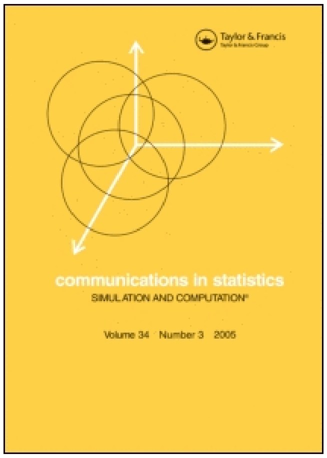

Communications in Statistics - Simulation and Computation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597237

Tolerance Factors in Multiple and Multivariate Linear Regressions

K. Krishnamoorthy [a]; Sumona Mondal [b]
[a] Department of Mathematics, University of Louisiana at Lafayette, Lafayette, Louisiana, USA
[b] Department of Mathematics and Computer Science, Clarkson University, Potsdam, New York, USA

Online Publication Date: 01 March 2008
To cite this Article: Krishnamoorthy, K. and Mondal, Sumona (2008) 'Tolerance Factors in Multiple and Multivariate Linear Regressions', Communications in Statistics - Simulation and Computation, 37:3, 546 - 559
To link to this article: DOI: 10.1080/03610910701812444
URL: http://dx.doi.org/10.1080/03610910701812444

Communications in Statistics—Simulation and Computation®, 37: 546–559, 2008

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910701812444

Regression Analysis

Tolerance Factors in Multiple and Multivariate Linear Regressions

K. KRISHNAMOORTHY[1] AND SUMONA MONDAL[2]

[1]Department of Mathematics, University of Louisiana at Lafayette, Lafayette, Louisiana, USA
[2]Department of Mathematics and Computer Science, Clarkson University, Potsdam, New York, USA

In this article, an improved method of computing tolerance factors for constructing tolerance regions in a multivariate linear regression model is proposed. The method is based on a chi-square approximation to the distribution of a linear function of noncentral chi-square variables and simulation. The merits of the proposed approach and the usual simulation method considered in Lee and Mathew (2004) are evaluated using Monte Carlo simulation. The study indicates that the proposed approach is stable and accurate even for small samples, and better than available methods. For constructing two-sided tolerance intervals in multiple linear regression, coverage level adjusted one-sided tolerance factors are shown to be better than available approximate tolerance factors. The results based on the coverage level adjusted one-sided tolerance factors are as good as the ones based on the exact two-sided tolerance factors in many cases.

Keywords Confidence; Content; Coverage probability; Wallis' approximation; Wishart Distribution.

Mathematics Subject Classification 62F25; 62J05.

1. Introduction

Tolerance limits are routinely used, among others, in acceptance sampling plan, for setting tolerance specification of engineering products, and assessing pollution level in a workplace. In general, tolerance intervals are used to assess the proportion of the population within some specification limits, and one-sided tolerance limits are used to estimate the proportion of the population below or above a threshold value. A tolerance interval in a multiple linear regression is constructed so that it would contain a specified percentage of all future measurements for a given set of explanatory variable values. A tolerance interval that is constructed so that it would

Received May 28, 2006; Accepted July 17, 2007
Address correspondence to K. Krishnamoorthy, Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504, USA; E-mail: krishna@louisiana.edu

contain at least α proportion of the population with confidence γ is referred to as (α, γ) tolerance interval or α content – γ coverage tolerance interval.

Methods for constructing tolerance intervals for some continuous univariate populations and for univariate linear regression models are widely available in the literature (e.g., see Graybill and Iyer, 1994, Sec. 6.13). However, only limited results are available for the multivariate case. Wald (1942) and Guttman (1970) provided large sample approach for computing tolerance regions for a multivariate normal population. John (1963) developed the theoretical framework for the problem of constructing tolerance region and provided some simple approximations for computing the tolerance factors for small samples. Using the framework of John (1963), Krishnamoorthy and Mathew (1999) developed an approximate method and compared it with several approximate methods including those given in John's article via Monte Carlo simulation. Comparison studies by these authors showed that their approximate method is the best among other methods even though it is not satisfactory for all situations. Their overall recommendation was to use "two-stage simulation" (simulation involving two nested do loops) to compute tolerance factors for all sample size – dimension configurations.

Haq and Rinco (1976), Khan and Haq (1994), Kibria and Haq (1999), and Krishnamoorthy (2006) considered the problems of constructing tolerance region subject to the criterion considered in Guttman's (1970) article. Lee and Mathew's (2004) article seems to be one of the earlier work addressed the problem of constructing tolerance region in a multivariate regression model with normal error vectors subject to the probability requirements given in the first paragraph of this section. To describe the problem, consider the multivariate linear regression model

$$\mathbf{y}_i = \boldsymbol{\beta}_0 + \mathbf{B}\mathbf{x}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $\mathbf{y}_i$ is a $p \times 1$ response vector, $\mathbf{x}_i$ is a known $m \times 1$ explanatory vector, $\boldsymbol{\beta}_0$ is a $p \times 1$ intercept vector and $\mathbf{B}$ is a $p \times m$ matrix of slope parameters. It is assumed that $\boldsymbol{\epsilon}_i$'s are independent $p \times 1$ random vectors with $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite matrix. Let $\mathbf{y}_h$ be a future observation independent of $\mathbf{y}_i$'s with known $\mathbf{x}_h$. That is,

$$\mathbf{y}_h = \boldsymbol{\beta}_0 + \mathbf{B}\mathbf{x}_h + \boldsymbol{\epsilon}_h, \tag{1.2}$$

where $\boldsymbol{\epsilon}_h \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ independently of $\boldsymbol{\epsilon}_i$'s in (1.1). Let

$$\mathbf{Y}_{p\times n} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \quad \text{and} \quad \mathbf{X}_{m\times n} = (\mathbf{x}_1, \ldots, \mathbf{x}_n), \tag{1.3}$$

and assume that the rank of $\mathbf{X}$ is m. In terms of these matrices, we can write the above models with the distributional assumptions as

$$\mathbf{Y} \sim N(\boldsymbol{\beta}_0 \mathbf{1}_n' + \mathbf{B}\mathbf{X}, I_n \otimes \boldsymbol{\Sigma}) \quad \text{independently of } \mathbf{y}_h \sim N_p(\boldsymbol{\beta}_0 + \mathbf{B}\mathbf{x}_h, \boldsymbol{\Sigma}), \tag{1.4}$$

where $\mathbf{I}_n \otimes \boldsymbol{\Sigma}$ is the variance-covariance matrix of the $np \times 1$ vector obtained by stacking the columns of $\mathbf{Y}$ one below another. The problem of interest here is to obtain tolerance region for the distribution of $\mathbf{y}_h$ in (1.4) based on the model data $(\mathbf{Y}, \mathbf{X})$ and $\mathbf{x}_h$.

The problem of constructing tolerance region for the distribution of $\mathbf{y}_h$ in (1.4) is similar to that of constructing tolerance region for a multivariate normal distribution except that in the former case the tolerance factors depend on $\mathbf{x}_h$ and the known design matrix, whereas in the latter case the tolerance factors do not depend on any data. Noting this similarity and following the idea of Krishnamoorthy and Mathew (1999), Lee and Mathew (2004) developed two approximate methods for the present problem. Their simulation studies showed that neither of the approximate method dominates the other for all (n, p) configurations. They recommended a two-stage simulation for situations where their approximations are not satisfactory.

This article is a followup work by Lee and Mathew (2004) and Krishnamoorthy and Mondal (2006). Specifically, using a chi-square approximation for the distribution of a linear combination of independent noncentral chi-square random variables, Krishnamoorthy and Mondal (2006) proposed a single-stage simulation for computing tolerance factors for a multivariate normal population. Their simulation studies indicated that this single-stage simulation is more stable and accurate than the two-stage simulation for computing tolerance factors for a multivariate normal distribution. We extend the single-stage simulation to the present setup, and evaluate its merits and other approximate methods using Monte Carlo simulation.

This article is organized as follows. In the following section, we give some preliminary results and describe the tolerance region in multivariate linear regression. In Sec. 3, we outline the exact method of computing tolerance factors in multiple linear regression by Wald and Wolfowitz (1946) and two approximate methods. We also study the accuracies of the tolerance intervals based on coverage level adjusted one-sided tolerance factors which are easier to compute than the exact two-sided factors by Wald and Wolfowitz approach. In Sec. 4, we present two approximate methods due to Lee and Mathew, and a new single-stage simulation approach for computing tolerance factors in the multivariate linear regression model. In Sec. 5, the validity of the approximate methods are judged by comparing them with the exact ones for multiple linear regression; for the multivariate case, the merits of the proposed approach and the two-stage simulation are evaluated by Monte Carlo simulation. Some concluding remarks are given in Sec. 6.

2. Preliminaries and the Tolerance Region

For the sake of completeness and convenience, we shall give some preliminaries along the lines of Lee and Mathew (2004).

Let

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i, \quad \text{and} \quad \mathbf{P} = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right), \tag{2.1}$$

where $\mathbf{x}_i$'s and $\mathbf{y}_i$'s are as defined in (1.3). The least square estimators of $\mathbf{B}$ and $\boldsymbol{\beta}_0$ are given, respectively, by

$$\widehat{\mathbf{B}} = \mathbf{YPX}'(\mathbf{X}'\mathbf{PX})^{-1} \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_0 = \bar{\mathbf{y}} - \widehat{\mathbf{B}}\bar{\mathbf{x}}. \tag{2.2}$$

Let $\mathbf{A}$ denote the residual sum of squares and cross-product matrix based on model (1.1). Then

$$\begin{aligned}\mathbf{A} &= (\mathbf{Y} - \widehat{\boldsymbol{\beta}}_0' - \widehat{\mathbf{B}}\mathbf{X})(\mathbf{Y} - \widehat{\boldsymbol{\beta}}_0' - \widehat{\mathbf{B}}\mathbf{X})' \\ &= \mathbf{Y}[\mathbf{I}_n - \mathbf{PX}'(\mathbf{XPX}')^{-1}\mathbf{XP}]\mathbf{Y}',\end{aligned} \tag{2.3}$$

where $\mathbf{P}$ is as defined in (2.1). Furthermore, $\bar{\mathbf{y}}$, $\widehat{\mathbf{B}}$ and $\mathbf{A}$ are independent with

$$\bar{\mathbf{y}} \sim N_p\left(\boldsymbol{\beta}_0 + \mathbf{B}\bar{\mathbf{x}}, \frac{1}{n}\boldsymbol{\Sigma}\right), \quad \widehat{\mathbf{B}} \sim N(\mathbf{B}, (\mathbf{X}P\mathbf{X})^{-1}), \quad \text{and} \quad \mathbf{A} \sim W_p(n-m-1, \boldsymbol{\Sigma}), \tag{2.4}$$

where $W_p(n-m-1, \boldsymbol{\Sigma})$ is the Wishart distribution with df $= n-m-1$ and the scale parameter matrix $\boldsymbol{\Sigma}$. Let us assume that $n-m-1 \geq p$ so that $\mathbf{A}^{-1}$ exists with probability one.

As the predictor of $\mathbf{y}_h$ is $\widehat{\boldsymbol{\beta}}_0 + \widehat{\mathbf{B}}\mathbf{x}_h = \bar{\mathbf{y}} + \widehat{\mathbf{B}}(\mathbf{x}_h - \bar{\mathbf{x}})$, and $\frac{1}{n-m-1}\mathbf{A}$ is an unbiased estimator of $\boldsymbol{\Sigma}$, we would like to construct the tolerance region of the form

$$\left\{\mathbf{y}_h : f[\mathbf{y}_h - \bar{\mathbf{y}} - \widehat{\mathbf{B}}(\mathbf{x}_h - \bar{\mathbf{x}})]'\mathbf{A}^{-1}[\mathbf{y}_h - \bar{\mathbf{y}} - \widehat{\mathbf{B}}(\mathbf{x}_h - \bar{\mathbf{x}})] \leq k(\mathbf{x}_h)\right\}, \tag{2.5}$$

where $f = n-m-1$ and the tolerance factor $k(\mathbf{x}_h)$ is to be determined subject to the content and coverage requirement. Let the independent random quantities

$$\mathbf{V} \sim W_p(f, \mathbf{I}_p), \quad \mathbf{q} \sim N_p(\mathbf{0}, d^2\mathbf{I}_p), \quad \text{and} \quad \mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I}_p), \tag{2.6}$$

where

$$d^2 = \left(\frac{1}{n} + c^2\right) \quad \text{and} \quad c^2 = (\mathbf{x}_h - \bar{\mathbf{x}})'(\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}(\mathbf{x}_h - \bar{\mathbf{x}}). \tag{2.7}$$

Using the distributional results in (1.4) and (2.4), and invariance argument, Lee and Mathew (2004) showed that the tolerance factor is a function of d^2, say, $k(d^2)$, and satisfies

$$P_{\mathbf{q},\mathbf{V}}\{P_{\mathbf{z}}[f(\mathbf{z}-\mathbf{q})'\mathbf{V}^{-1}(\mathbf{z}-\mathbf{q}) \leq k(d^2) \,|\, \mathbf{q}, \mathbf{V}] \geq \alpha\} = \gamma. \tag{2.8}$$

Furthermore, using the fact that the quadratic form in (2.8) is invariance under orthogonal transformation, it can be easily verified that (2.8) is equivalent to

$$P_{\mathbf{q},\mathbf{L}}\left\{P_{\mathbf{z}}\left[f\sum_{i=1}^{p}\frac{(z_i - q_i)^2}{l_i} \leq k(d^2) \,\middle|\, \mathbf{q}, \mathbf{L}\right] \geq \alpha\right\} = \gamma, \tag{2.9}$$

where $f = n-m-1$, $\mathbf{z} = (z_1, \ldots, z_p)$, $\mathbf{q} = (q_1, \ldots, q_p)$, and $\mathbf{L} = \text{diag}(l_1, \ldots, l_p)$, $l_1 > \cdots > l_p$ are the eigenvalues of $\mathbf{V}$.

3. Tolerance Factors in Multiple Linear Regression

In general, it is difficult to obtain explicit expression for $k(d^2)$ satisfying (2.9) or evaluate numerically except for the case of $p = 1$ (multiple linear regression). In the following, we shall outline some methods for computing $k(d^2)$ in the multiple linear regression model.

3.1. *An Exact Method*

For $p = 1$, the probability requirement in (2.8) simplifies to

$$P_{h,V}\left(P_z\left(\frac{(z-q)^2}{V} \leq k_1 \,\middle|\, h, V\right) \geq \alpha\right) = \gamma, \tag{3.1}$$

where $z \sim N(0, 1)$, $q \sim N(0, d^2)$, and $V \sim \chi^2_f/f$ with $f = n - m - 1$. It can be easily seen that the problem is equivalent to the one for computing tolerance factors for a normal distribution; the only difference is that for the normal case $q \sim N(0, 1/n^2)$ whereas in the present problem it follows $N(0, d^2)$. Therefore, the tolerance factor $k(d^2)$ can be computed following the lines of Wald and Wolfowitz's (1946) approach. The Fortran program due to Eberhardt et al. (1989) uses Wald and Wolfowitz's approach for computing the tolerance factors for a normal distribution.

Even though $k(d^2)$ can be computed (for the case of $p = 1$) using Wald and Wolfowitz's integral equation, we present a slightly different version of their integral equation as follows. For convenience, let $k = k(d^2)$. It is easy to see that the expression in (3.1) can be written as

$$P_{q,V}(\Phi(q + \sqrt{kV}) - \Phi(q - \sqrt{kV}) \geq \alpha) = \gamma, \tag{3.2}$$

where Φ is the standard normal cumulative distribution function (cdf). Note that, for fixed y, $\Phi(y + x) - \Phi(y - x)$ is an increasing function of x. Therefore, $\Phi(y + x) - \Phi(y - x) > \alpha$ if and only if $x > r$, where r is the solution of the equation $\Phi(y + r) - \Phi(y - r) = \alpha$. So $\Phi(q + \sqrt{kV}) - \Phi(q - \sqrt{kV}) \geq \alpha$ if and only if $\sqrt{kV} > r$ or $V > \frac{r^2}{k}$. It is not difficult to check that the solution of

$$\Phi(q + r) - \Phi(q - r) = \alpha \tag{3.3}$$

is given by $r^2 = \chi^2_{1,\alpha}(q^2)$, where $\chi^2_{m,\alpha}(\delta)$ denotes the αth quantile of a noncentral chi-square distribution with df $= m$ and the noncentrality parameter δ. Thus, (3.2) is equivalent to

$$E_q\left[P_V\left(V \geq \frac{\chi^2_{1,\alpha}(q^2)}{k} \,\middle|\, q\right)\right] = \gamma,$$

or, equivalently,

$$\frac{1}{\sqrt{2\pi}d}\int_{-\infty}^{\infty} P\left(\chi^2_f > \frac{f\chi^2_{1,\alpha}(q^2)}{k}\right)\exp\left(-\frac{q^2}{2d^2}\right)dq = \gamma, \tag{3.4}$$

where $f = n - m - 1$ and d^2 is defined in (2.7). It should be noted that Wald and Wolfowitz's integral equation is the one in (3.4) with $\chi^2_{1,\alpha}(q^2)$ replaced by r^2, where r is the solution of the equation $\Phi(q + r) - \Phi(q - r) = \alpha$. As we already noted, $r = \sqrt{\chi^2_{1,\alpha}(q^2)}$, and it can be easily computed using available function routines (e.g., CSNIN in IMSL). Therefore, for programing convenience, we prefer (3.4) to Wald and Wolfowitz's integral equation.

3.2. *Wallis' Method*

Wallis (1951) derived an approximation to the tolerance factor for the multiple linear regression by extending Wald and Wolfowitz (1946) method for the univariate normal distribution, and is given by

$$k_W(d^2) = \frac{n-m-1}{\chi^2_{n-m-1;\alpha}} z^2_{\frac{1+\alpha}{2}} \left[1 + \frac{d^2}{2} - \frac{d^4\left(2z^2_{\frac{1+p}{2}} - 3\right)}{24}\right]^2. \tag{3.5}$$

The above approximation is satisfactory only when d^2 is small. Lee and Mathew (2004) proposed the following approximation which is a special case of two approximations given for the multivariate linear regression (see Sec. 4.1), and is given by

$$k_{11}(d^2) = \frac{1+d^2}{1+\delta}\chi^2_{1,\alpha} F_{e,n-m-1,\gamma}, \tag{3.6}$$

where $e = (1+d^2)^2/d^4$, $\delta = d^2\left[\frac{3d^2+\sqrt{9d^4+6d^2+3}}{2d^2+1}\right]$ and $F_{m_1,m_2,\gamma}$ is the γth quantile of an F distribution with dfs m_1 and m_2.

3.3. *Coverage Level Adjusted One-Sided Tolerance Factors*

Exact tolerance factors based on noncentral t percentiles are available to compute one-sided tolerance limits for the response variable in multiple linear regression. This one-sided tolerance factor with an adjustment to confidence level can be used as an approximation to the factor for constructing two-sided tolerance limits. In particular, we consider one-sided tolerance factor with confidence level $(1+\gamma)/2$ (instead γ) as an approximation to $k(d)$. That is,

$$k_{1o}(d) \simeq d \times t_{n-m-1;\frac{1+\gamma}{2}}\left(\frac{z_p}{d}\right), \tag{3.7}$$

where $t_{m;p}(\delta)$ denotes the pth quantile of a noncentral t distribution with df $= m$ and noncentrality parameter δ.

4. Tolerance Factors in Multivariate Linear Regression

We shall first present the two approximations to the tolerance factor $k(d^2)$ by Lee and Mathew (2004), and then we describe the new approach.

4.1. *Lee and Mathew's Approach*

Lee and Mathew (2004) proposed two approximations, namely, $k_{11}(d^2)$ and $k_{12}(d^2)$ for the factor $k(d^2)$ satisfying (2.9). To express them, we first need to define the following quantities. Let

$$e = \frac{p(1+d^2)^2}{d^4}, \quad f = \frac{d^4}{1+d^2}, \quad \text{and}$$
$$\delta = d^2\left[\frac{d^2(p+2) + \sqrt{d^4(p+2)^2 + (2d^2+1)p(p+2)}}{2d^2+1}\right]. \tag{4.1}$$

Then

$$k_{11}(d^2) = \frac{(n-m-1)ef}{(n-m-p)(p+\delta)}\chi^2_{p,\alpha}(\delta)F_{e,n-m-p,\gamma}, \tag{4.2}$$

where $\chi^2_{p,\alpha}$ denotes the αth quantile of a noncentral chi-square distribution with df $= p$ and noncentrality parameter δ. To express the second approximation $k_{12}(d^2)$ by Lee and Mathew, let

$$\begin{aligned} e_{22} &= \frac{4p(n-m-p-1)(n-m-p) - 12(p-1)(n-m-p-2)}{3(n-m-2) + p(n-m-p-1)}, \\ e_{12} &= \frac{n-m-p-2}{e_{22}-2}. \end{aligned} \tag{4.3}$$

Then

$$k_{12}(d^2) = \frac{(n-m-1)ef}{e_{12}e_{22}(p+\delta)}\chi^2_{p,\alpha}(\delta)F_{e,e_{22},\gamma}. \tag{4.4}$$

It is easy to check that $k_{11}(d^2)$ and $k_{12}(d^2)$ are the same when $p = 1$, and is given in (3.6).

4.2. *Two-Stage Simulation*

Lee and Mathew (2004) simulation studies show that there is no clear-cut winner between the two approximations, and neither of them is satisfactory for all situations. They recommended the following two-stage simulation for the situations where their approximations are not satisfactory.

Algorithm 4.1. For a given d^2, $n-m-1$, p, α and γ:

For $i = 1, m_1$
  generate a $\mathbf{q} \sim N_p(0, d^2 I_p)$ and a $\mathbf{V} \sim W_p(n-m-1, \mathbf{I}_p)$
  For $j = 1, m_2$
    generate a $\mathbf{z} \sim N_p(0, \mathbf{I}_p)$
    compute $Q_j = (n-m-1)(\mathbf{z}-\mathbf{q})'\mathbf{V}^{-1}(\mathbf{z}-\mathbf{q})$
  end j loop
  compute $T_i = 100\alpha$th percentile of the Q_j's
end i loop

The 100γth percentile of the T_i's is a Monte Carlo estimate of the tolerance factor.

4.3. *The New Method*

The new method is essentially the one in Krishnamoorthy and Mondal (2006), and it can be described as follows. Using Imhof's (1961) approximation to a linear function of noncentral chi-square random variables, Krishnamoorthy and Mondal (2006) showed that

$$\sum_{i=1}^{p} \frac{(z_i - q_i)^2}{l_i} \sim \sqrt{\frac{c_2}{a}}(\chi^2_a - a) + c_1 \quad \text{approximately}, \tag{4.5}$$

where

$$c_j = \sum_{i=1}^{p} \frac{1 + jq_i^2}{l_i^j}, \quad j = 1, 2, 3 \text{ and } a = \frac{c_2^3}{c_3^2}. \tag{4.6}$$

Using the approximation (4.5), we see that the inner probability inequality in (2.9) holds

$$P\left[f\left(\sqrt{\frac{c_2}{a}}(\chi_a^2 - a) + c_1\right) \le k(d^2) \middle| \mathbf{q}, \mathbf{L}\right] \ge \alpha. \tag{4.7}$$

The probability inequality in (4.7) holds if and only if $k(d^2) \ge f(\sqrt{\frac{c_2}{a}}(\chi_{a,\alpha}^2 - a) + c_1)$, where $\chi_{a,\alpha}^2$ denotes the αth quantile of the χ_a^2 distribution, and hence it follows from (2.9) that an approximation to the tolerance factor $k(d^2)$ satisfies

$$P_{\mathbf{q},\mathbf{L}}\left((n - m - 1)\left(\sqrt{\frac{c_2}{a}}(\chi_{a,\alpha}^2 - a) + c_1\right) \le k(d^2)\right) = \gamma. \tag{4.8}$$

To compute $k(d^2)$ using Monte Carlo simulation, we need to generate only the eigenvalues $l_1, \ldots, l_p$ and $q_1^2, \ldots, q_p^2$. It follows from (2.6) that $q_1^2, \ldots, q_p^2$ are independent and identically distributed as $d^2\chi_1^2$ random variable. Using these results, we can obtain Monte Carlo estimates of $k(d^2)$ as shown in the following algorithm.

Algorithm 4.2. For a given n, p, $d^2 = (1/n + c^2)$, α and γ:

For $i = 1, NR$
  generate $q_1^2 \sim d^2\chi_1^2, \ldots, q_p^2 \sim d^2\chi_1^2$
    $\mathbf{V} \sim W_p(n - m - 1, \mathbf{I}_p)$
  compute the eigenvalues $l_1, \ldots, l_p$ of $\mathbf{V}$
  compute c_1, c_2, c_3 and a using (4.6)
  set $T_i = (n - m - 1)(\sqrt{\frac{c_2}{a}}(\chi_{a,\alpha}^2 - a) + c_1)$
end i loop

The 100γth percentile of the T_i's is a Monte Carlo estimate of the tolerance factor $k(d^2)$. We also note that the only difference between Algorithm 2 of Krishnamoorthy and Mondal (2006) and the above algorithm is that in the former $q_i^2 \sim \chi_1^2/n$ and the df associated with $\mathbf{V}$ is $n - 1$ whereas in the latter $q_i^2 \sim d^2\chi_1^2$ and the df associated with $\mathbf{V}$ is $n - m - 1$.

5. Monte Carlo Evaluation of the Approximations

For the case of $p = 1$, we first compare the approximate factors with the exact ones based on the integral equation in (3.4). We computed (1) the exact tolerance factors, (2) the coverage level adjusted one-sided tolerance factors in (3.7), (3) the ones based on Algorithm 4.2, (4) the approximate factors $k_{11}(d^2)$ in (3.6) by Lee and Mathew (2004), and (5) the approximate factors due to Wallis (1951) for $n - m - 1 = 10$. Mee et al. (1991) and Lee and Mathew (2004) argued that, in practical applications, the values of d^2 range from 0 to 1. So we compared the factors for $d^2 = .1, .3, .5, .8$, and 1. These factors are reported in Table 1 for all combinations of (α, γ) from the set $\{.9, .95, .99\}$. We see from Table 1 that all methods, except the one based on the one-sided tolerance factors, are producing factors close to the exact ones for small values of d^2. For $d^2 \geq .3$, the factors $k_{11}(d^2)$ are larger than the exact ones and the Wallis factors are much smaller than the exact ones. In general, Wallis' tolerance intervals are expected to be very liberal, and the ones by Lee and Mathew are very conservative. The factors based on Algorithm 4.2 seem to be very satisfactory for all values of d^2. Considering simplicity and accuracies, the coverage level adjusted one-sided tolerance factors seem to be the best for values of $d^2 \geq .3$. We also made comparisons when $n - m - 1 = 20$ and 40. These values are not reported here (they

Table 1
Comparison of the exact and approximate tolerance factors in multiple linear regression $n - m - 1 = 10$

		$p = .90$			$p = .95$			$p = .99$		
		γ			γ			γ		
d^2	Methods	.90	.95	.99	.90	.95	.99	.90	.95	.99
.1	1	2.49	2.77	3.45	2.95	3.29	4.09	3.86	4.30	5.35
	2	2.29	2.56	3.21	2.83	3.15	3.92	3.87	4.28	5.30
	3	2.48	2.76	3.42	2.94	3.28	4.08	3.86	4.30	5.37
	4	2.49	2.77	3.45	2.96	3.29	4.10	3.87	4.31	5.36
	5	2.47	2.75	3.41	2.94	3.27	4.06	3.86	4.29	5.33
.3	1	2.73	3.08	3.90	3.21	3.60	4.55	4.14	4.62	5.80
	2	2.64	2.99	3.82	3.15	3.54	4.48	4.13	4.61	5.77
	3	2.72	3.04	3.87	3.20	3.60	4.51	4.16	4.66	5.84
	4	2.77	3.11	3.93	3.25	3.65	4.60	4.15	4.66	5.88
	5	2.69	2.99	3.71	3.18	3.54	4.39	4.14	4.60	5.70
.5	1	2.96	3.36	4.31	3.44	3.88	4.95	4.36	4.89	6.18
	2	2.90	3.31	4.28	3.40	3.85	4.92	4.35	4.88	6.16
	3	2.93	3.32	4.28	3.43	3.87	4.95	4.40	4.94	6.26
	4	3.02	3.43	4.38	3.49	3.96	5.07	4.37	4.96	6.35
	5	2.88	3.20	3.97	3.39	3.76	4.67	4.35	4.84	6.00
.8	1	3.25	3.72	4.85	3.73	4.24	5.47	4.64	5.24	6.67
	2	3.22	3.70	4.84	3.70	4.22	5.46	4.64	5.23	6.66
	3	3.22	3.71	4.86	3.72	4.24	5.48	4.71	5.33	6.80
	4	3.33	3.82	4.97	3.78	4.35	5.65	4.64	5.33	6.93
	5	3.13	3.48	4.31	3.64	4.05	5.03	4.62	5.13	6.37
1	1	3.42	3.94	5.17	3.90	4.46	5.78	4.81	5.44	6.96
	2	3.40	3.93	5.16	3.88	4.44	5.77	4.80	5.44	6.96
	3	3.39	3.91	5.14	3.90	4.44	5.77	4.89	5.54	7.10
	4	3.50	4.05	5.31	3.95	4.57	5.99	4.79	5.54	7.27
	5	3.28	3.64	4.52	3.79	4.22	5.23	4.77	5.30	6.58

1. Exact; 2. One-sided tolerance factor with confidence $(1 + \gamma)/2$; 3. New Method; 4. Lee-Mathew; 5. Wallis' method

are reported in Mondal, 2007) because the performances of the methods are similar to those for the case of $n - m - 1 = 10$.

For $p \geq 2$, we shall compare the factors based on Algorithms 4.1 and 4.2. We computed the tolerance factors using Algorithm 4.2 with $NR = 100{,}000$. The chi-square random numbers are generated using IMSL subroutine RNCHI, the chi-square percentiles are computed using IMSL subroutine CHIIN and Wishart random matrices are generated using an algorithm similar to the one due to Smith and Hocking (1972). The eigenvalues of the generated Wishart matrices are computed using IMSL routine EVLSF. We do not include the approximate factors $k_{11}(d^2)$ and $k_{12}(d^2)$ in our comparison studies because Lee and Mathew simulation studies show that there is no clear-cut winner between them and none of them is satisfactory for all situations.

Table 2
Mean and standard deviation of tolerance factors based on 10 repetitions; A2. Algorithm 4.2 and A1. Algorithm 4.1 $n - m - 1 = 12$

			$p = 4$			$p = 7$			$p = 10$		
(α, γ)	Method	d^2	Mean	SD	CP	Mean	SD	CP	Mean	SD	CP
(.90,.90)	A1	.1	23.32	0.22	0.896	75.81	0.71	0.898	508.40	18.90	0.896
	A2	.1	23.67	0.06	0.901	76.56	0.19	0.900	515.34	2.95	0.899
	A1	.4	30.27	0.39	0.899	96.99	1.90	0.899	643.83	19.15	0.905
	A2	.4	30.73	0.07	0.903	97.79	0.40	0.902	647.87	2.43	0.905
	A1	.9	41.90	0.54	0.903	131.07	2.48	0.903	838.28	27.80	0.899
	A2	.9	41.84	0.16	0.902	131.11	0.30	0.903	832.53	5.22	0.898
(.90,.95)	A1	.1	27.55	0.25	0.950	95.66	1.33	0.950	815.77	40.55	0.951
	A2	.1	28.13	0.07	0.955	97.15	0.36	0.953	826.36	4.54	0.952
	A1	.4	37.27	0.75	0.952	124.59	3.05	0.950	1018.20	50.05	0.948
	A2	.4	37.22	0.14	0.952	125.99	0.36	0.952	1039.27	10.11	0.951
	A1	.9	51.96	0.88	0.948	171.71	4.34	0.952	1385.14	93.68	0.954
	A2	.9	52.12	0.26	0.948	171.76	0.80	0.952	1352.61	10.65	0.951
(.95,.90)	A1	.1	30.72	0.45	0.907	100.90	1.73	0.908	713.53	22.46	0.895
	A2	.1	31.08	0.05	0.910	100.60	0.22	0.907	709.99	5.42	0.894
	A1	.4	39.10	0.51	0.901	125.80	2.63	0.901	872.01	27.32	0.897
	A2	.4	39.33	0.08	0.904	125.89	0.53	0.901	866.97	5.61	0.897
	A1	.9	51.50	1.01	0.896	163.23	2.21	0.903	1094.81	24.72	0.901
	A2	.9	51.72	0.12	0.897	163.07	0.85	0.903	1089.33	6.29	0.900
(.95,.99)	A1	.1	54.75	1.49	0.990	229.03	11.27	0.992	3533.88	285.05	0.991
	A2	.1	55.14	0.47	0.990	227.37	2.980	0.992	3413.30	78.62	0.991
	A1	.4	73.24	1.84	0.990	294.91	20.11	0.991	4210.20	475.71	0.987
	A2	.4	73.03	0.51	0.990	291.60	3.026	0.990	4217.53	136.92	0.987
	A1	.9	102.36	3.96	0.992	397.45	17.33	0.990	5343.09	453.81	0.990
	A2	.9	102.11	0.52	0.992	392.49	4.015	0.989	5350.78	93.37	0.990
(.99,.90)	A1	.1	48.52	0.53	0.900	161.16	2.53	0.897	1184.89	45.02	0.901
	A2	.1	48.70	0.10	0.901	161.61	0.61	0.897	1188.74	9.26	0.901
	A1	.4	59.55	0.91	0.900	193.38	3.57	0.897	1383.84	43.09	0.907
	A2	.4	59.44	0.16	0.899	194.65	0.74	0.900	1400.38	8.25	0.909
	A1	.9	74.62	1.05	0.901	242.26	3.58	0.895	1671.66	52.87	0.889
	A2	.9	74.56	0.28	0.900	240.22	0.67	0.893	1680.70	13.76	0.889
Average	A1			73.4			147			286	
CPU in Sec.	A2			5.5			9.4			17.5	

Our preliminary simulation studies showed that the coverage probabilities of the tolerance regions based on Algorithms 4.1 and 4.2 are comparable, and they are in general close to the specified nominal level. So, it is of interest to compare Algorithms 4.1 and 4.2 with respect to speed and accuracy. We recorded the CPU time to judge the speed, and computed the standard deviation of the factors based

Table 3
Tolerance factors $k(d^2)$ based on Algorithm 4.2 when confidence levels are $\gamma_1 = .90$, $\gamma_2 = .95$, and $\gamma_3 = 0.99$

	$\alpha = .90$			$\alpha = .95$			$\alpha = .99$		
d^2	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
				$n - m - 1 = 12$, $p = 2$					
0.1	10.53	12.50	17.86	14.06	16.74	24.04	22.30	26.71	38.94
0.2	11.56	13.81	19.79	15.35	18.44	26.90	24.18	29.05	42.60
0.3	12.63	15.26	22.37	16.65	20.16	29.33	25.92	31.34	46.42
0.4	13.72	16.68	24.98	17.73	21.49	31.72	27.82	33.74	50.64
0.5	14.77	18.14	27.50	19.06	23.29	35.28	29.38	35.91	54.08
0.6	15.79	19.49	29.97	20.34	24.91	38.20	30.78	37.70	57.15
0.7	16.75	20.75	32.03	21.58	26.80	41.20	32.33	39.88	60.91
0.8	17.83	22.18	35.21	22.58	28.14	43.92	33.98	42.04	63.98
0.9	18.87	23.64	37.54	23.69	29.55	45.65	35.10	43.48	67.00
1	19.82	24.86	39.47	24.98	31.36	49.26	36.58	45.48	70.39
					$p = 3$				
0.1	16.36	19.29	27.44	21.23	25.24	36.73	32.54	39.20	58.52
0.2	17.94	21.37	31.00	23.14	27.72	40.80	35.51	42.87	64.40
0.3	19.52	23.40	34.65	25.07	30.10	45.10	37.89	46.22	69.81
0.4	21.14	25.49	38.07	26.95	32.73	48.86	40.58	49.52	75.84
0.5	22.68	27.61	41.36	28.80	35.01	52.95	42.81	52.42	79.96
0.6	24.07	29.38	44.89	30.65	37.55	58.03	45.09	55.57	84.66
0.7	25.66	31.71	48.56	32.35	39.85	61.91	47.44	58.20	89.45
0.8	27.24	33.69	51.92	34.10	42.37	66.39	49.74	61.71	96.77
0.9	28.89	35.75	55.89	35.86	44.36	69.49	51.69	64.01	99.67
1	30.44	38.03	59.93	37.49	46.69	73.22	53.88	66.85	104.94
					$p = 4$				
0.1	24.02	28.52	41.63	30.66	36.79	54.98	46.75	56.67	86.92
0.2	26.45	31.49	46.89	33.58	40.38	60.88	50.37	61.44	95.55
0.3	28.71	34.58	51.37	36.36	44.09	67.18	54.30	66.82	102.66
0.4	30.71	37.22	56.05	38.99	47.21	72.34	57.72	71.28	111.24
0.5	33.23	40.53	62.66	41.88	51.31	79.22	61.23	75.55	117.99
0.6	35.28	43.23	67.08	44.25	54.57	85.85	64.34	79.65	125.16
0.7	37.66	46.50	72.81	46.86	57.85	91.58	67.88	84.37	134.62
0.8	39.81	49.15	76.98	49.11	60.74	97.80	70.80	88.30	139.45
0.9	42.10	52.32	83.58	51.71	64.38	101.44	73.82	92.07	147.39
1	44.25	55.21	88.94	54.23	67.60	106.91	76.68	96.37	154.52

(continued)

Table 3
Continued

d^2	$\alpha = .90$			$\alpha = .95$			$\alpha = .99$		
	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
				$n-m-1=20;\ p=2$					
0.1	8.31	9.34	11.92	10.97	12.35	15.85	17.24	19.48	25.15
0.2	9.21	10.50	13.68	12.10	13.77	18.03	18.79	21.44	28.18
0.3	10.15	11.70	15.69	13.24	15.27	20.35	20.35	23.48	31.22
0.4	11.05	12.87	17.62	14.34	16.82	22.92	21.76	25.22	34.29
0.5	11.98	14.15	19.67	15.42	18.16	25.06	23.29	27.20	37.20
0.6	12.94	15.42	21.77	16.44	19.43	27.05	24.51	28.82	39.85
0.7	13.78	16.57	23.49	17.46	20.77	29.26	25.80	30.48	42.19
0.8	14.69	17.69	25.46	18.49	22.05	31.42	27.15	32.17	44.76
0.9	15.63	18.98	27.14	19.44	23.42	33.39	28.34	33.62	47.36
1	16.48	20.00	29.31	20.46	24.78	35.60	29.69	35.46	50.07
				$p=3$					
0.1	11.90	13.23	16.53	15.22	16.99	21.34	22.70	25.52	32.47
0.2	13.17	14.77	18.94	16.72	18.86	24.14	24.80	28.07	36.60
0.3	14.46	16.45	21.51	18.21	20.67	27.17	26.93	30.80	40.71
0.4	15.75	18.08	24.34	19.73	22.73	30.45	28.92	33.25	44.39
0.5	16.94	19.77	26.80	21.25	24.58	33.25	30.79	35.70	48.21
0.6	18.29	21.44	29.40	22.72	26.53	36.21	32.60	37.87	51.93
0.7	19.62	23.19	32.17	24.16	28.28	39.01	34.42	40.26	55.36
0.8	20.87	24.62	34.32	25.46	30.01	41.80	36.29	42.43	58.40
0.9	22.03	26.21	36.92	26.96	31.78	44.12	37.80	44.59	61.53
1	23.18	27.86	39.78	28.39	33.67	47.48	39.61	46.77	65.63
				$p=4$					
0.1	15.82	17.58	21.75	19.84	22.14	27.72	29.06	32.52	41.30
0.2	17.47	19.53	24.79	21.79	24.50	31.18	31.63	35.76	46.45
0.3	19.14	21.63	28.02	23.80	27.03	35.37	34.29	38.97	51.31
0.4	20.84	23.81	31.30	25.74	29.37	38.79	36.99	42.34	56.29
0.5	22.52	25.95	34.68	27.66	31.98	43.14	39.23	45.25	61.08
0.6	24.25	28.05	38.01	29.60	34.20	46.34	41.67	48.42	65.14
0.7	25.82	30.15	41.54	31.46	36.68	49.99	43.90	51.04	69.33
0.8	27.56	32.22	44.41	33.27	38.89	53.85	46.18	53.96	74.70
0.9	29.06	34.27	47.40	35.16	41.34	57.11	48.39	56.87	78.81
1	30.73	36.44	50.98	36.98	43.62	60.56	50.23	59.21	81.56
				$p=5$					
0.1	20.35	22.52	27.78	25.12	27.97	34.92	36.25	40.50	51.59
0.2	22.41	25.09	31.69	27.71	31.10	39.54	39.46	44.57	58.03
0.3	24.56	27.66	35.86	30.18	34.28	44.54	42.74	48.52	64.36
0.4	26.69	30.35	39.84	32.52	37.06	48.84	45.87	52.69	70.23
0.5	28.86	33.17	44.20	35.06	40.31	53.94	48.99	56.54	76.34
0.6	30.85	35.72	48.30	37.35	43.19	58.93	51.83	60.07	81.63
0.7	32.99	38.31	52.39	39.73	46.23	63.05	54.87	63.87	88.33
0.8	35.01	40.72	55.63	41.97	49.02	67.64	57.70	67.12	92.16
0.9	37.14	43.58	60.22	44.40	52.03	71.44	60.25	70.58	97.47
1	39.23	46.21	64.51	46.43	54.66	76.45	62.97	74.03	102.62

on 10 repetitions to appraise the accuracies. Specifically, for a given $(\alpha, \gamma, n - m - 1, p)$, we computed tolerance factors using Algorithm 4.2 with $NR = 100{,}000$ (Algorithm 4.1 with $m_1 = m_2 = 5000$), and repeated the simulation 10 times. The average and standard deviation of the factors of these 10 repetitions are given in Table 2. We observe from Table 2 that, on average basis, both methods produce approximately the same factor with coverage probabilities close to the nominal level. However, the standard deviation of the factors computed using Algorithm 4.1 is at least four times as large as the standard deviation of the factors based on Algorithm 4.2. This clearly implies that Algorithm 4.2 is not stable, and the factors based on it are heavily depending on the initial seed used for the simulation. Furthermore, average CPU time for Algorithm 4.1 is approximately 14 times as large as that of Algorithm 4.2 (see Table 2). Thus, Algorithm 4.2 is more stable and faster than Algorithm 4.1, and so the former is certainly preferable to the latter to compute tolerance factors for practical use.

6. Concluding Remarks

In this article, we have shown that available approximate tolerance factors in multiple linear regression are satisfactory only for small values of d^2. The tolerance factors based on Wallis' method are unsatisfactory, and the one by Lee and Mathew (2004) are conservative for $d^2 \geq .3$. The factors based on Algorithm 4.2 are slightly conservative when the values of d^2 are large. Our recommendation for the multiple linear regression case is that to use the exact approach for constructing tolerance factors if accuracy is important. If $d^2 \geq 0.5$, then coverage level adjusted one-sided tolerance factors can be used. The required noncentral t percentiles (to compute the one-sided tolernace factors) can be obtained using the online calculator by the Department of Statistics, UCLA, at the website http://calculators.stat.ucla.edu/cdf or by the PC calculator that accompanies the book by Krishnamoorthy (2006) computes the one-sided limits and exact tolerance intervals for a normal distribution. This calculator is available at http://www.ucs.louisiana.edu/~kxk4695. For the multivariate linear regression, the simulation method given in Algorithm 4.2 seems to be the only satisfactory approach for constructing tolerance factors. Even though the method based on Algorithm 4.2 is computationally involved, it is as simple as other procedures once it is programed.

It is not feasible to tabulate tolerance factors for all values of $(n - m - 1, p, d^2)$. The tolerance factors based on Algorithm 4.2 with $NR = 100{,}000$ are given in Table 3 for $n - m - 1 = 10$ and 20, and some values of p. These factors are provided so that they can be used as benchmarks for future results.

Acknowledgments

The authors are grateful to the reviewer of their earlier article (Krishnamoorthy and Mondal, 2006) who brought their attention to the present multivariate regression problem. They are also thankful to a reviewer for providing useful comments and suggestions.

References

Eberhardt, K. R., Mee, R. W., Reeve, C. P. (1989). Computing factors for exact two-sided tolerance limits for a normal distribution. *Communications in Statistics – Simulation and Computation* 18:397–413.

Graybill, F. A., Iyer, H. K. (1994). *Regression Analysis: Concepts and Applications*. North Scituate, MA: Duxbury Press.

Guttman, I. (1970). Construction of β-content tolerance regions at confidence level γ for large sample from the k-variate normal distribution. *The Annals of Mathematics and Statistics* 41:376–400.

Haq, M. S., Rinco, S. (1976). β-expectation tolerance regions for a generalized multivariate model with normal error variables. *Journal of Multivariate Analysis* 6:414–421.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* 48:419–426.

John, S. (1963). A tolerance region for multivariate normal distributions. *Sankhya, Series A* 25:363–368.

Khan, S., Haq, M. S. (1994). β-expectation tolerance region for the multilinear model with matrix-t error distribution. *Communications in Statistics – Theory and Methods* 23:1935–1951.

Kibria, B. M. G. (2006). The matric t distribution and its applications in predictive inference. *Journal of Multivariate Analysis* 97:785–795.

Kibria, B. M. G., Haq, M. S. (1999). Predictive inference for the elliptical linear model. *Journal of Multivariate Analysis* 68:235–249.

Krishnamoorthy, K., Mathew, T. (1999). Comparison of approximation methods for computing tolerance factors for a multivariate normal population. *Technometrics* 41:234–249.

Krishnamoorthy, K., Mondal, S. (2006). Improved tolerance factors for multivariate normal distributions. *Communications in Statistics – Simulation Computation* 35:461–478.

Krishnamoorthy, K. (2006). *Handbook of Statistical Distributions with Applications*. Boca Raton: Chapman & Hall/CRC.

Lee, Y., Mathew, T. (2004). Tolerance regions in multivariate linear regression. *Journal of Statistical Planning Inference* 126:253–271.

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

Mondal, S. (2007). Constructions of Tolerance Regions for Some Multivariate Linear Models. Ph.D. Dissertation, Department of Mathematics, University of Louisiana: Lafayette.

Smith, W. B., Hocking, R. R. (1972). Wishart variates generator (Algorithm AS 53). *Applied Statistics* 21:341–345.

Wald, A. (1942). Setting of tolerance limits when the sample is large. *The Annals of Mathematical Statistics* 13:389–399.

Wald, A., Wolfowitz, J. (1946). Tolerance limits for a normal distribution. *The Annals of Mathematics and Statistics* 17:208–215.

Wallis, W. A. (1951). Tolerance Intervals for Linear Regression. Proceedings of the Second Berkeley Symposium. Berkeley: University of California Press.