Assignment # 2, Stat 8053, Fall 2013

The reading for this week is Chapters 5–6 of Faraway and the LCA handout. The following problems are due on Wednesday, October 2, 2013, in class.

As with all problems, start by checking the data for obvious problems.

1. Page 112, problem 1. Why is it reasonable to not use transformations in this problem? Also consider the need for interactions, and provide useful graphical summaries. See `?hsb` for definition of the variables.

2. Page 113, problem 5. The response `ccarduse` is an ordered category. For simplicity, use only the 304 cases that are fully observed (that is, textttdebt1 <- an.omit(debt). Use graphical summaries where appropriate. See `?debt` for definition of the variables.

3. The data for this problem come from the 2005 Youth Risk Behavior Survey, `http://www.cdc.gov/healthyyouth/yrbs/data/index.htm`. The Youth Risk Behavior Surveillance System (YRBSS) is an epidemiologic surveillance system established by the Centers for Disease Control and Prevention (CDC) to monitor the prevalence of youth behaviors that most influence health. The YRBSS focuses on priority health-risk behaviors established during youth that result in the most significant mortality, morbidity, disability, and social problems during both youth and adulthood. These include: behaviors that result in unintentional and intentional injuries; tobacco use; alcohol and other drug use; sexual behaviors that result in HIV infection, other sexually-transmitted diseases (STDs), and unintended pregnancies; dietary behaviors; and physical activity, plus overweight and asthma.

   The target population for the survey consisted of all US students in grades 9 through 12. The data were collected using a cluster sample: 203 schools were selected with probability proportional to enrollment, and then essentially all students in grades 9 to 12 were surveyed. In the end 159/203 or 78% of the schools participated, and in these schools 13,917 out of 16,262 eligible students completed surveys for a response rate of 86%. The overall response rate was therefore 67%. Analysis should in principle account for unequal sampling probabilities and non-response, but we will ignore this issue for this problem.

   We will use in this example the following 14 variables.

QN29	Smoked first cigarette before age 13 (1=yes, 2=no)
QN34	Ever smoked daily for 30 consecutive days (1=yes, 2=no)
QN11	Driven while drinking in last 30 days (1=yes, 2=no)
QN40	First drank alcohol before age 13 (1=yes, 2=no)
QN42	At least 5 drinks in one day in last 30 days (1=yes, 2=no)
QN45	Tried marijuana before age 13 (1=yes, 2=no)
QN48	Ever used cocaine (1=yes, 2=no)
QN50	Ever sniffed glue (1=yes, 2=no)
QN52	Ever used methamphetamines (1=yes, 2=no)
QN53	Ever used ecstacy (1=yes, 2=no)
QN58	Had sex before age 13 (1=yes, 2=no)
QN59	Had sex with 4 or more partners in lifetime (1=yes, 2=no)
grade	factor with levels 9, 10, 11, 12
gender	factor with levels F and M

The variable names are the same as in the data file provided by CDC. The first 5 questions are concerned with drinking and driving, the next 5 with other drug use, and the last two with sexual behavior. Also, QN29, QN40, QN44 and QN57 have a time component.

You can load the data in the file `yrbs05` as follows:

```
loc <- "http://tinyurl.com/yrbs05-rda"
load(url(loc))
```

(a) Obtain the proportion responding Yes to each of the 12 risky behaviors.

(b) Use `poLCA` to fit latent class models for the 12 behavior indicators. First fit the model with `nclass=1`. Recall this is the model of complete independence for the 12 indicators in a 2^{12} contingency table. Verify that the G^2 or χ^2 test for this model are enormous relative to the df. This should be expected because (1) complete independence doesn't make any sense, and (2) the sample size is enormous and so power is high for any test.

(c) Fit a two-class latent variable for the 12 behavior indicators, ignoring the two covariates. You should use several random starts, say `nrep=5`. Provide a summary of the output. This should include: (1) comparison of this model to the one-class model probably via comparing G^2 values, AIC or BIC; (2) a summary of the fraction of respondents in each group; (3) examine the estimated probabilities; (4) finally if your model is called `p2`, then `p2$posterior` gives the posterior probabiliy of assignment of each of the subjects to classes. If 2 classes were adequate, then the probabilities of assignment to class 1 should all be close to 0 or 1. Look at a histgram of these and summarize.

(d) Add the covariates to the LCA fit and summarize the results. This means you need to interpret parameter estimates, examine change in G^2 and the like.

(e) Fit with k classes, for $k = 2, 3, \ldots, 6$. Use either AIC or BIC or both to verify that a 4 class solution is might be preferred. Summarize the findings of the 4 class solution by (1) describing the membership proportions; (2) summarizing the estimated probabilities, and come up with labels that give meaning to the classes; for example, if one class has very low probabilities of "yes" for all 12 questions, you might call that group the "risk avoiders"; (3) interpret the effects of the covariates given the 4 class model. You might find it helpful to use the `poLCA.reorder` function to make the "risk avoiders" class the baseline for understanding the estimates for the covariate effects.