Simultaneous confidence tubes for comparing several multivariate linear regression models

Jianan Peng | Wei Liu | Frank Bretz | A. J. Hayter

1 Department of Mathematics and Statistics, Acadia University, Wolfville, NS, Canada
2 S3RI and School of Mathematics, University of Southampton, UK
3 Novartis Pharma, Basel, AG, Switzerland
4 Department of Statistics and Operations Technology, University of Denver, Denver, CO, USA

Abstract

Much of the research on multiple comparison and simultaneous inference in the past 60 years or so has been for the comparisons of several population means. Spurrier seems to have been the first to investigate multiple comparisons of several simple linear regression lines using simultaneous confidence bands. In this paper, we extend the work of Liu et al. for finite comparisons of several univariate linear regression models using simultaneous confidence bands to finite comparisons of several multivariate linear regression models using simultaneous confidence tubes. We show how simultaneous confidence tubes can be constructed to allow more informative inferences for the comparison of several multivariate linear regression models than the current approach of hypotheses testing. The methods are illustrated with examples.

KEYWORDS
multiple comparisons, multivariate linear regression, simultaneous confidence bands, simultaneous inference, statistical simulation

1 | INTRODUCTION

The bulk of the work on simultaneous inference and multiple comparisons to date is for comparing the means of \( k (\geq 3) \) populations, following the work of Tukey (1953) on pairwise comparisons of \( k \) population means, of Dunnett (1955) on comparisons of several means with a control mean, and of Scheffé (1953) on all-contrast comparisons among the population means. Miller (1981), Hochberg and Tamhane (1987), Westfall and Young (1993), Hsu (1996), and Bretz et al. (2011) are excellent references of the work in this area. Spurrier (1999) seems to have been the first to work on the simultaneous comparison of several simple linear regression lines by using a set of simultaneous confidence bands. Since then, Spurrier (1999) work has been extended in several directions; see, for example, Spurrier (2002), Bhargava and Spurrier (2004) and Spurrier (2004), Liu et al. (2004), and Lu and Kuriki (2017). In particular, Liu et al. (2004) use simultaneous confidence bands for finite comparisons of several univariate linear regression models, which is directly applicable for pooling batches in drug stability study (Ruberg & Hsu, 1992) among many other applications. A review of the related works is given in Liu (2010). The purpose of this paper is to extend the work of Liu et al. (2004) on univariate linear regression models to multivariate linear regression models, which have wide applications.
To motivate our research, we consider the example from Raykov and Marcoulides (2008, p. 192). The study involves three intelligence measures, inductive reasoning (ir), figural relations (fr), and culture-fair tests (cf). To evaluate whether there is an effect of the training after accounting for possible group differences at pretest on the inductive reasoning measure, 248 seniors were recruited in a two-group (training and no-training) training program in which two assessments pretest and posttest of an induction reasoning test were obtained. Raykov and Marcoulides fit a bivariate linear model of the responses on the only covariate pretest measurement on inductive reasoning for each of the two groups. Note that their data file uses ir.1 for pretest for measurement on inductive reasoning, ir.2 for posttest measurement on inductive reasoning (which we do not use in this paper), as well as fr.2 and cf.2 for posttest for figural relations and culture-fair tests, respectively. We use ir, fr, and cf throughout this paper to simplify the notation.

Much of the recent work on simultaneous confidence bands for linear regression is almost exclusively for univariate linear regression models; see, for example, Al-Saidy et al. (2003), Nitcheva et al. (2005), Piegorsch et al. (2005), Deutsch and Piegorsch (2012), Peng, Robichaud, and Alsüb (2015), and Dette et al. (2018). Liu et al. (2016) consider simultaneous confidence band for one multivariate linear regression model over the whole covariate region, and call the confidence band “confidence tube” to reflect its true shape.

Assume that the $i$th multivariate linear regression model, corresponding to the $i$th treatment group, is given by

$$Y_i = X_i B_i + E_i, \quad i = 1, \ldots, k,$$

where $Y_i = (y_{i,1}, \ldots, y_{i,n_i})^T$ with $y_{i,l} = (y_{i,l,1}, \ldots, y_{i,l,m})$ being the observations on the $m$ response variables of the $j$th individual in the $i$th treatment group, $X_i$ is an $n_i \times (p+1)$ full rank design matrix with the first column given by $(1, \ldots, 1)^T$ and the $(l \geq 2)$th column given by $(x_{i,l,1}, \ldots, x_{i,l,n_i})^T$, $B_i = (b_{i,1}, \ldots, b_{i,m})^T$ being the regression coefficients for the $i$th response variable in the $i$th treatment group, and $E_i = (e_{i,1}, \ldots, e_{i,n_i})^T$ with all the $e_{i,j}$ being i.i.d. multivariate normal $N_m(0, \Omega)$ random vectors. As $X_i^T X_i$ is nonsingular, the least squares estimator of $B_i$ is given by $\hat{B}_i = (X_i^T X_i)^{-1} X_i^T Y_i$. Let the pooled estimator of the unknown error-vector covariance matrix $\Omega$ be $\hat{\Omega} = \sum_{i=1}^k (Y_i - X_i \hat{B}_i)^T (Y_i - X_i \hat{B}_i) / \nu$ with $\nu = \sum_{i=1}^k (n_i - p - 1)$. Then $\text{vec}(\hat{B}_i) \sim N(\text{vec}(B_i), \Omega \otimes (X_i^T X_i)^{-1})$ where vec($A$) denotes the resultant vector from stacking the columns of the matrix $A$, $\nu \hat{\Omega}$ has the Wishart distribution $W(\Omega, \nu)$, and $\hat{B}_1, \ldots, \hat{B}_k$ and $\hat{\Omega}$ are independent. All these results, which are generalizations of the well-known results for univariate regression ($m = 1$) to the multivariate cases ($m \geq 1$), can be found in the excellent book of Anderson (2003).

Our objective is to construct a set of simultaneous confidence tubes (SCTs) for

$$x^T B_i - x^T B_j = (1, x_1, \ldots, x_p) B_i - (1, x_1, \ldots, x_p) B_j, \quad (i, j) \in \Lambda$$

over a given covariate range $x_l \in [a_l, b_l], l = 1, \ldots, p$, where $\Lambda$ is an index set that determines the comparisons of interest. For example, if all pairwise comparisons are of interest, then $\Lambda = \{(i, j) : 1 \leq i \neq j \leq k\}$; if the comparison with a control, say, the second to $k$th regression models with the first regression model is of interest, then $\Lambda = \{(i, j) : 2 \leq i \leq k, j = 1\}$; if the successive comparison of the $k$ regression models is of interest, then $\Lambda = \{(i, i+1) : 1 \leq i \leq k - 1\}$. We construct the following set of SCTs:

$$[(x^T B_i - x^T B_j) - (x^T \hat{B}_i - x^T \hat{B}_j) (\nu \hat{\Omega})^{-1} [x^T B_i - x^T B_j] - (x^T \hat{B}_i - x^T \hat{B}_j)]^T$$

$$\leq c x^T \left( (X_i^T X_i)^{-1} + (X_j^T X_j)^{-1} \right) x, \quad \forall x_l \in [a_l, b_l] \text{ for } l = 1, \ldots, p \text{ and } \forall (i, j) \in \Lambda,$n

(1)

where $c$ is the critical constant suitably chosen so that the confidence level of this set of SCTs is equal to $1 - \alpha$.

When there is only $m = 1$ response variable, the set of SCTs in (1) becomes the set of simultaneous confidence bands given in Liu et al. (2004). The textbook approach to the comparison of $k$ multivariate regression models (Anderson, 2003; Raykov & Marcoulides, 2008) is to perform a hypotheses test of $H_0 : B_1 = \cdots = B_k$ against the alternative $H_a : B_i \neq B_j$. It is shown in this paper that the SCTs in (1) allow more detailed and informative inferences than the dichotomous inference, rejection, or nonrejection of $H_0$, of a hypotheses test.

The outline of the paper is as follows. Section 2 discusses the computation of the critical constant $c$ using simulation. Section 3 focuses on the comparison of two models (i.e., $k = 2$), and illuminates the relationship between Roy (1953) test
and the SCT in (Equation 1) in this case. We provide a real data example to illustrate the advantages of the SCTs approach over hypothesis testing. Section 4 considers the comparison of \( k(\geq 3) \) models. Again, we use an example to illustrate the versatile and informative inferences the SCTs approach allows. Finally, Section 5 contains concluding remarks.

2 DETE RMINATION O F T H E CRITICAL CONSTANT C

Note that the confidence level of the SCTs in (1) is given by \( P\{T < c\} \) where

\[
T = \sup_{(i,j)\in\Lambda} \sup_{x_l\in[a_l,b_l],l=1,...,p} T_{i,j}(x) \tag{2}
\]

with

\[
T_{i,j}(x) = \frac{[(x^T B_i - x^T B_j) - (x^T \hat{B}_i - x^T \hat{B}_j)] (\nu \hat{\Omega})^{-1} [(x^T B_i - x^T B_j) - (x^T \hat{B}_i - x^T \hat{B}_j)]^T}{x^T \left[(X_i^T X_i)^{-1} + (X_j^T X_j)^{-1}\right] x}.
\]

Denote \( U_i = (X_i^T X_i)^{1/2} (\hat{B}_i - B_i) \hat{\Omega}^{-1/2}, i = 1, ..., k. \) Then straightforward manipulation shows that \( vec(U_i) \sim N(0, I(p+1)m) \), and it is clear that \( U_1, ..., U_k \) are independent because \( \hat{B}_1, ..., \hat{B}_k \) are independent. Now, \( T_{i,j}(x) \) can be rewritten as

\[
x^T \left[(X_i^T X_i)^{-1/2} U_i - (X_j^T X_j)^{-1/2} U_j \right] \left(\Omega^{-1/2} \nu \hat{\Omega}^{-1/2}\right)^{-1} \left[U_i^T (X_i^T X_i)^{-1/2} - U_j^T (X_j^T X_j)^{-1/2}\right] x
\]

By noting that \( \Omega^{-1/2} \nu \hat{\Omega}^{-1/2} \sim W(I_r, v) \), the distributions of \( T_{i,j}(x) \) and so \( T \) do not depend on the unknown parameters \( B_1, ..., B_k \) and \( \Omega \) of the \( k \) regression models.

The critical value \( c \) is therefore the \( (1 - \alpha) \) quantile of \( T \) and can be computed by simulation involving the following steps:

1. Step 1: For a given data set, obtain the design matrices \( X_i, 1 \leq i \leq k. \)
2. Step 2: Generate pseudorandom data from Normal distribution \( R > 0 \) times. Select \( R \) as large as is feasible for the available computing resources.
3. Step 3: Calculate the statistic \( T_r \) using the expressions (2) and (3) from each of the \( r = 1, ..., R \) pseudorandom data sets.
4. Step 4: The critical value \( c \) is the \( (1 - \alpha) \) quantile of the \( R \)-independent replicates of \( T_1, ..., T_R \) of \( T \) in Step 3.

It is known that the sample \( (1 - \alpha) \) quantile \( \hat{c} \) converges to the population \( (1 - \alpha) \) quantile \( c \) almost surely as \( R \) approaches infinity (cf. Serfling, 1980). This means that \( \hat{c} \) can be as close to \( c \) as required by using a sufficiently large number \( R \) of simulations. For a finite \( R \), the accuracy of \( \hat{c} \) can be assessed by using the variance of the large sample approximate normal distribution of \( \hat{c} \); see, for example, Liu et al. (2005) for details.

For the examples given in this paper, there is only one covariate, \( x_1 \), in the regression models, and therefore, both the numerator and the denominator of \( T_{i,j}(x) \) in (2) are polynomials of \( x_1 \). As a result, \( \sup_{x_1\in[a_1,b_1]} T_{i,j}(x) \) in (Equation 2) can be computed very fast by using the method of Liu et al. (2008) in each simulation of \( T \). When the regression models have more than one covariate, generic and therefore less efficient algorithms for the maximization have to be used and are available in most numerical software such as R and Matlab.

In the examples in this paper, we use \( R = 1,000,000 \) and it takes about 500 s on an ordinary Window’s PC (Core(TM2) Due CPU P8400@2.26GHz) to compute one \( c \). The R code for the simulated critical value for the example in Section 3 is provided in the journal’s website. For \( R = 1,000,000 \) and \( \alpha = .05 \), the standard error (i.e., the square root of the variance) of \( \hat{c} \) is about 0.00004, and therefore, the critical constant is most likely accurate to the fourth decimal place at least. Alternatively, one can use the method of Edwards and Berry (1987) to assess how close the random variable \( P\{T < \hat{c}|\hat{c}\} \) is to \( 1 - \alpha \). This random variable has a Type I beta distribution with parameters \( R - (1 - \alpha)R - 1 \) and \( (1 - \alpha)R \), and is approximately normal for a large \( R \) (Edwards & Berry, 1987). For \( R = 1,000,000 \) and \( \alpha = .05 \), the standard error of
this random variable is s.e. = 0.00022, and therefore, the true confidence level by using \( \hat{c} \) is almost certainly in the range \( 1 - \alpha \pm 3 \times \text{s.e.} = [0.94934, 0.95066] \). All these indicate that the critical value \( \hat{c} \) based on \( R = 1,000,000 \) simulations should be accurate enough for most applications.

3 | COMPARISON OF TWO MODELS

For comparing \( k = 2 \) models, the set of \( 1 - \alpha \) SCTs in (1) contains just one SCT and is given by

\[
\left[ (\mathbf{x}^T \mathbf{B}_1 - \mathbf{x}^T \mathbf{B}_2) - (\mathbf{x}^T \mathbf{\hat{B}}_1 - \mathbf{x}^T \mathbf{\hat{B}}_2) \right] (\mathbf{\hat{\Omega}})^{-1} \left[ (\mathbf{x}^T \mathbf{B}_1 - \mathbf{x}^T \mathbf{B}_2) - (\mathbf{x}^T \mathbf{\hat{B}}_1 - \mathbf{x}^T \mathbf{\hat{B}}_2) \right]^T \\
\leq c \mathbf{x}^T \left[ (\mathbf{X}_1^T \mathbf{X}_1)^{-1} + (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \right] \mathbf{x}, \quad \forall \mathbf{x}_l \in [a_l, b_l] \text{ for } l = 1, \ldots, p. \tag{4}
\]

This SCT quantifies the magnitude of difference between the two models \( \mathbf{x}^T \mathbf{B}_1 \) and \( \mathbf{x}^T \mathbf{B}_2 \) over the covariate region \( \mathbf{x}_l \in [a_l, b_l] \) for \( l = 1, \ldots, p \). In particular, if \( \mathbf{B}_1 = \mathbf{B}_2 = \mathbf{B} \), then the zero line \( \mathbf{x}^T (\mathbf{B}_1 - \mathbf{B}_2) \) is completely contained in the SCT with probability \( 1 - \alpha \), which induces the following size \( \alpha \) test of \( H_0 : \mathbf{B}_1 = \mathbf{B}_2 \) against the alternative \( H_a : \text{not } H_0 \): \( H_0 \) is rejected if and only if the zero line is not completely contained in the SCT.

For the special covariate region \( \mathbf{x}_l \in [-\infty, \infty] \) for \( l = 1, \ldots, p \), that is, the whole covariate space \( \mathbb{R}^p \), the rejection region of \( H_0 \) of this induced test becomes \( L > c \) where

\[
L = \sup_{\mathbf{x}_l \in \mathbb{R}^l, l = 1, \ldots, p} \left( \mathbf{x}^T \left[ (\mathbf{X}_1^T \mathbf{X}_1)^{-1} + (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \right] \mathbf{x} \right)^{\mathbf{x}^T (\mathbf{B}_1 - \mathbf{B}_2) (\mathbf{\hat{\Omega}})^{-1} (\mathbf{B}_1 - \mathbf{B}_2)^T} \\
\quad = \text{the largest eigenvalue of } \left[ (\mathbf{X}_1^T \mathbf{X}_1)^{-1} + (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \right]^{-1} (\mathbf{B}_1 - \mathbf{B}_2) (\mathbf{\hat{\Omega}})^{-1} (\mathbf{B}_1 - \mathbf{B}_2)^T \\
\quad = \text{the largest eigenvalue of } (\mathbf{\hat{B}}_1 - \mathbf{\hat{B}}_2)^T \left[ (\mathbf{X}_1^T \mathbf{X}_1)^{-1} + (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \right]^{-1} (\mathbf{\hat{B}}_1 - \mathbf{\hat{B}}_2) (\mathbf{\hat{\Omega}})^{-1}, \tag{5}
\]

where the equality in (5) follows directly from Mardia et al. (1979, Theorem A.9.2) and the equality in (6) follows directly from Mardia et al. (1979, Theorem A.6.2). A few lines of manipulation show that this test is just Roy (1953) test of \( H_0 : \mathbf{B}_1 = \mathbf{B}_2 \) against \( H_a : \text{not } H_0 \); see Anderson (2003, Sections 8.4 and 8.6) for the construction of Roy’s test and other commonly used tests. This shows that Roy’s test is implied by the SCT over the whole covariate space. On the other hand, when \( a_l = b_l \) for \( l = 1, \ldots, p \) and so the covariate region contains just one point, the SCT in (1) becomes the pointwise band.

Direct manipulation, utilizing the generalized \( T^2 \)-statistic (see, e.g., Anderson, 2003, Theorem 5.2.2) shows that the critical constant \( c \) in this case is given by \( c = \frac{m}{\nu} f_{m, \nu, 1 - \alpha} \) where \( f_{m, \nu, 1 - \alpha} \) denotes the \( 1 - \alpha \) quantile of the \( F \) distribution with \( m \) and \( \nu \) degrees of freedom.

Of course, the magnitude of differences between the two models quantified by the SCT is more informative than either a rejection or a nonrejection of \( H_0 \) of a test. When \( H_0 \) is rejected, the SCT allows us to assess over what covariate region the two models are significantly different and the direction of the difference. Even when \( H_0 \) is not rejected, which can mean anything but the two models are the same, the magnitude of difference between the two models derived from the SCT still provides useful information.

Furthermore, it has been argued by numerous statisticians that statistical models often provide good approximations only over a certain covariate region (Naiman, 1987; Piegorsch & Casella, 1988). The SCT in (1) uses this information in the form of the covariate region \( \mathbf{x}_l \in [a_l, b_l] \) for \( l = 1, \ldots, p \) in its construction. This SCT is narrower and hence provides sharper inference over the covariate region of interest \( \mathbf{x}_l \in [a_l, b_l] \) for \( l = 1, \ldots, p \) than the SCT over the whole covariate region \( \mathbb{R}^p \).

**Example 1.** For the data mentioned in Section 1, we follow Raykov and Marcoulides (2008) and fit a bivariate linear model of the responses \( y_1 = fr \) and \( y_2 = cf \) on the covariate pretest measurement on inductive reasoning, which is the only predictor \( x = ir \) for each of the two groups of students: untrained (group 1) and trained (group 2). The purpose of having \( ir \) in the
TABLE 1  Critical values and p-values at \( \alpha = .05 \) for Roy’s test and SCT in Example 1

|               | Roy’s test | SCT       |
|---------------|------------|-----------|
| Critical value| 0.0360     | 0.0357    |
| Test statistic| 0.0876     | 0.0876    |
| p-value       | .000205    | .00019    |

**FIGURE 1**  The 95% SCT: given by the union of all the elliptic discs; the estimate \((x^T \hat{B}_2 - x^T \hat{B}_1)\): given by the straight line in the center of the SCT; and the zero line \(x^T(B - B)\): given by the other straight line.

The 95% SCT is formed by a collection of elliptic discs, one at each \(x \in [0, 78.6]\). The center of the SCT is given by the straight line \((x^T \hat{B}_2 - x^T \hat{B}_1)\), which is the estimate of \((x^T B_2 - x^T B_1)\) and also plotted in Figure 1. As the zero line \(x^T (B_2 - B_1)\) with \(B_1 = B_2\), plotted in Figure 1 by the other straight line, is not completely contained in the SCT over \(x \in [0, 78.6]\), \(H_0 : B_1 = B_2\) is also rejected by the SCT. But the SCT provides more information on \((x^T B_2 - x^T B_1)\). For example, by looking at the projection of the SCT to the \((y_1, x)\)-plane, plotted in Figure 2, one can conclude that the trained group (corresponding to \(x^T B_2\)) has higher \(y_1\) (i.e., \(f r\)).

**model is to have a common “baseline” between groups 1 and 2, that is, to adjust for potential baseline imbalance of subjects between groups 1 and group 2. Therefore, the comparison would not be biased due to group 2 subjects having higher pretest measurement on inductive reasoning \(ir\) than group 1 subjects. Note that we do not assume that the two bivariate linear regression models for the two groups to have the same slopes; otherwise, the problem becomes the comparison of the intercepts only and can be dealt with by a simpler multivariate ANCOVA (Anderson, 2003). Based on the observations on 248 students (with 87 students in group 1 and 161 students in group 2), one can easily compute the estimates

\[
\hat{B}_1 = \begin{pmatrix} 48.404 \\ 0.590 \end{pmatrix}, \quad \hat{B}_2 = \begin{pmatrix} 53.740 \\ 0.637 \end{pmatrix}, \quad \hat{\Omega} = \begin{pmatrix} 33075.6 & 20024.8 \\ 20024.8 & 37408.5 \end{pmatrix}
\]

with \(m = 2, p = 1, \) and \( \nu = 244 \).

The textbook approach for the comparison of the two models for the two groups is to test \(H_0 : B_1 = B_2\) against \(H_a : \) not \(H_0\). \(H_0\) is rejected by Roy (1953) test from Table 1 below. Other commonly used tests (Anderson, 2003, Sections 8.4 and 8.6) also reject \(H_0\) with comparable p-values.

The SCT for \((x^T B_2 - x^T B_1)\) over the observed covariate range of the pretest measurement on inductive reasoning (ir) for the two groups (trained and untrained) \(x \in [0, 78.6]\) in (4) is plotted in Figure 1. The SCT is formed by a collection of elliptic discs, one at each \(x \in [0, 78.6]\). The center of the SCT is given by the straight line \((x^T \hat{B}_2 - x^T \hat{B}_1)\), which is the estimate of \((x^T B_2 - x^T B_1)\) and also plotted in Figure 1. As the zero line \(x^T (B_2 - B_1)\) with \(B_1 = B_2\), plotted in Figure 1 by the other straight line, is not completely contained in the SCT over \(x \in [0, 78.6]\), \(H_0 : B_1 = B_2\) is also rejected by the SCT. But the SCT provides more information on \((x^T B_2 - x^T B_1)\). For example, by looking at the projection of the SCT to the \((y_1, x)\)-plane, plotted in Figure 2, one can conclude that the trained group (corresponding to \(x^T B_2\)) has higher \(y_1\) (i.e., \(f r\)).
FIGURE 2 The projection of the three-dimensional plot in Figure 1 to the $(y_1, x)$-plane

on average than the untrained group (corresponding to $x^T B_1$) among those students with the $ir$ score in $[17, 50]$. But the difference between the two groups is not significant among those students with the $ir$ score not in the interval $[17, 50]$. Similar observations can be made from the projection to the $(y_2, x)$-plane. From the SCT, one can also bind the largest possible difference of $(x^T B_2 - x^T B_1)$ over $x \in [0, 78.6]$.

It is noteworthy that, in this example, the 95% SCT over the whole covariate range $x \in [-\infty, \infty]$ has $c = 0.0360$, which is almost the same as the $c$ for the SCT over $x \in [0, 78.6]$ given above. On the other hand, the point-wise band uses $c = \frac{m}{n} f_{m,n,1-\alpha} = 0.0249$, which is $(0.0360 - 0.0249)/0.0360 = 31\%$ smaller than the $c = 0.0360$. This indicates the extent to which an SCT over a finite covariate range can potentially be narrower than the SCT over the whole covariate range.

One can download from http://www.personal.soton.ac.uk/wl/SCTsForMultComp/ the R codes for the computation of the results and the Matlab codes for drawing the graphs of this and the next sections.

4 | COMPARISON OF MORE THAN TWO MODELS

For comparison of $k (\geq 3)$ models, the SCTs in (1) allow one to assess which models are different and, if two models are different, over what covariate region and in which direction the models differ. In comparison, a hypotheses test, such as Roy’s test, only concludes whether or not the $k$ models are different. In this case, there is no clear relationship between the SCTs and Roy’s test. This is not surprising because, even in the simpler situation of univariate regression, there is no direct relationship between simultaneous confidence bands and the usual $F$-test for comparing $k (\geq 3)$ models (Liu, 2010, Section 6.2).

Example 2. Continue with the example considered in Section 3. Now the 161 students in the “trained” group have actually gone through one of two different training methods: 80 students were on training method 1 and the other 81 students were on training method 2. And we are interested in whether the three groups, group 1—untrained, group 2—training method 1, and group 3—training method 2, are different in terms of how the responses $fr$ and $cf$ depend on the covariate pretest measurement $ir$. Hence, we fit a bivariate linear model of the responses $y_1 = fr$ and $y_2 = cf$ on the only covariate $x = ir$ to each of the three groups of students, and we are interested in assessing whether the three models $x^T B_1$, $x^T B_2$, and $x^T B_3$ are the same or not. Based on the observations of the 248 students, one can easily compute the estimates
TABLE 2  Critical values and p-values at \( \alpha = .05 \) for Roy’s test and SCT in Example 2

|          | Roy’s test | SCT      |
|----------|------------|----------|
| Critical value | 0.0536     | 0.0462   |
| Test statistic  | 0.0899     | 0.0766   |
| p-value       | .0017      | .0020    |

FIGURE 3  The SCT for \((x^T B_1 - x^T B_3)\)

\[
\hat{B}_1 = \begin{pmatrix} 48.404 & 14.689 \\ 0.590 & 1.103 \end{pmatrix}, \quad \hat{B}_2 = \begin{pmatrix} 51.682 & 23.964 \\ 0.681 & 0.995 \end{pmatrix}, \\
\hat{B}_3 = \begin{pmatrix} 56.176 & 21.292 \\ 0.582 & 1.092 \end{pmatrix}, \quad \mathbb{V} = \begin{pmatrix} 32,908.6 & 20,090.3 \\ 20,090.3 & 37,323.4 \end{pmatrix},
\]

with \( m = 2, p = 1, \) and \( v = 242. \)

The textbook approach for comparing the three models for the three groups is to test \( H_0 : B_1 = B_2 = B_3 \) against \( H_a : \) not \( H_0. \) From Table 2, \( H_0 \) is rejected by Roy (1953) test. But this is all a test can tell us.

To get more information on how the three models differ between themselves, one can use the SCTs in (1) for all pairwise comparisons with \( \Lambda = \{(i, j) : 1 \leq i \neq j \leq 3\}. \) For \( \alpha = .05 \) and the observed covariate range \( x \in [0, 78.6], \) the critical constant \( c \) is 0.0462. Thus, one can plot the three SCTs for \((x^T B_2 - x^T B_1), (x^T B_3 - x^T B_1),\) and \((x^T B_3 - x^T B_2)\) over \( x \in [0, 78.6], \) respectively, in order to assess whether or how any two models differ. For example, Figure 3 plots the SCT for \((x^T B_3 - x^T B_1),\) the straight line \((x^T \hat{B}_3 - x^T \hat{B}_1)\) that is the center of the SCT, and the zero straight line \((x^T B_3 - x^T B_1)\). Since the zero straight line is not included in the SCT completely, the two models \( x^T B_3 \) and \( x^T B_1 \) are significantly different. By looking at the SCT from different angles, one can observe how the two models differ. For example, by looking at the projection of the SCT for \((x^T B_3 - x^T B_1)\) in the \((y_1, x)\)-plane, given in Figure 4, one can conclude that training method 2 produces significantly higher \( f^2 \) scores than untrained for the students with \( x = i r \) measure in the range [20, 41]. Similarly, by inspecting the SCT for \((x^T B_2 - x^T B_1)\) (see Figures 5 and 6), one can also conclude that training method 1 is also
significantly different from untrained because this SCT does not include the zero line completely. However, the SCT for $(\mathbf{x}^T \mathbf{B}_3 - \mathbf{x}^T \mathbf{B}_1)$ contains the zero line over $x \in [0, 78.6]$, and so, there is no significant difference between the two training methods. As all these inferences are based on the three SCTs with a simultaneous confidence level 95%, one can claim that all the inferences made are correct simultaneously with confidence level 95%.

Now suppose that one is only interested in whether and how the two training methods are different from the untrained method. If one uses Roy’s test for this purpose, then the same test, as given above, has to be used with the same conclusion that $H_0$ is rejected. On the other hand, one can use the SCTs in (1) with $\Lambda = \{(i, j) : i = 1, 2 \leq j \leq 3\}$ specifically for the inferences about $(\mathbf{x}^T \mathbf{B}_2 - \mathbf{x}^T \mathbf{B}_1)$ and $(\mathbf{x}^T \mathbf{B}_3 - \mathbf{x}^T \mathbf{B}_1)$ only. For $\alpha = .05$ and the covariate range $x \in [0, 78.6]$, the critical constant $c$ is computed to be 0.0424, which is smaller than the critical constant 0.0462 for pairwise comparisons as expected. Again, one can look at the two SCTs for $(\mathbf{x}^T \mathbf{B}_2 - \mathbf{x}^T \mathbf{B}_1)$ and $(\mathbf{x}^T \mathbf{B}_3 - \mathbf{x}^T \mathbf{B}_1)$ over $x \in [0, 78.6]$, respectively, to make appropriate inferences. As in this case, one is interested in two comparisons only, the corresponding SCTs are
FIGURE 6  The projection of the SCT for $(x^T B_2 - x^T B_1)$

(0.0462 − 0.0424)/0.0424 = 9% narrower and so allow sharper inferences than the SCTs for the three pairwise comparisons given above. This demonstrates how more informative SCTs can be constructed for particular inferences of interest.

In this particular example, there are two responses $y_1$ and $y_2$, and therefore, all the SCTs can be plotted in the three-dimensional space. By inspecting these SCTs directly, inferences about the comparisons of the models can be made. When there are more than two responses, the SCTs cannot be plotted in the three-dimensional space. This is, of course, due to the multivariate nature of the problem as with many other multivariate statistical techniques. On the other hand, if one is only interested in judge whether the zero line is completely contained in an SCT, then one can use the multiplicity-adjusted $p$-values of Westfall and Young (1993) in a way similar to what is used in Liu (2010, pp. 166–168) for the univariate regression case. Specifically, one first computes the observed value $t_{i,j}$ of

$$T_{i,j} = \sup_{x_l \in [a_l, b_l], l=1, \ldots, p} \left( \frac{(x^T \hat{B}_i - x^T \hat{B}_j) \nu^{-1} - (x^T \hat{B}_i - x^T \hat{B}_j)^T}{x^T (X^T_i X_i)^{-1} + (X^T_j X_j)^{-1}} \right).$$

One then computes $p_{i,j} = P\{T > t_{i,j}\}$ for $(i, j) \in \Lambda$ by simulating a large number of replicates of $T$, using the expressions (2) and (3) as before. Now the SCT for $(x^T B_i - x^T B_j)$ contains the zero line over the given covariate region if and only if $p_{i,j} > \alpha$. This allows one to judge whether an SCT contains the zero line without looking at the plot of the SCT.

For example, for the SCTs for pairwise comparisons of the three models, our R program has computed $p_{1,2} = 0.0195$, $p_{1,3} = 0.0020$, and $p_{2,3} = 0.8420$ (based on $R = 1,000,000$ simulations). From these, one can conclude directly that the SCTs for $(x^T B_2 - x^T B_1)$ and $(x^T B_3 - x^T B_1)$ do not contain the zero line and that the SCT for $(x^T B_1 - x^T B_2)$ does contain the zero line, which agrees with what one can see from the plots of the SCTs as expected. For the SCTs for comparisons of the two training methods with the untrained method, our R program has computed $p_{1,2} = 0.0129$ and $p_{1,3} = 0.0013$. From these, one can conclude directly that the SCTs for $(x^T B_2 - x^T B_1)$ and $(x^T B_3 - x^T B_1)$ do not contain the zero line.

In this section, we have considered the comparison of $k \geq 3$ models and demonstrated that the proposed SCTs are more informative than the classical hypotheses testing approach. It is tempting to compare the distances (or areas) between $y_1$ and 0 from the projection plots of SCTs for $x^T (\hat{B}_2 - \hat{B}_1) - x^T (B_2 - B_1)$ and $x^T (\hat{B}_3 - \hat{B}_1) - x^T (B_3 - B_1)$ to decide which method (group) would be better with respect to the untrained group. We think that using just one projection of the confidence tube, in $(y_1, x)$ plane or in $(y_2, x)$ plane, to judge whether one method is better than the other is risky as the
projection in \((y_1, x)\) plane may be above the zero line and the projection in \((y_2, x)\) plane may be below the zero line, and therefore, the two projections give opposite conclusions.

5 CONCLUSIONS

Much of the research on multiple comparison and simultaneous inference in the past 60 years or so has been for the comparison of several population means. Spurrier (1999) studies the multiple comparison of several simple linear regression lines by using simultaneous confidence bands. In this paper, the work of Liu et al. (2004) for finite comparisons of several univariate linear regression models by using simultaneous confidence bands has been extended to finite comparison of several multivariate linear regression models by using SCTs. We have demonstrated how the critical constants for many types of comparison can be easily computed by Monte Carlo simulation as in Liu et al. (2004, 2005).

An SCT provides useful information on the difference between two multivariate linear regression models over a given range of the explanatory variables. This information can be used to detect differences between the two models over the range, as illustrated in the examples provided. Potentially, it can also be used to establish the maximum difference and hence equivalence of the two models over the range. A set of SCTs is certainly more informative than the current textbook approach of a hypothesis test for comparing several multivariate linear regression models, which allows only two decisions, rejection or nonrejection of \(H_0\), and does not take into consideration of specific comparisons, such as the comparisons of several treatments with one control, which may be of interest in a problem.

It is also pointed out in this paper that Roy (1953) test for comparing two multivariate linear regression models is implied by an SCT for the difference of the two models over the whole covariate space. But often, comparison over a finite covariate space is of interest in applications because regression models are good approximations usually over a finite covariate space only. SCTs utilize this finite covariate space restriction naturally in its construction.

Many problems warrant further research, for example, the construction of SCTs of different shapes for different inferential purposes, assessing the nonsuperiority of one multivariate linear regression model to the other and the equivalence of two multivariate linear regression models by extending ideas from Liu et al. (2009) in the univariate regression case. When \(p = 1\) and \(m = 2\), we have used examples to show how to make further inferences by plotting the SCTs in a three-dimensional space, or the projection of the SCT to a specific two-dimensional plane. For \(p = 2\) (with covariates \(x_1\) and \(x_2\)) and \(m = 2\), we can fix the value of one covariate, say \(x_2\), then plot the three-dimensional tube as a function of \(x_1\) (which is just one slice). For more complicated cases, where \(p > 1\) and/or \(m > 2\), plotting is not possible and it is of interest to study how one can make informative inferences using the SCTs.

ACKNOWLEDGMENTS

We are grateful to the AE and two referees’ constructive comments. The first author’s research was partially supported by NSERC of Canada.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Jianan Peng https://orcid.org/0000-0002-4220-7515
Frank Bretz https://orcid.org/0000-0002-2008-8340

REFERENCES

Al-saidy, O. M., Piegorsch, W. W., West, R. W., & Nitcheva, D. K. (2003). Confidence bands for low-dose risk estimation with quantal response data. Biometrics, 59, 1056–1062.
Anderson, T. W. (2003). An introduction to multivariate statistical analysis. New York: Wiley.

Bhargava, P., & Spurrier, J. D. (2004). Exact confidence bounds for comparing two regression lines with a control regression line on a fixed interval. Biometrical Journal, 46, 720–730.

Bretz, F., Hothorn, T., & Westfall, P. (2011). Multiple comparisons using R. New York: Chapman and Hall/CRC.

Dette, H., Möllenhoff, K., Volgushev, S., & Bretz, F. (2018). Equivalence of regression curves. Journal of the American Statistical Association, 113, 711–729.

Deutsch, R. C., & Piegorsch, W. (2012). Benchmark dose profiles for joint-action quantal data in quantitative risk assessment. Biometrics, 68, 1313–1322.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50, 1096–1121.

Edwards, D., & Berry, J. J. (1987). The efficiency of simulation-based multiple comparisons. Biometrics, 43, 913–928.

Hochberg, Y., & Tamhane, A. C. (1987). Multiple comparison procedures. New York: John Wiley & Sons, Inc.

Hsu, J. C. (1996). Multiple Comparisons: Theory and Methods. Boca Raton, FL: Chapman and Hall/CRC.

Liu, W. (2010). Simultaneous inference in regression. New York: CRC Press.

Liu, W., Bretz, F., Hayter, A. J., & Wynn, H. P. (2009). Assessing nonsuperiority, noninferiority, or equivalence when comparing two regression models over a restricted covariate region. Biometrics, 65, 1279–1287.

Liu, W., Han, Y., Wan, F., Bretz, F., & Hayter, A. J. (2016). Simultaneous confidence tubes in multivariate linear regression. Scandinavian Journal of Statistics, 43, 879–885.

Liu, W., Jamshidian, M., & Zhang, Y. (2004). Multiple comparison of several linear regression models. Journal of the American Statistical Association, 99, 395–403.

Liu, W., Jamshidian, M., Zhang, Y., & Donnelly, J. (2005). Simulation-based simultaneous confidence bands in multiple linear regression with predictor variables constrained in intervals. Journal of Computational and Graphical Statistics, 14, 459–484.

Liu, W., Wynn, H. P., & Hayter, A. J. (2008). Statistical inferences for linear regression models when the covariates have functional relationships: Polynomial regression. Journal of Statistical Computation and Simulation, 78, 315–324.

Lu, X., & Kuriki, S. (2017). Simultaneous confidence bands for contrasts between several nonlinear regression curves. Journal of Multivariate Analysis, 155, 83–104.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). Multivariate analysis. New York: Academic Press.

Miller, R. G. (1981). Simultaneous statistical inference. New York: Springer New York.

Naiman, D. Q. (1987). Simultaneous confidence bounds in multiple regression using predictor variable constraints. Journal of Statistical Computation and Simulation, 82, 214–219.

Nitcheva, D. K., Piegorsch, W. W., West, R. W., & Kodell, R. L. (2005). Multiplicity-adjusted inferences in risk assessment: Benchmark analysis with quantal response data. Biometrics, 61, 277–286.

Peng, J., Robichaud, M., & Alsulibie, A. Q. (2014). Simultaneous confidence bands for low-dose risk estimation with quantal data. Biometrical Journal, 57, 27–38.

Piegorsch, W. W., & Casella, G. (1988). Confidence bands for logistic regression with restricted predictor variables. Biometrics, 44, 739–750.

Piegorsch, W. W., West, R. W., Pan, W., & Kodell, R. L. (2005). Low dose risk estimation via simultaneous statistical inferences. Journal of the Royal Statistical Society: Series C, 54, 245–258.

Raykov, T., & Marcoulides, G. A. (2008). An introduction to applied multivariate analysis. New York: Routledge.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. Annals of Mathematical Statistics, 24, 220–238.

Ruberg, S. J., & Hsu, J. C. (1992). Multiple comparison procedures for pooling batches in stability studies. Technometrics, 34, 465–472.

Schefé, H. (1953). A method for judging all contrasts in the analysis of variance*. Biometrika, 40, 87–110.

Serfling, R. J. (1980). Approximation theorems of mathematical statistics. New York: John Wiley & Sons, Inc.

Spurrier, J. D. (1999). Exact confidence bounds for all contrasts of three or more regression lines. Journal of the American Statistical Association, 94, 720–730.

Spurrier, J. D. (2002). Exact multiple comparisons of three or more regression lines: Pairwise comparisons and comparisons with a control. Biometrical Journal, 44, 801–812.

Tukey, J. W. (1953). The problem of multiple comparisons. Dittoed manuscript of 396 pages, Department of Statistics, Princeton University.

Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for P-value adjustment. New York: Wiley.

How to cite this article: Peng J, Liu W, Bretz F, Hayter AJ. Simultaneous confidence tubes for comparing several multivariate linear regression models. Biometrical Journal. 2022;64:290–300.

https://doi.org/10.1002/bimj.202000148

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of the article.