A comparison of simultaneous tolerance intervals in a simple linear regression model

[1]M. Chvosteková

[1]Institute of Measurement Science SAS, Bratislava, Slovakia
Email: chvosta@gmail.com

Abstract. *The derivation of a multiple use confidence interval in the statistical calibration problem can be solved by inverting simultaneous tolerance interval, see Lieberman, Miller, and Hamilton (1967) and Mee, Eberhardt, and Reeve (1991). The simultaneous tolerance intervals in a regression have been recognized and considered in various settings by many authors, but all existing intervals are approximate. We offer numerical comparison of the known methods for constructing simultaneous tolerance intervals for a linear regression. In particular, we compare the Lieberman-Miller method, the Wilson method, the Limam-Thomas method, the modified Wilson method, and the LRTW method based on the estimated of confidence in the specified simple linear regression model.*

Keywords: Linear Regression Model, Tolerance Factor, Simultaneous Tolerance Intervals

1. Introduction

Statistical calibration problem can be accomplished using a simultaneous tolerance intervals (STI), see e.g. [3], [5]. STI in regression are constructed using the vector of observations $Y = (Y_1, \ldots, Y_n)^T$ corresponding to n known independent predictors $\mathbf{x}_1, ..., \mathbf{x}_n$, so that with a confidence level $1-\alpha$, at least a γ proportion of the future observation $Y(\mathbf{x})$-distribution is to be contained in the corresponding tolerance interval, simultaneously for all possible values of predictors $\mathbf{x}$. In simultaneous statistical calibration, sometimes rather called inverse regression, the n pairs $(\mathbf{x}_i, Y_i)$, $i = 1, ..., n$ referred to as calibration data are used to construct confidence intervals for a sequence of unobserved independent predictor values $\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, ...$ corresponding to an infinite sequence of observable variables $Y_{n+1}, Y_{n+2}, \ldots$. Multiple use confidence interval constructed by inverting simultaneous tolerance interval in a linear regression cover the true predictor value with a probability γ and the probability of constructing the interval, based on the same calibration data, is $1-\alpha$. The simultaneous tolerance intervals in a regression have been recognized and considered in various settings by many authors. Lieberman and Miller (SW) in [2] presented an approximation for the case of a simple linear regression. Further suggested methods for computing STI in a linear regression, the Wilson (W) method in [6], Limam-Thomas (PS) method in [4], modified Wilson (MW) method in [4] and the LRTW method (LRTW) in [1], are based on the general confidence-set (GCS) approach. Mee, Eberhardt, and Reeve in [5] obtained the narrowest tolerance intervals, but they considered the STI for limited range of possible values of predictors. All known STI in a regression are derived using various approximations, there is no procedure satisfying the definition of the STI exactly. We present a numerical comparison of SW, W, PS, MW, and LRTW method. For the specified case of a simple linear regression we state estimates of confidence levels of the method for four combinations of pairs $\alpha = \{0.01, 0.05\}, \gamma = \{0.9, 0.95\}$.

2. Simultaneous tolerance intervals in a linear regression model

In this article we study STI for a multiple linear regression. Random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ of n independent observations is represented as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{Z}, \tag{1}$$

where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ is an $n \times q$ $(n > q)$ matrix of rank q, with known constant elements. Vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{q-1})^T$ and standard deviation $\sigma > 0$ represent unknown parameters of the regression model and $\mathbf{Z}$ is an $n \times 1$ vector of standard normal errors, i.e. $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$. Under the assumptions, the least squares estimators of $\boldsymbol{\beta}, \sigma$ are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \qquad \text{and} \qquad S^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - q}. \tag{2}$$

Note that $\hat{\boldsymbol{\beta}} \sim N_q(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ and $(n-q)S^2/\sigma^2 \sim \chi^2_{n-q}$, where χ^2_{n-q} denotes a central chi-square random variable with $n - q$ degrees of freedom. Random variables $\hat{\boldsymbol{\beta}}$ and S^2 are independent.

A tolerance interval is specified by its *content (coverage)* and *confidence level,* denoted $0 < \gamma < 1$ and $0 < 1 - \alpha < 1$, respectively. In practical applications the values are close to one. A future observation of a response at the predictor $\mathbf{x}^T = (1, x_1, \ldots, x_{q-1})^T$ is written as $Y(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta} + \sigma Z$, where $Z \sim N(0,1)$ and $Y(\mathbf{x})$ is assumed to be independent of $\mathbf{Y}$.

For a fixed predictor $\mathbf{x}$, a $(\gamma, 1-\alpha)$ two-sided tolerance interval for a future observation $Y(\mathbf{x})$ is considered in the following (general) form

$$\left\langle \mathbf{x}^T\hat{\boldsymbol{\beta}} - \lambda(\mathbf{x} \mid \gamma, 1-\alpha, \mathbf{X})S,\ \mathbf{x}^T\hat{\boldsymbol{\beta}} + \lambda(\mathbf{x} \mid \gamma, 1-\alpha, \mathbf{X})S \right\rangle, \tag{3}$$

where $\lambda(\mathbf{x} \mid \gamma, 1-\alpha, \mathbf{X})$ is a *tolerance factor* for the given content γ, confidence level $1-\alpha$ and $\mathbf{X}$, for simplification we will use a shorter notation $\lambda(\mathbf{x})$.

The simultaneous $(\gamma, 1-\alpha)$ two-sided tolerance intervals of the form (3) in a linear regression model with normally distributed errors are constructed using vectors of observations $\mathbf{Y}$ such that, with the confidence level $1-\alpha$, at least the proportion γ of the $Y(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta} + \sigma Z$ distribution is to be contained in the corresponding interval, simultaneously for all $\mathbf{x} \in \mathsf{R}^q$.

Let $C(\mathbf{x}; \hat{\boldsymbol{\beta}}, S) = P_{Y(\mathbf{x})}(\mathbf{x}^T\hat{\boldsymbol{\beta}} - \lambda(\mathbf{x})S \le Y(\mathbf{x}) \le \mathbf{x}^T\hat{\boldsymbol{\beta}} + \lambda(\mathbf{x})S \mid \hat{\boldsymbol{\beta}}, S)$ denote the content for the tolerance interval (3), given $\hat{\boldsymbol{\beta}}$ and S. The tolerance factors for all possible predictor values are determined subject to the content and confidence level requirements

$$P_{\hat{\boldsymbol{\beta}},S}(C(\mathbf{x}; \hat{\boldsymbol{\beta}}, S) \ge \gamma \quad \forall \mathbf{x} \in \mathsf{R}^{q\times 1}) = 1 - \alpha. \tag{4}$$

The probability $1-\alpha$ is associated with uncertainty of the outcome of the designed experiment and the probability γ is associated with uncertainty that can be attributed to errors in the future measurements.

Lieberman and Miller in [2] proposed to formulate the tolerance factors for all possible predictor values in a simple form $\lambda(\mathbf{x}) = \lambda^*\delta(\mathbf{x})$, where $\delta(\mathbf{x})$ is the standard error of $\mathbf{x}^T\hat{\boldsymbol{\beta}}$ and they derived the procedure for computing scalar λ^*. The general confidence set (GCS) approach to construction of STI for a linear regression model consists in defining a certain form of the $(1-\alpha)$-level pivotal set $G(\mathbf{X})$ for the pivotal quantities

$$\mathbf{b} = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma} \sim N(0, (\mathbf{X}^T\mathbf{X})^{-1}) \quad \text{and} \quad u = \frac{S}{\sigma},\ (n-q)u^2 \sim \chi^2_{n-q}. \tag{5}$$

That is, it holds $P((\mathbf{b}, u) \in G(\mathbf{X})) = 1 - \alpha$ and the distributions of $\mathbf{b}, u$ are free of unknown parameters $\boldsymbol{\beta}, \sigma$, dependent only on the design matrix $\mathbf{X}$. Furthermore, quantities $\mathbf{b}, u$ are independently distributed. The function $\lambda(\mathbf{x})$ that satisfies Eq. 4 based on a set $G(\mathbf{X})$ is determined to satisfy equation

$$\lambda(\mathbf{x}) = \min\{\lambda : C(\mathbf{x}; \mathbf{b}, u) \geq \gamma \quad (\mathbf{b}, u) \in G(\mathbf{X})\}. \tag{6}$$

The pivotal sets proposed in the Wilson [6], the Limam-Thomas [4] and the modified Wilson [4] methods are constructed with approximate confidence $1-\alpha$, only the pivotal set used in the LRTW method [1] is exact. In addition, the formulas for computing the tolerance factors by the methods were derived using the further approximations, in the Lieberman-Miller [2] and the LRTW methods too. There is no procedure to compute the tolerance factor $\lambda(\mathbf{x})$ satisfying the Eq. 4. In the next section we provide numerical comparison of the mentioned method for the case a simple linear regression, i.e. the future observation at the predictor $\mathbf{x}^T = (1, x)$ is expressed in the form $Y(\mathbf{x}) = \beta_0 + \beta_1 x + \sigma Z$, where β, σ will be specified and $Z \sim N(0,1)$.

3. Simulations

We compare the estimated confidence levels of STI constructed by the Lieberman-Miller (SW), the Wilson (W), the Limam-Thomas (PS), the modified Wilson (MW), and the LRTW (LRTW) method. The approximate values of the confidences are determined based on the 10 000 simulated samples. For case a simple linear regression the first column of design matrix $\mathbf{X}$ consists of ones and the second consist of possible various constants. In particulary, for each sample we obtained the 19-dimensional vectors of observation of normal $N(\mathbf{X}(2, 1.7)^T, 1.5^2 I_n)$ distribution, where the second column of design matrix consists of $n = 19$ values x_i, $i = 1,...,19$ randomly chosen from the range [1,10]. We computed coverage of the tolerance intervals determined by the SW, the W, the PS, the MW, the LRTW method for each sample at the x values -3.0, -2.9, - 2.8, ..., 14.0 as

$$coverage\ (\mathbf{x}) = \Phi\left(\frac{\mathbf{x}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \lambda(\mathbf{x})S}{\sigma}\right) - \Phi\left(\frac{\mathbf{x}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \lambda(\mathbf{x})S}{\sigma}\right),$$

where $\mathbf{x}^T = (1, x)$, $\hat{\boldsymbol{\beta}}, S$ are estimated from the designed experiment and Φ denotes cumulative distribution function of the standard normal distribution.

Table 1. The estimates of the confidence STI computed by the five methods for the four combinations of the content and the confidence values based on 10000 samples.

	$\alpha = 0.01$		$\alpha = 0.05$	
	$\gamma = 0.9$	$\gamma = 0.95$	$\gamma = 0.9$	$\gamma = 0.95$
SW	0.9989	0.9988	0.9933	0.9938
W	0.9989	0.9998	0.9954	0.9942
PS	0.9985	0.9988	0.9894	0.9896
MW	0.9989	0.9988	0.9877	0.9882
LRTW	0.9973	0.9965	0.9809	0.9814

The percentage of the samples, where coverage was at least γ over the grid of 181 predictor values is the estimate of the confidence. Table 1 contains the numerical results for combinations of $\gamma = \{0.90, 0.95\}$, $\alpha = \{0.05, 0.01\}$.

4. Conclusions

STI are defined to cover simultaneously for all $\mathbf{x} \in \mathsf{R}^q$ at least γ proportion of all corresponding $Y(\mathbf{x})$ -distributions with confidence $1-\alpha$. In this simulation study, we have checked the coverages over the finite subset of possible predictor values, therefore the confidence levels presented in Table 1 are only approximate. Based on the results, all the known procedures exceed the nominal level, they are conservative. For the same confidence level and different values of the contents the estimates are close, as we expected. In spite of the different value of confidence the estimates are changed minimally for the same content, noticeable differences are for the PS, the MW, the LRTW method in ascending order. STI determined by the LRTW method are the closest to satisfy the specified requirements for the confidence level and the content.

Acknowledgements

The paper was supported by the Slovak Research and Development Agency (APVV), grant SK-AT-0003-08 and by the Scientific Grant Agency of the Slovak Republic (VEGA), grant 1/0077/09 and 2/0019/10.

References

[1] Chvosteková, M. Determination of Two-sided Tolerance Interval in a Linear Regression Model. *Forum Statisticum Slovacum*, 6 (5): 79 – 84, 2010.

[2] Lieberman, G. J., Miller, R. G., Jr. Simultaneous Tolerance Intervals in Regression. *Biometrika*, 50 (1/2): 155 – 168, 1963.

[3] Lieberman, G. J., Miller, R. G., Jr., and Hamilton, M. A. Simultaneous Discrimination Intervals in Regression. *Biometrika*, 54 (1/2), 133– 145, 1967.

[4] Limam, M. M. T., Thomas, R. Simultaneous Tolerance Intervals for the Linear Regression Model. *Journal of the American Statistical Association*, 83 (403): 801 – 804, 1988.

[5] Mee R. W., Eberhardt, K. R., Reeve, C. P. Calibration and Simultaneous Tolerance Intervals for Regression. *Technometrics,* 33 (2): 211– 219, 1991.

[6] Wilson, A. L. An Approach to Simultaneous Tolerance Intervals in Regression. *The Annals of Mathematical Statistics*, 38 (5): 1536 – 1540, 1967.