Final Exam, Stat 8053, Fall 2009

This exam is due on Wednesday, December 23, at noon. Either hand your exam to me in 312 Ford or if I am not in my office to Myrna in 313 Ford. Do not leave your exam in my mailbox. Be sure you have a copy of your exam for yourself. If you cannot finish the exam on time, please contact me; late exams may not be graded until January. Graded exams and course grades will be ready by Tuesday, December 29 by the end of the day. Good luck! All work must be your own. You may not discuss the exam with anyone except for the instructor. If there are any changes or corrections to the exam, I will send email. I will be available by email and by appointment throughout the exam period.

What to turn in

It will be helpful to me, and may also help you organize your results, if you follow the following outline for each problem.

1. Provide an executive summary of your results. This should be no more than about a half page, and should give your conclusions. The summary will usually be supported by a small number of graphs and tables, but this is not necessary, as tables and graphs can appear in the next section.

2. Outline the analysis you did to reach the conclusions you stated in the summary. It is unlikely that you will include everything you did, as you are likely to try things that turn out to be irrelevant or not useful. Explain briefly why the methods you present are relevant. Computer output should illustrate what you did, but what you write is much more important than the computer output. You will lose credit for including computer output that you do not discuss. Apart from tables and graphs, this part of you solution should be no longer than two pages, and less may suffice. Please provide enough detail that I can figure out what you did and what you found without actually having to reproduce your computations. For example, if you use hierarchical clustering with complete linkage, you either need to write something like “I used hierarchical clustering with complete linkage”, or “I used hierarchical clustering using the R function hclust(x,method="complete")”.

3. You may feel the need for completeness to include additional computer output, such as fitted models, lots and lots of graphs,... This is unnecessary and is unlikely to help me understand the analysis you did, and so I suggest you do NOT do this. If you decide you must, you must explain why you have done so, and describe what the additional output reveals.

Educational Data

The data used for this example contains test scores on 418 eighth graders on six different tests:

- English composite score $Eng$
• Grades, overall grades
• Verbal VR reasoning score
• Nonverbal NVR reasoning score
• Arithmetic Arith score
• Comp Computational score

The first two scores Eng and Grades are thought to be observable or manifest variables that reflect the student's latent ability called "achievement" $Y$. Verbal and nonverbal ability are manifest variables thought to depend on a latent variable $F_1$ called "general reasoning ability", while the Arith and Comp scores are manifest variables for a latent $F_2$ "quantitative reasoning ability". If $Y$, $F_1$ and $F_2$ were observable, we might suspect that

$$E(Y|F_1, F_2) = \alpha_0 + \alpha_1 F_1 + \alpha_2 F_2$$  \hspace{1cm} (1)$$

at least as a starting point.

The data can be read into R with the following command

```r
> cvdata <- read.table(url("http://tinyurl.com/ltoa3f"), header = TRUE)
> head(cvdata)

    NVR VR Arith Comp Eng Grades
1     16 16   23    16  58    51
2     22 12   11    8  33    24
3     23 11   13    11  50    44
4     24 18   20   11  41    40
5     15 14   17   11  39    41
6     26 20   21   14  61    51
> dim(cvdata)

[1] 418   6
```

All six variables are the scaled results of tests given to eighth graders.

Your assignment is to learn about these data using any of the methods we have learned in this course. In particular, does (1) seem to make any sense for these data?

**CIA Factbook Data**

The data for this example is from the CIA Factbook [http://tinyurl.com/s8053CIA](http://tinyurl.com/s8053CIA), and can be loaded using

```r
loc<="http://www.stat.umn.edu/~sandy/courses/8053/Data/CIA/cia1.csv"
cia <- read.csv(url(loc), header=TRUE, row.names=1)
```
The data are for 195 countries, mostly for 2008, and include only variables that are fully observed on these countries. The variables in the file are:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>Land area kmsq</td>
</tr>
<tr>
<td>PopGrowthPct</td>
<td>Population</td>
</tr>
<tr>
<td>Pgrowth</td>
<td>PopGrowth Percent</td>
</tr>
<tr>
<td>BirthRate</td>
<td>Birth rate per 1000 women</td>
</tr>
<tr>
<td>DeathRate</td>
<td>Death rate per 1000 population</td>
</tr>
<tr>
<td>InfMort</td>
<td>Infant Mortality per 1000</td>
</tr>
<tr>
<td>LifeExp</td>
<td>Life Expectancy, years</td>
</tr>
<tr>
<td>Fertility</td>
<td>Number of children per woman</td>
</tr>
<tr>
<td>GDPgrowth</td>
<td>GDP growth rate, percent</td>
</tr>
<tr>
<td>GPDpp</td>
<td>GDP per person</td>
</tr>
<tr>
<td>InternetUsers</td>
<td>Number of internet Users</td>
</tr>
<tr>
<td>InternetHosts</td>
<td>Number of internet Hosts</td>
</tr>
<tr>
<td>LandLines</td>
<td>Number of land line telephones</td>
</tr>
<tr>
<td>CellPhones</td>
<td>Number of cell phones</td>
</tr>
<tr>
<td>RoadKm</td>
<td>Km of roads</td>
</tr>
<tr>
<td>Airports</td>
<td>Number of airports</td>
</tr>
</tbody>
</table>

You have great latitude in interpreting the questions below. Almost anything you do that uses the methods we learned in this course is acceptable.

1. First, exclude the variable *CellPhones*, and then explore the remaining variables using appropriate multivariate methods and summarize your results.

2. Then, treating *CellPhones*, or perhaps some transformation of it, as a response, explore using the methods we used in this course the conditional distribution of this variable given all the others.