Listy Biometryczne – Biometrical Letters
Vol. 41 (2004), No. 1, 25-35

Multiple comparisons of slopes of regression lines

Jolanta Wojnar[1], Wojciech Zieliński[2]

[1]Institute of Statistics and Econometrics, University of Rzeszów, ul. Ćwiklińskiej 2, 35-601 Rzeszów
[2]Department of Econometrics and Computer Sciences, Agricultural University, ul. Nowoursynowska 166, 02-787 Warszawa

SUMMARY

The problem of comparing K simple regression lines is considered. A statistical procedure for finding groups of parallel lines is proposed.

KEY WORDS: multiple comparisons, simultaneous inference, regression analysis.

1. Introduction

Consider K regression lines $Y = \alpha_k + \beta_k x + \varepsilon$, $k = 1, \ldots, K$ and assume that ε's are i.i.d. random variables distributed as $N(0, \sigma^2)$. The problem is to divide a set of regression coefficients $\{\beta_1, \ldots, \beta_K\}$ into homogeneous groups. A subset $\{\beta_{i_1}, \ldots, \beta_{i_m}\}$ is called the homogenous group if $\beta_{i_1} = \cdots = \beta_{i_m}$ and any other $\beta \in \{\beta_1, \ldots, \beta_K\}$ is not equal to β_{i_1}.

This problem is similar to the problem of extracting homogeneous groups of means in the ANOVA. Some of classical multiple comparison procedures (cf Miller, 1982, Hochberg and Tamhane, 1988) such as *Tukey, Scheffé, Bonfferroni* can be adopted to the above problem. In what follows, the W procedure of multiple comparison proposed by Zieliński (1992) is used. As a criterion of the procedure quality the probability of the correct decision is taken.

2. Statistical model

Assume that for each of the regression functions we have n_k observations (x_{ki}, Y_{ki}). Than the overall number of observations is $N = \sum_k n_k$. Hence we have the joint model

$$Y_{ki} = \alpha_k + \beta_k x_{ki} + \varepsilon_{ki},\ i = 1, \ldots, n_k,\ k = 1, \ldots, K, \tag{1}$$

where α's and β's are unknown regression coefficients and ε's are independent normally distributed random variables with mean zero and variance σ^2. In matrix notation the model may be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the vector $\boldsymbol{\varepsilon}$ is distributed as $N_N(\mathbf{0}, \sigma^2\mathbf{I})$ and

$$\mathbf{y}' = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{K1}, \ldots, Y_{Kn_K}),$$

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & 0 & \ldots & 0 & x_{11} & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 & x_{12} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & \ldots & 0 & x_{1n_1} & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & x_{21} & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & x_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 1 & \ldots & 0 & 0 & x_{2n_2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & x_{K1} \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & x_{K2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & x_{Kn_K} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \ldots \\ \alpha_K \\ \beta_1 \\ \beta_2 \\ \ldots \\ \beta_K \end{pmatrix}.$$

Let $\mathbf{A}$ be a given $q \times 2K$ matrix of rank r and $\mathbf{c}$ be a given $q \times 1$ vector. On the basis of the general theory of linear models we obtain the following test statistics for the hypothesis $H : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$,

$$F = \frac{(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}']^{-}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})}{\mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}')\mathbf{y}} \cdot \frac{N - \operatorname{rank}(\mathbf{X})}{r}, \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is the *LSE* of $\boldsymbol{\beta}$. Its null distribution is F with $(r, N-\operatorname{rank}(\mathbf{X}))$ degrees of freedom and the hypothesis is rejected at a significance level α if $F > F^{\alpha}_{r;N-\operatorname{rank}(\mathbf{X})}$, where $F^{\alpha}_{r;N-\operatorname{rank}(\mathbf{X})}$ is an appropriate critical value.

If for each $k = 1, ..., K$ there exist at least two different x'_{ki}s, then the rank$(\mathbf{X}) = 2K$ and there exists $(\mathbf{X}'\mathbf{X})^{-1}$ of the form $\begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_2 \\ \mathbf{W}_2 & \mathbf{W}_3 \end{bmatrix}$, where

$$\mathbf{W}_1 = \text{diag}\left\{\frac{\sum x_{1i}^2}{n_1 SS_1}, ..., \frac{\sum x_{Ki}^2}{n_K SS_K}\right\},$$

$$\mathbf{W}_2 = \text{diag}\left\{\frac{-\sum x_{1i}}{n_1 SS_1}, ..., \frac{-\sum x_{Ki}}{n_K SS_K}\right\},$$

$$\mathbf{W}_3 = \text{diag}\left\{\frac{1}{SS_1}, ..., \frac{1}{SS_K}\right\}.$$

Here diag$\{a_1, ..., a_K\}$ denotes the diagonal matrix with diagonal elements $a_1, ..., a_K$ and $SS_k = \sum_{i=1}^{n_k}(x_{ki} - \bar{x}_k)^2$. Note that

$$\frac{1}{N-2K}\mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}$$

is the least square unbiased estimator of the variance σ^2.

3. Procedure

The procedure of comparison of regression coefficients is based on the statistic (2) with $\mathbf{c} = \mathbf{0}$ and is stepwise. In the first step, it is verified whether $\beta_1 = \cdots = \beta_K$. The matrix $\mathbf{A}$ is then of the form

$$\mathbf{A} = \left[\mathbf{0}_{K\times K} \quad \vdots \quad \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}'_K\right],$$

where $\mathbf{0}_{K\times K}$ is $K \times K$ zero matrix, $\mathbf{I}_K$ denotes the identity matrix of order K and $\mathbf{1}'_K$ denotes the $K \times 1$ vector of ones. The explicite form of the nominator of (2) is

$$\sum_{k=1}^{K}\left(\widehat{\beta}_k - \widehat{\beta}\right)^2,$$

where

$$\widehat{\beta}_k = \frac{\sum_{i=1}^{n_k}(Y_{ik} - \bar{Y}_k)(x_{ik} - \bar{x}_k)}{\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2}, \qquad \widehat{\beta} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ik} - \bar{Y}_k)(x_{ik} - \bar{x}_k)}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2}.$$

Note that $\widehat{\beta}_k$ is the *LSE* of β_k for k-th regression line and $\widehat{\beta}$ is the *LSE* of the regression coefficient under assumption $\beta_1 = \cdots = \beta_K = \beta$. If the value of statistic (2) is less than $F^{\alpha}_{K-1;N-2K}$, the procedure stops, and regression coefficients are considered as equal. Elsewhere we go to the second step.

On the p-th step we consider a division of the set of regression coefficients into p disjoint homogenous groups. Let $I_1, \ldots, I_p$ be a division of $\{1, \ldots, K\}$ into p disjoint subsets. Let $\mathcal{J}(p)$ denote this division. The corresponding matrix $\mathbf{A}$ (after appropriate permutation of regression coefficients) takes on the form

$$\mathbf{A}_{\mathcal{J}(p)} = \left[\begin{array}{ccccccc} \mathbf{0}_{m_1 \times K} & \vdots & \mathbf{I}_{m_1} - \frac{1}{m_1}\mathbf{1}_{m_1}\mathbf{1}'_{m_1} & \mathbf{0}_{m_1 \times m_2} & \cdots & \mathbf{0}_{m_1 \times m_p} \\ \mathbf{0}_{m_1 \times K} & \vdots & \mathbf{0}_{m_2 \times m_1} & \mathbf{I}_{m_2} - \frac{1}{m_2}\mathbf{1}_{m_2}\mathbf{1}'_{m_2} & \cdots & \mathbf{0}_{m_2 \times m_p} \\ \cdot & \vdots & \cdot & \cdot & \cdots & \cdot \\ \mathbf{0}_{m_1 \times K} & \vdots & \mathbf{0}_{m_p \times m_1} & \mathbf{0}_{m_p \times m_2} & \cdots & \mathbf{I}_{m_p} - \frac{1}{m_p}\mathbf{1}_{m_p}\mathbf{1}'_{m_p} \end{array}\right]$$

where m_i is the cardinality of the subset I_i. Let $F_{\mathcal{J}(p)}$ denote the statistic (2) with the matrix $\mathbf{A}_{\mathcal{J}(p)}$. The nominator of (2) equals to

$$\sum_{j=1}^{p} \sum_{k \in I_j} \left(\widehat{\beta}_k - \widehat{\beta}_{I_j}\right)^2,$$

where

$$\widehat{\beta}_k = \frac{\sum_{i=1}^{n_k}(Y_{ik} - \bar{Y}_k)(x_{ik} - \bar{x}_k)}{\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2}, \qquad \widehat{\beta}_{I_j} = \frac{\sum_{k \in I_j}\sum_{i=1}^{n_k}(Y_{ik} - \bar{Y}_{I_j})(x_{ik} - \bar{x}_{I_j})}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_{I_j})^2},$$

$$\bar{Y}_{I_j} = \frac{\sum_{k \in I_j}\sum_{i=1}^{n_k} Y_{ki}}{\sum_{k \in I_j} n_k}, \qquad \bar{x}_{I_j} = \frac{\sum_{k \in I_j}\sum_{i=1}^{n_k} x_{ki}}{\sum_{k \in I_j} n_k}.$$

The estimator $\hat{\beta}_{I_j}$ is the *LSE* of regression coefficient under assumption that all β_k for $k \in I_j$ are equal.

Let $\mathcal{J}^*(p)$ be a division into p groups such that

$$F_{\mathcal{J}^*(p)} = \min F_{\mathcal{J}(p)}.$$

If $F_{\mathcal{J}^*(p)} < F^{\alpha}_{K-p;N-2K}$, then we stop the procedure and accept the division $\mathcal{J}^*(p)$. Otherwise we consider divisions into $p+1$ groups. If $p = K-1$ and $F_{\mathcal{J}^*(p)} > F^{\alpha}_{K-p;N-2K}$ holds, we decide that we have K groups, i.e. all coefficients are distinct.

4. Criterion

Let $\Theta = \{\theta_1, \theta_2, \ldots\}$ denote the set of all possible divisions of the set of regression coefficients into homogenous groups. Elements of the set Θ are disjoint subsets of $\mathbf{R}^K$ and for every $(\beta_1, \ldots, \beta_K) \in \mathbf{R}^K$ there exists only one $\theta \in \Theta$ such that $(\beta_1, \ldots, \beta_K) \in \theta$. Note that Θ is a finite set. The elements of the set Θ are commonly called *states* *of nature.* For example consider $K = 3$. The set Θ consists of the following elements:

$$\begin{aligned}
\theta_1 &= \{(\beta_1,\beta_2,\beta_3) \in \mathbf{R}^3 : \beta_1 = \beta_2 = \beta_3\},\\
\theta_2 &= \{(\beta_1,\beta_2,\beta_3) \in \mathbf{R}^3 : \beta_1 = \beta_2, \beta_3 \neq \beta_1\},\\
\theta_3 &= \{(\beta_1,\beta_2,\beta_3) \in \mathbf{R}^3 : \beta_1 = \beta_3, \beta_2 \neq \beta_1\},\\
\theta_4 &= \{(\beta_1,\beta_2,\beta_3) \in \mathbf{R}^3 : \beta_2 = \beta_3, \beta_1 \neq \beta_2\},\\
\theta_5 &= \{(\beta_1,\beta_2,\beta_3) \in \mathbf{R}^3 : \beta_1 \neq \beta_2, \beta_1 \neq \beta_3, \beta_2 \neq \beta_3\}.
\end{aligned}$$

The aim of any multiple comparison procedure is to "detect" the true state of nature. Let $\mathcal{D}$ be a set of all decisions which can be made on the basis of observations. The elements of the set $\mathcal{D}$ are called *decisions.* We assume that $\mathcal{D} = \Theta$.

We define the loss function in the following manner

$$L(d,\theta) = \begin{cases} 0, & \text{if } d = 0, \\ 1, & \text{if } d \neq 0, \end{cases} \quad \text{for } d \in \mathcal{D} \text{ and } \theta \in \Theta.$$

This loss function gives penalty of one when our decision is not correct.

If we denote by $\mathcal{X}$ the space of all observations, then the function $\delta : \mathcal{X} \ni \mathbf{x} \to d \in \mathcal{D}$ is called a *decision rule.* The considered procedure of multiple comparisons may be described as a decision rule.

A decision rule δ is characterized by its risk function, i.e., average loss. Let $(\beta_1, \ldots, \beta_K) \in \theta$. Then the risk function of the rule δ equals

$$R_\delta(\beta_1, \ldots, \beta_K) = E_{(\beta_1,\ldots,\beta_K)} L(\delta(\mathbf{x}), \theta) = P_{(\beta_1,\ldots,\beta_K)}\{\delta(\mathbf{x}) \neq \theta\}.$$

Note that in general, the risk depends on the differences of the values of the parameters $(\beta_1, \ldots, \beta_K)$. For example if we assume $K = 3$ and $\sigma^2 = 1$, then it is easier to make misclassification for $\beta_1 = \beta_2 = 1$, $\beta_3 = 1.1$ than for $\beta_1 = \beta_2 = 1$, $\beta_3 = 5$, though both situations are the same state of nature. Only in the case $\beta_1 = \cdots = \beta_K = \beta$ the risk does not depend on the value of β.

The risk of the rule δ is the probability of the false decision. This probability should be as small as possible. In our investigation we are interested in a probability of the correct decision which is equal to $1 - R_\delta$.

The most common approach to the problem under consideration is via theory of multiple hypothesis testing. In that framework, different criterions of goodness are considered, such as a Familywise Error Rate or Per Comparison Error Rate connected with controlling the risk of committing an error of type I (see Gather et al. 1996). Those criterions may be considered as a generalizations of the notion of the significance level in the Neyman–Pearson theory of testing hypotheses. According to that terminology, we may say that the probability of the correct decision is the criterion which simultaneously takes into account the possibility of committing the errors of type I and type II, as Wald decision theory does. Note that the imposed criterion, as opposed to the theory of multiple hypotheses testing, does not keep the Familywise Error Rate at a fixed level. Thus it is advisable to consider rules δ with R_δ for $\beta_1 = \cdots = \beta_K$ equal to the value of the significance level of the hypothesis $H_0 : \beta_1 = \cdots = \beta_K$. As a consequence, there is no possibility that the results obtained by the rule contradicts that obtained for the above hypothesis. The weak point of the presented approach is that there is no possibility to obtain the uniformly best procedure. But, on the other hand, we avoid the obvious disadvantage that the constructed procedures will be too conservative (i.e. giving too large homogenous groups).

5. Experiment

The probability of the correct decision was estimated on the basis of a simulation experiment. In the experiment we choose $K = 5$ regression functions and each of it was observed 20 times. Random errors were normal with $\sigma = 0.1$. Parameters $\alpha_1, \ldots, \alpha_5$ were zero. The regression functions were considered on the interval $[-1, 1]$.

For five regression functions there are 67 states of nature but it is enough to consider seven of them. The considered states are given in Table 1. Notation $\{\beta_1 = \beta_2 = \beta_3, \beta_4, \beta_5\}$ means the following subset: $\{(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) : \beta_1 = \beta_2 = \beta_3, \beta_4 \neq \beta_1, \beta_5 \neq \beta_1, \beta_4 \neq \beta_5\} \subset \mathbf{R}^5$.

Table 1. The states of nature considered for five regression functions

Number of groups	State of nature	Notation
1	$\{\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5\}$	$\{5\}$
2	$\{\beta_1 = \beta_2 = \beta_3, \beta_4 = \beta_5\}$	$\{2,3\}$
	$\{\beta_1 = \beta_2 = \beta_3 = \beta_4, \beta_5\}$	$\{1,4\}$
3	$\{\beta_1 = \beta_2 = \beta_3, \beta_4, \beta_5\}$	$\{1,1,3\}$
	$\{\beta_1 = \beta_2, \beta_3 = \beta_4, \beta_5\}$	$\{1,2,2\}$
4	$\{\beta_1 = \beta_2, \beta_3, \beta_4, \beta_5\}$	$\{1,2,2,2\}$
5	$\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$	$\{1,1,1,1,1\}$

For each state of nature, regression coefficients were generated from the interval $(0.5, 1.5)$ according to uniform distribution. For example, for the state $\{2,3\}$ two numbers x, y were generated from the distribution $U(0.5, 1.5)$ and it was set $\beta_1 = \beta_2 = \beta_3 = x$ and $\beta_4 = \beta_5 = y$. Such generation was repeated one hundred times.

For each generated regression coefficients $(\beta_1, \ldots, \beta_5)$, 1000 samples were drawn of (x_{ki}, Y_{ki}) for $k = 1, \ldots, 5$ and $i = 1, \ldots, 20$, such that $Y_{ki} = \beta_k x_{ki} + \varepsilon_{ki}$. The described procedure was applied to each sample and it was noted if the obtained division of regression coefficients is consistent with the state of nature. The probability of the correct decision was estimated by the fraction of divisions consistent with the state of nature.

It is obvious that the probability of the correct decision depends on a plan of experiment, i.e., on the choice of values of regressors. Three plans were considered. In the first case (random plan) twenty values of x's were chosen randomly from the $[-1, 1]$ interval (under the uniform distribution) for every regression function separately. The second plan was a naive one, i.e. values of regressor were $-1+2i/19$ for $i = 0, 1, \ldots, 19$. The third plan was the G–optimal plan in which $x = -1$ or $x = 1$ and at each x ten observations were taken. The second plan, as well as the third one, were common for all regression functions.

6. Results

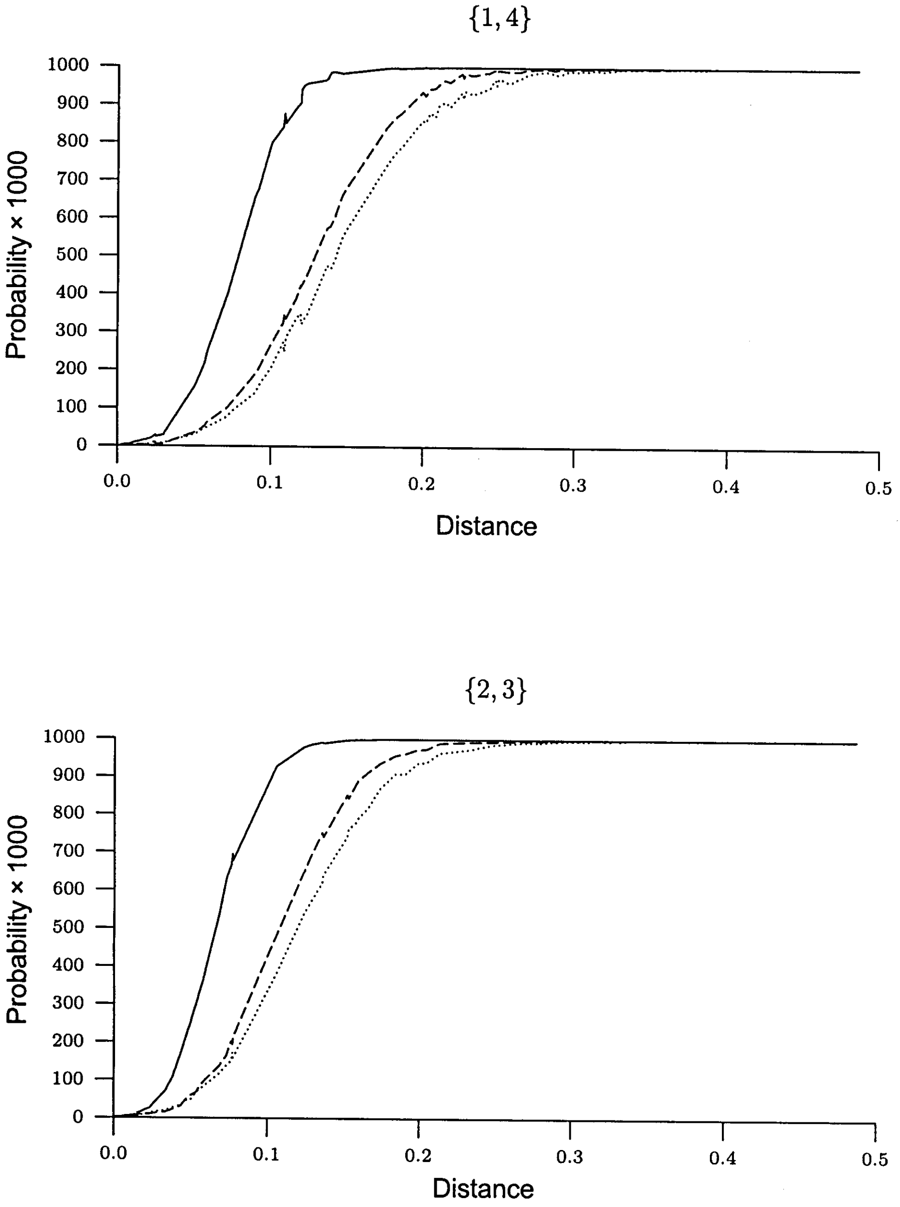
The results are presented graphically (Fig. 1a, b, c). On y axis there is the estimated probability (multiplied by 1000) of the correct decision, while on the x axis there is the minimal distance between groups. The solid line represents the probability of the correct decision for the G–optimal plan, dashed line — for the naive plan, and the dotted line — for the random plan.

On the basis of simulations we may formulate the following conclusions.

1. The proposed procedure of detecting the division of regression coefficients closely corresponding to the true states of nature is more precise when there is a small number of groups of coefficients. When we increase the number of groups of coefficients the probability of detecting differences between them is decreasing.

2. In the case of each division of regression coefficients we can conclude that the best plan of experiment was the G–optimal plan. The probability of taking the correct decision is very high even for small differences between the sets of regression coefficients.

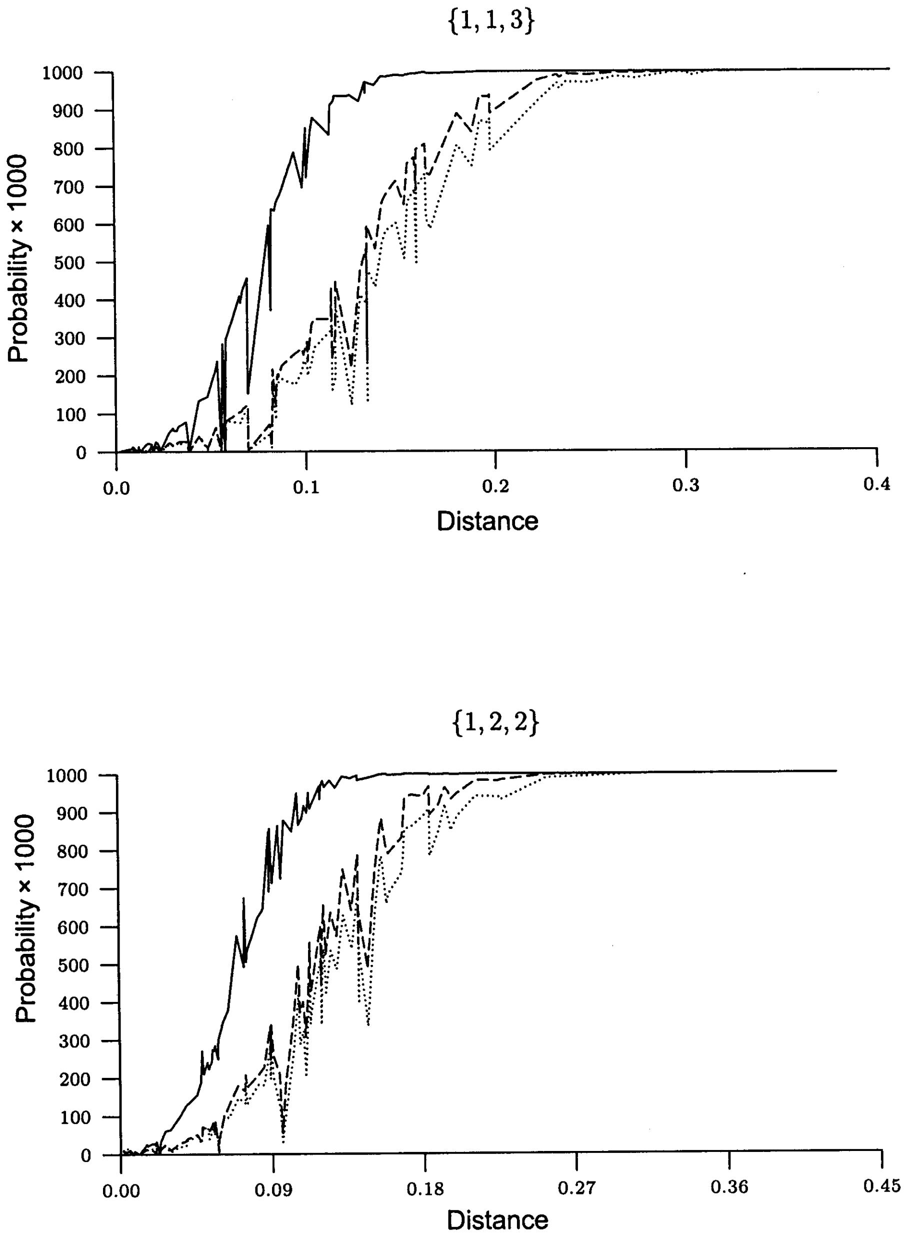
Above conclusions are true for five regression functions, but it may be expected that similar conclusions may be formulated for the higher number of functions.



{1,4}

{2,3}

Fig. 1a. Simulation results for the states of nature $\{1,4\}$ and $\{2,3\}$; see text for explanation



$\{1,1,3\}$

$\{1,2,2\}$

Fig. 1b. Simulation results for the states of nature $\{1,1,3\}$ and $\{1,2,2\}$; see text for explanation

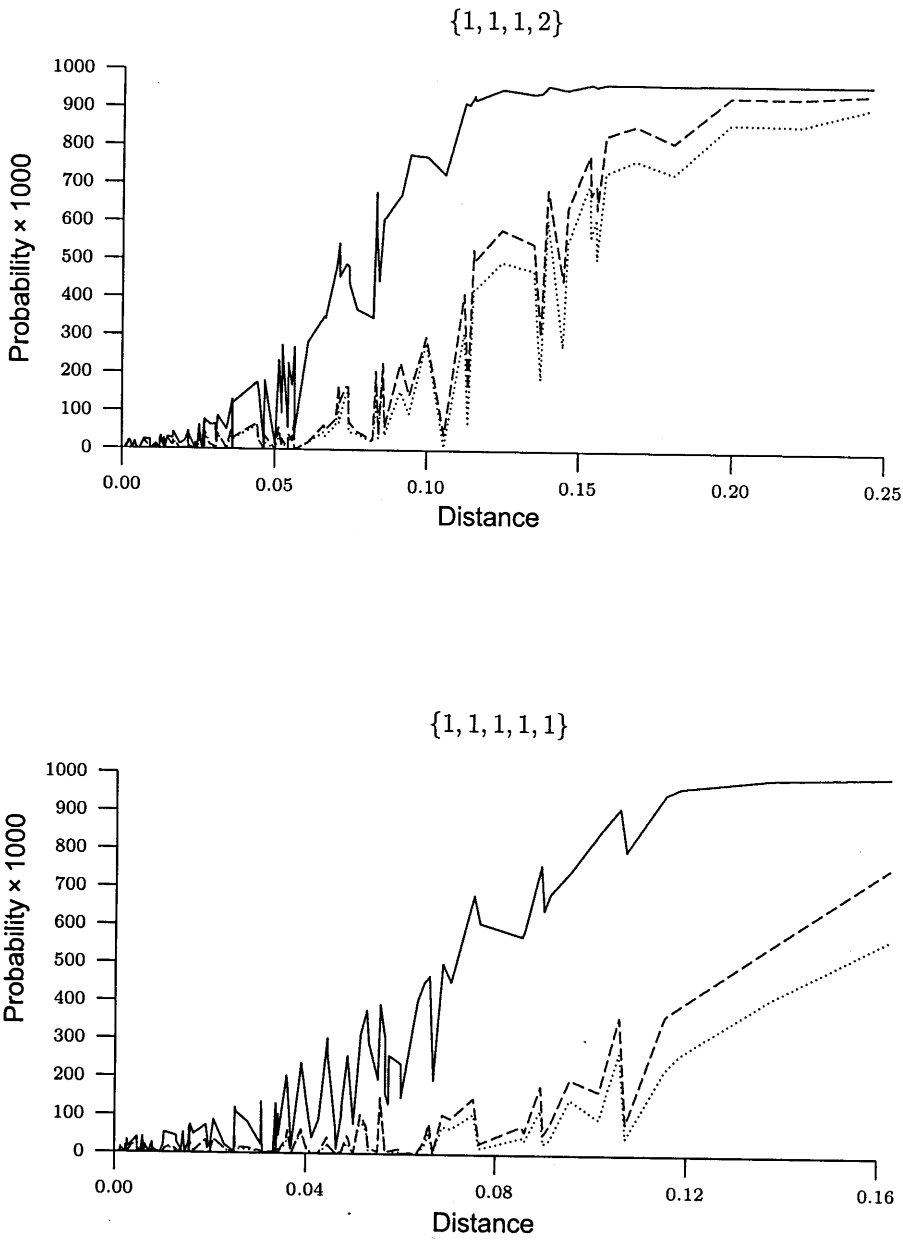


Fig. 1c. Simulation results for the states of nature $\{1, 1, 1, 2\}$ and $\{1, 1, 1, 1, 1\}$; see text for explanation

References

Gather U., Pawlitschko J., Pigeot I. (1966). Unbiasedness of multiple tests. *Scandinavian Journal of Statistics* **23**, 117-127.

Hochberg Y., Tamhane A.C. (1988). *Multiple Comparison Procedures.* John Wiley & Sons.

Miller Jr. R.G. (1982). *Simultaneous Statistical Inference*, Springer Verlag, 2nd ed.

Zieliński W. (1992). Monte Carlo comparison of multiple comparison procedures. *Biometrical Journal* **34**, 291–296.

Received 2 January 2004; revised 15 September 2004

Porównania wielokrotne współczynników kierunkowych prostych regresji

Streszczenie

W pracy rozważane jest zagadnienie porównania K prostych regresji. Zaproponowana została procedura znajdowania grup równoległych prostych.

Słowa kluczowe: porównania wielokrotne, wnioskowanie jednoczesne, analiza regresji.