Homework 2, Linear mixed models

Methods of Applied Statistics II

Due 28 Feb 2018 at the beginning of the lecture

Covariance (5 marks)

Consider the following model where Y_{ijk} is a observation made at farm i on animal j at time t_{ijk}. The animal's weight at time t_{ijk} is denoted W_{ijk} and X_{ijk} is a vector of additional covariates. The random effects A, U and V are all Normally distributed.

$$
\begin{aligned}
Y_{ijk}|U, V, A \overset{\text{ind}}{\sim} & \text{N}(\mu_{ijk}, \tau^2) \\
\mu_{ijk} =& X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk} \\
\text{cov}(A_{ijk}, A_{\ell mn}) =& \begin{cases} 0 & i \neq \ell \text{ or } j \neq m \\ \sigma_A^2 \exp(-|t_{ijk} - t_{\ell mn}|/\phi) & i = \ell \text{ and } j = m \end{cases} \\
\begin{pmatrix} V_{ij1} \\ V_{ij2} \end{pmatrix} \sim & \text{MVN}(0, \Gamma) \\
U_i \overset{\text{ind}}{\sim} & \text{N}(0, \sigma_U^2)
\end{aligned}
$$

1. Were we to estimate these model parameters using Maximum Likelihood Estimation, which model parameters would need to be estimated using a numerical optimizer and which parameters have closed-form expressions for their MLE's given this first set of parameters.
2. Derive expressions (and show your work) for the following.
   - $\text{cov}(Y_{ijk}, A_{ijn})$, $n \neq k$
   - $\text{cov}(Y_{ijk}, V_{ij1})$
   - $\text{cov}(Y_{ijk}, Y_{imn})$, $(m, n) \neq (j, k)$
   - $\text{cov}(Y_{ijk}, Y_{ijn})$, $n \neq k$
   - $\text{cov}(Y_{ijk}, Y_{\ell jk})$, $\ell \neq i$
   - $\text{var}(Y_{ijk}|A, V)$
   - $\text{E}[\exp(Y_{ijk})]$

3. Suppose you obtained the following parameter estimates: $\hat{\tau}^2 = 2$, $\hat{\sigma}_U^2 = 0.001$, $\hat{\phi} = 4$, $\hat{\sigma}_A^2 = 0.00002$, and
$$\hat{\Gamma} = \begin{pmatrix} 1 & 0.05 \\ 0.05 & 0.002 \end{pmatrix}$$
Further, suppose W_{ijk} is weight measured in ounces, and a typical animal in this study weights between 1000 and 3000 ounces. Write down a model which would be more suitable for these data than the model above, where only the random effects you feel are important are included.

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. Discuss how the decision whether to treat `School` as a fixed or a as a random effect might depend on the purpose of the study? Carry out an analysis that treats `School` as a random effect. Are differences between schools greater than can be explained by within-school variation?

Cystic Fybrosis data (15 marks)

You have been asked to help a medical scientist interpret their findings regarding the effect of the F508 gene on the decline in lung function in individuals with cystic fibrosis over time. The data set below can be obtained from pbrown.ca/teaching/astwo/data/CF.RData and was produced using code in the Appendix. The variables below are as follows

- ID: subject identifier
- FEV1: lung function
- AGE: age of the subject when the lung function measurement was taken
- GENDER: male or female
- PSEUDOA: infection with Pseudo Aeruginosa, and established confounder variable
- F508: genotype homozygous, heterozygous or none
- PANCREAT: don't use.

```
head(x)
```

```
      ID   FEV1    AGE GENDER PSEUDOA       F508 PANCREAT
1 100073 113.80  8.452 female     yes homozygous      yes
2 100073  98.18  8.783 female     yes homozygous      yes
3 100073  98.73  9.785 female     yes homozygous      yes
4 100073 101.79 10.538 female     yes homozygous      yes
5 100073  98.04 12.329 female     yes homozygous      yes
6 100073  94.32 13.306 female     yes homozygous      yes
```

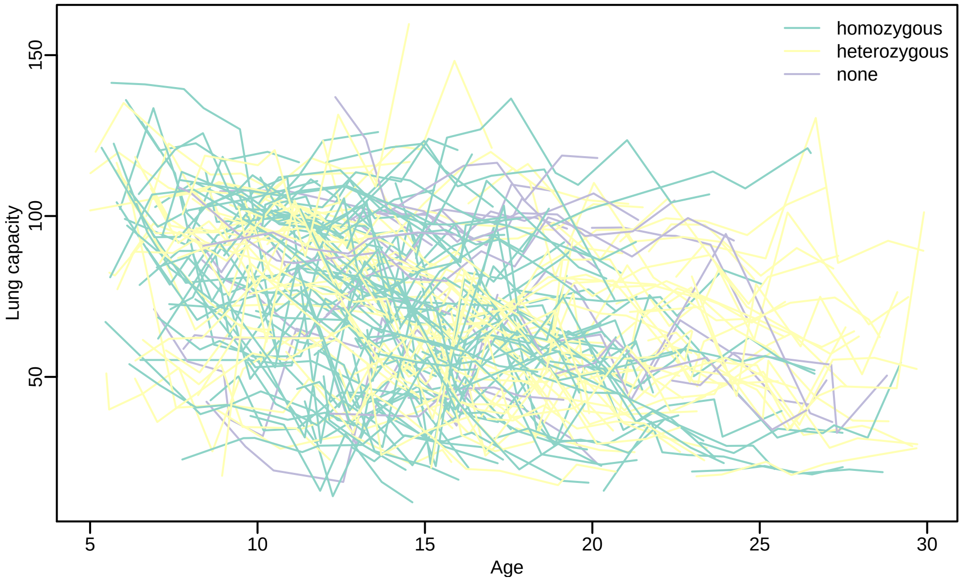


Figure 1: Lung function over time

The research hypotheses are:

1. the rate at which lung function declines for CF patients depends on the F508 gene; and
2. the effect of the F508 gene on lung function decline differs for females and males.

The medical scientist arrives with the R output below and is pleased that these hypotheses appear to be supported. However, they recall being told that random slope models (where the age coefficient is different for each subject) and models with serial correlation are worth considering with data of this type and would like confirmation from a Statistician that the results do indeed provide evidence in support of their hypotheses.

```
library("nlme")
x$ageC = x$AGE - 18
resS = lme(FEV1 ~ GENDER * F508 * ageC + PSEUDOA, random = ~1 |
  ID, data = x)
knitr::kable(summary(resS)$tTable, digits = 3)
```

	Value	Std.Error	DF	t-value	p-value
(Intercept)	66.877	3.749	1306	17.840	0.000
GENDERfemale	-2.203	4.996	194	-0.441	0.660
F508heterozygous	6.269	5.001	194	1.254	0.212
F508none	7.157	7.348	194	0.974	0.331
ageC	-1.754	0.251	1306	-6.989	0.000
PSEUDOAyes	-2.159	1.059	1306	-2.038	0.042
GENDERfemale:F508heterozygous	-6.275	7.031	194	-0.892	0.373
GENDERfemale:F508none	2.014	11.032	194	0.183	0.855
GENDERfemale:ageC	0.071	0.352	1306	0.200	0.841
F508heterozygous:ageC	0.744	0.344	1306	2.165	0.031
F508none:ageC	1.610	0.534	1306	3.014	0.003
GENDERfemale:F508heterozygous:ageC	-0.998	0.495	1306	-2.018	0.044
GENDERfemale:F508none:ageC	-1.889	0.810	1306	-2.333	0.020

Prepare a short report (roughly 2 pages of writing) accessible to a numerate clinical scientist.

- Write down (in mathematical notation, not R code) the model fit above, as well as a random slope model and a model with serial correlation which could be considered for use with these data. Explain the terms in the models.
- Outline the differences between the models, and identify which model or models are making the strongest assumptions and what these assumptions are.
- Show results from fitting a random slope model and a serially correlated model to this dataset. Explain important features of the results and how they differ from the initial model.
- Reach a formal conclusion about the research hypotheses using whichever model or models you feel are most appropriate. Explain why your conclusions are different from (or similar to) the medical scientist's initial conclusions.

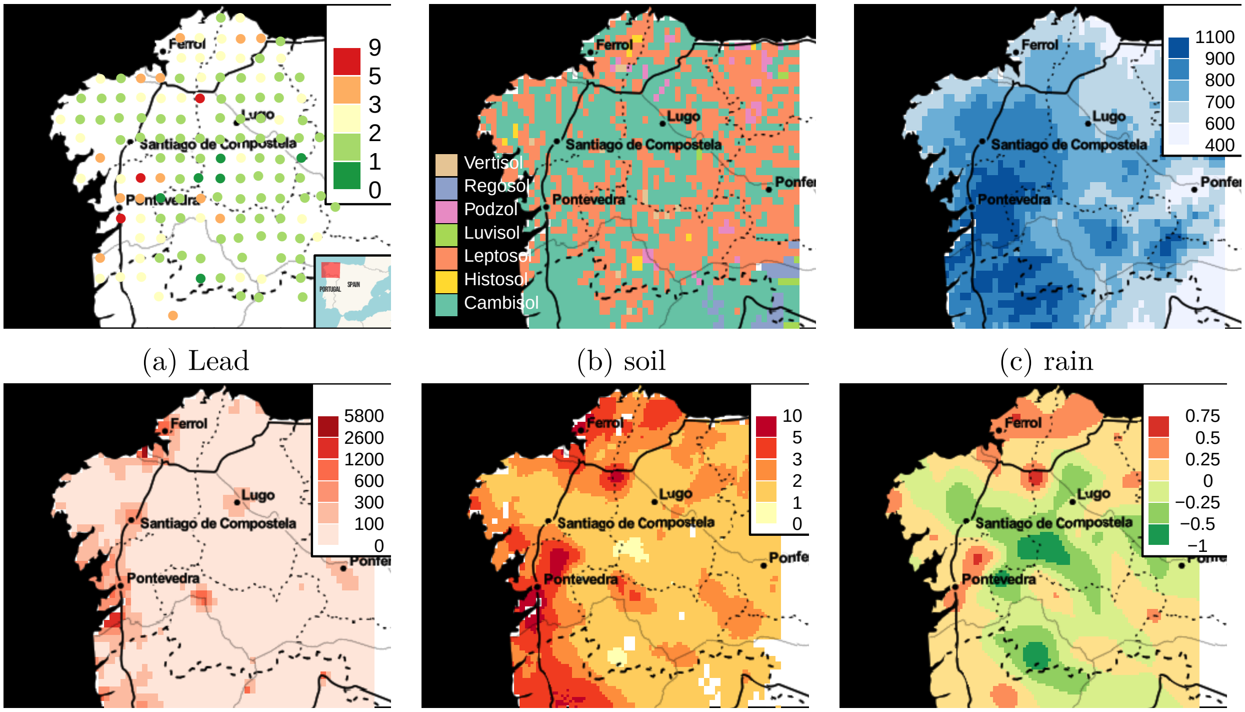
Write in prose and make a coherent report (rather than homework-style short answers) although use section headings or itemized lists as you see fit.

Moss in Galicia (15 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.

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

```r
library(geostatsp)
covariates$logPop = log(covariates$pop)
```



(a) Lead (b) soil (c) rain

(d) population (e) fitted lead (f) random effect

Figure 2: Lead in Galicia, Spain. Background ©Stamen Design

```
mossRes = lgm(lead ~ logPop + rain + soil, grid = extend(squareRaster(moss,
  100), 10), data = moss, covariates = covariates,
  boxcox = 0, fixBoxcox = FALSE)
```

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$summary[, c("estimate", "ci0.025",
  "ci0.975", "Estimated")], digits = 2)
```

	estimate	ci0.025	ci0.975	Estimated
(Intercept)	-0.14	-2.25	1.97	TRUE
logPop	0.06	-0.11	0.22	TRUE
rain	0.00	0.00	0.00	TRUE
soilCambisol	0.00	NA	NA	FALSE
soilLeptosol	-0.04	-0.31	0.23	TRUE
soilPodzol	0.02	-0.76	0.80	TRUE
soilVertisol	-0.22	-1.14	0.69	TRUE
sdNugget	0.00	NA	NA	TRUE
sdSpatial	0.29	NA	NA	TRUE
range/1000	34.22	NA	NA	TRUE
shape	1.00	NA	NA	FALSE
anisoRatio	1.00	NA	NA	FALSE

	estimate	ci0.025	ci0.975	Estimated
anisoAngleRadians	0.00	NA	NA	FALSE
anisoAngleDegrees	0.00	NA	NA	FALSE
boxcox	-0.52	NA	NA	TRUE

Below is some code, the results from which may or may not prove interesting or useful.

```
mossResFull = lgm(lead ~ logPop + rain + soil, grid = 100,
  data = moss, covariates = covariates, reml = FALSE,
  boxcox = 0, fixBoxcox = FALSE)
mossResSub = lgm(lead ~ logPop, grid = 100, data = moss,
  covariates = covariates, reml = FALSE, boxcox = 0,
  fixBoxcox = FALSE)
mossResNoCov = lgm(lead ~ 1, grid = 100, data = moss,
  covariates = covariates, reml = FALSE, boxcox = 0,
  fixBoxcox = FALSE)
```

```
lmtest::lrtest(mossResFull, mossResSub, mossResNoCov)
```

```
Likelihood ratio test

Model 1: mossResFull
Model 2: mossResSub
Model 3: mossResNoCov
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  10 -122.50
2   6 -126.49 -4 7.9917    0.09188 .
3   5 -130.34 -1 7.6974    0.00553 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
knitr::kable(mossResSub$summary[, c("estimate", "ci0.025",
  "ci0.975", "Estimated")], digits = 2)
```

	estimate	ci0.025	ci0.975	Estimated
(Intercept)	0.34	-0.37	1.04	TRUE
logPop	0.06	-0.11	0.23	TRUE
sdNugget	0.04	0.00	65.22	TRUE
sdSpatial	0.25	0.18	0.35	TRUE
range/1000	30.91	17.67	54.09	TRUE
shape	1.00	NA	NA	FALSE
anisoRatio	1.00	NA	NA	FALSE
anisoAngleRadians	0.00	NA	NA	FALSE
anisoAngleDegrees	0.00	NA	NA	FALSE
boxcox	-0.62	-0.93	-0.31	TRUE

1. Write down a model corresponding to that contained in the `mossRes` object, explaining the terms in this model.
2. Does it appear that the environmental variables (soil type and rain) influence the lead content of moss in Galicia? Justify your answer, if possible quote a p-value or probability and explain what it means.
3. Does it appear that humans have an influence on the lead content of moss in Galicia? Justify your answer, if possible quote a p-value or probability and explain what it means.
4. Write a paragraph describing the second order statistical properties of lead in Galicia, using the estimated model parameters when appropriate.

Appendix

CF data

```
cUrl = paste("https://faculty.washington.edu/heagerty/Courses/VA-longitudinal/private/",
  c("NewCFkids.data", "NewCFkids.txt"), sep = "")
cFile = file.path("..", "data", basename(cUrl))
for (D in 1:length(cFile)) {
  if (!file.exists(cFile[D])) {
    download.file(cUrl[D], cFile[D])
  }
}
x = read.table(cFile[1], header = FALSE)
header = scan(cFile[2], skip = 11, n = ncol(x), what = "a",
  sep = "\n")
header = matrix(trimws(unlist(strsplit(header, " = "))),
  ncol = 2, byrow = TRUE)
colnames(x) = header[, 1]
factors = header[grep("=", header[, 2]), ]
for (D in 1:nrow(factors)) {
  levels = unlist(strsplit(gsub("^.*\\(|)", "", factors[D,
    2]), ","))
  noEq = grep("=", levels, invert = TRUE)
  levels[noEq] = paste(levels[noEq], "=never")
  levels = matrix(unlist(strsplit(levels, "=")),
    ncol = 2, byrow = TRUE)
  x[, factors[D, 1]] = factor(x[, factors[D, 1]],
    levels = as.numeric(levels[, 1]), labels = trimws(levels[,
      2]))
}
x$ID = factor(x$ID)
save(x, file = "../data/CF.Rdata")
```

```r
Scol = rep_len(RColorBrewer::brewer.pal(12, "Set3"),
  nlevels(x$F508))
names(Scol) = levels(x$F508)
plot(x$AGE, x$FEV1, type = "n", xlab = "Age", ylab = "Lung capacity")
junk = by(x, x$ID, function(qq) {
  lines(qq$AGE, qq$FEV1, col = Scol[as.character(qq$F508)])
})
legend("topright", lty = 1, col = Scol, legend = names(Scol),
  bty = "n")
```

Moss data

```r
library("rgdal")
library("mapmisc")
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)
library("rgdal")
moss = SpatialPointsDataFrame(coords = as.matrix(x[,
  c("x", "y")]), data = data.frame(lead = x[, "z"]),
  proj4string = CRS("+init=epsg:25829"))
mossLL = spTransform(moss, mapmisc::crsLL)
```

```r
library(raster)
gzUrl = c(soil = "http://worldgrids.org/lib/exe/fetch.php/stghws1a.tif.gz",
  rain = "http://worldgrids.org/lib/exe/fetch.php?media=pregsm1a.tif.gz",
  pop = "http://worldgrids.org/lib/exe/fetch.php?media=pdmgpw1a.tif.gz")
gzFile = file.path("/store/patrick/spatialData", basename(gzUrl))
tFile = rep(NA, length(gzUrl))
names(tFile) = names(gzFile) = names(gzUrl)
covariates = list()
mossExtent = extend(extent(mossLL), 0.4)
sBorder = raster::getData("GADM", country = "ESP",
  level = 1)
pBorder = raster::getData("GADM", country = "PRT",
  level = 1)
border = bind(sBorder, pBorder)
for (D in names(gzFile)) {
  if (!file.exists(gzFile[D]))
    download.file(gzUrl[D], gzFile[D])
  tFile[D] = R.utils::gunzip(gzFile[D], remove = FALSE,
    skip = TRUE)
  covariates[[D]] = crop(raster(tFile[D]), mossExtent)
  covariates[[D]] = mask(covariates[[D]], border)
}
lUrl = "http://worldgrids.org/lib/exe/fetch.php?media=stghws1.txt"
lFile = file.path("/store/patrick/spatialData", basename(lUrl))
if (!file.exists(lFile)) download.file(lUrl, lFile)

soilLevels = read.table(lFile, header = TRUE)
soilLevels$ID = soilLevels$MINIMUM
# sort out some ID values matching to MAXIMUM
# column
uSoil = unique(covariates[["soil"]])
matchToMax = which(soilLevels$MAXIMUM %in% uSoil &
  !(soilLevels$MINIMUM %in% uSoil) & !soilLevels$MAXIMUM %in%
  soilLevels$MINIMUM)
soilLevels[matchToMax, "ID"] = soilLevels[matchToMax,
  "MAXIMUM"]
soilTable = colourScale(x = covariates[["soil"]], labels = soilLevels,
  style = "unique", breaks = 7, col = "Set2")
covariates[["soil"]]@legend@colortable = soilTable$colortable
levels(covariates[["soil"]]) = list(soilTable$levels)

library("mapmisc")
bg = tonerToTrans(openmap(moss, buffer = 50000, path = "stamen-toner"))
bgLL = tonerToTrans(openmap(mossLL, buffer = 50000,
  path = "stamen-toner", crs = crsLL))

mCol = colourScale(moss$lead, col = "RdYlGn", breaks = 6,
  style = "equal", transform = "log", dec = 0, rev = TRUE)
map.new(moss, buffer = 10000)
plot(bg, add = TRUE)
legendBreaks("topright", mCol, cex = 1.1, inset = 0)
insetMap(moss, cropInset = 1000 * c(100, 800, 800,
  100), zoom = 4, map = "waze")
plot(moss, add = TRUE, col = mCol$plot, pch = 16)
map.new(mossLL, buffer = 0.1)
plot(covariates[["soil"]], add = TRUE)
plot(bgLL, add = TRUE)
legendBreaks("bottomleft", covariates[["soil"]], bty = "n",
  text.col = "white", inset = 0, cex = 0.9)
colRain = colourScale(covariates[["rain"]], opacity = 1,
  breaks = 6, col = "Blues", style = "equal", dec = -2)
par(cex = 0.6)
map.new(mossLL, buffer = 0.1)
plot(covariates[["rain"]], add = TRUE, legend = FALSE,
  breaks = colRain$breaks, col = colRain$colOpacity)
plot(bgLL, add = TRUE)
legendBreaks("topright", colRain, cex = 0.5, inset = 0)
colPop = colourScale(covariates[["pop"]], breaks = 12,
  col = "Reds", style = "equal", dec = -2, transform = "log")
map.new(mossLL, buffer = 0.1)
plot(covariates[["pop"]], add = TRUE, legend = FALSE,
  breaks = colPop$breaks, col = colPop$colOpacity)
plot(bgLL, add = TRUE)
legendBreaks("topright", colPop, cex = 0.5, inset = 0)
colPred = colourScale(mossRes$predict[["predict"]],
  breaks = c(0, 1, 2, 3, 5, 10), col = "YlOrRd",
  style = "fixed", dec = 0, inset = 0)
map.new(moss, buffer = 10000)
plot(mossRes$predict[["predict"]], add = TRUE, legend = FALSE,
  breaks = colPred$breaks, col = colPred$colOpacity)
plot(bg, add = TRUE)
legendBreaks("topright", colPred, cex = 0.5, inset = 0)
colR = colourScale(mossRes$predict[["random"]], breaks = 8,
  col = "RdYlGn", rev = TRUE, style = "equal", dec = -log10(0.25))
map.new(moss, buffer = 10000)
plot(mossRes$predict[["random"]], add = TRUE, legend = FALSE,
  breaks = colR$breaks, col = colR$colOpacity)
plot(bg, add = TRUE)
legendBreaks("topright", colR, cex = 0.5, inset = 0)
```