ADJUSTED CONFIDENCE INTERVALS FOR THE SLOPE OF A SIMPLE REGRESSION LINE

Sa-aat Niwitpong* and Chanaphun Chananetr

Department of Applied Statistics, Faculty of Applied Science
King Mongkut's Instite of Technology North Bangkok, Bangkok, Thailand.

ABSTRACT

Nanayakkara and Cressie [1] have presented a confidence interval for the slope (β_1) in the regression model (1) when the error term has non constant variance. For the case of a single predictor and the error terms are normally distributed with variance which is a function of the predictor variable. Their simulation results show that the proposed confidence interval, is preferable to the confidence interval based on the ordinary least squares, OLS method. Later, Wilcox [2] found that confidence intervals based on bootstrap techniques can improve Nanayakkara and Cressie's (NC) method. However, the simulation results show only a comparison of coverage probabilities of confidence intervals for the slope. Unlike Nanayakkara and Cressie and Wilcox, this paper presents a comparison of confidence intervals for the slope of a simple regression using OLS and NC methods when these confidence intervals have minimum coverage probability 1-α. The ratio of the expected lengths of these confidence intervals are compared using Monte Carlo simulation. The conclusion is that the confidence interval for β_1 based on the NC method is less efficient than the confidence interval for β_1 based on the OLS method for some cases. These results are not similar to those of Nanayakkara and Cressie and Wilcox.

KEYWORDS : Simple linear regression model, Heteroscedasticity, Ordinary least-squares method, Coverage probability, Expected Length

1. INTRODUCTION

Consider a simple regression model

$$Y_i \quad = \quad \beta_0 + \beta_1 X_i + \varepsilon_i . \qquad (1)$$

where the set $[(X_i, Y_i); i = 1,2,...,n]$ is a random sample of n pairs of observations and ε_i is the random error which are independent and identically distributed.

Nanayakkara and Cressie [1] constructed a confidence interval for β_1 using a smaller standard error of β_1 th† an the OLS method's. Their method gives a confidence interval which has better coverage probability than a standard OLS's method, when the error term has non-constant variance.

Wilcox [2] extended Nanayakkara and Cressie's results by considering the following cases a) the predictor and error terms are normally distributed, b) the predictor and error terms are non-normally distributed. Wilcox found that a modified percentile-bootstrap technique gives a fairly good coverage probability which is close to a nominal value of 1-α. Nevertheless Wilcox did not consider the expected length of their proposed confidence intervals.

* Corresponding author. Tel.: 066-02-913-2500-24 ext 4910 ; Fax: 066-02-585-6105.
E-mail address: snw@kmitnb.ac.th , chanaphun@yahoo.com

To compare two confidence intervals, Kabaila [3], [4] argued that two confidence intervals can be compared solely based on the ratio of their expected length when these confidence intervals have minimum coverage probabilities 1-α where α is the level of significance.

In this paper, we therefore adjust two confidence intervals for the slope of a regression line based on the NC and the OLS methods, when the error term has non-constant variance, to have minimum coverage probabilities 1-α. Then, the two expected lengths are compared by looking at their ratio. In section 2, we review methods to construct two confidence intervals for β_1. Section 3 gives a simple method to adjust the confidence intervals in order to have minimum coverage probabilities 1-α. Numerical examples are shown in Section 4. The conclusion is presented in Section 5.

2. CONFIDENCE INTERVAL FOR β_1

This section reviews two methods of constructing confidence intervals for β_1, when the distribution of the independent variable and the error terms have non-constant variance.

2.1 OLS Method

We shall find b_1, the estimator of β_1 from the ordinary least-squares method. This estimator is

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

and the (1-α) confidence interval for β_1 is

$$CI_{OLS} = [b_1 - c_1\, s.e.(ols)\, ,\, b_1 + c_1\, s.e.(ols)]\, .$$

where the quantity c_1 is a constant value of confidence interval for which CI_{OLS} has a minimum coverage probability $1-\alpha$. The quantity $s.e.(ols)$ stands for the standard error of b_1 from OLS method and it is calculated from

$$s.e.(ols) = \sqrt{\frac{S^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}.$$

where $S^2 = \sum_{i=1}^{n} e_i^2/(n-2)$ is the usual estimator of σ^2 based on the error $e_i = Y_i - b_1 X_i - b_0$.

where $b_0 = \overline{Y} - b_1\overline{X}$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

The expected length of CI_{OLS} from OLS method can be calculated by

$$E(Length)_{OLS} = E(2c_1\, s.e.(ols)).$$

2.2 NC Method

This method was proposed by Nayakkara and Cressie [1]. It has a smaller standard error of β_1 than the OLS method's. The 1-α confidence interval for the parameter β_1 is

$$CI_{NC} = [b_1 - c_2\, s.e.(NC)\, ,\, b_1 + c_2\, s.e.(NC)]\, .$$

where the quantity c_2 is a constant value of confidence interval for which CI_{NC} has a minimum coverage probability $1-\alpha$. The quantity $s.e.(NC)$ is the standard error of b_1 from NC method and is calculated by first computing

$$d_i = \frac{(n-1)\sum_{i=1}^{n} X_i^2 - (n\overline{X})^2 + 2X_i n\overline{X} + nX_i^2}{n\sum_{i=1}^{n} X_i^2 - (n\overline{X})^2},$$

and

$$s.e.(NC) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 e_i^2 / d_i}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}.$$

The expected length of CI_{NC} from NC method can be calculated by

$$E(Length)_{NC} = E\left(2c_2\, s.e.(NC)\right).$$

3. A SIMPLE METHOD TO ADJUSTED CONFIDENCE INTERVALS

In this section, we propose a simple method to adjust the confidence interval for the slope of a regression line for CI_{OLS}, to have a minimum coverage probability 1-α. The estimated $f(c_{ij})$ were a function of a coverage probability for CI_{OLS} where $c_{ij} = \{1.3, 1.4, \ldots, 4.2\}$; $i = 1,2$, $j = 1,2,\ldots,30$. We carried out Monte Carlo simulations to find the estimated coverage probability for CI_{OLS} where we fixed the values of $n = 20$, $\beta_1 = 1$ and varied over the grid of values c_{ij}. 10,000 samples from the g-and-h distribution were generated for the independent variable and the error term in the four cases specified in Section 4. The resulting coverage probabilities of CI_{OLS}, for four variance configurations; VC1- VC4 (described in section 4), are shown in Figure 1.

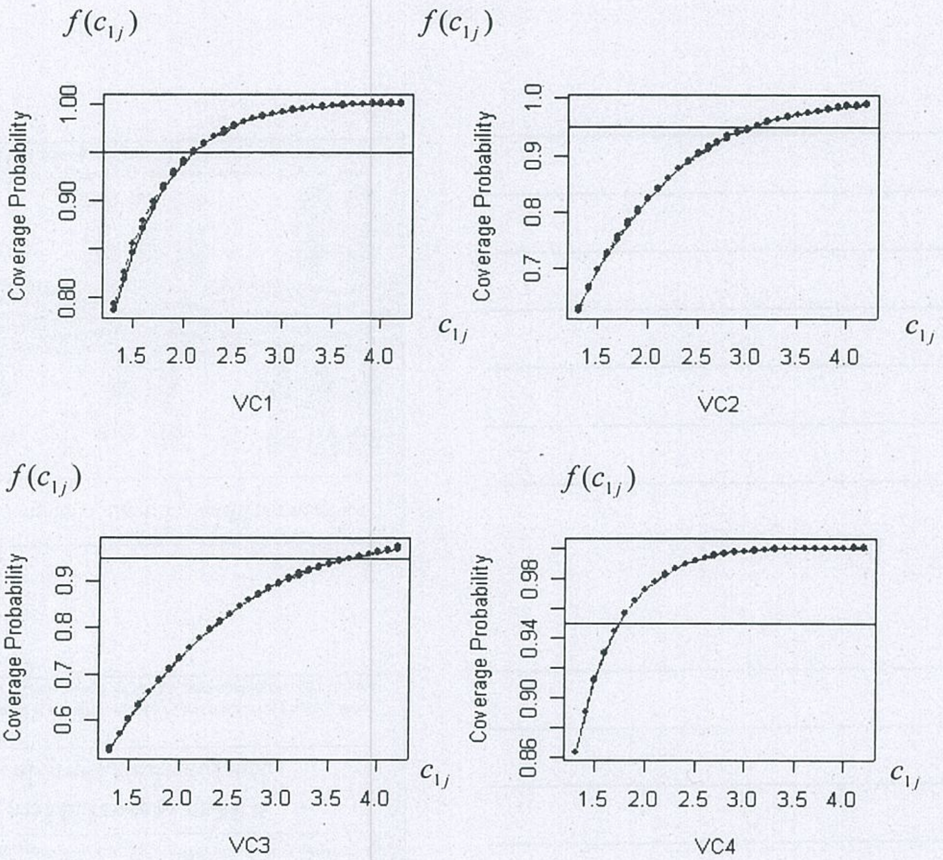

Figure 1. Coverage probabilities of CI_{OLS} method when the independent variable and the error term were a symmetric distribution

Figure 1 shows that the estimated coverage probability for CI_{OLS} is a monotonic increasing function. We therefore fit these estimated coverage probabilities by using the following non-linear regression model

$$f(\hat{c}_{ij}) \quad = \quad \hat{\theta}_1 + \hat{\theta}_2 \exp(\hat{\theta}_3 \, \hat{c}_{ij}) \ ,$$

where $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ are the estimator of θ_1, θ_2 and θ_3 obtained by using non-linear least squares method. A minimum coverage probability 1-α for CI_{OLS} can be found by setting $f(\hat{c}_{ij})$ = 0.95 and solving for $\hat{c}_{ij}$. The results of all simulations and values of $\hat{c}_{ij}$ are partly tabulated in Tables 1 and 2 particularly in some cases of g-and-h distribution. Wilcox [2]. A complete set of results can be obtained upon request to the authors.

4. NUMERICAL EXAMPLES

Wilcox [2] has pointed out that the g-and-h distribution is suited for examining the non-normal distribution since it shows the distribution with skewness and heavy-tailedness.

Following Wilcox [2], we explain how to generate X_i and ε_i in regression model (1). The observations were generated from the g-and-h distribution (Hoaglin [5]) with $g = 0.0, 0.5$ and $h = 0.0, 0.5$. As expressed in Wilcox [2], as g increases skewness increases, and as h increases heavy-tailedness increases. Figure 2 shows the characteristics of the g-and-h distribution. Using g-and-h values, we simulated X_i and ε_i values, where

$$X_i \;=\; \left(\frac{\exp(gZ_i)-1}{g}\right)\exp(hZ_i^2/2),$$

and Z_i are generated from a standard normal distribution.

When $g=0$. The last expression is taken to be

$$X_i \;=\; Z_i\exp(hZ_i^2/2),$$

which has a standard normal distribution if $h=0$.

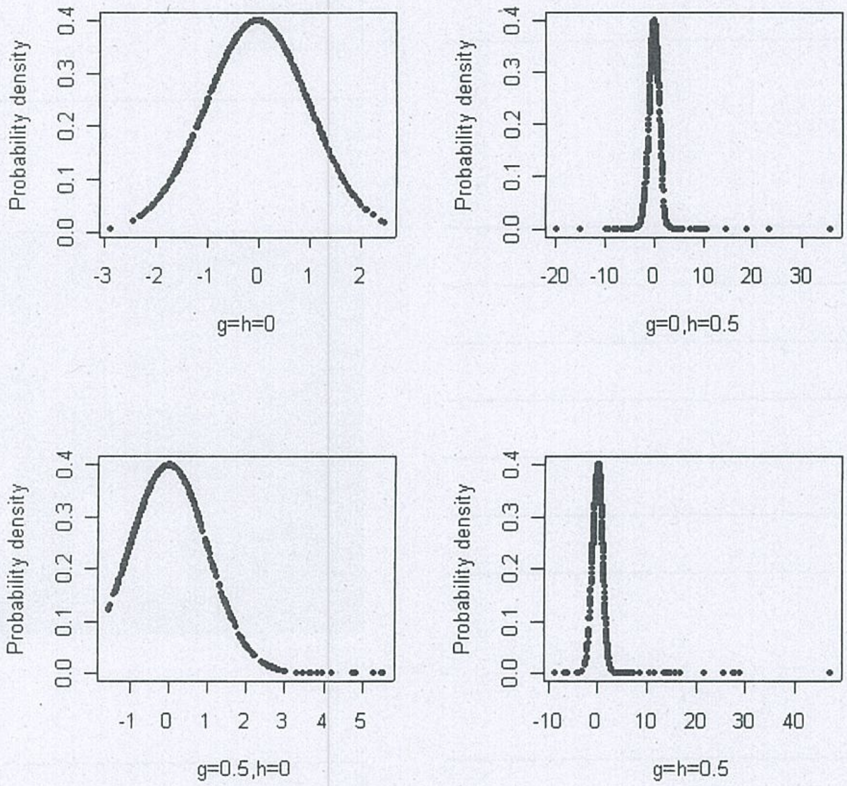

Figure 2 The shape of X_i's when we generated by each g-and-h distribution.

As in Wilcox [2], we now simulate X_i and ε_i from one of the four distributions $[g=h=0$, $g=0, h=0.5$, $g=0.5, h=0$ and $g=h=0.5]$, $i=1,2,\ldots,n$ with X_i independent of ε_i, and then setting

$$Y_i \;=\; \beta_1 X_i + f(X_i)\varepsilon_i\,.$$

when $f(X_i)\varepsilon_i$ is the function of variance configurations, which have four types of variance

For convenience, the four variance configurations (VC) will be referred to as VC1, VC2, VC3 and VC4 respectively. The four VC used were as follows; VC1 is $f(X)=1$, VC2 is $f(X)=\sqrt{|X|}$, VC3 is $f(X)=|X|$ and VC4 is $f(X)=1+\frac{2}{(|X|+1)}$.

The two situations VC2 and VC3 correspond to large error variances when the value of X is in the tails of its distribution, while VC4 is the reverse, an example of which can be seen in Figure 3.

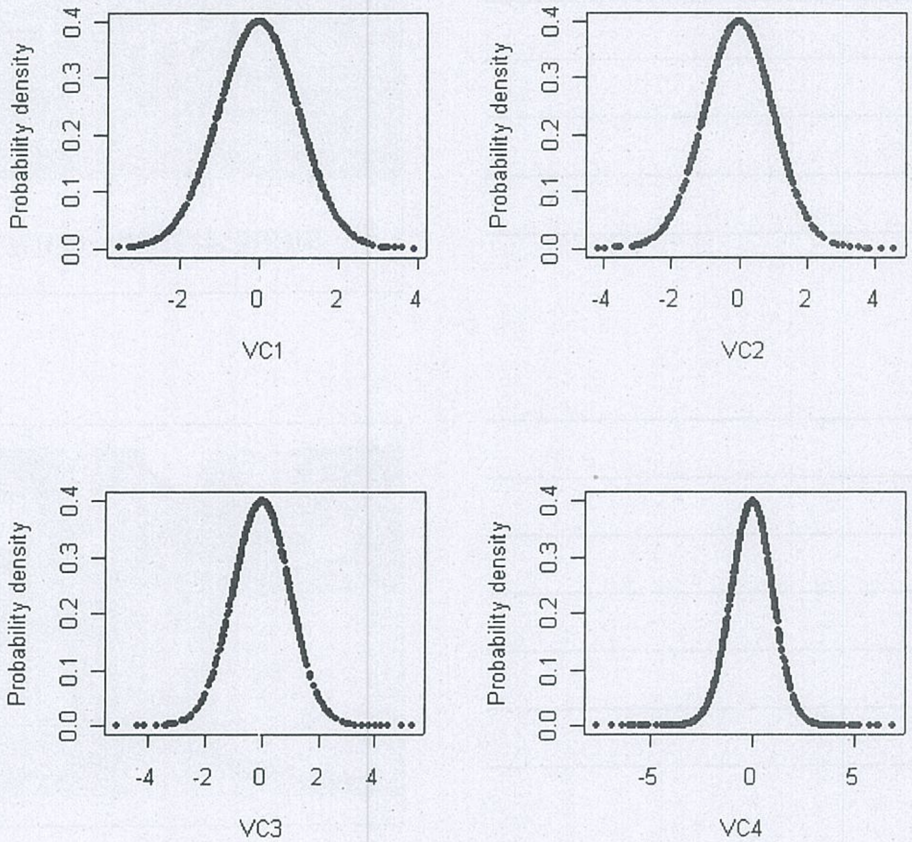

Figure 3 Examples of Y for each VC when X_i and ε_i are distributed with $g=h=0$.

Table 1: Coverage probabilities and expected lengths for CI_{OLS} and CI_{NC} when the g-and-h distribution of X and ε are a symmetric, $g = h = 0$.

VC	Method	$y \sim \hat{\theta}_1 + \hat{\theta}_2 * \exp(-\hat{\theta}_3 * x)$				c_{ij}	Cov.Prob	E (Length)	Ratio #
		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	S.E.				
1	OLS	1.002434	-2.077822	1.763315	0.001845	2.0867	0.9500	0.9829	0.9161
	NC	1.000878	-1.577622	1.296540	0.001044	2.6488	0.9502	1.0729	
2	OLS	1.010890	-1.485754	1.039382	0.002960	3.0736	0.9516	1.2333	0.9290
	NC	0.999262	-1.468490	1.131999	0.001881	2.9980	0.9517	1.3275	
3	OLS	1.028643	-1.346620	0.764695	0.003495	3.7145	0.9518	1.5752	0.9248
	NC	1.001765	-1.442287	1.038579	0.002531	3.2037	0.9518	1.7032	
4	OLS	1.000709	-2.583228	2.253329	0.000555	1.7444	0.9500	1.8745	0.9455
	NC	1.001771	-1.713179	1.426813	0.001580	2.4525	0.9518	1.9825	

is Ratio between E(Length) of OLS and E(Length) of NC

The results from Table 1 show that the CI_{OLS} is preferable to the CI_{NC} in terms of their ratio of expected lengths in the case $g = h = 0$.

When X has a symmetric distribution and ε has increasing heavy-tailedness and VC1, VC2, VC3 and VC4 the NC method gives the better expected lengths and the results are shown in Table 2.

In Table 2 when $g = 0, h = 0.5$ and $g = 0.5, h = 0.5$ X has a symmetric distribution and ε is asymmetric. We found that CI_{NC} is preferable to CI_{OLS}.

However, in Table 3, there are some cases which CI_{OLS} perform better than CI_{NC} in terms of their ratio of expected lengths. These results are different from these of Nanayakkara and Cressie [1].

Table 2: Coverage probability, expected of length and the ratio of the expected lengths between OLS method and NC method when g-and-h distribution of X is a symmetric and ε is a asymmetric, $g = 0, h = 0.5$ and $g = 0.5, h = 0.5$ respectively

VC	Method	$y \sim \hat{\theta}_1 + \hat{\theta}_2 * \exp(-\hat{\theta}_3 * x)$				c_{ij}	Cov.Prob	E (Length)	Ratio
		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	S.E.				
1	OLS	1.002796	-1.884451	1.652927	0.001496	2.1628	0.9528	3.0621	1.1532
	NC	1.001478	-2.922532	1.789181	0.001658	2.2575	0.9520	2.6553	
2	OLS	1.010955	-1.729989	1.207264	0.002382	2.7713	0.9529	3.2033	1.0371
	NC	1.001447	-2.528524	1.568422	0.001377	2.4833	0.9525	3.0887	
3	OLS	1.023260	-1.651356	0.961231	0.003658	3.2410	0.9529	3.9862	1.0655
	NC	1.001081	-2.438134	1.454900	0.001523	2.6569	0.9527	3.7413	
4	OLS	1.001397	-1.946379	1.909163	0.001153	1.9033	0.9526	6.1861	1.1987
	NC	1.001188	-3.245683	1.939006	0.001800	2.1400	0.9526	5.1608	

Table 3: The ratio of the expected coverage probability of length between OLS method and NC method

X		E		Expected lengths of CI_{OLS} / Expected lengths of CI_{NC}			
g	h	g	h	VC1	VC2	VC3	VC4
0	0	0	0	0.9161	0.9290	0.9248	0.9455
0	0.5	0	0	0.7375	0.5664	0.5340	0.8319
0	0.5	0.5	0	0.8221	0.5873	0.5408	0.9261
0.5	0	0	0	0.9020	0.8651	0.8732	0.9538
0.5	0	0.5	0	0.9480	0.8853	0.8775	0.9970
0.5	0.5	0.5	0	0.7073	0.4832	0.4236	0.9066

5. CONCLUSION

As shown in Section 4, the efficiency of OLS method is better than that of the NC method in some cases when the CI_{OLS} and the CI_{NC} are adjusted to have minimum coverage probabilities $1-\alpha$. These results are not the same as those of the Nanayakkara and Cressie [1] and Wilcox [2]'s studies in which they did not adjust confidence intervals to have minimum coverage probabilities 1- α .

Therefore, based on the results in Table 3. The simple and easy, CI_{OLS} is preferable to CI_{NC} in some cases of g -and- h distributions.

REFERENCES

[1] Nanayakkara, N. and N.Cressie, **1991** Robustness to unequal scale and other departures from the classical linear model, in W. Stahel and S. Weisberg Eds, *Directions in robust statistics and diagnostics*, New York, Springer, pp. 65-113.

[2] Wilcox, R. **1996** Confidence Intervals for the Slope of a Regression Line when the error term has nonconstant variance, *Computational Statistics & Data Analysis*, *22*,89-98.

[3] Kabaila, P. **1995** The effect of model selection on confidence region and prediction region, *Econometric Theory, 11*, 537-549.

[4] Kabaila, P. **1998** Valid confidence intervals in regression after variable selection, *Econometric Theory*, *14*, 463-482.

[5] Hoaglin, D.C., **1985** Summarizing shape numerically: the g and h distributions,in:D. Hoaglin, F. Mosteller and J. Tukey Eds , *Exploring data tables, trends and shapes.*, New York, Wiley.