Fixing a Problem Detected with the 1970 Malaysian Data, July 6, 2006, Patty Glynn

With permission from the Malaysian Government, Charles Hirschman turned the Malaysian Census data over to Robert McCaa of Minneapolis Population Center so that the data can be distributed by IPUMS. Professor McCaa wrote in May of 2006 letting us know that the 1970 data did not fit the code book. I found out that the problem was that the data were not formatted properly, probably because they had not been copied from tape correctly (without blocking). All of the data were on a single line.

I found a copy of the file that I had been copied from tape correctly. For reasons I do not understand, there are 748 records with a zero in the first column and nothing in other columns. And, the file was variable length. Both of these made the data a little more difficult to read. I created versions that are easier to use.

Professor Hirschman asked that I not delete older versions. To avoid future confusion, I am storing older versions in a web subdirectory named "Fix1970Problem". The files will not be accessible from the web. Information pertaining to this correction (including this document) and programs used to create different versions will be stored there. All versions of data sets will be stored zipped files. A password will be required to access them.

The file that is formatted incorrectly and is therefore hard to use with SAS or SPSS is now named "MAL70BADFMT.DAT". Prior to discovering that it is not formatted properly, it was named Malay70.dat. It was created in June of 1999. I do not have information about how it was copied or created.

MAL70FMT is the name of the file that I originally copied from tape on the UNIX when I was backing up Charlie's data. It has a variable length, which makes it a little harder to read. I have named "MAL70VARLNG.DAT" it the following to store on the web. More information about how the data were copied from tape to disk can be found in the file MAL70VARLINGinfo.txt.

I created a file "MAL70FMTpad.DAT" with the program called "MAL70FMTpad.sas". It is the same as MAL70VARLNG.DAT except that the records are padded so that every line extends 80 columns.

Finally, I created the file MALAY70R.DATA with the program MALAY70R.sas. This is the same as MAL70FMTpad.DAT except I removed the 748 records with a zero in the first column and nothing in other columns. This final data set is the one that is linked on the website because it is most usable.

This is the first time that I have worked with these data, and I was puzzled by the setup. Perhaps you have already worked through these puzzles, but in case I can save you some time, I will explain what have found.

The file has three types of records in it.

1. Household Level
2. Person records for persons under 10 years of age
3. Person records for persons 10 and over years of age

Household Level

For the household level data, the variable indicating record type can be read in column 8, but a selection must also be made in for column 9.

```
if rectype = 4 ;
if col9 ne 5  ;
```

Person Level

The variable indicating record type can be read in column 9. Values of 5 should be selected. Age in columns 28 - 30 must be read to determine if the person is under or over 10.

Remaining Puzzles

On page 9 of the documentation there are frequencies for the variable "STATE". I replicated the frequency for the household level, but I have not been able to replicate the frequencies at the person level. I find 175,882 rather than 176,287 people.

I find that "age" ranges from 0 to 199 but are concentrated in reasonable values. I have not checked all frequencies.