochrance-Orcutt method is a way to transform the variable data to solve the autocorrelation error problems. It has three main steps.

Estimation of the "ρ" value

The first-Order autoregression type error in regression form can be considered in the linear regression through the origin pattern.

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

When "ε_t" is the dependent variable, "ε_{t-1}" is the independent variable which is the error and has "ρ" as the linear slope through the origin.

Because we do not know the value of "ε_t" and "ε_{t-1}", so, we use residual "e_t" and "e_{t-1}" instead [4]. With the same method, the dependent and independent variable are estimated for the slope of the linear regression equation through the origin that has a formula as below. We can estimate the slope of "ρ" and replace it with "r" in the formula:

$$r = \frac{\sum_{t=2}^{n} e_{t-1}e_t}{\sum_{t=2}^{n} e_{t-1}^2}$$

Creating the regression equation from the transformed data

Use the estimator of "ρ"calculated from the transformed variable formula by "$Y_t^{'} = (Y_t - \rho Y_{t-1})$" and "$X_t^{'} = (X_t - \rho X_{t-1})$" as the formula, and then creates the regression equation by the least square form the transformed data

Autocorrelation test

Testing for the error value in the transformed regression equation pattern, if it has correlation or not, by using the Durbin-Watson method. If the test shows independent of the error, the work will then finish.

Durbin's Two – Step method

From the equation pattern

$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + u_t$ It can be rewritten as:-

$Y_t = \beta_0(1-\rho) + \beta_1 X_t - \rho\beta_1 X_{t-1} + \rho Y_{t-1} + u_t$ Durbin suggests the estimated method for "ρ" using two-step method as:-

Define the equation as multiple regression pattern. Create the multiple regression "Y_t" on "X_t, X_{t-1}" and "Y_{t-1}"; and then define the estimator of the regression coefficient of "Y_{t-1}" as the estimator of "ρ (=r)" which, although is lean, it will still be the consistent estimator for "ρ".

When the "r" gained and the data is already transformed by define:

$Y_t^{'} = (Y_t - rY_{t-1})$ and $X_t^{'} = (X_t - rX_{t-1})$

Then create the regression equation by least square (OLS) from transformed data as same as the following equation:

$$Y_t^{'} = \beta_0^{'} + \beta_1 X_t^{'} + u_t^{'}$$

Durbin – Watson test for autocorrelation

Examination method with statistical test for the independent of error has stated the hypothesis that the error posses the first-order autoregressive form which the independent variable is fixed. The examination will consider in the point that the autocorrelation parameter, $\rho = 0$ will gain $\varepsilon_t = u_t$. Thus, the error "ε_t" is independent because "u_t" is independent. Because of the business and economic application, the error with correlation is usually positive correlation. Thus, it is normally test as positive autocorrelation pattern [5].

Test hypothesis:

$$H_o : \rho = 0$$
$$H_1 : \rho > 0$$

or H_o : The error has no correlation

H_1 : The error has positive correlation

Test statistic is the statistic "d" of Durbin – Watson that define as:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

When $e_t = Y_t - \hat{Y}_t$, t= 1,2,….n

n is the observation numbers

Probability distribution of "d" depends on the matrix "X". However, the Durbin and Watson demonstrated that "d" is between "d_L" and "d_U" by the value of "d_L" and "d_U" as shown in the Durbin-Watson statistic table.

If d < d_L reject $H_o : \rho = 0$

If d > d_U accept $H_o : \rho = 0$

If $d_L \le d \le d_u$ the test can not be concluded [6]. The "d" with small value can be described as "$\rho > 0$" because the nearby error are "ε_t" and "ε_{t-1}" that the amount are nearly similar in case of positive relation. Thus, the difference of residual "$e_t - e_{t-1}$" will be small and the statistical top line of "d" will also be small. But if the errors have no relation, then "$e_t - e_{t-1}$" will be large value and the statistical top line will also become large. Therefore, if "d" is small, the "H_1" will be consistent to the conclusion while if the statistical "d" is large, the "H_0" will be consistent to the conclusion instead.

Normally, the negative autocorrelation hardly occurs, but if it necessary to test, the negative autocorrelation can be created by using the statistical "4 – d" instead of "d" and set the hypothesis of $H_o : \rho = 0$ contrast with "$H_1 : \rho < 0$" and do the same test conclusion as the positive autocorrelation. The test operation can explain this phenomenon better. It can be seen that the limit of "d" is between 0 and 4 which can be proven by extending the formula of "d"

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Which gained

$$d = \frac{\sum e_t^2 + \sum e_{t-1}^2 - 2\sum e_t e_{t-1}}{\sum e_t^2}$$

Because $\sum e_t^2$ and $\sum e_{t-1}^2$ have only one different observation, so, both are approximately the same. Therefore, let "$\sum e_{t-1}^2 = \sum e_t^2$" which will be

$$d = 2\left(1 - \frac{\sum e_t e_{t-1}}{\sum e_t^2}\right) \quad \text{approximately}$$

Let the estimated "ρ" defined by

$$r = \frac{\sum e_t e_{t-1}}{\sum e_t^2}$$

Replace "r" in the equation "d"

$$d = 2(1-r)$$

Owing to "$-1 \le \rho \le 1$" it will gain

$$0 \le d \le 4$$

III. Results and Discussion

Where n=30 the autocorrelation level is increased continuously, the percentage of autocorrelation problem remedial ability in each method tends to decrease. The Durbin's Two-Step method gave the best average problem remedial result with the percentage of 97.87, followed by the Cochrane-Orcutt and the generalized differencing methods that gave the percentage of 95.98 and 92.33, respectively.

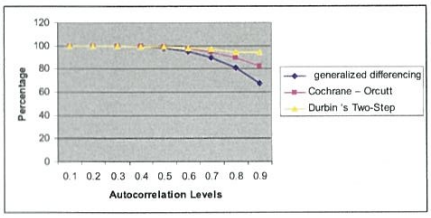

Figure 2 Percentage comparison of the data sets with no error autocorrelation presented when determined by three mentioned methods (n=30)

Where n=50 the autocorrelation level increased continuously, the percentage of autocorrelation problem remedial ability in each method tended to decrease. The Cochrane-Orcutt method gave the best average problem remedial result with the percentage of 99.33, followed by the generalized differencing methods and the Durbin's Two-Step methods that gave the percentage of 98.74 and 96.02, respectively.

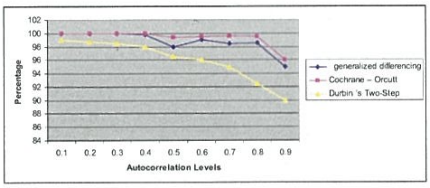

Figure 3 Percentage comparison of the data sets with no error autocorrelation presented when determined by three mentioned methods (n=50)

The average MSE values for each problem remedial in each autocorrelation level indicated that the Durbin's Two-Step method was the best forecast method as shown in the figure 4.

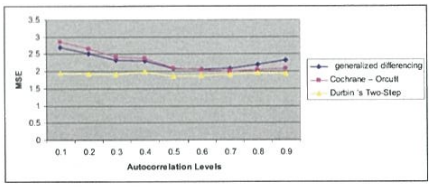

Figure 4 Comparison of MSE from each autocorrelation problem remedial method for the error levels from 0.1 to 0.9 and all sample size

IV. Conclusion

Defining of autocorrelation level (ρ) where the autocorrelation level was continuously increased in each testing time found that the percentage of autocorrelation problem remedial ability in each method data set tended to decrease.

By considering each best problem remedial method in each autocorrelation level, it was found that the Durbin's Two-Step and Cochrane – Orcutt methods were usually the best problem remedial method in all autocorrelation levels with the sample size 30 while the generalized differencing and Cochrane – Orcutt methods were the best remedial method in all autocorrelation levels instead when the sample was 50. In considering the most suitable forecast way, it could be concluded that the

Durbin 's Two-Step was the best one in all sample sizes at the autocorrelation levels (ρ) 0.1-0.7 . In considering for the best problem remedial method which also was the best forecast method at the same time, it was found that the Cochrane – Orcutt method gave the best autocorrelation problem Remedial method but not the best forecast method at the autocorrelation levels(ρ)0.1-0.7. However, forecast values given form this method were not significantly different from the Durbin's Two-Step method.

Acknowledgements

I would like to thank my family and my colleague for their encouragement and kindness.

References

[1]Anderson, David R., Dennis J. Sweeney and Thomas A. Williams.1994. Statistics for Business and Economics. 5th. Ed. WestPublishingcompany.

[2]Berk, Richard A. 2003. Regression analysis: A constructivecritique.SagePublications.

[3]Berry, William D. 1993. Understanding Regression Assumptions. Series: Quantitative Applications in the Social Sciences. Sage Publications.

[4]Berenson, Mark L. and David M. Levine. 1996. Basic Business Statistics : Concepts and Applications. 6th ed. Prentice-Hall.

[5]Chatterjee, S. and Price, B.1977.Regression Analysis by Example. New York: John Wiley & Sons.

[5]Drapper, N.R., and H. Smith.1981.Applied Regression Analysis.2nd ed. New York: John Wiley & Sons.

[6]Montgomery Douglas C. and Elizabeth A. Peck.1982. Introduction to Linear Regression Analysis. New York: John Wiley & Sons.