Homework 2, Mixed effects models

Methods of Applied Statistics II

Due 30 Oct 2020

School leaver's data (20 marks)

Consider the school leaver's data obtained below.

```
sUrl = "http://www.bristol.ac.uk/cmm/media/migrated/jsp.zip"
dir.create(file.path("..", "data"), showWarnings = FALSE)
(Pmisc::downloadIfOld(sUrl, file.path("..", "data")))
```

```
Loading required namespace: R.utils
```

```
[1] "../data/jsp.zip" "../data/JSP.ASC" "../data/JSP.WS"  "../data/JSP.DAT"
```

the file `JSP.ASC` describes the variables in the dataset. The data can be read in as follows

```
school = read.fwf("../data/JSP.DAT", widths = c(2,
  1, 1, 1, 2, 4, 2, 2, 1), col.names = c("school",
  "class", "gender", "socialClass", "ravensTest",
  "student", "english", "math", "year"))
school$socialClass = factor(school$socialClass, labels = c("I",
  "II", "IIIn", "IIIm", "IV", "V", "longUnemp", "currUnemp",
  "absent"))
school$gender = factor(school$gender, labels = c("f",
  "m"))
school$classUnique = paste(school$school, school$class)
school$studentUnique = paste(school$school, school$class,
  school$student)
school$grade = factor(school$year)
```

Below is a Generalized Linear Mixed Model fit to the data

```
schoolLme = glmmTMB::glmmTMB(math ~ gender + socialClass +
  grade + (1 | school) + (1 | classUnique) + (1 |
  studentUnique), data = school)
summary(schoolLme)
```

and a historgram

```r
hist(40 - school$math, breaks = 100)
```

You are contracted to write a short report for a school board about which factors affect performance on math tests. Your client tells you that

- they want a Bayesian analysis, because they heard that Bayesian things are trendy at the moment
- they don't want a Normal distribution for the response, but the number of questions the students gets wrong look like a Poisson distribution.
- what are the most important influences on student performance on math tests? Social Class, or grade the student is in, or differences between schools or something else?
- … more specifically do the data suggest that identifying poorly performing schools and providing them extra funding would be worthwhile? Or would it be better to find individual class rooms which perform poorly and give the teachers in those classes extra training? Or do neither school or classroom have much of an influence on scores, it would be better to identify individual weak students and give them extra attention?

Write a self-contained report, roughly a page of writing plus figures and/or tables. Most of the document must be comprehensible to a school administrator with a basic knowledge of statistics, but a statistician will be reading the 'methods' section and want to know more details of your model and any prior distributions.

Smoking (20 marks)

This task concerns the 2014 American National Youth Tobacco Survey. On the pbrown.ca/teaching/astwo/data page there is an R version of the 2014 dataset `smoke2014.RData`, a pdf documentation file `NYTS2014-Codebook-p.pdf`, and the code used to create the R version of the data `smokingData2014.R`.

Cigarette smoking amongst children is known to be more common for males than females, in rural areas than urban areas, and to vary by age and ethnicity. It is likely that significant variation amongst the US states exists, and that there is variation from one school to the next.

The hypotheses to be investigated are:

1. Geographic variation (between states) in the rate of students smoking cigarettes is substantially greater than variation amongst schools. As a result, tobacco control programs should target the states with the most smoking and not concern themselves with finding particular schools where smoking is a problem.

2. Rural-urban differences are much greater than differences between states.

A secondary task is to convey the differences in the effect of age on smoking for white, Black, and Hispanic americans. The effect is expected to be different by sex and by rurality.
Use one or more figures or tables convey how age affects smoking for these groups.

The collaborating scientists have provided the following prior information

- The variability in the rate of smoking between states substantial, with some states having double or triple the rate of smoking update compared other states for comparable individuals. It's not expected to see the 'worst' states having five or 10 times the rate of the 'healthiest' states.

- Within a given state, the 'worst' schools are expected to have at most 50% greater rate than the 'healthiest' schools, and differences of 10% to 20% in smoking rates is more typical.
- When pressed on what is meant by 'worst' or 'unlikely', your collaborators suggest that 10th percentile or 10% probability are of the right order of magnitude.

Write a short consulting report addressing these hypotheses and the secondary problem. Some additional notes:

- Show graphs of prior and posterior densities of model parameters related to the research questions.
- Interpret your model parameters in the context of the smoking problem, transforming model parameters to a more 'natural' scale as necessary.
- It is important to state precisely what your prior distributions are (i.e. a Gamma(0.4, 3.1) distribution for the log of the intercept parameter), but also show how these distributions are consistent with the prior assumptions by showing quantiles or means or tail probabilities.
- You're given three confounders (sex, rural/urban, ethnicity), though the secondary research question is implying there should be interactions. This is a large dataset, if you can comfortably include all interactions then do so. You should consider including age as a categorical variable.
- You might want to fit more than one model, either as exploratory work or sensitivity assessments, but you should use a single 'best' model to answer the research questions. Fitting two models and selecting one of them with a fairly *ad hoc* explanation is fine, comparing 10 models without some sort of formal assessment (a topic we haven't covered) wouldn't be.

- do some data cleaning, for example the data from 9 and 10 year olds looks suspicious and could be removed

Appendix

Smoking

```
dataDir = "../data"
smokeFile = file.path(dataDir, "smoke2014.RData")
if (!file.exists(smokeFile)) {
  download.file("http://pbrown.ca/teaching/appliedstats/data/smoke2014.RData",
    smokeFile)
}
```

```
load(smokeFile)
smoke[1:3, c("Age", "ever_cigarettes", "Sex", "Race",
  "state", "school", "RuralUrban")]
```

```
  Age ever_cigarettes Sex    Race state       school RuralUrban
1  18            TRUE   F   white    AL mdr_00013045      Rural
2  18            TRUE   M pacific    AL mdr_00013045      Rural
3  16            TRUE   M   white    AL mdr_00013045      Rural
```

for some reason, I want to see if the effect of age on smoking is different depending on how harmful a student believes chewing tobacco is (in comparison to cigarettes).

```
forInla = smoke[smoke$Age > 10, c("Age", "ever_cigarettes",
  "Sex", "Race", "state", "school", "RuralUrban",
  "Harm_belief_of_chewing_to")]
forInla = na.omit(forInla)
forInla$y = as.numeric(forInla$ever_cigarettes)
forInla$ageFac = factor(as.numeric(as.character(forInla$Age)))
forInla$chewingHarm = factor(forInla$Harm_belief_of_chewing_to,
  levels = 1:4, labels = c("less", "equal", "more",
    "dunno"))
library("INLA")
toPredict = expand.grid(ageFac = levels(forInla$ageFac),
  RuralUrban = levels(forInla$RuralUrban), chewingHarm = levels(forInla$chewingHarm),
  Sex = levels(forInla$Sex))
forLincombs = do.call(inla.make.lincombs, as.data.frame(model.matrix(~Sex +
  ageFac * RuralUrban * chewingHarm, data = toPredict)))
fitS2 = inla(y ~ Sex + ageFac * RuralUrban * chewingHarm +
  f(state, model = "iid", hyper = list(prec = list(prior = "pc.prec",
    param = c(99, 0.05)))), data = forInla, family = "binomial",
  control.inla = list(strategy = "gaussian"), lincomb = forLincombs)
```

```
Warning in writeChar(lines[i], fp, nchars = nchar(lines[i]), eos = NULL):
problem writing to connection
```

```
rbind(fitS2$summary.fixed[, c("mean", "0.025quant",
  "0.975quant")], Pmisc::priorPostSd(fitS2)$summary[,
  c("mean", "0.025quant", "0.975quant")])
```

```
# create matrix of predicted probabilities
theCoef = exp(fitS2$summary.lincomb.derived[, c("0.5quant",
  "0.025quant", "0.975quant")])
theCoef = theCoef/(1 + theCoef)
# create an x axis, shift age by chewing harm group
toPredict$Age = as.numeric(as.character(toPredict$ageFac))
toPredict$shiftX = as.numeric(toPredict$chewingHarm)/10
toPredict$x = toPredict$Age + toPredict$shiftX
# only plot rural males
```

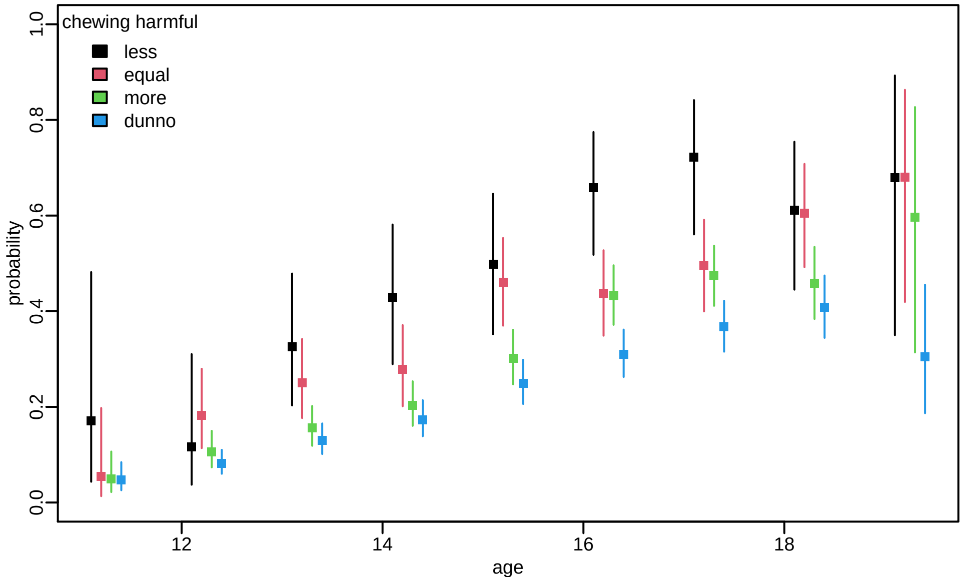

Figure 1: Rural males

```
toPlot = toPredict$Sex == "M" & toPredict$RuralUrban ==
  "Rural"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"],
  xlab = "age", ylab = "probability", ylim = c(0,
    1), pch = 15, col = toPredict[toPlot, "chewingHarm"])
segments(toPredict[toPlot, "x"], theCoef[toPlot, "0.025quant"],
  y1 = theCoef[toPlot, "0.975quant"], col = toPredict[toPlot,
    "chewingHarm"])
legend("topleft", fill = 1:nlevels(toPredict$chewingHarm),
  legend = levels(toPredict$chewingHarm), bty = "n",
  title = "chewing harmful")
```