STAT 798c, HW Problem Set #2, due Wednesday 2/26/03

Suppose that 48 demographic strata are defined in the lexicographic order of the five demographic factor variables, related to Census block-groups:

FB with levels Lo, Hi
FOWN with levels Lo, Hi
FSPOU with levels Lo, Hi
SNGF with levels 0, 1
HTYP with levels Mob, Sng, Apt

Note that in the first three covariates, ”Lo” comes before ”Hi”, so the first 4 of the 48 strata are:

```
FB=Lo, FOWN=Lo, FSPOU=Lo, SNGF=0, HTYP=Mob
FB=Lo, FOWN=Lo, FSPOU=Lo, SNGF=0, HTYP=Sng
FB=Lo, FOWN=Lo, FSPOU=Lo, SNGF=0, HTYP=Apt
FB=Lo, FOWN=Lo, FSPOU=Lo, SNGF=1, HTYP=Mob
```

Suppose further that you are given the Splus object **StratMR** which is available in the **/usr/local/StatData/SplusCrs/.Data** public directory or which you can get in the form of three appended 51x48 ASCII tables in the Data link of the web-page as **StratMR.asc**. (The three tables correspond successively to ”Predicted”, ”Observed”, and ”Cellct”. Each line of each table begins with an abbreviated state-name followed by 48 numbers separated by blanks.) **StratMR** is an array with dimension-vector c(51,48,3) , where the first dimension corresponds to States and DC (with appropriate abbreviated *dimnames*); the second dimension corresponds to the 48 covariate-combinations ordered lexicographically by the five factors indicated above; and the third dimension corresponds to the three levels ”Predicted”, ”Observed”, and ”Cellct”. The array entries for ”Predicted” and ”Observed” 3rd-dimension levels are respectively model-fitted and actual fractions of Households in the cross-classified State and Demographic neighborhood-category who responded to the 1990 Census by mail. The models were fitted by me using Census Bureau data, and my overall objective is to understand which State/Demographic combinations are associated with discrepancies between the ”Predicted” and ”Observed” numbers, but it is likely that the sizes of discrepancies is strongly associated with ”Cellct” entries, which are the total numbers of Households enumerated nationally in each State-by-Demographic group.

(A) Create from the array **StratMR** a data-frame with 48 rows and one column for each of the five factor-variables FB, FOWN, FSPOU, SNGF, and HTYP together with the columns "Nprdict", "Nobs", and "Cellct", where the "Nprdict" column contains the sum over all states of the product of "Predicted" and "Cellct" entries from **StratMR** ; and similarly the "Nobs" contains the sum over all states of the product of "Observed" and "Cellct" entries from **StratMR** .

(B) Produce two plots on the same page, one for strata with FOWN=LO and one for strata with FOWN=HI, of the ratios Nprdict/Cellct versus Nobs/Cellct, with the following plotting-characters:

squares for cases with FOWN=LO and SNGF=0
hollow diamonds for cases with FOWN=LO and SNGF=1
circles for cases with FOWN=HI and SNGF=0
filled diamonds for cases with FOWN=HI and SNGF=1

Also provide captions, a heading-title, and legend-boxes for your plots.

(C) Define a smaller data-frame starting with the data-frame you have constructed in (A), deleting all rows corresponding to national Cell-counts (i.e., **Cellct** summed over all states and DC) less than 5000, and ordering the remaining rows in increasing order of **Nobs/Cellct**.

(D) There are several ways to group States into regions. Here is one:

NewEng: MAI, MA, RI, CT, VT, NH, NY
MidAtl+: NJ, PA, DE, MD, VA, DC, HA
No.Cen: OH, IN, WV, IL, MICH, WI, MN, ND, SD
MidW: IO, NEB, KS, OK, MO
South: FL, MISS, ALAB, GA, NC, SC, LA, TENN, TX, ARK, KY
West: CA, OR, WASH, IDA, MON, WY, UT, AZ, NM, ALAS, NV, CO

Provide plots or tables to answer the question: are the discrepancies between **Nobs/Cellct** and **Nprdict/Cellct** closely associated with any of the Regional groupings of States ? *This is an atheoretical question: the issue is descriptive data-display to show similarity (or lack of it) by region.*

(E) Can you find any patterns at all in these data which have not already been suggested by the questions above ? This question is open-ended: you could try to form your own grouping of states and/or strata, e.g. to differentiate Urban from Rural; you could try to see whether the ordering of **Nobs/Cellct** or of **(Nprdict-Nobs)/Cellct** remains the same among (the larger) strata from one state to another; or you could try anything else you want. *Don't hand in more than your best one or two displays for this part. See Ripley & Venables' discussion of Boxplots — which is what you get when you plot a quantitative response-variable versus a Factor — and other Descriptive Statistics for further ideas of displays.*