Advances and Applications in Statistics
Volume 9, Number 2, 2008, Pages 293-301
Published Online: September 17, 2008
This paper is available online at http://www.pphmj.com

A NOTE ON THE COMPARISON OF SEVERAL LINEAR REGRESSION MODELS

WEI LIU*, WAN-KAI PANG, PING-KEI LEUNG and SHUI-HUNG HOU

*School of Mathematics
University of Southampton
Southampton, SO17 1BJ
U. K.

Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong
e-mail: maleung@inet.polyu.edu.hk

Abstract

Construction of simultaneous confidence bands for all the contrasts of the multiple linear regression over the entire real line has been studied in the past. However if the independent variables are bounded, the critical values under normal theory for multiple comparison will be inappropriate since the exact form of the sampling distribution is intractable. In this article, we use simulation to generate the critical values of the sampling distribution for comparing simultaneous confidence bands for bounded variables. These critical values have not been available in existing literatures.

2000 Mathematics Subject Classification: 62J15.

Keywords and phrases: linear regression, simultaneous inference, multiple comparisons, statistical simulation.

This research work is supported by the Research Committee of The Hong Kong Polytechnic University (Grant code: A-PH55).

Communicated by Wai Cheung Ip

Received May 28, 2007; Revised March 17, 2008

1. Introduction

Consider the problem of comparing $k(\geq 3)$ linear regression models. Suppose the ith linear regression model is given by

$$\mathbf{Y}_i = \mathbf{X}_i \mathbf{b}_i + \mathbf{e}_i, \quad i = 1, \ldots, k, \tag{1}$$

where $\mathbf{Y}_i^T = (y_{i1}, \ldots, y_{in_i})$, $\mathbf{X}_i$ is an $n_i \times (p+1)$ full column rank matrix with the first column given by $(1, \ldots, 1)^T$ and the pth column (≥ 2) given by $(x_{1,p-1}^i, \ldots, x_{n_i,p-1}^i)$, $\mathbf{b}_i^T = (\mathbf{b}_0^i, \ldots, \mathbf{b}_p^i)$, and $\mathbf{e}_i^T = (e_{i1}, \ldots, e_{in_i})$ with all the $\{e_{ij}, i = 1, \ldots, k, j = 1, \ldots, n_i\}$ being i.i.d. $N(0, \sigma^2)$. p is the number of independent variables in each regression model. Since $\mathbf{X}_i^T\mathbf{X}_i$ is non-singular, the least squares estimator of $\mathbf{b}_i$ is given by $\hat{\mathbf{b}}_i = (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{Y}_i$, $i = 1, \ldots, k$. Let $\hat{\sigma}^2$ denote the pooled error mean square with degrees of freedom $\nu = \sum_{i=1}^k (n_i - p - 1)$; $\hat{\sigma}^2$ is independent of the $\hat{\mathbf{b}}_i$'s.

Let $\mathbf{x} = (1, x_1, \ldots, x_p)^T$. Liu et al. [4] proposed the following set of simultaneous confidence bands for the comparison of the regression models

$$\mathbf{x}^T\mathbf{b}_i - \mathbf{x}^T\mathbf{b}_j \in \mathbf{x}^T\hat{\mathbf{b}}_i - \mathbf{x}^T\hat{\mathbf{b}}_j \pm \gamma_\alpha \hat{\sigma}\sqrt{\mathbf{x}^T\Delta_{ij}\mathbf{x}}, \tag{2}$$

for all $x_i \in [a_l, b_l]$ and for $l = 1, \ldots, p$, where a_l and b_l are two real constants with $a_l < b_l$, and for all i and j belong in Λ. Here $\Delta_{ij} = (\mathbf{X}_i^T\mathbf{X}_i)^{-1} + (\mathbf{X}_j^T\mathbf{X}_j)^{-1}$, Λ is a given index set that determines the comparison of interest (e.g., if the pairwise comparison is of interest, then $\Lambda = \{(i, j) : 1 \leq i \neq j \leq k\}$, if the comparisons of the second to kth regression models with the first regression model are of interest, then $\Lambda = \{(i, j) : 2 \leq i \neq k, j = 1\}$; if the successive comparison of the kth regression model is of interest, then $\Lambda = \{(i, i+1) : 1 \leq i \leq k-1\}$, $x_l \in [a_l, b_l]$ for $l = 1, \ldots, p$ is a given range over which the comparison of linear regression models $\mathbf{x}^T\mathbf{b}_i$'s is of interest, and γ_α is the critical constant chosen so that the confidence level of this set of simultaneous confidence bands is equal to $1 - \alpha$. Liu et al. [4] provided some simulation methods to determine γ_α.

For the special case of $p = 1$ and $X_1 = \cdots = X_k$, Spurrier [6] provided a set of simultaneous confidence bands for all contrasts of the regression lines $\sum_{i=1}^{k} c_i\mathbf{x}^T\mathbf{b}_i$ over the entire range of the explanatory variable $x_l \in (-\infty, \infty)$, where $\sum_{i=1}^{k} c_i = 0$. Spurrier [7] considered the pairwise comparison of the regression models over the entire range of the explanatory variables under the assumption $X_1 = \cdots = X_k$, and the comparison of the first $k-1$ regression lines with the kth regression line over the entire range of the explanatory variables under certain assumptions on the X_i's.

Note that the set of $1 - \alpha$ simultaneous confidence bands in (2) can be used to test the hypotheses

$$H_0 : \mathbf{b}_1 = \cdots = \mathbf{b}_k \text{ against } H_a : \text{not } H_0 \tag{3}$$

by rejecting H_0 if and only if $T > \gamma_\alpha$, where

$$T = \sup_{1 \le i \ne j \le k} \sup_{x_l \in [a_l, b_l],\, l=1, \ldots, p} \frac{|\mathbf{x}^T((\hat{\mathbf{b}}_i - \mathbf{b}_i) - (\hat{\mathbf{b}}_j - \mathbf{b}_j))|}{\hat{\sigma}\sqrt{\mathbf{x}^T\Delta_{ij}\mathbf{x}}}. \tag{4}$$

This test is of size α. On the other hand, it is well known that the hypotheses in (3) can be tested by an F-test (see Section 2). A natural question is to compare the values of d_α and γ_α, where d_α is the usual value from the normal theory of unbounded variables. To shed some lights on this question, we carry out a simulation study to generate the values of γ_α which are unavailable in the existing literature. This is discussed in Section 3. In order to understand better the F-test, we derive in Section 2 the set of conservative simultaneous confidence bands for all contrasts of the regression models that is associated with the F-test. Finally, we provide some concluding remarks in Section 4.

2. The F-test and Associated Confidence Bands

The most familiar form of the F-test for testing the hypotheses (3) is to define $k-1$ zero-one dummy variables to represent all the observations $(\mathbf{Y}_i, X_i)$; $i = 1, \ldots, k$ by one overall linear regression model and then to apply a partial F-test to test that certain coefficients of this overall model are equal to zero (see Kleinbaum et al. [3]). An equivalent but less familiar form of the F-test is given by the following (see Scheffé [5]).

Now if we represent all the observations $(\mathbf{Y_i}, \mathbf{X_i})$; $i = 1, \ldots, k$ by

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e}, \tag{5}$$

where $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_k^T)^T$, $\mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_k^T)^T$, $\mathbf{e} = (e_1^T, \ldots, e_k^T)^T$, and $\mathbf{X}$ is the partition diagonal matrix $\operatorname{diag}(X_\mathbf{i})$. The least squares estimator of $\mathbf{b}$ is clearly given by $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1^T, \ldots, \hat{\mathbf{b}}_k^T)^T$. The hypotheses (3) now become

$$H_0 : \mathbf{Hb} = 0 \quad \text{against} \quad H_a : \text{not } H_{0,} \tag{6}$$

where the partition matrix $\mathbf{H}$ is given by

$$\mathbf{H} = \begin{pmatrix} I_{p+1} & -I_{p+1} & & & \\ & I_{p+1} & -I_{p+1} & & \\ & & \ddots & \ddots & \\ & & & I_{p+1} & -I_{p+1} \end{pmatrix}.$$

Note that $\mathbf{H}\hat{\mathbf{b}} \sim N(\mathbf{Hb}, \sigma^2\mathbf{H}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{H}^T)$. The F-test rejects H_0 if and only if

$$\frac{(\mathbf{H}\hat{\mathbf{b}})^T\{\mathbf{H}(X^T\mathbf{X})^{-1}\mathbf{H}^T\}^{-1}(\mathbf{H}\hat{\mathbf{b}})}{\sigma^2} > d_\alpha, \tag{7}$$

where $d_\alpha = (k-1)(p+1)F^{\alpha}_{(k-1)(p+1), N-k(p+1)}$ with $F^{\alpha}_{(k-1)(p+1), N-k(p+1)}$ being the upper α point of an F distribution with $(k-1)(p+1)$ and $N-k(p+1)$ degrees of freedom.

Now we derive a set of simultaneous confidence bands, associated with the F-test, for all contrasts of the regression models $\sum_{i=1}^{k} c_i\mathbf{x}^T\mathbf{b}_i$ for all $\mathbf{c} = (c_1, ..., c_k)^T$ satisfying $\sum_{i=1}^{k} c_i = 0$ over the entire range of all the explanatory variables $x_l \in (-\infty, \infty)$, $l = 1, ..., p$. Note the following confidence set for $\mathbf{b}$ that underlies the F-test in (7):

$$P\{[\mathbf{H}(\hat{\mathbf{b}} - \mathbf{b})]^T \{\mathbf{H}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{H}^T\}^{-1}[\mathbf{H}(\hat{\mathbf{b}} - \mathbf{b})] < d_\alpha\hat{\sigma}^2\} = 1 - \alpha. \tag{8}$$

Let the square matrix $\mathbf{Q}$ satisfy $\{\mathbf{H}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{H}^T\}^{-1} = \mathbf{Q}^T\mathbf{Q}$. Then we have

$$[\mathbf{H}(\hat{\mathbf{b}} - \mathbf{b})]^T \{\mathbf{H}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{H}^T\}^{-1}[\mathbf{H}(\hat{\mathbf{b}} - \mathbf{b})] < d_\alpha\hat{\sigma}^2,$$

$$[\mathbf{QH}(\hat{\mathbf{b}} - \mathbf{b})]^T[\mathbf{QH}(\hat{\mathbf{b}} - \mathbf{b})] < d_\alpha\hat{\sigma}^2,$$

$$-\sqrt{d_\alpha}\,\hat{\sigma} < \frac{\mathbf{v}^T}{\|\mathbf{v}\|}\mathbf{QH}(\hat{\mathbf{b}} - \mathbf{b}) < \sqrt{d_\alpha}\,\hat{\sigma} \quad \text{for all} \quad \mathbf{v} \in R^{(k-1)(p+1)}, \tag{9}$$

where the last equivalence follows from a simple geometric projection result (see, e.g., Hsu [1, pp. 231-233]). Now let $\mathbf{w}^T = \mathbf{v}^T\mathbf{Q}.$ Then (9) is further equivalent to

$$-\sqrt{d_\alpha}\,\hat{\sigma} < \frac{\mathbf{w}^T\mathbf{H}(\hat{\mathbf{b}} - \mathbf{b})}{\sqrt{\mathbf{w}^T\mathbf{H}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{H}^T\mathbf{w}}} < \sqrt{d_\alpha}\,\hat{\sigma}, \tag{10}$$

for all $\mathbf{w} \in R^{(k-1)(p+1)}$. For $\mathbf{w}$ of the form $\mathbf{w} = (c_1\mathbf{x}^T, (c_1 + c_2)\mathbf{x}^T, ..., (c_1 + \cdots + c_{k-1})\mathbf{x}^T)^T$ and $\sum_{i=1}^{k} c_i = 0$ or $\sum_{i=1}^{k-1} c_i = 0$, we have $\mathbf{w}^T\mathbf{H}(\hat{\mathbf{b}} - \mathbf{b}) = \sum_{i=1}^{k} c_i(\mathbf{x}^T\hat{\mathbf{b}} - \mathbf{x}^T\mathbf{b})$ and so (10) implies

$$-\sqrt{d_\alpha}\,\hat{\sigma} < \frac{\sum_{i=1}^{k} c_i(\mathbf{x}^T\hat{\mathbf{b}} - \mathbf{x}^T\mathbf{b})}{\sqrt{\sum_{i=1}^{k} c_i^2\mathbf{x}^T(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{x}}} < \sqrt{d_\alpha}\,\hat{\sigma}, \tag{11}$$

for all $x_i \in (-\infty, \infty)$, $i, ..., p$ and for all $\mathbf{c}$ satisfying $\sum_{i=1}^{k} c_i = 0$. Therefore confidence statement (8) implies, with a probability of at least $1 - \alpha$ that

$$\sum_{i=1}^{k} c_i\mathbf{x}^T\mathbf{b}_i \in \sum_{i=1}^{k} c_i\mathbf{x}^T\mathbf{b}_i \pm \sqrt{d_\alpha}\,\hat{\sigma}\sqrt{\sum_{i=1}^{k} c_i^2\mathbf{x}^T(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{x}} \tag{12}$$

for all $x_i \in (-\infty, \infty)$, $i, ..., p$ and for all $\mathbf{c}$ satisfying $\sum_{i=1}^{k} c_i = 0$. This provides a set of conservative simultaneous confidence bands for all contrasts of the regression models.

Note that the simultaneous confidence bands (12) are of level $1 - \alpha$ exactly for $k = 2$ and are strictly conservative for $k \geq 3$. It is not clear what critical value should be in the place of $\sqrt{d_\alpha}$ in (12) so that the simultaneous confidence level of the bands in (12) is equal to $1 - \alpha$ for $k \geq 3$. Spurrier [6] provided the answer to this question for the special situation of $p = 1$ and $\mathbf{X}_1 = \cdots = \mathbf{X}_k$.

The bands (12) include the bands for the pairwise comparisons of any two regression models $\mathbf{x}^T\mathbf{b}_i$ and $\mathbf{x}^T\mathbf{b}_j$ (by choosing the ith and jth elements of $\mathbf{c}$ to be one and the rest to be zero),

$$\mathbf{x}^T\mathbf{b}_i - \mathbf{x}^T\mathbf{b}_j \in \mathbf{x}^T\hat{\mathbf{b}}_i - \mathbf{x}^T\hat{\mathbf{b}}_j \pm \sqrt{d_\alpha}\,\hat{\sigma}\sqrt{\mathbf{x}^T\Delta_{ij}\mathbf{x}} \tag{13}$$

for all $x_i \in (-\infty, \infty)$, $i = 1, \ldots, p$ and for all $1 \leq i \neq j \leq k$. In comparison with the simultaneous confidence bands (2) with $\Lambda = \{(i, j) : 1 \leq i \neq j \leq k\}$, these confidence bands are conservative for two reasons. Firstly, the critical value $\sqrt{d_\alpha}$ is conservative. By conservative, we mean that the critical region given by $\sqrt{d_\alpha}$ is smaller than the other critical values under other sampling distributions. Secondly, these bands are over the entire range $x_i \in (-\infty, \infty)$, $i = 1, \ldots, p$, on which it is inconceivable the regression models will be true for any real problems. The bands (13) should be used as a conservative substitute only if the critical value γ_α of the bands (2) is not easily available.

3. Critical Values for the Bounded *T*-test

Before we could carry-out a simulation study on the power of the T-test for bounded variables using simultaneous confidence bands, we need to obtain the critical values for this test. The critical values of this test can only be obtained via simulation as the sampling distribution is mathematically intractable. The algorithm for generating these critical values can be found in Liu et al. [4]. Here we simulate three linear contrasting regression models with bounded explanatory $x_i \in [10, 20]$, each of size n. The linear contrasting regression models are

$$Y_{1i} = 1.0 + 2.0X_{1i} + e_{1i}, \quad i = 1, ..., n, \tag{14}$$

$$Y_{2i} = 2.0 + 4.0X_{2i} + e_{2i}, \quad i = 1, ..., n, \tag{15}$$

$$Y_{3i} = -3.0 + 6.0X_{3i} + e_{3i}, \quad i = 1, ..., n. \tag{16}$$

The error component e_j, $j = 1, 2, 3,$ follows $N(0, \sigma^2)$. We then used the algorithm provided by Liu et al. [4] to simulate the γ_α critical values for various sample sizes n. A Fortran program is written specifically for this algorithm and IMSL [2] subroutines are used in the program for various random number generations. We conducted the simulation run for one million times and obtained the 90th, 95th and 99th percentiles as the critical values of the T-test with bounded explanatory variable at $\alpha = 0.1, 0.05,$ and 0.01 levels of significance. In Table 1, we present the simulated γ_α critical values at 5% level of significance so as to compare with the critical values given by Spurrier [6] as well as the d critical values.

Table 1. Critical values at $\alpha = 5\%$ level of significance

Sample size n	γ_α	Spurrier$_\alpha$	d_α	Sample size n	γ_α	Spurrier$_\alpha$	d_α
3	6.078	5.712	6.039	12	2.791	3.118	3.280
4	3.944	4.036	4.256	15	2.731	3.074	3.231
5	3.418	3.617	3.811	20	2.676	3.034	3.189
6	3.183	3.428	3.611	25	2.647	3.011	3.165
7	3.050	3.321	3.498	30	2.630	2.997	3.149
8	2.957	3.251	3.423	40	2.607	2.979	3.131
9	2.898	3.203	3.388	50	2.596	2.969	3.120
10	2.855	3.167	3.335	60	2.584	2.963	3.113
11	2.817	3.140	3.304	100	2.565	unavailable	3.100

In Table 2, we present the simulated γ_α critical values at 10% and 1% level of significances.

Table 2. Simulated γ_α critical values

Sample size n	10% Level	1% Level	Sample size n	10% Level	1% Level
3	4.570	11.021	12	2.438	3.536
4	3.221	5.845	15	2.392	3.433
5	2.871	4.758	20	2.355	3.334
6	2.710	4.272	25	2.335	3.288
7	2.617	4.025	30	2.325	3.244
8	2.553	3.862	40	2.307	3.205
9	2.512	3.752	50	2.301	3.185
10	2.482	3.658	60	2.292	3.168
11	2.456	3.594	100	2.291	3.159

Spurrier [6] only presented the 5% critical values. We have in fact obtained the γ_α critical values for 1% and 10% levels. The γ_α critical values with bounded explanatory variable are uniformly smaller than the unbounded case (see Spurrier [6]) except when $n = 3$. Also we have tried several other sets of linear contrasting regression models with $k = 3$ to simulate the γ_α critical values, the results are of no difference with those presented in Tables 1 and 2. Therefore we are quite confident with our results and it will be straightforward to extend this method to the contrasting regression models involving two or more bounded explanatory variables. The algorithms were also given in Liu et al. [4].

4. Conclusion

In conclusion, we have obtained the γ_α-critical values for the regression model with bounded variables by simulation. These γ_α-critical values are not yet available in the existing literatures. Our algorithm can be extended to obtain other critical values for $p \geq 2$ and $k \geq 4$ and the simulation work is straightforward. Our results are important since one should use the γ_α-critical values for testing contrasts in multiple comparison of regression models if the independent variables are bounded. This is often the case in many applications.

References

[1] J. C. Hsu, Multiple Comparisons: Theory and Methods, Chapman & Hall, London, 1996.

[2] IMSL, Fortran Subroutines for Mathematical and Statistical Analysis, IMSL, Inc., Texas, 1992.

[3] D. G. Kleinbaum, L. K. Kupper and K. E. Muller, Applied Regression Analysis and other Multivariable Methods, 2nd ed., PWS-Kent Pub. Co., Boston, 1988.

[4] W. Liu, M. Jamshidian and Y. Zhang, Multiple comparison of several linear regression models, J. Amer. Statist. Assoc. 99 (2004), 395-403.

[5] H. Scheffé, A method for judging all contrasts in the analysis of variance, Biometrika 40 (1953), 87-104.

[6] J. D. Spurrier, Exact confidence bounds for all contrasts of three or more regression lines, J. Amer. Statist. Assoc. 94 (1999), 438-488.

[7] J. D. Spurrier, Exact multiple comparisons of three or more regression lines: pairwise comparisons and comparisons with a control, Biometrical J. 44 (2002), 801-812.

■