Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 29

11-1-2005

JMASM21: PCIC_SAS: Best Subsets Using Information Criteria

C. Mitchell Dayton
University of Maryland, cdayton@umd.edu

Xuemei Pan
University of Maryland, xpan1@umd.edu

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 621-626

1538 – 9472/05/$95.00

JMASM21: PCIC_SAS: Best Subsets Using Information Criteria

C. Mitchell Dayton Xuemei Pan
Department of Measurement & Statistics
University of Maryland

PCIC_SAS is a SAS program for identifying optimal subsets of means based on independent groups. All possible configurations of ordered subsets of groups are considered and a best model is identified using both the AIC and BIC information criteria. Results for models with homogeneous variances as well as models with heterogeneity of variance in the same pattern as the means are reported.

Key words: PCIC_SAS, information criteria, AIC, BIC, paired-comparisons

Introduction

Researchers often use analysis of variance (ANOVA) to investigate mean differences among several groups. If the null hypothesis of equality of means is rejected, it is common practice to employ multiple comparison techniques to further study the pattern of differences among the means. For example, Kirk (1995) described 22 multiple comparison procedures including nine pairwise comparisons such as the Tukey honestly significantly different (HSD) procedure and Dunnett's T3 test. Statistical packages often include a variety of competing procedures with, for example, SAS 8.1 allowing the user to choose among 12 distinct methods for pairwise comparisons. Often, these procedures depend upon interpreting multiple significance tests. As detailed in the next section, Dayton (1998, 2003) advocated replacing these procedures by a holistic model selection approach based on information criteria. The purpose of this article is to describe and make available to applied researchers a SAS program, PCIC_SAS, that implements this modern information theoretic approach for comparisons among means.

C. Mitchell Dayton is Professor and Chair of the Department of Measurement, Statistics & Evaluation. His research interests include experimental design and latent class modeling. Email him at cdayton@umd.edu. Xuemei Pan is a Ph D candidate in the Department of Measurement, Statistics & Evaluation. Her research interests include latent class modeling and model comparison procedures. Email her at xpan1@umd.edu. The program mentioned in this article is available at www.education.umd.edu/EDMS/Latent/PCIC.txt

Application of Information Criteria to the Paired-Comparisons of Means

The widely-used Tukey Honestly Significantly Different (HSD) procedure for K independent group means involves the computation of 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. Like similar pair-wise comparison procedures, Tukey HSD entails testing K(K – 1)/2 hypotheses of the form $\mu_k = \mu_{k'}$ for $k \neq k'$. Often this is done subsequent to testing the omnibus hypothesis of equality of means (i.e., $\mu_k = \mu$ for k = 1,…,K) using analysis of variance techniques. Theoretically, the omnibus test is not required since the K-range pairwise comparison is an equivalent, although less powerful, test. There are many optional procedures based on modifications to the Tukey procedure or based on related notions using stepwise procedures. See, for example, the Kirk (1995) reference cited above for details of many of these procedures.

Among the problems with pairwise comparison procedures cited by Dayton (1998, 2003) 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 because 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 11.5 for Windows offers eighteen distinct procedures to choose among).

For K independent groups, there is a total of 2^{K-1} patterns of ordered subsets with equal means within subsets. For example, with four groups with means ranked and labeled 1, 2, 3, 4, the $2^3 = 8$ distinct ordered subsets are {1234}, {1,234}, {12,34}, {123,4}, {1,2,34}, {1,23,4}, {12,3,4} and {1,2,3,4}, where a comma is used to separate subsets with unequal means. Dayton (1998, 2003) proposed using model-selection criteria such as the Akaike (1973) AIC statistic for selecting the most appropriate ordering of subsets of means for purposes of interpretation. In particular, this approach avoids many of the objections that can be raised with respect to conventional pairwise comparison procedures. Information criteria such as AIC are based on the logarithm of the likelihood of the data, Log_e(likelihood). Sclove (1987) noted that AIC represents a penalized log-likelihood function of the general form:

$$-2Log_eL(likelihood) + a(n)p$$

where a(n) is a function that may depend upon the total sample size, n, and p is the number of independent parameters estimated in fitting the model to the data. Akaike's AIC is equal to

$$-2Log_eL(likelihood) + 2p$$

which does not directly depend upon sample size. Various adaptations of or alternatives to AIC have been suggested that, unlike AIC, are explicitly dependent upon sample size. In particular, the Schwarz (1978) BIC statistic and the Bozdogan (1987) CAIC statistic use penalty terms equal to $Log_e(n)$ and $Log_e(n) + 1$, 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. It is beyond the scope of this article to discuss the basis for selecting among alternative information criteria. However, these issues are discussed in Dayton (2003).

In practice, AIC (or, BIC) is computed for all competing models that the researcher wishes to compare. Then, from an information theoretic perspective, the model satisfying a min(AIC) (or, min(BIC)) criterion is selected as the best approximating model for the data being analyzed. Note that the min(AIC) (or, min(BIC)) strategy does not suggest that the selected model either fits or does not fit the data but that, among the models being compared, it is, in the information sense, the best choice. If additional models were added to the basis of comparison, a different selection might occur although the previously computed AIC values would not be altered.

The program, PCIC_SAS, 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, in practice, become quite large (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 (e.g., 5), are printed out. There is no limit to the number of groups that can be analyzed but, of course, execution time can become relatively

long for large K. In PCIC_SAS, it is assumed that the observations arise from normal densities.

Note, that 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. PCIC_SAS calculates AIC and BIC based on the usual assumption of homogeneity of variance as well as based on a restricted heterogeneous variance model in 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, 2003) and is the procedure incorporated into PCIC_SAS.

For a model with T subsets of means, the number of independent parameters, p, is equal to T+1 for the homogeneous case and 2T for the restricted heterogeneous case. Because $Log_e(n)$ is greater than 2 for n greater than 7, AIC and BIC may, and often do, result in different orderings of subsets of means with, predictably, simpler models being favored by BIC because of the larger penalty term. In Dayton (1998), results of a limited simulation with AIC and CAIC (the slightly different criterion than BIC with penalty term $Log_e(n+1)p$ suggested by Bozdogan (1987)), 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." For a more extensive simulation providing favorable results for PCIC, see Cribbie and Keselman (2003).

Using the PCIC_SAS Program

PCIC_SAS is written in the SAS programming language. For general-purpose analysis with a major statistical computer package, there is no other program that computes AIC and/or BIC for the models available in PCIC_SAS. 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). For users of the matrix-language, Gauss (Aptech Systems, 1997), appropriate code that provides input from spreadsheets such as Microsoft Excel is available (Dayton, 2001).

Data for analysis with PCIC_SAS can be in a SAS data base or imported into SAS from a spreadsheet or database program. It is conventional to code the groups with names, or 1, 2, etc., or A, B, etc. but PCIC_SAS rearranges the groups in rank order of means, from smallest to largest, and presents groups in ranked order, 1, 2, etc., in the output. Results are directed to the SAS output screen that can be printed and/or saved.

Example

Summary statistics for five ethic groups, based on a 5% random sample of cases from the NELS88 database, are presented below (see //nces.ed.gov/surveys/nels88/ for information about the longitudinal study of youth). The dependent variable is mathematics achievement on a standardized scale with population mean of about 50 and standard deviation of about 10. The five groups, as documented with the database, are: (1) API (Asian/Pacific Islander), (2) Hispanic, (3) Black-Non-Hispanic, (4) White-Non-Hispanic, and (5) American Indian. In rank order of means from low to high on the output these become: (3) Black-Non-Hispanic, (2) Hispanic, (5) American Indian, (4) White-Non-Hispanic and (1) API. The PCIC_SAS summary table and output for the five smallest values of AIC and BIC are summarized below:

Summary Table - group means in original order

Obs	race	_FREQ_	mean	sd	n	varunb	varmle	sum	ss
1	1	75	53.25	10.26	75.00	105.19	103.79	3993.45	7783.89
2	2	139	47.00	8.28	139.00	68.50	68.01	6532.98	9453.36
3	3	153	45.63	8.37	153.00	70.09	69.63	6981.58	10654.00
4	4	798	52.96	10.14	798.00	102.78	102.65	42258.81	81913.54
5	5	44	47.21	7.22	44.00	52.15	50.96	2077.40	2242.25
		1209						112047.04	

Summary Table - group means in rank order

Obs	race	_FREQ_	mean	sd	n	varunb	varmle	sum	ss
1	3	153	45.63	8.37	153.00	70.09	69.63	6981.58	10654.00
2	2	139	47.00	8.28	139.00	68.50	68.01	6532.98	9453.36
3	5	44	47.21	7.22	44.00	52.15	50.96	2077.40	2242.25
4	4	798	52.96	10.14	798.00	102.78	102.65	42258.81	81913.54
5	1	75	53.25	10.26	75.00	105.19	103.79	3993.45	7783.89

AIC and BIC for Homogeneous Case

Rank of AIC, value of AIC and ordered subsets for homogeneous variance case:

AIC_HOMOG

1	8914.598	1	1	1	2	2
2	8914.785	1	2	2	3	3
3	8916.240	1	1	2	3	3
4	8916.535	1	1	1	2	3
5	8916.722	1	2	2	3	4

Rank of BIC, value of BIC and ordered subsets for homogeneous variance case:

BIC_HOMOG

1	8929.890	1	1	1	2	2
2	8935.175	1	2	2	3	3
3	8936.630	1	1	2	3	3
4	8936.926	1	1	1	2	3
5	8942.210	1	2	2	3	4

AIC and BIC for Heterogeneous Case

Rank of AIC, value of AIC and ordered subsets for patterned heterogeneous variance case:

AIC_HETEROG

1	8895.898	1	1	1	2	2
2	8897.075	1	2	2	3	3
3	8897.724	1	1	2	3	3
4	8899.729	1	2	3	4	4
5	8899.838	1	1	1	2	3

Rank of BIC, value of BIC and ordered subsets for patterned heterogeneous variance case:

BIC_HETEROG

1	8916.288	1	1	1	2	2
2	8927.660	1	2	2	3	3
3	8928.309	1	1	2	3	3
4	8930.423	1	1	1	2	3
5	8936.311	1	1	2	2	2

Interpretation

For AIC, all five reported heterogeneous-variance models have smaller values than the best homogeneous-variance model and for BIC this is true for the first three heterogeneous models. Thus, models with variances that differ among subsets of means are favored over homogeneous-variance models. Based on both AIC and BIC, the preferred model is reported as: 1, 1, 1, 2, 2. This suggests that there are two subsets of means comprised of the groups with the three smallest means in one subset and the groups with the two largest means in the second subset. This corresponds to the pattern {Black-Non-Hispanic, Hispanic, American Indian} in the subset with smaller means and {White-Non-Hispanic, API} in the subset with larger means. Note that the conclusion should not be drawn that, for example, the means are equal for the White-Non-Hispanic and API groups but, rather that the data are not sufficiently reliable to permit an ordering within that subset. The variances for the two subsets are not reported but can be easily computed from the output (see Dayton, 1998) and are equal to 67.02 and 102.75, respectively.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & 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.

Cribbie, R. A. & Keselman, H. J. (2003). A power comparison of pairwise multiple comparison procedures: A model testing approach versus stepwise procedures. *British Journal of Statistical & Mathematical Psychology*, *56*, 157-182.

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

Dayton, C. M. (2001). SUBSET: Best subsets using information criteria. *Journal of Statistical Software, 6*(2).

Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods, 8, 61-7*.

Greeley, A. M., McCready, W. C. & Theisen, G. (1980). *Ethnic drinking subcultures*. New York: Praeger.

Kirk, R. E. (1995). *Experimental design* (3^{rd} ed.). Brooks/Cole.

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

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.

Appendix

The theoretical background for AIC derives from information-theoretic concepts originally presented by Kullback and Leibler (1951). The mathematical material presented in this section is supplementary to that presented above and can be skimmed or omitted without any serious loss of understanding of the PCIC technique.

Adapting the notation of Akaike (1973, 1974, 1987) for univariate data, the Kullback-Leibler information for the true distribution, $g_t(x)$, of random variable x, relative to some other distribution, $g_o(x)$, is:

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

where all expectations are taken with respect to $g_t(x)$. In statistical applications making use 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, $\boldsymbol{\theta}$, and let $\boldsymbol{\theta}_x$ be the usual maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ found by maximizing $g(\mathbf{x} \mid \boldsymbol{\theta})$ over the sample by treating $\boldsymbol{\theta}$ as variable. Assuming p independent parameters, 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.

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:

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

Noting that the first term in Equation (1) is constant for any model, Akaike defines the AIC estimator of Kullback-Leibler information as:

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

For M different models for the same data, the Akaike min(AIC) procedure involves using Equation (5) to calculate AIC_m, m = 1,…,M, for the models and selecting 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.

Sclove (1987) notes that AIC represents a penalized log-likelihood function of 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 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.