

Assignment # 1, Stat 8053, Fall 2011

Reading

The reading for this week is Chapter 10 of Faraway.

Problems

The following problems are due on Wednesday, September 21, 2011, in class. A group of three of you, to be named on Monday, September 12, will be responsible for preparing solutions to the problems and one of the three will present in class that day.

1. The data for this problem are from the 1982 “High School and Beyond” survey, and pertain to 7185 high school students from 160 schools. The data can be read into R with the following command

```
> load(url("http://www.stat.umn.edu/~sandy/courses/8053/Data/Achiv.Rda"))
```

The variables in the data are:

school A factor with 160 levels, the school numbers.

ses Socio-economic status. The ses is measured on a scale that has mean of almost zero in these data.

mathach the student’s score on a math-achievement test.

sector a factor coded ‘Catholic’ or ‘Public’. This is a school-level variable and hence is identical for all students in the same school.

The goal is to answer: (1) is math achievement related to socioeconomic status; (2) does the relationship, if any, depend on sector, and (3) how do the relationships vary across schools within the same sector.

Your answer should include several steps. First, look at the data to decide if it makes any sense. This will include standard 1D summary statistics, appropriate 2D statistics and 2D plots. You might want to look at graphs of achievement vs. ses separately for each of several of the schools (probably not all 160 of them...) to decide if linear fits make any sense here. In this stage, you might find the `xypplot` function in the `lattice` package and the `lmList` function in the `nlme` package helpful. Then you should turn to using a mixed-model that seems appropriate to you to answer the questions posed, and obtain and describe appropriate graphical and numerical summaries of the models. Explain how these summaries answer the questions posed.

2. As part of the on-going study of wolves, a subset of the new pups born each year are captured and various physical measurements of the pups are taken. In this example we are concerned with the presence ($Y = 1$) or absence ($Y = 0$) of a particular virus in the pup’s blood. Of interest is whether or not the prevalence of the virus varies over time and if pack membership is important.

The data are available to be read into R as follows:

```
loc <- "http://www.stat.umn.edu/~sandy/courses/8053/Data/simwolf.txt"
wolfdata <- read.table(url(loc),header=TRUE)
```

The relevant variables are Year, the year of measurement, Pack, the number of the animal’s pack, and Y the indicator of presence of the virus. (Don’t forget that both Year and Pack are numeric variables so you will need to convert them to factors.)

Provide a justification for setting Year to be a fixed effect and Pack to be a random effect. Explain why Year should be a factor, not a continuous variable. Use GLMMs (and `lme4`) to estimate year and pack effects. Provide appropriate and useful numerical and graphical summaries.

3. Continuing with the last problem, suppose the data had consisted both of pups and of older animals. If a pup is observed and is infected, we know for sure is that it was infected in the year of measurement. If an age a animal is observed and is infected all we know is that it was infected *sometime in the last a years*, but

we do not know which year. We assume that once an animal is infected it stays infected forever. Suppose y_{ki} is the response of the i -th measured animal in the k -th pack, b_{ki} is the birth year of the animal, and m_{ki} is the capture year of the animal. We have that

$$y_{ki} = \sum_{t=b_{ki}}^{m_{ki}} z_{ki}(t)$$

where $z_{ki}(t)$ is the generally unobservable random variable that has value one if the animal is initially infected in year t and zero otherwise.

Using these definitions, write down a likelihood function for the data, and indicate how you would go about computing the estimates of year and pack effects. As far as I can figure out, you cannot use a standard program like lmer or SAS proc nlmixed to do this calculation.