First-Order Autoregressive Error in Simple Linear Regression : Comparison of Three Remedial Method

อัตตสหสัมพันธ์ของความคลาดเคลื่อนอันดับที่หนึ่งในการถดถอยเชิงเส้นอย่างง่าย : เปรียบเทียบวิธีการแก้ไขสามวิธี

Atsavin Saneechai (อัศวิน เสนีชัย)* Dr. Dechavudh Nityasuddhi (ดร.เดชาวุธ นิตยสุทธิ)**
Piangchan Rojanavipart (เพียงจันทร์ โรจนวิภาต)*** Dr. Chutatip Tansathit (ดร.จุฑาธิป ตัณสถิตย์)****

ABSTRACT

The objective of this study was to compare the remedial methods of first-order autocorrelation in simple linear regression analysis with 9 levels of autocorrelation (ρ): 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 defined. Three compared autocorrelation remedial methods were the generalized differencing, the Cochrane-Orcutt, and the Durbin's Two-Step. The given sample sizes which data obtained from simulation technique were 30 and 50. Each case was generated by using autocorrelation's Durbin-Watson test and carried out 500 times run repeatedly. It could be concluded that Cochrane-Orcutt method was the most suitable solution for the autocorrelation problem solving in all cases. However, the Durbin's Two-Step was also the most appropriate method in forecasting of most autocorrelation levels.

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการแก้ปัญหาอัตตสหสัมพันธ์ของความคลาดเคลื่อนอันดับที่ 1 ในการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย เมื่อกำหนดระดับอัตตสหสัมพันธ์ (ρ) 9 ระดับ คือ 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, และ 0.9 โดยทำการเปรียบเทียบวิธีการแก้ปัญหาอัตตสหสัมพันธ์ 3 วิธี ได้แก่ วิธีที่ 1 generalized differencing วิธีที่ 2 Cochrane–Orcutt วิธีที่ 3 Durbin 's Two-Step โดยกำหนดขนาดตัวอย่างเท่ากับ 30 และ 50 โดยใช้ข้อมูลที่ได้จากการจำลองขึ้นซึ่งในแต่ละกรณีได้ทำซ้ำ 500 ครั้ง โดยใช้การทดสอบของ Durbin-Watson ตรวจสอบว่าความคลาดเคลื่อนมีอัตตสหสัมพันธ์กันหรือไม่ ซึ่งสามารถสรุปได้ว่า วิธี Cochrane-Orcutt เป็นวิธีที่เหมาะสมในการแก้ปัญหาอัตตสหสัมพันธ์แทบทุกกรณี แต่ วิธี Durbin's Two-Step เป็นวิธีที่เหมาะสมในการพยากรณ์เกือบทุกระดับอัตตสหสัมพันธ์

Key Words : First-order autocorrelation, Simple Linear Regression
คำสำคัญ : อัตตสหสัมพันธ์อันดับที่หนึ่ง การถดถอยเชิงเส้นอย่างง่าย

---

* *Master degree student, Department of Biostatistics, Faculty of Public Health, Mahidol University.*
** *Associate Professor, Department of Biostatistics, Faculty of Public Health, Mahidol University.*
*** *Associate Professor, Department of Biostatistics, Faculty of Public Health, Mahidol University.*
**** *Assistant Professor, Department of Applied Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang.*

Introduction

One concept about the error patterns in regression states that the certain number of error termed (ε_i) will not correlate among each others and being the random variable corresponding to normal distribution with average zero and deviate constant (Anderson & David et al., 1994). Data collected from health science or medical areas are usually involved in time series and their dependent and independent variables were a type so called "time series data" in regression equations. These types of data always lack of free error value such as "ε_i" and "$\varepsilon_j \ (i \neq j)$". Some correlations among time series data were defined as "autocorrelation" or "serial correlation" (Berk & Richard et al., 2003). 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 $X_t \sim N(0,1)$ according to each sample size and defining the dependent variable pattern from the equation $Y_t = \beta_o + \beta_1 X_t + \varepsilon_t$ with defined $\beta_o = 0$ and $\beta_1 = 1$. All processes were simulated using Visual Studio software.

Research methods

Regression model

In defining of the regression model used in the study, the following pattern of simple regression model was used:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \quad ; \ t = 1,2,...,n$$

Error term (ε_t)

In creating of the error term (ε_t), the first-order autoregressive pattern contained 9 correlation level errors (ρ) (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9) were used under 500 runs repeated for each level of the 30 and 50 units sample size.

Variable pattern

The pattern of this research was independent variable defining by randomizing way which provided normal distribution of $X_t \sim N(0,1)$ according to each sample size and defining the dependent variable pattern from the equation $Y_t = \beta_o + \beta_1 X_t + \varepsilon_t$ with defined $\beta_o = 0$ and $\beta_1 = 1$.

Durbin – Watson Test

The equation $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$ obtained from the simulation contained the error (ε_t) with the first-order autoregressive pattern that its error correlation was within defined level. The equation was examined for the data if it contained correlation or not by using the Durbin – Watson test with the confidence level $\alpha = 0.05$.

Transformation methods

The autocorrelation problem was then solved by using 3 variable transformation methods: generalized differencing, Cochrane – Orcutt, and Durbin's Two-Step methods.

The property of error

From the definition of first-order autoregressive for the error term "ε_t"

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

The average and deviation of "ε_t" for the first-order autoregressive were:

$$E(\varepsilon_t) = o$$

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

It was suggested that average of "ε_t" was zero and the deviation was constant. Covariance between "ε_t" and "ε_{t-1}"could be replaced by "$\rho(\varepsilon_t, \varepsilon_{t-1})$" which was:

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

And that the correlation coefficient between "ε_t" and "ε_{t-1}" could be replaced with "$\rho(\varepsilon_t, \varepsilon_{t-1})$" that was:

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

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

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

That was autocorrelation parameter "ρ" represented the reliable relation between nearby error. The covariance between "ε" with the distances "S" away from each other was:

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

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

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

Thus, when "ρ" was positive, all error should be relative. However, if the time period was more far away, the error relation would be decrease.

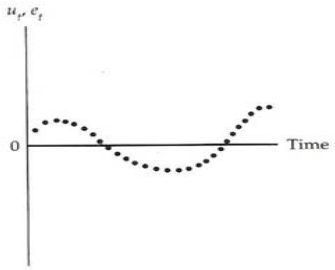

Figure 1 Demonstrated the character of Autocorrelation in positive pattern

Solving the problem of autocorrelation with variable's data transformation

From the regression equation like:

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

If this equation was true at the time "t", then it would 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" would gains

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

Minus (3) out of (1) would gain

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

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 could 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" was free random variable, thus when the independent and dependent data was transformed, the linear regression profile contained free error was obtained. Therefore, the least square could be used to estimate the regression equation and also transformed variables as "$X^{'}$" and "$Y^{'}$". However, it was shown that the data transformation caused 2 problems. Firstly, the parameter was absence. Secondly, the "ρ" become unknown (Berry & William, 1993).

Because the "ρ" value was unknown, thus, it needed to be estimated before data was transformed. The variable transformation required "ρ" estimation step because the actual value was generally unknown. The "r" was defined as an estimated value for "ρ". The variable which was transformed using "r" would be:

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

After data transformation, it was taken to establish the regression equation by least square and gained regression equation was:

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

If the regression equation with the error autocorrelation removed was successfully obtained, the equation then could be transformed back to establish the regression contained 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 could be calculated from:

$$S_{b_0} = \frac{S_{b_0^{'}}}{1-r} \quad \text{and} \quad S_{b_1} = S_{b_1^{'}}$$

Autocorrelation problem solving methods

Generalized differencing method

To solve the error autocorrelation problem by this method was to transform the data into generalized difference equation form which was the regression equation that the equation's variable data and the error were transformed to make difference between the time variable values at present and before. Then, the parameters were estimated using the least square which also needed to estimate the "ρ" by correlation of "r" between OLS-Residual to use in data transformation (Berenson & Mark et al., 1996).

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

From OLS – Residual, $e = y - x\hat{\beta}$, when $e = (e_1, e_2, ..., e_n)$, it could separated 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 the linear correlation between "e_{t-1}" and "e_t" was calculated.

Here, it was 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 obtained equation (4) so called generalized difference model for new error was "$\varepsilon_t - \rho\varepsilon_{t-1}$" or "$u_t$". Then the "$\rho$" together with "$\hat{\rho}$" were taken to substitute in the generalized difference equation and rearranged in more simple form as:

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

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

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

Finally, the parameters $\beta_0^{'}$ and $\beta_1^{'}$ were presented in the model with least square method .

Cochrane-Orcutt method

The Cochrance-Orcutt method was a way to transform the variable data in autocorrelation error problem solving. It composed of three main steps.

Estimation of the "ρ" value

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

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

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

Because we did not know the value of "ε_t" and "ε_{t-1}", so, the residual "e_t" and "e_{t-1}" were used instead (Chatterjee & Price, 1977). With the same method, the dependent and independent variable were estimated for the slope of the linear regression equation through the origin having a formula as below. The slope of "ρ" could be estimated and replaced 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 of the regression equation from the transformed data

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

Autocorrelation test

In testing for the error value in the transformed regression equation pattern for the existence of correlation error using the Durbin-Watson method. If the result indicated the error independence, then the process was 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 could 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 suggested the estimated method for "ρ" using two-step method as:

Defining the equation as multiple regression pattern. The multiple regression "Y_t" on "X_t, X_{t-1}" and "Y_{t-1}" were created; and then the estimator of the regression coefficient of "Y_{t-1}" was defined as the estimator of "ρ (=r)" which, although it leaned, it was still the consistent estimator for "ρ".

When the "r" was gained and the data was already transformed as it defined:

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

Then the regression equation was created 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

Examination method with statistical test for the independent of error had stated the hypothesis

that the error possessed the first-order autoregressive form which the independent variable was fixed. The examination would be considered in the point that the autocorrelation parameter, $\rho = 0$ would gain $\varepsilon_t = u_t$. Thus, the error "ε_t" was independent because "u_t" was independent. Since it was applied in business and economic areas, the error with correlation was usually positive correlation. Thus, it was normally test as positive autocorrelation pattern (Drapper & Smith, 1981).

Test hypothesis:

$$H_o : \rho = 0$$

$$H_1 : \rho > 0$$

Or H_o : The error had no correlation

H_1 : The error had positive correlation

Test statistic was the statistic "d" of Durbin – Watson that defined 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 was the observation numbers

Probability distribution of "d" depended on the matrix "X". However, the Durbin and Watson demonstrated that "d" was 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 could not be concluded (Montgomery and Elizabeth 1982). The "d" with small value could be described as "$\rho > 0$" because the nearby errors were "ε_t" and "ε_{t-1}" that the amount were nearly similar in case of positive relation. Thus, the difference of residual "$e_t - e_{t-1}$" would be small and the statistical top line of "d" would also be small. But if the errors had no relation, then "$e_t - e_{t-1}$" would be large value and the statistical top line would also become large. Therefore, if "d" was small, the "H_1" would be consistent to the conclusion while the statistical "d" was large, the "H_0" would be consistent to the conclusion instead.

Normally, the negative autocorrelation was hardly occurred, but if it necessary to be test, the negative autocorrelation could be created by using the statistical "4 – d" instead of "d" and set the hypothesis of $H_o : \rho = 0$ contrasted with "$H_1 : \rho < 0$" and did the same conclusion as the positive autocorrelation. The test operation could be explained as better phenomenon it was. It could be seen that the limit of "d" was between 0 and 4 which could be proven by extended 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 $\sum e_{t-1}^2$ contained only one different observation, so, both were approximately the same. Therefore, let "$\sum e_{t-1}^2 = \sum e_t^2$" which would be

$$d = 2\left(1 - \frac{\sum e_t e_{t-1}}{\sum e_t^2}\right)$$ 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 would gain

$$0 \le \rho \le 4$$

Results and Discussion

When n=30 the autocorrelation level increased continuously, the percentage of autocorrelation problem remedial ability in each method tended to decrease. The Durbin's Two-Step method gave the best average problem remedial result by 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.

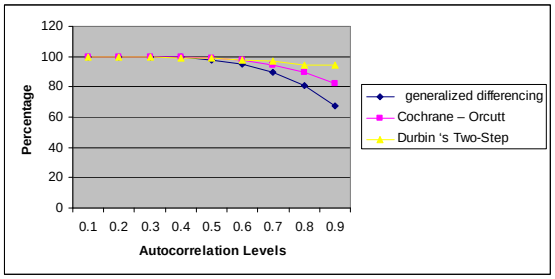

Figure 2 Percentage comparison of the data sets with no error autocorrelation presented when determined by three methods mentioned above (n=30)

When 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 by 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.

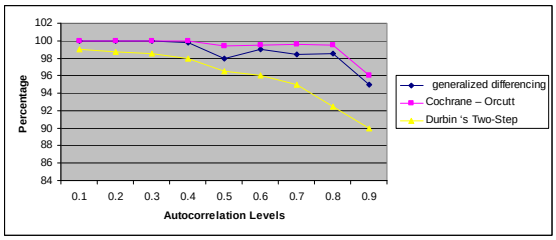

Figure 3 Percentage comparison of the data sets with no error autocorrelation presented when determined by three methods mentioned above (n=50)

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

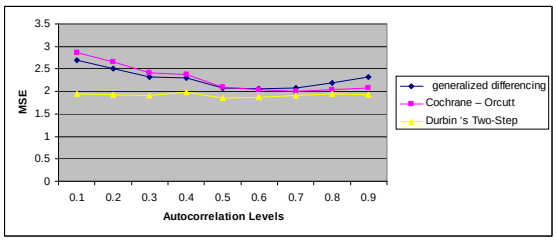

Figure 4 Comparison of MSE from each autocorrelation problem remedial method for the error levels ranged from 0.1 to 0.9 and all of sample size

Conclusion

In defining of autocorrelation level (ρ) where the autocorrelation levels which were continuously increased in each testing time, it was 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 most suitable methods in problem solving for all autocorrelation levels with the sample size 30 while the generalized differencing and Cochrane – Orcutt methods were the most suitable methods in solving for all autocorrelation levels instead when the sample was 50. By considering the most suitable forecast way, it could be concluded that the Durbin's Two-Step was the best one for 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 in the same time, it was found that the Cochrane – Orcutt method gave the best autocorrelation problem solving 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 am also grateful to the Department of Biostatistics, Faculty of Public Health, Mahidol University, particularly Assoc. Prof. Dr. Dechavudh Nityasuddhi, staffs, senior and junior students, or my beloved friends. I would like to say "thank you" to all of them for their assistances all the way through since my first day in the department.

References

Anderson, David R., Dennis J. Sweeney and Thomas A. Williams.1994. Statistics for Business and Economics. 5th ed. West Publishing company.

Berk, Richard A. 2003. Regression analysis: A constructivecritique.SagePublications.

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

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

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

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

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