'agesmk' :

Age Started Smoking

This variable is the response to the question: “How old was [ person] when he/she started smoking cigarettes fairly regularly?”

'race': An expanded version of race which separates the American Indian, etc. and Asian, etc. out of the Other category in early CPS files.

|  |  |
| --- | --- |
| **Codes** | |
| . | Unknown or missing |
| 1 | White |
| 2 | Black |
| 3 | American Indian or Alaskan Native |
| 4 | Asian or Pacific Islander |
| 5 | Other nonwhite |
|  |  |

'age': Age at Time of Interview

The age of the person at the time of the survey. If age was not available on the CPS record, then age was taken from CPS control file data. If age was not available from either of these sources but month and year of birth were available, then age was calculated using month and year of birth. Age is top coded at 90 years. Records of full NLMS persons of unknown age are not included in this file

'sex':

|  |
| --- |
| **Codes** |
| 1. Male |
| 1. Female |

'ms': Marital status at time of survey(in our study, it coverted to 1 or 0, where 1 means married.)

|  |  |
| --- | --- |
| **Codes** | |
| . | Unknown, persons LT 15 years old |
| 1 | Married |
| 2 | Widowed |
| 3 | Divorced |
| 4 | Separated |
| 5 | Never married |

'wt': Adjusted Weight

These public-use files consist of records from many separate CPS or ASEC Surveys. All records have been weighted to give population totals for the non-institutionalized population of the U.S. on April 1, 1983 for **File 11** and File **6a**, April 1, 1993 for File **6b**, and April 1, 2002 for File **6c**, and April 1, 2000 for **File tu**. Weights were obtained by raking age-sex-race group totals by state totals for each survey. Survey weights were then readjusted to the target dates of the associated U.S. population and accounting for the different numbers of records by survey in this each of the four files.

'hhnum': The number of persons residing in the household at the time of the interview

'follow': The length of follow-up period in days.

Persons who were alive at the end of the 11 year follow-up for File 11 period are given a value of 4018, the maximum follow-up period considered. Persons who were alive at the end of the 6 year follow-up period, Files 6a, 6b, 6c are given a value of 2192, the maximum follow-up considered. Persons in the tobacco use file, tu, are given the maximum of 1827 or five years of followup.

Get 5% of the data code

nlms <- read.csv(“...”)

dim(nlms) # 493282 41

# we need to have response y

require(plyr)

nlmsc <- nlms[!is.na(nlms$agesmk), ]

dim(nlmsc) # 193089 41

nlmsc<- nlmsc[ ,c('agesmk','race','age','sex','ms','wt','hhnum','follow')]

nlmsmk <- nlmsc[which(complete.cases(nlmsc)), ]

# complete cases for the selected variables

dim(nlmsmk) #192764

# change

nlmsmk$race <- factor(nlmsmk$race, levels = unique(nlmsmk$race))

nlmsmk$sex <- factor(nlmsmk$sex)

nlmsmk$ms <- ifelse(nlmsmk$ms==1, 1, 0) # ms to binary

nlmsmk$ms <- factor(nlmsmk$ms)

# select 5%

set.seed(101);

nlms5per <- nlmsmk[ sample(1:nrow(nlmsmk), 0.05\*nrow(nlmsmk) ) , ]

dim(nlms5per) # 9638 8

# the race distribution percentage

table(nlms5per$race)/sum(table(nlms5per$race))#5% data race distribution

1 3 5 2 4

0.892 0.013 0.005 0.072 0.018

table(nlmsmk$race)/sum(table(nlmsmk$race)) # full data race distribution

1 3 5 2 4

0.893 0.013 0.005 0.071 0.018

# save/read the data

write.table(nlms5per, "nlms5per.txt", quote = FALSE)

dat <- read.table("…/nlms5per.txt", sep="")