Second Mini-Project, 36-401, Spring 1998

Description of the problem

What factors affect childrens' perceptions of themselves, their life goals, etc? Chase and Dummer (Chase, M. A., and Dummer, G. M. (1992), *The Role of Sports as a Social Determinant for Children*, Research Quarterly for Exercise and Sport, 63, 418–424) surveyed school children in Michigan to find out.

The subjects consist of 478 students in grades 4–6 from three school districts in Ingham and Clinton Counties, Michigan. Chase and Dummer stratified their sample, selecting students from urban, suburban, and rural school districts with approximately 1/3 of their sample coming from each district. Students indicated whether good grades, athletic ability, or popularity was most important to them. They also ranked four factors: grades, sports, looks, and money, in order of their importance for popularity. The questionnaire also asked for gender, grade level, and other demographic information.

An extract of the data suitable for log-linear modeling is contained in the data file `catkids.dat` in `http://www.stat.cmu.edu/~brian/402/mp2`. The data looks like this

```
  Gender Grade  Race Urban  Goals Resp
1    boy     4 Other Rural Grades    0
2   girl     4 Other Rural Grades    0
3    boy     5 Other Rural Grades    0
4   girl     5 Other Rural Grades    1
5    boy     6 Other Rural Grades    0
6   girl     6 Other Rural Grades    2
```

where the variable names are:

- `Gender` is the gender of the child (boy/girl)
- `Grade` is the child's grade in school (4/5/6)
- `Race` is the child's race (White/Other)
- `Urban` is the location of the school district (Urban, Suburban or Rural)
- `Goals` are the child's goals (Grades/Popular/Sports)

  and
- `Resp` is the number of children with each combination of the other factors.

Your mission

Your mission is to explore and analyze these data, and write up a report of your conclusions. The main questions to keep in mind are:

- Which factors listed above are related to each other? What are the conditional independence relationships among the variables?

- Which factors seem to influence the "Goals" of the child?
- How do childrens' goals change depending on gender? race? urban setting? grade in school?
- Is there a causal story to tell about these variables?

You may use either (or both) of log-linear modeling and logistic regression, as well as any appropriate EDA, residual diagnostics, etc. For the log-linear modeling, please:

- Since the authors tried to stratify their sampling on "Urban", assume that the totals for Urban are fixed.
- In your final models please try to obey both
  - *the hierarchy principle* (including an interaction forces you to include all lower-order interactions and main effects for the same variable); and
  - *the graphical principle* (if every $(n-1)$-way interaction among n variables is in the model, tnen also put the n-way interaction in the model)

If you also decide to do some logistic regression,

- Note that logistic regression will require a reformulation of the data (see problems 2 and 3 on the third set of finger exercises for an example of such a reformulation).
- Try to obey the hierarchy principle in your final models

On Writeups

A good generic outline for project reports is:

0. Executive summary (state background, problem, and conclusions in one page or less);
1. Description of problem, problem background, and data;
2. Summary of analyses performed (what, why, how, results);
3. Conclusions;
4. Technical Appendices;
5. Bibliography and Credits

Parts 1–3 should have as little technical material as you can possibly get away with: a few well-chosen graphs, the names of the techniques you applied and what parts of the data were used, and perhaps one or two parameter estimates or p-values is the most you should think about. Part 4 should contain the "story" of your technical analyses. However, it should be possible to understand what you've done from parts 1–3, without looking at part 4 at all.

You are welcome to discuss this problem with each other, but I would like a list of people you discussed this with (in and out of class), as well as a list of any outside sources you consulted (books, Web pages, etc.) in Part 5 of your report.

Timeline, Deadines

I would like this to be a "miniproject", not a full-blown project. So I would like to discuss with you your progress after a week or so, probably in class, and then I would like you to hand in your final reports a week later. So the deadlines are:

- *Thu Mar 19.* I will take questions in class on the data, the project, etc.
- *Tue Mar 31.* Bring a draft of Part 4 of your report to class for me to check, in class.
- *Tue Apr 6.* Bring your final reports to class.

Appendix: the full data set

The file `kids.dat` in `http://www.stat.cmu.edu/~brian/402/mp2` is a more complete version of the data set. It looks like this

```
  Gender Grade Age  Race Urban.Rural School   Goals Grades Sports Looks Money
1    boy     5  11 White       Rural    Elm  Sports      1      2     4     3
2    boy     5  10 White       Rural    Elm Popular      2      1     4     3
3   girl     5  11 White       Rural    Elm Popular      4      3     1     2
4   girl     5  11 White       Rural    Elm Popular      2      3     4     1
```

The variable names are

1. Gender: Boy or girl
2. Grade: 4, 5 or 6
3. Age: Age in years
4. Race: White, Other
5. Urban/Rural: Rural, Suburban, or Urban school district
6. School: Brentwood Elementary, Brentwood Middle, Ridge, Sand, Eureka, Brown, Main, Portage, Westdale Middle
7. Goals: Student's choice in the personal goals question where options were 1 = Make Good Grades, 2 = Be Popular, 3 = Be Good in Sports
8. Grades: Rank of "make good grades" (1=most important for popularity, 4=least important)
9. Sports: Rank of "being good at sports" (1=most important for popularity, 4=least important)
10. Looks: Rank of "being handsome or pretty" (1=most important for popularity, 4=least important)
11. Money: Rank of "having lots of money" (1=most important for popularity, 4=least important)

You may find it useful to go back to this data set, and perhaps produce a different extract of the data for your particular analyses. You may also find that there are interesting questions using the full data set which simply can't be answered with the more compact data set `catkids.dat`. (For example we could build a logistic regression model for the outcome "Grades ranked first as a determinant of popularity"; this can be done with `kids.dat` but not with `catkids.dat`).

Please feel free to work with either data set `catkids.dat` and.or `kids.dat` depending on where your investigations lead you.