[6] D. Landers, *Sufficient and minimal sufficient σ-fields*, Z. f. Wahrscheinlichkeitstheorie 23 (1972) 3, 197–207.

[7] A. N. Shiryaev, *Statistical sequential analysis*, Nauka, Moscow 1976 (in Russian).

[8] D. O. Siegmund, *Some problems in the theory of optimal stopping rules*, Ann. Math. Statist. 38 (1967), pp. 1627-5640.

Presented to the semester
MATHEMATICAL STATISTICS
September 15–December 18, 1976

MATHEMATICAL STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 6
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1980

ON MULTIPLE TEST PROCEDURES

STURE HOLM

Chalmers University of Technology and the University of Göteborg,
Göteborg, Sweden

Introduction

In many applications the statistical analysis is characterized by the fact that a number of detail questions should be answered and an overall view should be created by the totality of answers to the detail questions. Here are some examples of such situations:

A. The distribution of a random variable depends on a number of background variables. For each background variable the detail question is if the distribution of the random variable is influenced by this background variable. And the totality of the answer to these questions creates a picture of the dependence on the background variables. This kind of problems appears in many contexts.

B. In a comparison of some multidimensional random variable for two cases (e.g. treated and non-treated patients in a medical investigation) we may be interested in differences in the different components of the variable. These are the detail questions. But we are also probably interested in the differences in general, i.e. the totality of differences in all the components.

C. In an analysis of a stationary time series we may be interested in detail questions concerning the correlation at different time distances. But we may also be interested in getting a general picture of the dependence.

More examples of the same kind from different fields of applications are easily found. The examples are illustrations of *multiple statistical inference problems* where we have to take into consideration that we both want to answer detail questions and get a general view by the totality of answers to the detail questions. To make a test with conventional level of significance for each detail question is not good from an overall point of view. If we, for instance, make 40 independent tests of different detail hypotheses with level 0.05, we have a probability of only $0.95^{40} \approx 0.13$ that all hypotheses would be accepted if they were true. And there would be difficulties in getting a general view of the investigation if just a few hypotheses were rejected. The aim of these notes is to study the problem of constructing tests in such a way that their totality will give a general view in a reasonable way.

2. Some notation

Let the interesting detail hypotheses be $H_1, H_2, \ldots, H_n$ and let $Y_1, Y_2, \ldots, Y_n$ be good test statistics for testing these hypotheses. We suppose for the sake of simplicity that the test statistics are constructed in such a way that they have a tendency of getting higher values when the hypotheses are not true. This means that the separate hypotheses should be rejected when large values of the test statistics occur. Further, let $F_1(x), F_2(x), \ldots, F_n(x)$ be the cumulative distribution functions of the test statistics. From a general philosophical point of view it is desirable that the distribution of a test statistics of a true hypothesesis is not affected by other hypotheses being true or false. This means that the test statistic in some sense measures deviations from its own hypothesis only. In practical applications it is often not possible to get such an experimental planning and one has to accept an influence from other hypotheses. But of course it is important that the influence is not to big, i.e. that the distribution of a test statistic of a true hypothesis is approximately independent of the other hypotheses being true or false. We denote by $F_{k,0}(x)$ the infimum of $F_k(x)$ over all distributions such that the hypothesis H_k is true. This means that if we want to make a separate test of the hypothesis H_k at the level α we would reject the hypothesis if $Y_k > F_{k,0}^{-1}(1-\alpha)$. For the discussions we are going to make later it will be convenient to have a notation of the obtained level of the test statistics for the different hypotheses. For this reason we introduce $R_k = 1 - F_{k,0}(Y_k)$ for $k = 1, 2, \ldots, n$.

3. Philosophical considerations and definitions

The philosophy of using a prescribed low level of significance in a test of a single hypothesis may be described as follows:

Let H be a hypothesis which we want to reject. By using a low level of significance (α) we give a latent opponent believing in the hypothesis a big probability $(1-\alpha)$ of getting the hypothesis accepted if he is right. And we give ourselves only a small probability (α) of getting the hypothesis rejected by chance. It is then our task to plan the experiment well and make enough of experiments to be able to reject the hypothesis if it is not true. Only rejection leads to a real discovery and acceptance only means that either the hypothesis is true or the hypothesis is false but we have not succeded in discovering this fact in our investigation.

In a multiple test problem, where we have several hypotheses, we have to consider the possibility that a latent opponent is of the opinion that a number of the hypotheses are true, and we must give him a great probability of getting all those hypotheses accepted. We could not defend ourselves against this latent opponent if we had constructed a statistical method not giving him a good chance to get 'his hypotheses' accepted if they were true. And our 'discoveries' in form of rejected hypotheses would not be well established unless this condition is satisfied. This leads us to the following definition.

DEFINITION 1. A multiple test procedure for test of the hypotheses $H_1, H_2, \ldots, H_n$, ending up with acceptance or rejection of each separate hypothesis, is said to have the *multiple level of significance α for free combination* if the probability of accepting all true hypotheses is at least $1-\alpha$ independent of how many and which the true hypotheses are. ∎

The words 'for free combination' are put into the definition in order to indicate that we have to 'protect ourselves' against a latent opponent regarding any combination of hypotheses being true. At some rare instants there may be a natural ordering of the hypotheses so that if a hypothesis is true the following hypotheses in the order are also true. Numbering the hypotheses in this order, we have to take into consideration only subsets of true hypotheses of the type $\{H_k, H_{k+1}, \ldots, H_n\}$ and we are led to the following definition.

DEFINITION 2. A multiple test procedure for test of the hypotheses $H_1, H_2, \ldots, H_n$, ending up with acceptance or rejection of each separate hypothesis, is said to have the *multiple level of significance α for ordered combination* if the probability of accepting all true hypotheses is at least $1-\alpha$ for any set of true hypotheses of the type $\{H_k, H_{k+1}, \ldots, H_n\}$. ∎

4. General multiple test procedures with prescribed levels

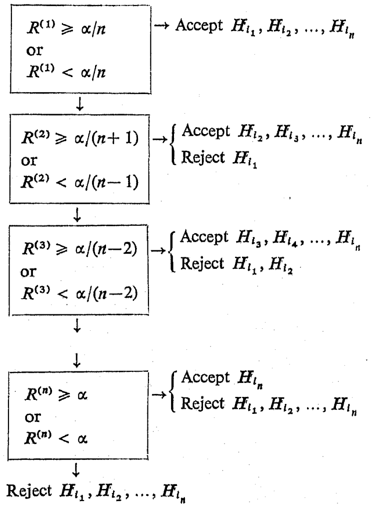
In the previous section we have given two different definitions of multiple level of significance for two different types of allowed combinations of hypotheses. Let us first study the simplest case where we allow only ordered combinations. We define a simple and general sequential procedure which is easily shown to have the multiple level of significance α for ordered combination in the following way:

We first make any level α test of the hypothesis that all detail hypotheses $H_1, H_2, \ldots, H_n$ are true. If this test leads to acceptance we accept all hypotheses. If it leads to rejection we continue to the second step. In the second step we make any level α test of the hypothesis that all the detail hypotheses $H_2, H_3, \ldots, H_n$ are true. If this test leads to acceptance we reject H_1 and accept $H_2, \ldots, H_n$. If it leads to rejection we continue to the third step. In the third step we make any test of the hypothesis that all the detail hypotheses $H_3, H_4, \ldots, H_n$ are true and so on.

Now if we have a set of true detail hypotheses of the type $H_k, H_{k+1}, \ldots, H_n$, there is a probability of at least $1-\alpha$ to accept all these detail hypotheses, since either we make a level α test on this combination or we stop at an earlier step where they are accepted (together with other detail hypotheses). Thus we have the following theorem.

THEOREM 1. *The above-described sequential test procedure has a multiple level of significance α for ordered combination.* ∎

Note that we are allowed to make the different tests of the hypotheses that $H_k, \ldots, H_n$ are all true in any way we want. We can use for instance a quadratic form in $Y_k, \dots, Y_n$ or the maximum of $Y_k, \dots, Y_n$ or the minimum of $R_k, \dots, R_n$ as a test statistic. In the last case we easily get a level α test by using the Boole inequality which is of the form that we accept when $\min(R_k, \dots, R_n) \geqslant \alpha/(n-k+1)$. The same type of test occurs in the steps in a general sequential test procedure with multiple test level α for free combination, which we are now going to introduce. In order to make a simple description of the procedure we use the notation $R^{(1)}, R^{(2)}, \dots, R^{(n)}$ for the ordered obtained levels and the notation $i_1, i_2, \dots, i_n$ for the indexes of the hypotheses where those ordered obtained levels occur. This means for instance that $R^{(2)}$ is the second smallest obtained level and i_2 the index of the hypothesis where it occurs. Now the general sequential procedure is most conveniently described by the following figure.


$R^{(1)} \geqslant \alpha/n$
or
$R^{(1)} < \alpha/n$
→ Accept $H_{i_1}, H_{i_2}, \dots, H_{i_n}$
$R^{(2)} \geqslant \alpha/(n+1)$
or
$R^{(2)} < \alpha/(n-1)$
Accept $H_{i_2}, H_{i_3}, \dots, H_{i_n}$
Reject H_{i_1}
$R^{(3)} \geqslant \alpha/(n-2)$
or
$R^{(3)} < \alpha/(n-2)$
Accept $H_{i_3}, H_{i_4}, \dots, H_{i_n}$
Reject H_{i_1}, H_{i_2}
$R^{(n)} \geqslant \alpha$
or
$R^{(n)} < \alpha$
Accept H_{i_n}
Reject $H_{i_1}, H_{i_2}, \dots, H_{i_n}$
Reject $H_{i_1}, H_{i_2}, \dots, H_{i_n}$

The total multiple procedure is easily carried through because it is completely determined by the obtained levels and their relations to some simple constants. For its multiple level of significance we have the following theorem.

THEOREM 2. *The multiple test defined by the above figure has the multiple level of significance α for free combination.* ∎

The proof consists of three steps. Let I denote the set of indexes of the true hypotheses and let $N(I)$ denote the number of elements in I. First the Boole inequality is used to prove that $P(R_i \geqslant \alpha/N(I)$ for all $i \in I) \geqslant 1-\alpha$. Then it is shown that the occurrence of this event implies stop at step $N+1-N(I)$ or earlier in the procedure. And finally it is shown that this implies acceptance of all hypotheses with indexes belonging to I.

An even simpler multiple test procedure having a miltiple test level α for free combination could of course be constructed by making comparisons of all R_i's with α/n, reject the hypotheses with obtained levels under α/n and accept the others. This is however a very conservative procedure and it is easily seen that the suggested sequential procedure always has a higher power, i.e. a higher probability of rejecting false hypotheses. With one exception (the first step) the obtained levels are compared with greater numbers in the sequential procedure than in the simple non-sequential procedure.

Examining the sequential procedure we find that each test in the procedure is performed with help of the smallest of the 'remaining' obtained levels. This is not a common way of constructing a test of the hypothesis that all detail hypotheses in a set are true. Using quadratic forms of the involved test statistics is probably much more common. But for the type of alternative we have here the use of smallest obtained level seems to have advantage. We should observe that in each step in the sequential procedure we want to reject *one* hypothesis that has not been rejected before. In order to see that such an aim could lead to a method based on smallest obtained level rather than a quadratic form we study a very simplified problem.

Suppose that $U_1, U_2, \dots, U_m$ are independent and that U_k is normally distributed with expectation μ_k and variance 1. We want to test the hypothesis that $\mu_1 = \mu_2 = \dots = \mu_m = 0$ against the alternative that any one of the μ:s is different from 0 while the other are equal to 0. A simple calculation now shows that the likelihood ratio test should be based on the U_k with the maximal modulus, i.e. on the minimal obtained level.

We have not yet been able to show any optimal properties of the procedure and there are even great difficulties in formulating optimal properties of multiple test procedures. Since the method is very general, we should not be surprised to discover that it is possible to make a better procedure for a more specified situation. In the next section we will present two special procedures, which are better than the general procedure in the special cases for which they are constructed.

5. Special multiple test procedures with prescribed levels

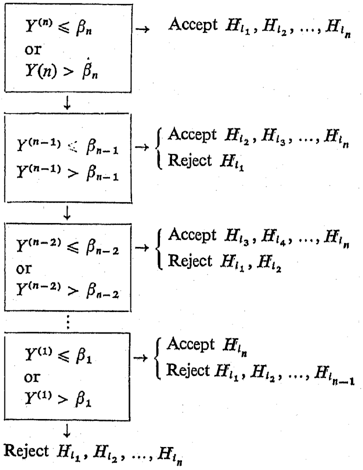
The first of the special cases we are going to study is the one where all the test statistics $Y_1, Y_2, \dots, Y_n$ are independent. In that case we could change the constants $\alpha/n, \alpha/(n-1), \dots, \alpha$ in the general procedure to the slightly greater constants $1-(1-\alpha)^{1/n}, 1-(1-\alpha)^{1/(n-1)}, \dots, 1-(1-\alpha)^{1/2}, 1-(1-\alpha)^1$ and still get the multiple test level of significance α. The proof of this follows the line in the general case but instead of using the Boole inequality in the first step we can now use the independence and make an exact calculation. Since the obtained levels are compared to greater constants in the special method than in the general method, the special method will have a greater power. We summarize in the following theorem.

THEOREM 3. *If the random variables $Y_1, Y_2, \ldots, Y_n$ are independent the above-described multiple test procedure has the multiple test level of significance α for free combination and it gives a higher power than the general procedure with the same multiple test level of significance.* ■

Our second special case is the analysis of variance situation. In this situation the test statistics for detail hypotheses are either t-statistics or F-statistics with a common variable appearing in the denominator. The appearing F-statistics could be regarded as quadratic form compositions of more primary t-statistics which could be thought of as test statistics for more detailed hypotheses. For the sake of simplicity we consider here only the case where all detail hypotheses are tested by use of t-statistics. This means that we test the hypotheses $H_1, H_2, \ldots, H_n$ by use of t-statistics

$$Y_1 = \frac{|U_1|}{W}, \quad Y_2 = \frac{|U_2|}{W}, \quad \ldots, \quad Y_n = \frac{|U_n|}{W}$$

where $U_1, U_2, \ldots, U_n$ and W are independent, U_k is normally distributed with parameters μ_k and 1 while W has the same distribution as the square root of a χ^2-distributed random variable divided by its degree of freedom. The different detail hypotheses $H_1, H_2, \ldots, H_n$ to be tested are that the different μ_k's are equal to 0. The procedure we suggest is in fact quite similar to the previous two procedures, but since we are defining the procedure by the test statistics Y_k themselves and not the obtained levels, it seems to be different from the previous ones. Let $Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}$ be the ordered random variables from the series $Y_1, Y_2, \ldots, Y_n$ and $i_n, i_{n-1}, \ldots, i_1$ the indexes of the hypotheses where statistics occur. Then the sequential procedure can easily be described by the following figure.


$Y^{(n)} \leqslant \beta_n$
or
$Y(n) > \dot{\beta}_n$
Accept $H_{i_1}, H_{i_2}, \ldots, H_{i_n}$
$Y^{(n-1)} \leqslant \beta_{n-1}$
$Y^{(n-1)} > \beta_{n-1}$
Accept $H_{i_2}, H_{i_3}, \ldots, H_{i_n}$
Reject H_{i_1}
$Y^{(n-2)} \leqslant \beta_{n-2}$
or
$Y^{(n-2)} > \beta_{n-2}$
Accept $H_{i_3}, H_{i_4}, \ldots, H_{i_n}$
Reject H_{i_1}, H_{i_2}
$Y^{(1)} \leqslant \beta_1$
or
$Y^{(1)} > \beta_1$
Accept H_{i_n}
Reject $H_{i_1}, H_{i_2}, \ldots, H_{i_{n-1}}$
Reject $H_{i_1}, H_{i_2}, \ldots, H_{i_n}$

The constants β_k for $k = 1, 2, \ldots, n$ are determined by the distribution of the studentized maximum modulus in a sample of independent observations from a normal distribution. This distribution has two parameters, the number of degrees of freedom in the denominator and the number of variables in the maximization. The constant β_k equals the $1-\alpha$ fractile in the distribution with the second parameter equal to k while the first parameter is equal to the number of degrees of freedom in the denominators of the Y's. Tables of β_k are found in Pillai and Ramachandran [3].

The suggested procedure can be shown to have a multiple test level of significance α for free combination and to have a higher power than the general procedure along the lines of the previous case although the proof is technically more complicated. It is also possible to reformulate the procedure in terms of obtained levels, but then there appear constants in the scheme which are determined as the value of the inverse of a c.d.f. of a t-distribution in a $1-\alpha$ fractile point of a studentized maximum modulus distribution. We again summarize in a theorem.

THEOREM 4. *Suppose we have the analysis of variance situation with normal basic distributions. Then the above-described multiple test procedure has the multiple test level of significance α for free combination and it gives higher power than the general procedure with the same multiple test level of significance.* ■

The two procedures we have suggested in this section are sequentially rejective procedures which could be thought of as improvements of two well-known older procedures. In the case of independence a simple old procedure consists of testing all hypothesis at the level $1-(1-\alpha)^{1/n}$ and our method is an improvement of this. In the analysis of variance situation an old procedure called the *studentized maximum modulus procedure* consists of comparing all t-statistics with the number β_n and our sequentially rejective procedure is an improvement of this.

It is also possible to make analogous improvements of the so-called *many-one-t-procedure* and *many-one-rank-procedure*. The older methods and their properties are extensively treated in Miller [4].

6. Some examples

EXAMPLE 1. *Decomposition of a two-sided test.* It is quite common in applications to make two-sided tests of simple hypotheses and in case of rejection it is not only stated that the hypothesis is not true but also on which 'side of the hypothesis' the real case is thought to be situated. In a medical investigation it may, for instance, be tested if two drugs have the same effect on a disease and in case of rejection it is also stated which drug has the greatest effect. Such a hypothesis testing procedure may formally be looked upon as a multiple test procedure with two detail tests of the same hypotheses (equal effects) and with test statistics differing in sign only. The first step in the sequential procedure is then to check the smallest obtained level with $\alpha/2$, and we get a procedure which is exactly the two sided test with equal tails supplemented with a statement of type of deviation in case of rejection.

EXAMPLE 2. *Multiple tests in a* 3×3 *table*. This example of application in medicine is taken from Armitage [1], p. 213, and arises from Medical Research Council.

Patients suffering from pulmonary tuberculosis are treated with PAS (99 patients), Streptomycin (84 patients) or a combination of both drugs (90 patients). The sputa from the patients are tested for degree of positivity and the following results are obtained.

Sputum Treatment	Positive smear	Negative smear positive culture	Negative smear negative culture
PAS	56	30	13
Streptomycin	46	18	20
Streptomycin and PAS	37	18	35

In this example it is interesting to make pairwise comparisons between treatments. The interesting division of the reactions ought to be positive contra negative smear and any positive reaction (positive smear or negative smear and positive culture) contra negative reaction. In all combinations it is motivated to make one-sided tests in both directions as in Example 1 and thus we get $3\times 2\times 2 = 12$ detail tests whose results constitute the totality. All tests are performed in 2×2 tables and they are one-sided tests of equality of probabilities in these tables. For the 12 tests we obtain the following ordered obtained levels:

0.000025 0.0164 0.0169 0.0306 0.0350 0.403 0.597
0.9650 0.9694 0.9831 0.9836 0.999975

If we use the multiple test level of significance 0.05, we can reject the hypothesis only in the first case. This is the comparison of PAS and the combined drug with respect to any positive reaction (positive smear or negative smear and positive culture) contra negative reaction (negative smear and negative culture) on the side where the combined drug gives tendency to negative reaction. It is to be noted that if we had used tests with conventional level at all places we should have rejected four other hypotheses as well. But a procedure using the conventional level 0.05 at each of the 12 tests would have a very low multiple test level of significance and a very high probability of rejecting some true hypotheses.

EXAMPLE 3. *A simple one-way classification*. The following example is taken from Miller [2] where it is used to illustrate the Newman–Keuls and Duncan multiple test procedures. There are taken 5 observations in each of 5 classes. The observations are supposed to be independent and normally distributed with the same variance in all classes. The pooled estimate of the common standard deviation was $1.2\cdot\sqrt{5}\approx 2.68$ and the means in the five classes were 16.1, 17.0, 20.7, 21.1, and 26.5. We can attack the problem by making $\binom{5}{2} = 10$ two-sided test for pairwise comparisons of means or $(5)_2 = 20$ one-sided tests for pairwise comparisons of means. We use here two-sided tests.

Denoting the classes corresponding to the above-ordered means by A, B, C, D, E we get the following outcomes of test statistics and obtained levels for pairwise comparisons.

Comparison	Outcome of t-statistic with 20 degrees of freedom	Obtained level		
A–E	6.12	0.0000056	$<$	0.0050
B–E	5.59	0.000018	$<$	0.0055
C–E	3.41	0.0028	$<$	0.0063
D–E	3.18	0.0047	$<$	0.0071
A–D	2.94	0.0081	$<$	0.0083
A–C	2.71	0.0135	$\geqslant$	0.0100
B–D	2.41	0.0257		
B–C	2.18	0.0414		
A–B	0.53	0.602		
C–D	0.24	0.813		

Using a multiple test level of significance 0.05 we can now reject the hypotheses of equality for the pairs A–E, B–E, C–E, D–E and A–D. Thus we can say that E is *separated from all the others* and within this group we can separate A and D only.

7. Comments

In these notes I have not made any attempt to review older multiple test methods. The reader interested in getting a background is referred to Miller [2] and Sverdrup [4]. Neither have I made any attempt to make any comparisons with older multiple test methods except the trivial comparisons between some old test procedures and the corresponding sequentially rejective procedures. In all these cases the sequentially rejective procedure is always better. A more extensive report including proofs, more about special methods, comparisons between methods and further applications is in preparation.

8. References

[1] P. Armitage, *Statistical methods in medical research*, John Wiley & Sons, New York, 1971.

[2] R. G. Miller, *Simultaneous statistical inference*, McGraw-Hill, New York 1966.

[3] K. G. S. Pillai and K. V. Ramachandran, *On the distribution of the ratio of the i-th observation in ordered sample from normal population to an independent estimate of the standard deviation*, Ann. Math. Stat. 25 (1954), pp. 565–572.

[4] E. Sverdrup, *Multiple decision theory*, Aarhus Universitet, Matematisk Institutt, Lecture Notes Series No. 11 (1969).

Added in proofreading

Another construction of multiple test procedures often leading to tests equivalent to those suggested here is given in:

[5] R. Marcus, E. Peritz and K. R. Gabriel, *On closed testing procedures with special reference to ordered analysis of variance*, Biometrika 63 (1976), pp. 655–660.

Further development of sequentially rejective multiple test procedures can be found in:

[6] S. Holm, *A simple sequentially rejective multiple test procedure*, Scand. Journ. of Stat. 6 (1979), pp. 65–70.

[7] J. P. Shaffer, *Control of directional errors with stagewise multiple test procedures*, Ann. Stat., to appear.

[8] S. Holm, *A stagewise directional test for the normal regression situation*, Conference report from The sixth conference on probability theory, Brasov, Romania, 11–15 Sept. 1979, to appear.

Presented to the semester
MATHEMATICAL STATISTICS
September 15–December 18, 1976

MATHEMATICAL STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 6
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1980

ROBUST ESTIMATION IN A LINEAR REGRESSION MODEL

JANA JUREČKOVÁ

Charles University, Prague, Czechoslovakia

1. Introduction

A detailed text concerning robust estimation in a linear model will appear in the monograph: K. M. S. Humak: *Statistische Methoden der Modellbildung*, Band II (Academic Verlag, Berlin). Here we shall only give a brief survey of the most usual types of robust estimates of the regression parameter vector and mention some of their asymptotic properties.

2. Robust alternatives to the method of least squares

We shall consider the problem of estimating the regression parameters of a linear model. We want to estimate β after observing $X_n' = (X_{1n}, \ldots, X_{nn})$ where

$$X_n = C_n\beta + E, \tag{2.1}$$

$\beta = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown regression parameters, $E = (E_1, \ldots, E_n)'$ is a vector of errors and $C_n = ((c_{ij}))_{i=1,\ldots,n}^{j=1,\ldots,p}$ is a matrix of known regression constants (design matrix) of the rank p. Most of our considerations will be asymptotic as the number of observations n grows and the number of regression parameters p remains fixed. Thus, the coordinates of X_n and of C_n depend on n; we shall not indicate explicitly this dependence provided no confusion arises.

We shall suppose throughout that E_i, $i = 1, \ldots, n$, are independent and identically distributed with a common distribution function F and density f with respect to the Lebesgue measure; F and f are generally unspecified.

If F is normal with the mean 0, the appropriate procedure is to minimize the sum of squares

$$\sum_{i=1}^{n}\Big(X_i - \sum_{j=1}^{p} c_{ij}\beta_j\Big)^2 = \min \tag{2.2}$$

or, equivalently, to solve the system of equations

$$\sum_{i=1}^{n}\Big(X_i - \sum_{k=1}^{p} c_{ik}\beta_k\Big)c_{ij} = 0, \quad j = 1, \ldots, p. \tag{2.3}$$