36-617: Applied Linear Models
Fall 2022
HW08 – Due Wed Nov 16, 11:59pm

- Please turn the homework in, as a single pdf, online in GradeScope using the link provided on the Assignment page on canvas.cmu.edu.
- This hw heavily depends on material in lectures 21 and 22, which I will give "just in time" on Mon Nov 14 and Wed Nov 16.
  - In order that you can start earlier than that (after all, it's due Wed Nov 16!), I have already put the lecture notes for lectures 21 and 22 in the week11 folder in the files area on canvas.
  - Also look closely at the R files, "`21 - mlm residuals.r`" and "`22 - estimation and model selection.r`"
- Weekly Reading & Quiz:
  - There is no new reading this week. Everything is in the lecture notes.
  - There is no quiz on Monday.
- Information about the final project will be available sometime on Fri Nov 11.
- There are four main exercises below...

Exercises

1. The file allvars.csv contains CD4 percentages (CD4PCT) for a set of 254 young children with HIV who were measured several times over a period of two years (A CD4 count measures the number of CD4 cells in your blood. It's used to check the immune system function in people with HIV[1]). The dataset also includes the ages of the children at each measurement. This is an example of *growth curve data* in which the Level 1 observations are CD4PCT's at each visit, and the Level 2 "group" is a child. How the CD4PCT's change over time for each child is the primary question of interest. Because of skewing, you should replace CD4PCT with sqrt(CD4PCT) for all of the following questions.

   (a) For the first 12 children only (`newpid` $\leq$ `12`), make a facet plot of CD4PCT vs VISIT number. Use different plotting point colors for children who have treatmnt=1 and treatmnt=0[2]. Do the same for CD4PCT vs. visage[3] (the age at the child on each visit). Which is a better measure of time?

   (b) Build a multilevel model with random slopes and intercepts for all the children, using a measure of time (VISIT or visage, whichever you answered in part (a)) as a Level 1 predictor, and treatmnt as a Level 2 predictor. Report the estimated fixed effects and variances from the model, and add the fitted regression lines to your plot from part (a).

   (c) Try to expand the model by adding an interaction between treatmnt and your measure of time[4]. Comment on changes in the estimated fixed effects and variances for this model, vs. the model in part (b). Add the fitted regression lines from this model to your previous facets plot. Does treatmnt account for an interesting amount of the variation in slopes (or intercepts) among children in the study?

[1] https://medlineplus.gov/lab-tests/cd4-lymphocyte-count/

[2] Presumably, "treatmnt" stands for some sort of treatment to improve childrens' HIV status.

[3] If you are using `ggplot()`, you may wish to set `scales="free_x"` in the `facet_wrap()` function.

[4] Beacuse treatmnt is a level-2 variable and the time measure is a level-1 variable, this is somethings called a *cross-level interaction*.

2. Now we will consider in more detail the model below for the CD4 data set:

```
lmer.1 <- lmer(sqrt.CD4PCT ~ 1 + VISIT * treatmnt + (1 + VISIT | newpid),
               data=cd4)
```

Bring together all the residuals (except random effects residuals) that we have talked about, using library(HLMdiag):

```
> r.1 <- hlm_resid(lmer.1,level=1,include.ls=F)
> r.1s <- hlm_resid(lmer.1,level=1,include.ls=F,standardize=T)
> r.2 <- hlm_resid(lmer.1,level="newpid",include.ls=F)
> r.2s <- hlm_resid(lmer.1,level="newpid",include.ls=F,standardize=T)
> names(r.1); names(r.1s); names(r.2); names(r.2s)
```

Read the help file for `hlm_resid` if necessary to understand what all the compenents are. (You can create all these residuals "by hand" as I have shown in lecture slides and in the accompanying R files, but using HLMdiag is simpler and more reliable.)

(a) Make a facets plot of the marginal residuals, as a function of the marginal fitted values (use `scales="free_x"` if needed to make the plot legible). Explain in a sentence or two why a facets plot is not very useful for assessing model fit for this problem, whether we look at the first 12 children, or all 251 children).

(b) Make an ungrouped (that is, no facets) scatter plot of marginal residuals as a function of marginal fitted values, using the full data set (not just the first 12 children). Color the points for `treatmnt=1` kids and treatmnt=2 kids with different colors. Overlay a smooth fit (`geom_smooth` is the easist to use here). Explain, in a couple of sentences (optionally with some math):

- What is causing the dominant structure in this plot, and why that dominant structure is essentially irrelevant for checking the relationship between `sqrt.CD4PCT` and `VISIT`; and
- What in this plot makes you happy or unhappy about having a linear relationship between `sqrt.CD4PCT` and `VISIT` in the model.

(c) Make an ungrouped (that is, no facets) scatter plot of conditional residuals as a function of conditional fitted values, using the full data set (not just the first 12 children). Color the points for `treatmnt=1` kids and `treatmnt=2` kids with different colors. Overlay a smooth fit (`geom_smooth` is the easist to use here). Explain, in a couple of sentences (optionally with some math):

- Why the dominant structure in the marginal residuals is not also present in this plot of conditional residuals
- What might be causing the trend you see in this plot to be different from the trend in the plot of the marginal residuals.

(d) Use standardized residuals and standardized random effects estimates to assess the normality of ϵ_i, η_{0j} and η_{1j} in the fitted model, and to check for any outliers. Include qq plots for each, and accompany each plot with a sentence or two describing what is good or bad in that plot.

3. Continuing with the CD4 data...

(a) Make a table giving values of AIC, BIC, DIC, and cAIC (you compared two of these on the last assignment,using just AIC, BIC and DIC):

```
sqrt.CD4PCT ~ 1 + visage +            (1+visage|newpid)
sqrt.CD4PCT ~ 1 + visage + treatmnt + (1+visage|newpid)
sqrt.CD4PCT ~ 1 + visage * treatmnt + (1+visage|newpid)
sqrt.CD4PCT ~ 1 + VISIT  +            (1+VISIT|newpid)
sqrt.CD4PCT ~ 1 + VISIT  + treatmnt + (1+VISIT|newpid)
sqrt.CD4PCT ~ 1 + VISIT  * treatmnt + (1+VISIT|newpid)
```

Comment briefly on any similarities or differences in how the different criteria choose fixed effects.

(b) Now let's look at the random effects in the model

```
sqrt.CD4PCT ~ 1 + VISIT + treatmnt + (1+VISIT|newpid)
```

Again, make a table giving values of AIC, BIC, DIC, and cAIC for the following models:

```
sqrt.CD4PCT ~ 1 + VISIT + treatmnt
sqrt.CD4PCT ~ 1 + VISIT + treatmnt + (1|newpid)
sqrt.CD4PCT ~ 1 + VISIT + treatmnt + (0+VISIT|newpid)
sqrt.CD4PCT ~ 1 + VISIT + treatmnt + (1+VISIT|newpid)
```

You'll fit the first model with lm(), and the others with lmer(). Comment briefly on any similarities or differences in how the different criteria choose random effects. (Note that only DIC and cAIC have a strong theoretical justification here).

(c) Repeat part (b) but with the interaction `VISIT * treatmnt` in each model instead of just the main effects `VISIT + treatmnt`.

4. IMRAD and IDMRAD.

(a) *(You do not need to turn anything in for this part.)* Compare the papers

- `COVID breakthrough rates in England - PIIS1473309921004606.pdf` and `COVID breakthrough... (appendix) 1-s2.0-S1473309921004606-mmc1.pdf`
- `menu pricing IDMRAD paper with technical appendix.pdf`

The COVID paper is a published IMRAD paper with a separate technical appendix. The "menu pricing" paper is a simple IDMRAD paper (with technical appendix attached) that you can use as a model for your final project paper. Note similarities and differences. In particular, when you write your paper, the technical appendix should show all the R code and results that you used to answer the questions posed for the data analysis[5]. This is a little different from the published COVID paper, which really just provides some supplementary details in the technical appendix that didn't fit in the main paper.

(b) *(You do need to turn something in for this part.)* Refer again to the two papers in part (a). The difference between an IMRAD paper and an IDMRAD paper is that the "Data" part of the **Methods** section of an IMRAD paper is taked out of **Methods** and made into its own **Data** section between **Introduction** and **Methods**.

Find the **Methods** section of the COVID paper, several pages in, and within the **Methods** section find the subsections that describe the data (rather than describing methods of analysis). Copy and paste these sections (including subsection titles) into a separate **Data** section, and turn that section in as an answer to this question.

[5] A small portion of the results in the appendix actually makes it into the main paper.