Simultaneous Tolerance Intervals for the Linear Regression Model

V. Witkovský, M. Chvosteková

Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia
Email: witkovsky@savba.sk

Abstract. *In this article, we address the problem of constructing the simultaneous tolerance intervals for linear regression. The suggested method is based on inverting the exact likelihood ratio test (LRT) for testing the simple null hypothesis on all parameters of the linear regression model with normally distributed errors, as proposed in Chvosteková and Witkovský (2009).*

Keywords: Linear Regression Model, Exact Likelihood Ratio Test, Simultaneous Tolerance Intervals

1. Introduction

The simultaneous tolerance intervals are important for many measurement procedures. The most common application for simultaneous tolerance intervals is the multiple-use calibration problem; see e.g. Scheffé (1973), Mee *et al.* (1991), and De Gryze et al. (2007). The tolerance intervals has been recognized and considered in various settings by many authors, see e.g. Wilks (1942), Wallis (1951), Wilson (1967), Lieberman and Miller (1963), Lieberman *et al.* (1967), Limam and Thomas (1988), Mee *et al*. (1991), and Krishnamoorthy and Mathew (2009). These simultaneous tolerance intervals are constructed such that with given confidence level $1-\alpha$ at least a specified proportion $1-\gamma$ of the population is contained in the tolerance interval for all possible values of the predictor variates. All known simultaneous tolerance intervals in regression are conservative in that the actual confidence level exceeds the nominal level $1-\alpha$.

Here we suggest a new method for constructing the simultaneous tolerance intervals for linear regression, based on inverting the exact likelihood ratio test (LRT) for testing the simple null hypothesis on all parameters of the linear regression model with normally distributed errors.

2. The New Simultaneous Tolerance Intervals for Liner Regression

Consider the linear regression model $Y = X\beta + \sigma Z$ with normally distributed errors, where Y represents the n-dimensional random vector of response variables, X is the $n \times k$ matrix of non-stochastic explanatory variables (for simplicity, here we assume that X is a full-ranked matrix), β is a k-dimensional vector of regression parameters, Z is n-dimensional vector of standard normal errors, i.e. $Z \sim N(0, I_n)$, and σ is the error's standard deviation, $\sigma > 0$.

Chvosteková and Witkovský (2009) suggested the likelihood ratio test for testing the simple null hypothesis $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ against the alternative $H_1 : (\beta, \sigma) \neq (\beta_0, \sigma_0)$ on all parameters of the linear regression model with normally distributed errors. The LRT rejects the null hypothesis for large values of $\lambda(y)$, the observed value of the test statistic

$$\lambda(Y) = (Y - X\beta_0)'(T - X\beta_0)/\sigma_0^2 - n\log(\hat{\sigma}^2_{ML}/\sigma_0^2) - n, \tag{1}$$

where $\hat{\sigma}^2_{ML} = (Y - X\hat{\beta})'(Y - X\hat{\beta})/n$ and $\hat{\beta} = (X'X)^{-1}X'Y$.

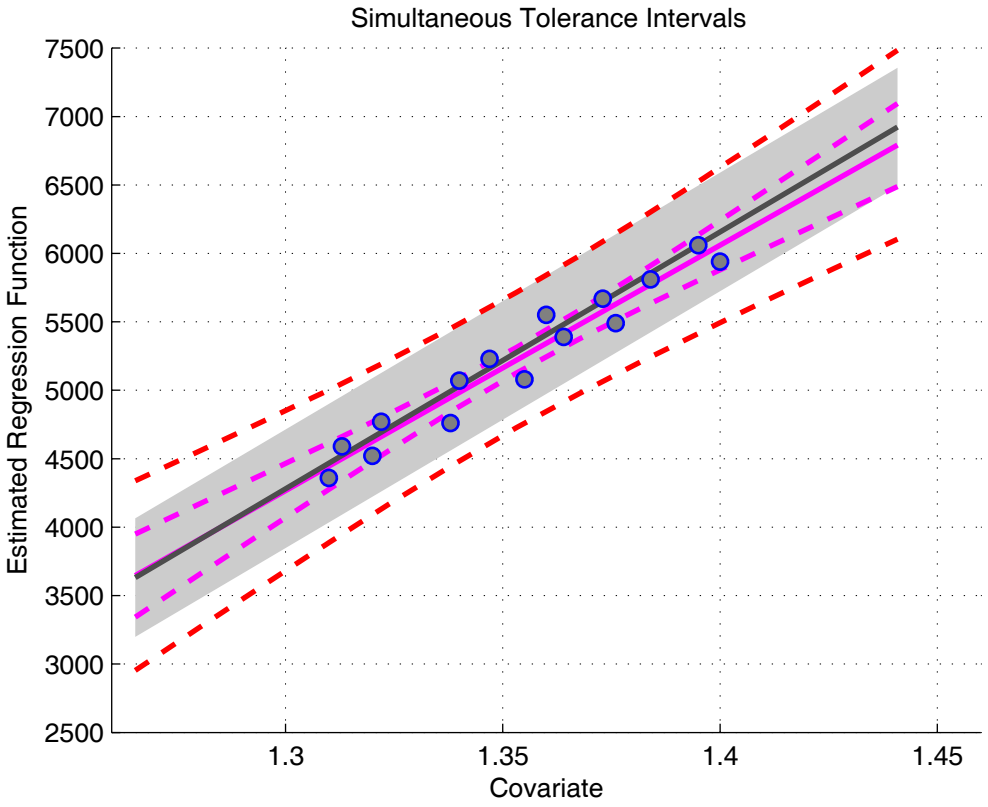

Fig. 1. Illustration of the simultaneous tolerance intervals. The circles represent the observed data. The outer dashed lines represent the simultaneous tolerance interval covering at least the $(1-\gamma)$-content about the true regression function $x'\beta$ with probability at least $(1-\alpha)$, simultaneously for all vectors $x = (x_1, \ldots, x_k)'$ of explanatory variables. The dark solid line represents the true regression function $x'\beta$, together with the $(1-\gamma)$-content region (shaded area) of all possible future observations. The light solid line represents the fitted regression function $x'\hat{\beta}$ together with the simultanoues $(1-\alpha)$-confidence region (light dashed lines) for the true regression line.

So, for the given significance level $\alpha \in (0, 1)$ the test rejects the null hypothesis if

$$\lambda(y) > \lambda_{1-\alpha}, \tag{2}$$

where $\lambda_{1-\alpha}$ is the $(1-\alpha)$-quantile of the distribution of the random variable $\lambda(Y)$, where

$$\lambda(Y) \sim Q_k + Q_{n-k} - n \log\left(Q_{n-k}\right) + n\left(\log(n) - 1\right), \tag{3}$$

and $Q_k \sim \chi^2_k$ and $Q_{n-k} \sim \chi^2_{n-k}$ are two independent random variables with chi-square distributions, with k and $n-k$ degrees of freedom, respectively. The critical values of the test could be easily estimated by Monte Carlo simulations, and/or computed exactly by numerical integration. For more details see Chvosteková and Witkovský (2009), where the critical values of the LRT test, at the significance level $\alpha = 0.05$, are given for normal linear regression models with $k = 1, \ldots, 10$ explanatory variables and $n = k+1, \ldots, 100$ observations.

The exact LR test for testing the simple null hypothesis $H_0 : (\beta, \sigma) = (\beta_0, \sigma_0)$ could be directly used to construct the exact confidence region for the parameters of the linear regression model. In particular, the exact $(1-\alpha)$-confidence region for the parameters β and σ is given as

$$\mathcal{C}_{1-\alpha}(Y) = \{(\beta, \sigma) : \lambda(Y) \leq \lambda_{1-\alpha}\}. \tag{4}$$

Based on that, we define the $(1-\alpha)$-simultaneous tolerance interval covering at least the $(1-\gamma)$-content about the true mean $x'\beta$, for any vector $x=(x_1,\ldots,x_k)'$ of explanatory variables, as

$$\mathcal{T}_{1-\alpha}^{1-\gamma}(x|Y)=\Big[\inf_{(\beta,\sigma)\in\mathcal{C}_{1-\alpha}(Y)}\left\{x'\beta+u_{\gamma_1}\sigma\right\};\sup_{(\beta,\sigma)\in\mathcal{C}_{1-\alpha}(Y)}\left\{x'\beta+u_{1-\gamma_2}\sigma\right\}\Big], \tag{5}$$

where u_{γ_1} and $u_{1-\gamma_2}$ are pre-specified quantiles of the standard normal distribition such that $\gamma=\gamma_1+\gamma_2$, with $\gamma\in(0,1)$. As a special case we get the one-sided tolerance intervals, if $\gamma_1=\gamma$ or $\gamma_2=\gamma$. However, typically the symmetric tolerance intervals about the estimated regression function are used most frequently, with $\gamma_1=\gamma_2=\gamma/2$.
Notice, that directly from the construction of the tolerance intervals $\mathcal{T}_{1-\alpha}^{1-\gamma}(x|Y)$ we get the following probability statement

$$\Pr\Big(\Pr\Big(x'\beta+\sigma Z\in\mathcal{T}_{1-\alpha}^{1-\gamma}(x|Y)\Big)\geq 1-\gamma,\ \text{for all } x \text{ and } Z\sim N(0,1), Z\perp Y\Big)\geq 1-\alpha, \tag{6}$$

where $Z\sim N(0,1)$ is a standard normal random variable stochastically independent of the random vector Y.

3. Monte Carlo Method for Approximate Derivation of the Tolerance Bounds

Derivation of the tolerance bounds given by Eq. (4) requires numerical optimization for given x, α and γ (in particular, γ_1 and/or γ_2 such that $\gamma=\gamma_1+\gamma_2$,), and the observed value y of Y. As an alternative, we suggest a simple algorithm for computing the approximate values of the simultaneous tolerance bounds, based on the Monte Carlo simulations method. The algorithm is based on generating the random sample of size N from the Fisher's fiducial distribution of the regression parameters β and σ, given by

$$\begin{aligned}\tilde{\sigma}^2&=\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{Q_{n-k}^{\star}}=\frac{(n-k)S^2}{Q_{n-k}^{\star}}=\frac{n\hat{\sigma}_{ML}^2}{Q_{n-k}^{\star}},\\ \tilde{\beta}&=\hat{\beta}-\tilde{\sigma}(X'X)^{-1}X'Z^{\star},\end{aligned} \tag{7}$$

with $Q_{n-k}^{\star}\sim\chi_{n-k}^2$ and $Z^{\star}\sim N(0,I)$, the stochastically independent random variables. For more details on applications of the fiducial inference see e.g. Fisher (1935), Hannig et al. (2006), and Hannig (2009).
For given observed data y, and the observed fiducial vector of parameters $(\tilde{\beta},\tilde{\sigma})$ the value of the LR test statistic $\lambda(y|(\tilde{\beta},\tilde{\sigma}))$ is evaluated

$$\begin{aligned}\lambda(y|(\tilde{\beta},\tilde{\sigma}))&=\frac{(y-X\tilde{\beta})'(y-X\tilde{\beta})}{\tilde{\sigma}^2}-n\log\left(\frac{\hat{\sigma}_{ML}^2}{\tilde{\sigma}^2}\right)-n\\ &=\frac{(y-X\hat{\beta}+\tilde{\sigma}X(X'X)^{-1}X'Z)'(y-X\hat{\beta}+\tilde{\sigma}X(X'X)^{-1}X'Z)}{\tilde{\sigma}^2}-n\log\left(\frac{\hat{\sigma}_{ML}^2}{\tilde{\sigma}^2}\right)-n\\ &=\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{\tilde{\sigma}^2}+\frac{\tilde{\sigma}^2Z'X(X'X)^{-1}X'Z}{\tilde{\sigma}^2}-n\log\left(\frac{\hat{\sigma}_{ML}^2}{\tilde{\sigma}^2}\right)-n\\ &=q_{n-k}^{\star}+q_k^{\star}-n\log\left(q_{n-k}^{\star}\right)+n(\log(n)-1),\end{aligned} \tag{8}$$

where $q_k^{\star}$ is the observed value of $Q_k^{\star}\sim\chi_k^2$ and $q_{n-k}^{\star}$ is the observed value of $Q_{n-k}^{\star}\sim\chi_{n-k}^2$, the two independent random variables with chi-square distributions, with k and $n-k$ degrees of freedom, respectively. Notice, that the fiducial confidence region

$$\tilde{\mathcal{C}}_{1-\alpha}(y)=\left\{(\tilde{\beta},\tilde{\sigma}):\lambda(y|(\tilde{\beta},\tilde{\sigma}))\leq\lambda_{1-\alpha}\right\}, \tag{9}$$

where $\lambda_{1-\alpha}$ is the $(1-\alpha)$-quantile of the distribution of the random variable $\lambda(Y)$, given by Eq. (3), is equal to the $(1-\alpha)$-confidence region $\mathcal{C}_{1-\alpha}(y)$, defined in (4). Finally, for any

vector $x = (x_1, \ldots, x_k)'$, chosen α and γ (in particular, γ_1 and/or γ_2 such that $\gamma = \gamma_1 + \gamma_2$), the $(1-\gamma)$-content $(1-\alpha)$-simultaneous tolerance interval for $x'\beta + \sigma Z$ could be approximately evaluated as

$$\mathcal{T}_{1-\alpha}^{1-\gamma}(x|y) = \Big[\min_{(\tilde{\beta},\tilde{\sigma})\in\tilde{\mathcal{C}}_{1-\alpha}(y)} \left\{ x'\tilde{\beta} + u_{\gamma_1}\tilde{\sigma} \right\}; \max_{(\tilde{\beta},\tilde{\sigma})\in\tilde{\mathcal{C}}_{1-\alpha}(y)} \left\{ x'\tilde{\beta} + u_{1-\gamma_2}\tilde{\sigma} \right\} \Big]. \qquad (10)$$

The MATLAB algorithm for computing the approximate values of the simultaneous tolerance bounds, based on the Monte Carlo simulations method, is available upon request from the authors.

Acknowledgements

The research was supported by the grants VEGA 1/0077/09, VEGA 2/7087/27, APVV-SK-AT-0003-09, and APVV-RPEU-0008-06.

References

[1] Chvosteková M. and Witkovský V. Exact likelihood ratio test for the parameters of the linear regression model with normal errors. *Measurement Science Review* 9(1): 1–8, 2009.

[2] Fisher R.A. The fiducial argument in statistical inference. Annals of Eugenics, 6:391–398, 1935.

[3] De Gryze S., Langhans I. and Vandebroek M. Using the correct intervals for prediction: A tutorial on tolerance intervals for ordinary least-squares regression. *Chemometrics and Intelligent Laboratory Systems*, 87: 147–154, 2007.

[4] Hannig J., Iyer H.K., Patterson P. Fiducial generalized confidence intervals. *Journal of the American Statistical Association*, 101:484–499, 2006.

[5] Hannig J.: On generalized fiducial inference. *Statistica Sinica*, submitted (2008). See also the working paper at http://www.stat.colostate.edu/research/2006_3.pdf.

[6] Krishnamoorthy K. and Mathew T. *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley, ISBN: 978-0-470-38026-0, 512 pages, 2009.

[7] Lieberman G. J. and Miller R. G. Jr. Simultaneous tolerance intervals in regression. *Biometrika*, 50:155-168, 1963.

[8] Lieberman G.J. and Miller R.G. Jr. and Hamilton A. Unlimited simultaneous discrimination intervals in regression. *Biometrika*, 54:133-145. Corrections in Biometrika 58:687, 1967.

[9] Limam M.M.T. and Thomas D.R. Simultaneous tolerance intervals for the linear regression model. *Journal of the American Statistical Association*, 83(403): 801-804, 1988.

[10] Mee R.W., Eberhardt K.R. and Reeve, C.P. Calibration and simultaneous tolerance intervals for regression. *Technometrics*, 33(2): 211-219, 1991.

[11] Scheffe H. A statistical theory of calibration. *Annals of Statistics* 1(1):1-37, 1973.

[12] Wallis, W.A. Tolerance intervals for linear regression. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley. CA: University of California Press. Berkeley, 1951.

[13] Wilks. S.S. Statistical prediction with special reference to the problem of tolerance limits. *The Annals of Mathematical Statistics*, 13: 400-409, 1942.

[14] Wilson A.L. An approach to simultaneous tolerance intervals in regression. *The Annals of Mathematical Statistics*, 38(5): 1536-1540, 1967.