SUBSET: Best Subsets using Information Criteria

C. Mitchell Dayton
University of Maryland

Abstract

SUBSET, written in the matrix language Gauss, is a program that identifies optimal subsets of means or proportions based on independent groups. All possible configurations of ordered subsets of groups are identified and the best model is selected using either the AIC or BIC information criterion. For means, both homogeneous and heterogeneous variance cases are considered. SUBSET offers an alternative approach to traditional post-hoc multiple-comparison procedures such as the Tukey test for pairwise comparisons. Major advantages of SUBSET over traditional pairwise comparison procedures include the fact that intransitive decisions are avoided and that issues related to type I error control, sample size and heterogeneity of variance do not arise.

1 Introduction

Researchers often use analysis of variance to investigate mean differences among several response groups. If the null hypothesis based on equality of means is rejected, it is common practice to employ multiple comparison techniques to study the patterns of differences among the means. For example, Kirk (1995) describes 22 multiple comparison procedures including pairwise comparisons such as the Tukey test. In general, these procedures depend upon interpreting multiple tests of significance. As detailed in Section 3, below, Dayton (1998) advocated replacing these procedures by a wholistic model selection approach based on information criteria. The program, SUBSET, implements this information theoretic approach for comparisons among means or among proportions from independent samples.

Section 2 presents a summary of the theory underlying the use of information criteria for model selection while Sections 3 and 4 consider applications of this theory to sample means and sample proportions, respectively. Section 5 describes how to use the SUBSET program and exemplary applications are presented in Section 6.

2 Information Criteria

Akaike (1973, 1974) developed a decision-making strategy based on the Kullback-Leibler (1951) information measure arguing that this measure provides a natural criterion for ordering alternate statistical models for data. Adapting the notation of Akaike (1987) for the case of univariate data, the Kullback-Leibler information for the true distribution, $g_t(x)$, of random variable x, relative to some other distribution, $g_o(x)$, can be written as:

$$(1) \qquad I(g_t; g_o) = E(\text{Log}_e[g_t(x)]) - E(\text{Log}_e[g_o(x)])$$

where expectations are taken with respect to $g_t(x)$. In the context of maximum likelihood estimation, let $\mathbf{x} = \{x_i\}$ be N values of an iid random variable, x, with true density

function $g(\cdot \mid \boldsymbol{\theta})$ based on the parameter vector, $\boldsymbol{\theta}$. Also, let $\boldsymbol{\theta}_x$ represent the usual maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ found as a result of maximizing the density, $g(\mathbf{x} \mid \boldsymbol{\theta})$, over the sample by treating $\boldsymbol{\theta}$ as variable. Then, assuming p independent parameters in $\boldsymbol{\theta}$, a large-sample result for the distribution of likelihood ratios is:

$$(2) \qquad L_1 = 2\{\mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_x)] - \mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_t)]\} = \chi_p^2$$

where χ_p^2 is central chi-square with p degrees of freedom.

Again, following Akaike let y be an additional observation from the same distribution as **x**. Akaike (1974) shows that, asymptotically:

$$(3) \qquad L_2 = 2\{E_y\mathrm{Log}_e[g(y \mid \boldsymbol{\theta}_x)] - E_y\mathrm{Log}_e[g(y \mid \boldsymbol{\theta}_t)]\} = -\chi_p^2$$

Then:

$$(4) \qquad E(L_1 - L_2) = 2\mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_x)] - 2E_y\mathrm{Log}_e[g(y \mid \boldsymbol{\theta}_x)] \cong 2p.$$

One-half of the second term in Equation (4), $E_y\mathrm{Log}_e[g(y \mid \boldsymbol{\theta}_x)]$, corresponds to the second term in the definition of Kullback-Leibler information, $E(\mathrm{Log}_e[g_o(\mathbf{x})])$. Also, note that the first term in Kullback-Leibler information is constant for any model. Akaike defines his AIC estimator of Kullback-Leibler information as:

$$(5) \qquad \text{Constant} - E_y\mathrm{Log}_e[g(y \mid \boldsymbol{\theta}_x)] \cong -2\mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_x)] + 2p = \mathrm{AIC}$$

When selecting among M competing models, Akaike uses Equation (5) to calculate AIC_m, m = 1,…,M, for the models and, then, selects the model with $\min(\mathrm{AIC}_m)$ as the preferred model. The conventional interpretation of AIC is as an estimate of the loss of precision (or, increase in information) that results when $\boldsymbol{\theta}_x$, the MLE, is substituted for the true parametric value, $\boldsymbol{\theta}_t$, in the likelihood function. Thus, by selecting the model with $\min(\mathrm{AIC}_m)$, the (estimated) loss of precision is minimized.

As noted by Sclove (1987), AIC represents a penalized log-likelihood function that can be written in the general form:

$$(6) \qquad -2\mathrm{Log}_e[g(\mathbf{x} \mid \boldsymbol{\theta}_x)] + a(N)p$$

where a(N) is a function that may depend upon the total sample size, N. Various adaptations of AIC have been suggested that, unlike AIC, make the statistic dependent upon sample size. In particular, the Schwarz (1978) BIC (or, SIC) statistic and the Bozdogan (1987) CAIC statistic use penalty terms equal to $\mathrm{Log}_e(N)$ and $\mathrm{Log}_e(N) + 1$, respectively. As noted by Bozdogan (1987), these latter procedures are asymptotically consistent in the sense that, when the null case is the true model, the probability of selecting the true model approaches one, rather than an arbitrary significance level, as is true for conventional hypothesis testing procedures.

3 Application of Information Criteria to the Paired-Comparisons of Means

Conventional pairwise-comparison procedures for means involve conducting a set of statistical tests. Often this is done subsequent to testing the omnibus hypothesis of equality of means for K independent groups (i.e., $\mu_k = \mu$ for k = 1,…,K) using analysis of variance techniques although this is not technically required for many procedures. One popular approach, the Tukey HSD procedure, sets up q statistics for the K(K – 1)/2 different pairs of means and refers these statistics to the appropriate null distribution of the studentized range statistic for a span of K means. Thus, K(K – 1)/2 hypotheses of the form $\mu_k = \mu_{k'}$ for $k \neq k'$ are tested. Among the problems with such procedures cited by Dayton (1998) are:

(1) Some arbitrary technique is utilized to control the family-wise type I error rate for the set of correlated pairwise tests;
(2) The issues of homogeneity of variance and differential sample size pose problems for many paired-comparison procedures;
(3) Intransitive decisions (e.g., outcomes suggesting mean 1 = mean 2, mean 2 = mean 3, but mean 1 < mean 3) are the rule rather than the exception with typical paired comparison procedures since they entail a series of discrete, pairwise significance tests.
(4) There exists a large variety of competing procedures that differ in how type I error is controlled and, consequently, in power (e.g., SPSS for Windows offers seven distinct procedures to choose among).

For means based on K independent groups, there is a total of 2^{K-1} patterns of ordered subsets with equal means within subsets. For example, with three groups for which the means are ranked and labeled 1, 2, 3, the $2^2 = 4$ distinct ordered subsets are {123}, {1,23}, {12,3} and {1,2,3}, where a comma is used to separate subsets that are unequal in mean value. Dayton (1998) proposed using model-selection criteria such as the AIC or BIC statistic for selecting the most appropriate ordering of subsets of means for purposes of interpretation. In particular, this approach was advocated as avoiding many of the objections were raised to conventional pairwise comparison procedures. The program, SUBSET, computes both the Akaike AIC and the Schwarz BIC statistics for all 2^{K-1} distinct ordered subsets. Since the number of ordered subsets can be quite large for practical problems (e.g., 512 for K = 10 groups but 524,288 for K = 20 groups), only the ordered subsets corresponding to the smallest AIC and BIC values, as specified by the user, are printed out.

Creating the patterns of ordered subsets of means within SUBSET is based on the recognition that digit inversions in the first 2^{K-1} binary equivalents of the integers from 0 through 2^K uniquely define these patterns. For example, for K=4 these eight binary equivalents are 0000, 0001, 0010, 0011, 0100, 0101, 0110 and 0111 and they correspond to the ordered subsets {1234}, {123,4}, {12,3,4}, {12,34}, {1,2,34}, {1,2,3,4}, {1,23,4} and {1,234}. In the program, SUBSET, once these binary equivalents are generated, the sub-matrix extraction and substitution features of the Gauss language are used to create the actual patterns of equivalent means (and variances, for the heterogeneous case). There is no limit to the number of groups that can be analyzed since the program only stores results for the S (specified by user) smallest AIC and BIC values at each iteration. Of

course, execution time can become relatively long for large K. Typical execution times on a 266mz notebook computer are:

K = 4 groups, $2^3 = 8$ patterns:	.06 seconds
K = 12 groups, $2^{11} = 2{,}048$ patterns:	6.97 seconds
K = 20 groups: $2^{19} = 524{,}288$ patterns	3049.74 seconds, or 50.83 minutes

Information criteria such as AIC or BIC are based on the log-likelihood of the data. In SUBSET, it is assumed that the observations arise from normal densities. Since the log-likelihood is maximized for any given model when variance estimates are computed using the sample size, n, rather than n-1, in the denominator, this conversion is made within the program. SUBSET calculates AIC and BIC based on the usual assumption of homogeneity of variance as well as based on a restricted heterogeneous variance model for which it is assumed that there is a unique population variance for each of the distinct subsets of means. For the homogeneous case, the conventional analysis of variance within-groups sum of squares, SS_w, is converted to a variance estimate, SS_w/N, where N is the total sample size. For the restricted, heterogeneous variance case, an estimated variance for a subset of means can be obtained (a) by pooling the estimates from the separate groups or (b) by computing the sample variance for the combined sample. The latter approach is illustrated in Dayton (1998) and is the procedure incorporated into SUBSET. For any given model, AIC is given by the expression $-2\text{Log}_e(\text{likelihood}) + 2p$, where p is the number of independent parameters estimated in calculating the likelihood for the observed data. Similarly, BIC is given by $-2\text{Log}_e(\text{likelihood}) + \text{Log}_e(N)p$. For a model with T subsets of means, p equals T+1 for the homogeneous case and 2T for the restricted heterogeneous case. For example, for the ordered subset {1,2,34} the values of T are 4 and 6, respectively, for AIC and BIC. Since $\text{Log}_e(N) > 2$ for $N > 7$, AIC and BIC may, and often do, result in different orderings of subsets of means with, predictably, simpler models being favored by BIC. In Dayton (1998), results of a limited simulation with AIC and CAIC (the slightly different criterion than BIC suggested by Bozdogan (1987) with penalty term $\text{Log}_e(N+1)p$), it was found that: “Overall…the accuracy of CAIC is always approximately equal to or superior to Tukey HSD but tends to be lower than AIC when there are relatively many clusters of means, especially with smaller sample sizes.” Accuracy, in this study, was stringently defined in terms of all-pairs power following Ramsey (1978).

4 Application of Information Criteria to the Paired-Comparisons of Proportions

A simple extension of the approach presented above for sample means allows the identification of optimal subsets for data in the form of proportions. Consider K groups of sizes $n_1,\ldots,n_K$ with sample proportions, $p_1,\ldots,p_K$, respectively. Assuming independent Bernoulli trials, the log-likelihood for the k^{th} (ordered) sample outcome is $n_k p_k \text{Log}_e(p_k) + n_k(1-p_k)\text{Log}_e(1 - p_k)$ and the log-likelihood for all samples is found by summing across the K groups. Note that the sample proportion, p_k, is the MLE for the corresponding population proportion and that omitting the combinatorial constant to take into account unordered samples only omits a constant term from the log-likelihood. Unlike the situation for sample means, there is no need to consider homogeneous and heterogeneous cases since each Bernoulli process is based on a single parameter, π_k, say. Otherwise,

model selection can be based on the same reasoning as for sample means. That is, there is a total of 2^{K-1} distinct patterns of subsets of proportions to evaluate. For each pattern, the log-likelihood is converted to AIC by the formula $-2\text{Log}_e(\text{likelihood}) + 2p$ and to BIC by the formula $-2\text{Log}_e(\text{likelihood}) + \text{Log}_e(N)p$, where $p = T$ for a model with T subsets of proportions.

5 Using the SUBSET Program

SUBSET is written in the microcomputer matrix programming language, Gauss for Windows NT/95 Version 3.2.32 (Aptech Systems, 1997). SUBSET is run in interpretive mode, which means that the Gauss system must be installed on the microcomputer. However, extensive knowledge of Gauss syntax is not required to run the program. The source code, SUBSET.E, as well as a compiled version, SUBSET.GCG, of the program are available but note that the Gauss system is required to run either version. For general-purpose analysis, there is no other program that computes AIC and/or BIC for the models available in SUBSET. For a small number of groups (e.g., 5 or less), it is reasonably easy to program the computations in a spreadsheet as was reported by Dayton (1998).

Data for analysis is imported into SUBSET from a spreadsheet or database program. The import routine in the Gauss program determines the nature of the spreadsheet/database from the file extension (e.g., file.XLS denotes a Microsoft Excel file whereas file.DB2 denotes a dBase II file). The general format for the spreadsheet/database file is:

Row 1, Columns A - D	Labels such as Group, Count, Mean, and Variance
Rows 2,..,K+1 Column A	Arbitrary group label
Rows 2,..,K+1 Column B	Group sample size (n)
Rows 2,..,K+1 Column C	Group sample mean {or sample proportion}
Rows 2,..,K+1 Column D	Group sample variance (unbiased estimate using n -1 in denominator); omit for case of proportions

It is conventional to code the groups with names, or 1, 2, etc., or A, B, etc. but SUBSET rearranges the groups in rank order of means {proportions}, from smallest to largest, and presents groups in ranked order, 1, 2, etc., in the output. Thus, in practice, it is most convenient to order the means {proportions} in this same manner in the spreadsheet/database prior to analysis. A sample data set for five groups clipped from a Microsoft Excel spreadsheet is shown in the Exemplary Output section, below.

The Gauss program can import data from spreadsheet formats such as Microsoft Excel, Lotus 123 or Quattro-Pro or from database programs such as dBase IV, Paradox or FoxPro or from a Gauss dataset. There are restrictions on the nature of the spreadsheet or database that can be imported. These restrictions can be found by referring to the description of the Gauss “import” command in Gauss Help. For example, for GAUSS for Windows NT/95 version 3.2.32 when using a Microsoft Excel spreadsheet, it must be

saved as version 7.0 or earlier (but no earlier than 2.1). In particular, spreadsheets created by later versions of Excel such as that found in Office 97 cannot be directly imported but must be saved as an earlier version (e.g., version 4.0). Actually, data can also be input from a character-delimited ASCII file but this is typically less convenient than using, for example, a spreadsheet.

To run the compiled version of SUBSET, follow these steps (assume SUBSET.GCG is located in the directory C:\Program):

Open the Gauss program to the Command Window;
At the command line, (gauss), enter: **run c:\program\subset.gcg**
The program prints the following queries:
Name of spreadsheet file? {provide a directory and file name (e.g., **c:\drink.xls)}**

Range in spreadsheet? {provide the spreadsheet cell locations (e.g., **A1:D6** or **A1..D6)}**

Number of AIC/BIC values to display (5 is recommended)? {provide an appropriate number}

Output is directed to the screen and to a default file named Subset.out in the directory in which the Gauss system is started. The output file can be changed by editing the appropriate line in the Gauss program. Note that only output from the current analysis is saved to the file.

6 Exemplary Output

Example 1: Assume the data below in cells A1:D6 of an Excel 4.0 spreadsheet (note that the groups have been sorted in ascending magnitude of means). The data are taken from the SPSS/PC+ manual (Norusis, 1986). The dependent variable is annual consumption of alcohol in pints by adult males as reported by Greeley et al. (1980) for the named ethnic groups.

Group	Count	Mean	Var(unbiased)
Jewish	41	9.250	467.641
Swedish	74	16.563	715.563
English	90	21.875	464.963
Irish	119	24.250	653.416
Italian	84	24.312	585.059

The input to SUBSET and the output generated by SUBSET are:

```
(gauss) run c:\program\subset.gcg
Default file for all printed output is Subset.out in the current directory
Program SUBSET for Ordered Subsets of Means or Proportions
Prepared by: C. Mitchell Dayton
Department of Measurement & Statistics
```

University of Maryland
E-Mail: CD4@UMAIL.UMD.EDU
Enter 1 for means or 2 for proportions ? 1
Number of AIC/BIC values to display (5 is recommended) ? 5
Name of spreadsheet file? c:\drink.xls
Range in spreadsheet? a1:d6

****
GAUSS Data Import Facility
****

Begin import...
Import completed
Number of AIC/BIC values to display (5 is recommended)? 5

NOTE: Means have been sorted from smallest to largest

Sorted means are:
9.250 16.563 21.875 24.250 24.312

Best models assuming Homogeneity of Variance

Smallest AIC values and ordered subsets:

3764.900	1.000	2.000	3.000	3.000	3.000
3765.319	1.000	1.000	2.000	2.000	2.000
3766.249	1.000	2.000	2.000	3.000	3.000
3766.284	1.000	2.000	3.000	4.000	4.000
3766.703	1.000	1.000	2.000	3.000	3.000

Smallest BIC values and ordered subsets:

3777.353	1.000	1.000	2.000	2.000	2.000
3779.862	1.000	2.000	2.000	2.000	2.000
3780.946	1.000	2.000	3.000	3.000	3.000
3782.165	1.000	1.000	1.000	2.000	2.000
3782.294	1.000	2.000	2.000	3.000	3.000

Best models assuming Heterogeneity of Variance

Smallest AIC values and ordered subsets:

3766.260	1.000	2.000	3.000	3.000	3.000
3766.919	1.000	1.000	2.000	2.000	2.000
3766.983	1.000	2.000	3.000	4.000	4.000
3767.643	1.000	1.000	2.000	3.000	3.000

3768.408	1.000	2.000	2.000	2.000	2.000

Smallest BIC values and ordered subsets:

3782.964	1.000	1.000	2.000	2.000	2.000
3784.036	1.000	1.000	1.000	1.000	1.000
3784.453	1.000	2.000	2.000	2.000	2.000
3787.759	1.000	1.000	1.000	2.000	2.000
3790.327	1.000	2.000	3.000	3.000	3.000

Execution time in seconds = 0.0600

Interpretation: For AIC, the first three homogeneous models have smaller values than the best heterogeneous model and for BIC all five reported values for homogeneous models have smaller values than for the heterogeneous models. Thus, we can focus on the relatively simpler models that assume all populations have the same variance. Based on AIC, the preferred model, 1,2,3,3,3, corresponds to the pattern {1,2,345}. That is, there are three distinct subsets of means for population 1, population 2 and a combination of populations 3, 4 and 5. However, BIC favors the somewhat simpler model {12,345}. In terms of the group labels, based on AIC there are three distinct clusters with Jewish lower than Swedish who are, in turn, lower than the cluster {English, Irish, Italian} that are indistinguishable given the present data. Based on BIC, the {Jewish, Swedish} pair is indistinguishable but is lower than the {English, Irish, Italian} cluster.

Example 2: The table below, clipped from an Excel 4.0 spreadsheet, contains fictional proportion data for 6 groups:

Group	Number	Proportion
1	30	0.60
2	30	0.65
3	40	0.70
4	40	0.80
5	50	0.81
6	50	0.82

The input to SUBSET and the output generated by SUBSET are:

(gauss) run c:\program\subset.gcg
Default file for all printed output is Subset.out in the current directory
Program SUBSET for Ordered Subsets of Means or Proportions
Prepared by: C. Mitchell Dayton
Department of Measurement & Statistics
University of Maryland
E-Mail: CD4@UMAIL.UMD.EDU

```
Enter 1 for means or 2 for proportions ? 2
Number of AIC/BIC values to display (5 is recommended)  ? 5
Name of spreadsheet file?  c:\prop6.xls
Range in spreadsheet?  a1:c7

*******************************************************************************
                    GAUSS Data Import Facility
*******************************************************************************

Begin import...
Import completed

NOTE: Proportions have been sorted from smallest to largest

Sorted proportions are:
    0.600      0.650      0.700      0.800      0.810      0.820

Smallest AIC values and ordered subsets:

  268.711      1.000      1.000      1.000      2.000      2.000      2.000
  270.109      1.000      1.000      2.000      3.000      3.000      3.000
  270.144      1.000      2.000      2.000      3.000      3.000      3.000
  270.251      1.000      1.000      2.000      2.000      2.000      2.000
  270.667      1.000      1.000      1.000      2.000      2.000      3.000

Smallest BIC values and ordered subsets:

  275.673      1.000      1.000      1.000      2.000      2.000      2.000
  277.212      1.000      1.000      2.000      2.000      2.000      2.000
  277.577      1.000      1.000      1.000      1.000      1.000      1.000
  278.620      1.000      1.000      1.000      1.000      2.000      2.000
  279.517      1.000      2.000      2.000      2.000      2.000      2.000
Execution time in seconds =       0.050
```

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csake (eds.), Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19,* 716-723.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317-332.

Aptech Systems, Inc. (1997). GAUSS for Windows NT/95: Version 3.2.32, Maple Valley, WA.

Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika,* 52, 345-370.

Dayton, C. M. (1998) "Information Criteria for the Paired-Comparisons Problem." *American Statistician, 52,* 144-151.

Greeley, A.M., McCready, W.C. & Theisen, G. (1980). Ethnic Drinking Subcultures. New York: Praeger.

Kirk, R.E. (1995). *Experimental Design (third edition).* Brooks/Cole.

Kullback, S. & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22,* 79-86.

Norusis, M.J. (1986). SPSS/PC+ for the IBM PC/XT/AT. SPSS, Inc.

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association, 73,* 479-485.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Sclove, S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52,* 333-343.