STAA 554 Homework 5

Contents

Libraries you may want:

```r
library(dplyr)
library(ggplot2)
library(tidyr)
library(lme4)
library(pbkrtest)
library(RLRsim)
library(lmerTest)
```

Q1 Earthquakes

The attenu.csv data gives peak accelerations measured at various observation stations for 23 earthquakes in California. The data has been used by various workers to estimate the attenuating effect of distance on ground acceleration.

(a) 1pt Plot

Plot lines showing how the acceleration changes with distance for each quake. Make transformations of both axes so that the relationship is easier to see and replot.

(b) 2pt Mixed Model

Fit a mixed effects model with the transformed variables which takes account of both events and stations as random effects; include magnitude as a fixed effect. Express the effect of magnitude on the acceleration.

(c) 2pt Quadratic Term

Does adding a quadratic term in distance improve the model?

(d) 2pt Station variance component

Can we remove the station variation term? *(hint: attenu2 <-na.omit(attenu) may be useful if you use REML methods for comparisons, recall methods based on likelihoods require the same data for all models)*

(e) 2pts Prediction

Using the model with a quadratic term and random effects for both station and event, as in part c: For a new magnitude 6 quake, predict the acceleration for up to a distance of 200 miles. Make a plot of the data and show your predicted curve on top of the data in a different color.

(f) 3pts Prediction for observation 1

Predict how the acceleration varied for the first event where only one observation was available. Show the predicted acceleration up to 200 miles in a plot. Add the actual observation to the plot. (just point predictions, no intervals needed)

Q2 Rat Drink Data

The ratdrink.csv data consist of five weekly measurements of body weight for 27 rats. The first 10 rats are on a control treatment while 7 rats have thyroxine added to their drinking water. Ten rats have thiouracil added to their water.

(a) 1pt Plots

Plot the data showing how weight increases with age on a single panel, taking care to distinguish the three treatment groups. Now create a three-panel plot, one for each group. Discuss what can be seen.

(b) 3pt Mixed Model Interpretation

Fit a linear longitudinal model that allows for a random slope and intercept for each rat. Each group should have a different mean line. Give interpretation for the following estimates:

i. The fixed effect intercept term.
ii. The interaction between thiouracil and week.
iii. The intercept random effect SD.

(c) 2pt Test

Check whether there is a statistically significant treatment effect.

(d) 2pt Diagnostic

Construct diagnostic plots showing the residuals against the fitted values and a QQ plot of the residuals. Interpret.

(e) 2pt Confidence Intervals

Construct confidence intervals for the parameters of the model.

i. Which random effect terms may not be significant?
ii. Is the thyroxine group significantly different from the control group?

(f) 2pt Covariance structure

Fit this same model from (b) using lme() and extract the marginal covariance matrix for observations from a particular rat. Describe the observed structure and comment on if it makes sense in this context.

(g) 3pt Covariance Structure

Compare the covariance matrix from (f) to the covariance matrix in each of the following models:

i. fit the same model in (b), but without the random slope and assume a compound symmetric matrix. Does the compound symmetric assumption make sense in this context?

ii. fit the same model in (b), but without the random slope and assume an unstructured covariance matrix. Describe any general trends in the structure.

iii. fit the same model in (b), but without the random slope and assume an autoregressive 1 structure in the covariance matrix. Why might we consider this structure in this context?

i.

ii.

iii.

(h) 2pt Compare using information criteria

Compare the 4 models from f and g using AIC and BIC. Which model appears best?

Q3

The National Youth Survey collected a sample of 11–17 year olds, 117 boys and 120 girls, asking questions about marijuana usage. The data is presented in potuse.csv.

Potuse levels: 1: “non-user” 2: “light” 3: “Heavy”

Sex: 1: male. 2: female

Each row represents a possible profile of pot usage over the years 1976 - 1980. For example, row 1 shows the number of males (sex= 1) who never used pot: 48 Row 2 show the number of males who never used pot except lightly in year 1980: 8.

```
  sex year.76 year.77 year.78 year.79 year.80 count
1   1       1       1       1       1       1    48
2   1       1       1       1       1       2     8
3   1       1       1       1       1       3     4
```

(a) 1pt Plot

Plot the total number of people falling into each usage category as it varies over time separately for each sex.

(b) 2pt Format and fit model

Condense the levels of the response into whether the person did or did not use marijuana that year. Turn the year into a numerical variable. Fit a GLMM for the now binary response with an interaction between sex and year as a predictor using Gauss-Hermite quadrature. Comment on the effect of sex.

Hint: The idea of this problem is to fit a GLMM modeling a binary response (i.e. , will a person with these characteristics be likely to use pot or not) to the data. The tricky part is that the potuse.csv data comes in the form of count data, not individual data, so you have to ‘wrangle’ it until it represents individuals. Consider the following functions:

```
tidyr::gather() to make the data long instead of wide (i.e., to start to turn the years into a numeric
tidyr::pivot_longer(), newer than gather(), possibly more intuitive?
dplyr::uncount() function to turn the count data into individual data... very cool little function.
```

(c) 2pt Test

Fit a reduced model without sex and use it to test for the significance of sex in the larger model.

(d) 3pt Linearity

Fit a model with year as a factor. (For simplicity, No sex term in the model.)

i. Should this model be preferred to the model with year as just a linear term?

ii. Interpret the estimated effects of the year in the factor version of the model.