AUSTRALIAN COUNCIL FOR EDUCATIONAL RESEARCH

Number 4

July, 1974

ACER DATA PROCESSING BULLETIN

Checking of punched data

Before passing any data through for processing it is important that the following checks be applied to the punched cards.

1. Count of cards - make sure that the number of cases punched equals the number of subjects in the study.
2. Line up of data columns/gaps/blanks.
3. Check of identification numbering system - it is very important that there are no errors in the identification system when data is being run with the OSIRIS package of programs.

Marge Corfe will undertake this basic checking of data cards. However Marge will not be able to carry out full checks of data preparation. Full checks of punched data should be organized in cooperation with Joan Tyson and Edith Cooper.

If any project wishes the punching of their data cards to be fully verified, would they discuss this proposal with Ken Ross.

IEBGENER COPY SYSTEM

Within the IBM standard library of programs there is a useful copying program which may be used to transfer a data set from one medium to another.

The following Job Control Language setups have been found to be quite useful at ACER.

(i) Copy punched data onto ACER's disk pack at ICI (in "card-image" format).

```
      job card
//GO EXEC PGM=IEBGENER
//SYSPRINT  DD  SYSOUT=A
//SYSIN  DD  DUMMY
//SYSUT1  DD  DDNAME=DATAIN
//SYSUT2  DD  DDNAME=DATAOUT
//GO.DATAOUT  DD  DSN=KENSDATA,UNIT=2314,VOL=SER=051
//  DCB=(RECFM=FB,LRECL=80,BLKSIZE=800),
//  DISP=(NEW,KEEP),SPACE=(CYL,(1,2),RLSE)
//GO.DATAIN  DD  *
      data cards
/*
```

For each dataset being copied a new name must be supplied. The name of the dataset should not exceed eight letters. In the above example the dataset has been named "KENSDATA".

(ii) Duplicate a deck of punched cards.

```
      job card
//GO EXEC PGM=IEBGENER
//SYSPRINT  DD  SYSOUT=A
//SYSIN  DD  DUMMY
//SYSUT1  DD  DDNAME=DATAIN
//SYSUT2  DD  DDNAME=DATAOUT
//GO.DATAOUT  DD  UNIT=00D,
//  DCB=(RECFM=F,BLKSIZE=80,LRECL=80)
//GO.DATAIN  DD  *
      data cards or program cards
/*
```

(iii) Copy a dataset from ACER's disk pack to the printer.

```
       job card
//GO EXEC PGM=IEBGENER
//SYSPRINT  DD  SYSOUT=A
//SYSIN  DD  DUMMY
//SYSUT1  DD  DDNAME=DATAIN
//SYSUT2  DD  DDNAME=DATAOUT
//GO.DATAOUT  DD  SYSOUT=A,
//  DCB=(RECFM=FB,LRECL=80,BLKSIZE=480)
//GO.DATAIN  DD  DSN=KENSDATA,UNIT=2314,
//  VOL=SER=051,DISP=(OLD,KEEP)
/*
```

OSIRIS system

The Automatic Interaction Detector program was recently "unwrapped" by Chris Slee.

The AID operates on data with one dependent variable (which may be dichotomous, continuous or equal-interval) and up to 40 categorical predictors.

The program successively sub-divides the sample by a series of dichotomies according to a pre-determined strategy. The first step is to split the sample into two groups based on the values of one predictor. The choice of predictor and split is made to maximise the sum of squares explained by the group difference. The next step is to select the subgroup with the greatest within-group sum of squares, and then split this group in a similar fashion. The outcome of this splitting process is a "tree" structure of subgroups.

At each step of the AID process all predictors and all allowable splits are examined. If the values of a predictor are ordered, this order can be preserved; otherwise all feasible splits are examined.

The AID program runs at 76K on the ICI computer. More information may be obtained from the OSIRIS manual of programs.

COOLEY and LOHNES Program : CANON

The Cooley and Lohnes Canonical Analysis program has been recently put into operation by John Thomson and Chris Slee, following receipt of advice from W.W. Cooley on modifications necessary for the program to run on an IBM 360 system.

The canonical correlation is the maximum correlation between linear functions of two vector variables. However after a pair of linear functions that maximally correlates has been located, there may be an opportunity to locate additional pairs of functions. that maximally correlate, subject to the restriction that the functions in each new pair must be uncorrelated with all previously located functions in both domains. Thus, whereas in factor analysis the factor model selects linear functions of tests that have maximum variances, subject to restrictions of orthogonality, in canonical analysis the canonical model selects linear functions that have maximum covariances between domains, subject again to restrictions of orthogonality.

Also included in the printout of this analysis is the "total redundancy" of each set of variables (eg. the proportion of the 1st battery variance that is redundant to the 2nd battery variance according to the rank n canonical model). This total redundancy measure is nonsymmetric. The redundancy measure is important because a very large canonical correlation coefficient could be the result of a very large zero-order correlation of just one variable of one set with just one variable of the other set, and the remainder of the two sets could be essentially uninvolved in the canonical structure.

The CANON program runs at 88K on the ICI Computer. More information may be obtained from:

Cooley, W.W. and Lohnes, P.R. Multivariate Data Analysis : New York, Wiley, 1971.