Homework 3, Generalized linear mixed models

Methods of Applied Statistics II

Due 30 March 2016

1 Short answers

1.1 Non-parametrics (10 marks)

The `co2s` dataset in the `gamair` package contains carbon dioxide concentrations measured in Antarctica over time.

```
data("co2s", package = "gamair")
co2s$date = ISOdate(1957 + floor((co2s$c.month - 1)/12),
  co2s$month, 1, 0, 0, 0, tz = "UTC")
dim(co2s)
```

```
[1] 507   4
```

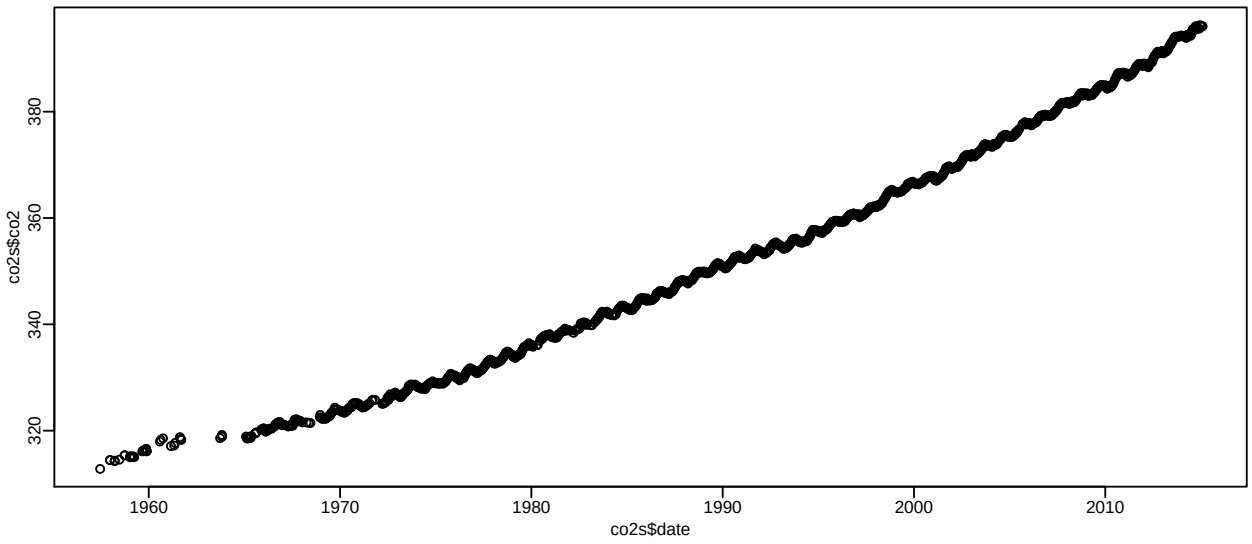
If that dataset seems too easy for you, try:

`http://scrippsco2.ucsd.edu/sites/default/files/data/`

`flask_co2_and_isotopic/daily_co2/fldav_spo.csv`

```
cUrl = paste("http://scrippsco2.ucsd.edu/sites/default/files",
  "/data/flask_co2_and_isotopic/daily_co2/fldav_spo.csv",
  sep = "")
cFile = basename(cUrl)
if (!file.exists(cFile)) download.file(cUrl, cFile)
co2s = read.table(cFile, header = FALSE, sep = ",", skip = 69,
  stringsAsFactors = FALSE)
co2s[co2s[, 6] > 0, 7] = NA
co2s = data.frame(date = strptime(co2s[, 1], format = "%Y-%m-%d",
  tz = "UTC"), co2 = co2s[, 7])
plot(co2s$date, co2s$co2)
```

Table 1: Figure: CO2 in Antarctica



a)

Write a short report addressing the following hypotheses:

- Although carbon in the atmosphere is still increasing, there are indications that the increase has slowed somewhat recently.
- The data are consistent with carbon slowing during the global economic recessions around 1980-1982 and 1990 (hint `?predict.gam` shows how to predict derivatives)
- Carbon tends to be higher in October than March.
- Carbon will likely exceed 400 parts per gallon by 2020

You should

- explain fully the model you are using and why you have chosen to use it
- make your graphs look nice

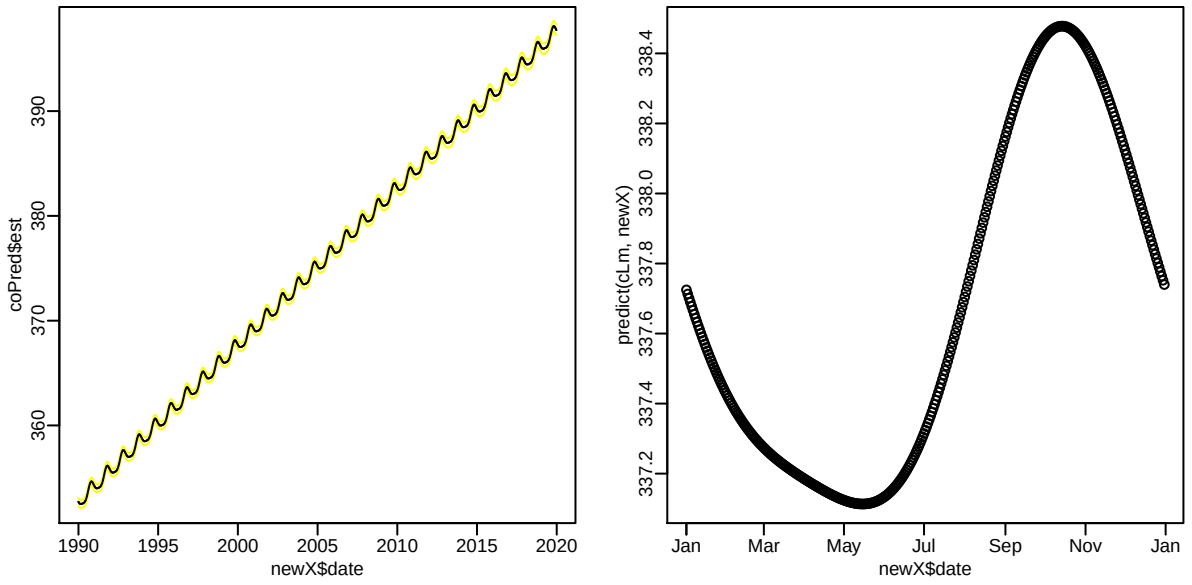
The code below might prove useful.

```
timeOrigin = ISOdate(1980, 1, 1, 0, 0, 0, tz = "UTC")
co2s$days = as.numeric(difftime(co2s$date, timeOrigin, units = "days"))
co2s$cos12 = cos(2 * pi * co2s$days/365.25)
co2s$sin12 = sin(2 * pi * co2s$days/365.25)
co2s$cos6 = cos(2 * 2 * pi * co2s$days/365.25)
co2s$sin6 = sin(2 * 2 * pi * co2s$days/365.25)
cLm = lm(co2 ~ days + cos12 + sin12 + cos6 + sin6, data = co2s)
summary(cLm)$coef[, 1:2]
```

```
                  Estimate   Std. Error
(Intercept) 337.671940928 9.997216e-02
days          0.004102301 1.513206e-05
cos12         0.202690509 1.183445e-01
sin12        -0.633093983 1.167055e-01
cos6         -0.143796734 1.180639e-01
sin6         -0.032543226 1.169413e-01
```

```
newX = data.frame(date = seq(ISOdate(1990, 1, 1, 0, 0, 0,
  tz = "UTC"), by = "1 days", length.out = 365 * 30))
newX$days = as.numeric(difftime(newX$date, timeOrigin, units = "days"))
newX$cos12 = cos(2 * pi * newX$days/365.25)
newX$sin12 = sin(2 * pi * newX$days/365.25)
newX$cos6 = cos(2 * 2 * pi * newX$days/365.25)
newX$sin6 = sin(2 * 2 * pi * newX$days/365.25)
coPred = predict(cLm, newX, se.fit = TRUE)
coPred = data.frame(est = coPred$fit, lower = coPred$fit -
  2 * coPred$se.fit, upper = coPred$fit + 2 * coPred$se.fit)
plot(newX$date, coPred$est, type = "l")
matlines(as.numeric(newX$date), coPred[, c("lower", "upper",
  "est")], lty = 1, col = c("yellow", "yellow", "black"))
newX = newX[1:365, ]
newX$days = 0
plot(newX$date, predict(cLm, newX))
```

Table 2: Figure: Results



a) forecast b) cycle

1.2 Math (5 marks)

```
data("MathAchieve", package = "MEMSS")
head(MathAchieve)
```

```
  School Minority    Sex    SES MathAch MEANSES
1   1224       No Female -1.528   5.876  -0.428
2   1224       No Female -0.588  19.708  -0.428
3   1224       No   Male -0.528  20.349  -0.428
4   1224       No   Male -0.668   8.781  -0.428
5   1224       No   Male -0.158  17.898  -0.428
6   1224       No   Male  0.022   4.583  -0.428
```

From Maindonald and Braun, ch 10 q 5. In the data set `MathAchieve` (`MEMSS` package), the factors `Minority` (levels `yes` and `no`), and the variable `SES` (socio-economic status) are clearly fixed effects. Carry out an analysis that treats `School` as a random effect. The data appear left skewed, is it plausible to assume the data are Normal? Are differences between schools greater than can be explained by within-school variation?
Explain your choice of model and the distribution you've used for math scores, as well as prior distributions (if you choose to be Bayesian)

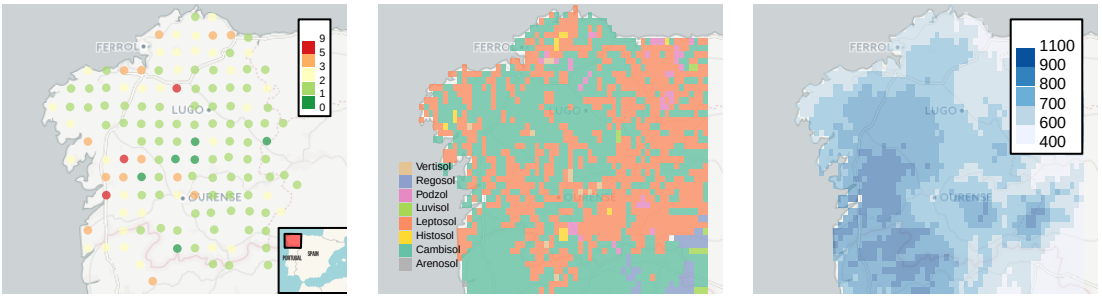
Note: you have been reliably informed by an un-named source that the negative math scores are contaminated data and you are justified in removing them.

Hint: find a 95% interval for test scores of students in the baseline category in a 'typical' school (with random effect of zero).
Compare this to a 95% interval for the school average test scores (on the natural scale, not log scale).

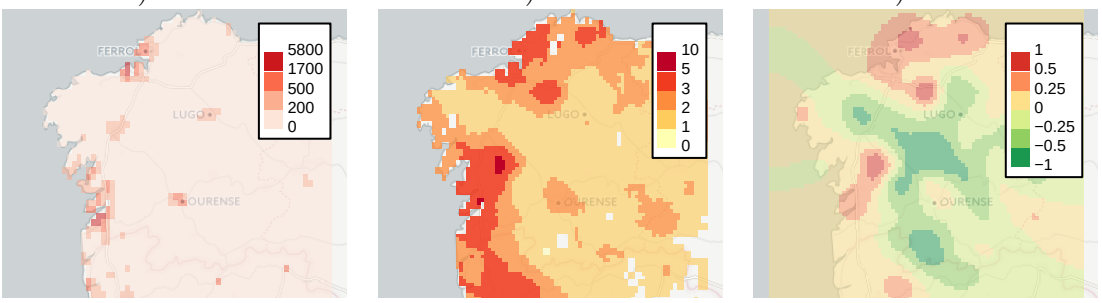
1.3 Moss in Galicia Redux (5 marks)

The Figure below shows (in Subfigure a) the locations of measurements of lead levels taken from moss growing in or near the province of Galicia in Spain. Code for downloading and displaying the data are in the Appendix, though you won't need to run code yourself for this question. Subfigures b, c, and d show dominant soil types, average annual rainfall, and population density for the region concerned.

Table 3: Figure: Lead in Galicia, Spain



a) Lead b) soil c) rain



d) population e) fitted lead f) random effect

A Generalized Linear Geostatistical Model has been fit to the data as follows.

```
library(geostatsp)
covariates$logPop = log(covariates$pop)
mossRes = glgm(lead ~ logPop + rain + soil, grid = extend(squareRaster(moss,
  70), 10), data = moss, covariates = covariates, family = "gamma",
  priorCI = list(range = c(20000, 1e+05), sd = c(u = 1,
    alpha = 0.05)), control.family = list(prior = "normal",
    param = c(1, 1^(-2))), control.mode = list(theta = c(2,
    2, 3), restart = TRUE))
```

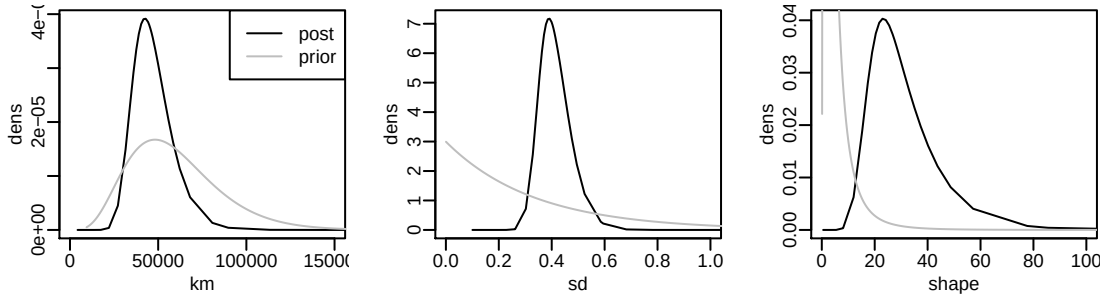
This produces the following parameter estimates. Spatial predictions of lead content in moss and of the residual spatial effects appear in Subfigures e and f above.

```
knitr::kable(mossRes$parameters$summary[, c("mean", "0.025quant",
  "0.975quant")], digits = 4)
```

	mean	0.025quant	0.975quant
(Intercept)	-0.2982	-1.2697	0.6609
logPop	0.0879	0.0113	0.1639
rain	0.0010	-0.0001	0.0022
soilLeptosol	-0.0595	-0.1876	0.0691
soilPodzol	0.1107	-0.2616	0.4908
soilVertisol	-0.2588	-0.7092	0.2008

	mean	0.025quant	0.975quant
range	47124.8523	29417.8226	73636.8211
gammaShape	31.1779	13.6654	65.7765
sd	0.4135	0.3159	0.5515

Table 5: Figure: Model parameters



a) range b) sd c) gamma shape

1. Write down a model corresponding to that contained in the `mossRes` object, explaining the terms in this model. Refering to Figure 5, give an approximate 95% prior interval for each of the three parameters shown (no need to calculate a precise value).
2. Does it appear that the environmental variables (soil type and rain) and human population influence the lead content of moss in Galicia? Justify your answer.
3. Make a figure or figures showing the Gamma and transformed-Normal density functions which this model and the model from Homework 2 ascribe to the lead content. Which model gives a distribution closer to the empirical distribution of lead in Galician moss?
4. Write a paragraph describing contrasting this model to that from Homework 2, and explain which model you would prefer to use in a research paper examining spatial variation in Galician lead.

Moss data:

```
sUrl = "http://www.lancaster.ac.uk/staff/diggle/APTS-data-sets/lead2000_data.txt"
sFile = file.path("..", "data", basename(sUrl))
if (!file.exists(sFile)) {
  download.file(sUrl, sFile)
}
x = read.table(sFile, header = TRUE, skip = 3)
hist(x[, "z"])
```

A hint

```
xNorm = rnorm(1e+05, mean = 0.34 + log(500) * 0.06, sd = 0.2)
xBc = (xNorm * (-0.4) + 1)^(-1/0.4)
hist(xBc[xBc < 10], breaks = 50)
```

2 Application (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 `smoke.RData`, a pdf documentation file `2014-Codebook.pdf`, and the code used to create the R version of the data `smokingData.R`.

The age at which children first try cigarette smoking is known to be earlier for males than females, earlier in rural areas than urban areas, and to vary by 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 mean age children first try cigarettes is substantially greater than variation amongst schools. As a result, tobacco control programs should target the states with the earliest smoking ages and not concern themselves with finding particular schools where smoking is a problem.

2. First cigarette smoking has a flat hazard function, or in other words is a first order Markov process. This means two non-smoking children have the same probability of trying cigarettes within the next month, irrespective of their ages but provided the known confounders (sex, rural/urban, etnicity) and random effects (school and state) are identical.

A secondary task is to convey the differences between white urban males and white rural males in their smoking uptake habits. Use one or more figures or tables to convey what should be expected if first smoking ages of a large number of white urban or rural males were obtained.

The collaborating scientists have provided the following prior information

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

- Within a given state, the 'worst' schools are expected to have at most 50% greater hazards than the 'healthiest' schools, and differences of 10% to 20% in hazards is more typical.
- Although a flat hazard function is expected, it's more likely that the hazard increases with age than decreases with age. The prior probability the hazard falls with age is less than 10%. It wouldn't be unusual to see a quadratic or cubic increase in the hazard with age, but polynomial increases with age involving 5th or 6th powers is improbable.
- 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, using the same criteria as Homework 1. 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), it's up to you if you'd like to include interactions. If you are able to comfortably fit a model with a large number of interactions, and the model parameters appear to be reasonably well identified, that would be a 'conservative' approach. Here 'conservative' means 'if in doubt, conclude it's all down to the covariates and not random effects or time-varying hazard'.

- 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.