# Problem Set 7: Complex models

S-043/Stat-151 (Fall 2023)

Don't panic :-)

This is a structured assignment. This means there are more words on our end because there is less for you to do. The assignment is mainly reading through things, tweaking some code, and reflecting on results. This will give you some hands-on exposure to some of the later concepts of the course.

## Overview

The PEDS (Pakistan Early Child Development Scale-Up) trial was a longitudinal single-blind cluster randomized 2x2 factorial trial conducted in rural Pakistan between 2010 and 2012.[1] It evaluated the effectiveness of The National Program for Family Planning and Primary Healthcare (known as the Lady Health Worker (LHW) program). Using a two-stage stratified random sampling strategy, 80 clusters (defined as the catchment areas of LHWs), were identified and randomized into one of four intervention arms.

In each intervention arm, LHWs conducted monthly home visits to deliver the intervention (described below). Mother-child pairs were enrolled at birth (child was <2.5 months old) and followed until the child reached 24 months of age. The main outcomes of the trial were child growth and development.

This is three-level data with time nested in individuals/children nested in clusters/LHWs. A total of 1,489 mother-child pairs were recruited. Some were lost to follow-up, so we have an unbalanced panel.

[1] Yousafzai AK, Rasheed MA, Rizvi A, et al. (2014) Effect of integrated responsive stimulation and nutrition interventions in the Lady Health Worker programme in Pakistan on child development, growth, and health outcomes: A cluster-randomised factorial effectiveness trial. Lancet 384, 1282–1293.

## Variables

### Level 1, time

- `time`: categorical variable for each survey wave, 1 = Baseline, 2 = 6 months in, 3 = 12 months, 4 = 18 months, 5 = 24 months.
- `ch_haz`: child height-for-age Z-score (HAZ), a continuous age- and sex-specific standardized Z-score, calculated according to the WHO Multicenter Growth Standards.[2]
- `ch_stunted`: binary indicator for whether the child is stunted ($HAZ < -2SD$) at a given moment in time (in theory, a child could move in and out of "stunted" status over time, but mainly we will be modeling the chances of a child "moving in" to stunted status). This is our outcome.

[2] HAZ is an age-standardized outcome that basically measures how many SDs away a given child is from a median reference child. It's externally standardized (think about growth charts used in hospitals to assess growth). If this is going down over time then the child is "losing ground" against their peers in terms of their size.

### Level 2, child

- `ch_id`: unique identifier for each child
- a host of demographic baseline characteristics (see margin note).

### Level 3, cluser/LHW

- `lhw_id`: unique identifier for each lady health worker or cluster
- `treatment`: categorical variable with the experimental condition. There were essentially 4 treatments: (1) responsive stimulation (RS), (2) enhanced nutrition (EN), (3) both, and (4) neither. The intervention arms were allocated an equal number of clusters.
  - 1 = Responsive stimulation (RS): Received a locally adapted version of the UNICEF and WHO Care for Child Development package, which promotes sensitive and responsive parenting through developmentally appropriate play and communication activities.
  - 2 = Enhanced nutrition (EN): Received enhanced nutrition education and supplementation with multiple micronutrient powders for children from 6 to 24 months of age.
  - 3 = RS+EN: Received both of the above.
  - 4 = Control: Received basic nutrition, health and hygiene education.

### Child baseline variables

- `ch_agemorec`: continuous variable for child age (in months) at recruitment into the trial (baseline)
- `ch_sex`: binary variable for child sex, 1 = boy, 0 = girl
- `ch_siblings`: count variable for the number of siblings at baseline, used as proxy for household size
- `hh_wealth`: continuous variable for household SES at baseline, a factor score calculated using PCA based on 44 items assessing property and livestock ownership, and water and electricity access.
- `mo_edugrade`: count variable for maternal education at baseline, number of completed grades of schooling
- `hh_foodsec`: binary variable for whether the household was food secure at baseline based on the Household Food Insecurity Access Scale, 1 = food secure, 0 = not food secure

## Goals

The goal of this assignment is to assess whether enhanced nutrition worked (we are going to ignore the second factor of responsive stimulation).[3]

Pedagogically, the goal is to learn about three-level models where we have growth nested in clusters.

Our primary outcomes of interest are `ch_haz` and `ch_stunted` and our overarching research questions are:

RQ1: Does the treatment of enhanced nutrition (vs no enhanced nutrition) impact rate of growth, as captured by the `ch_haz` variable?

RQ2: Does the treatment of enhanced nutrition impact the chance of being stunted over time, as captured by the `ch_stunted` variable?

The second RQ will be investigated in the next assignment. This assignment focuses on RQ1.

[3] Throughout this assignment, when we refer to "treatment," we mean the receipt of enhanced nutrition.

## What to turn in

Turn in a report as a PDF file made from, e.g., Microsoft Word or R Markdown. Your final PDF report should have any code and output from R along with your prose discussion of the results.

The TFs will focus on the results and discussion.[4]

[4] As usual, your code is there as a reference to help the TFs figure out where things may have gone wrong so they can give you useful feedback. Unless you submit a knitted R Markdown file, you should also turn in an R script that has all the code to load the data and run all your analyses and make all your plots, in a single file.

### Before you start

Please do the following:

1. Read through this assignment before starting so you have a sense of where you are going.
2. Remind yourself of the assignment guidelines.[5]
3. Play with the provided code[6] to make sure it makes sense to you.
4. Run `prepare_data.R`, which will clean your data and make a new file for analysis.
5. Examine the script `initial_exploration.R`, which loads the cleaned data and fits an initial model after making some exploratory plots. Read through the code and get familiar with it.

[5] Assignment Guidelines

[6] There are two files, `prepare_data.R` which takes the datafiles and makes a clean data file, and `initial_exploration.R` which makes some plots and fits a model that is discussed further below. The latter file is also embedded in the template Rmd file.

# The Problems

### 1. Some code review questions

a) For the cleaned and prepared data, what is the `EN` variable and how was it made?

b) What was done with children with missing outcome data?

c) What are the `mtime` and `time.f` variables and how are they different?

### 2. Evaluate potential model fit

Given the aggregate plots of average hazard and average proportion of children stunted, assess whether we can easily tell if a linear growth model is appropriate for our outcome of interest. Write a sentence or two of explanation.

### 3. Identify a mathematical model

The bottom of `initial_exploration.R` fits a multilevel model. Write out in mathematical notation the model being fit.

### 4. Make a prediction interval

Assess whether you would say there is a large or small degree of individual variation in child size in the data by making a 95% prediction interval of the baseline height-for-age Z-score in the control group (for a median LWH). Do

this by calculating predicted Z-score for a child with a random intercept 2 SD above average and 2 SD below average.[7]

[7] You do not need to take into account uncertainty in the estimated parameters when you do this. (I.e., follow practices we have been using in class.)

## Extension: Improving Model Fit

The linear growth model might not be an appropriate choice. But the exploratory graphs suggest the pattern of change over time has a similar shape for each treatment arm. One way forward is to add time fixed effects for each time point. We do this by adding `time.f` to our model (not interacted with treatment). This means each branch of treatment has a shared overall structure as captured by the `time.f` variable, plus a "tilt" defined by the linear growth part of the model. We will still have only a few parameters for each treatment group, so our model will still be interpretable.

### 5. Allowing for a curve

Add the `time.f` to your model, and drop the main effect of `mtime` so you only have `mtime` interacted with treatment.[8]

How many parameters are in your final model?

[8] Add `time.f` to the end of your model formula. You will get collinearity warnings (rank deficient means your covariates are not all fully independent), but you can ignore them because the final model is a valid modeling choice. We are just letting R make our dummy variables and throw away the extra ones automatically.

### 6. Plot the treatment curves

Make a plot of the treatment trajectories for each treatment arm by predicting HAZ for two hypothetical children, one in the treatment group and one in the control. You will run into a wrinkle of your model now having both `time` and `time.f`, and so the `expand.grid` won't quite work right. One way to fix this is the following:

```
groups = expand.grid( ch_id = -1,
                      lhw_id = -1,
                      time.f = unique( dat$time.f ),
                      EN = unique( dat$EN ) )
times = parse_number( levels( dat$time.f ) )
groups$mtime = times[ groups$time.f ]
```

At this point you can use `predict( MY_MODEL, newdata=groups )` to get model-predicted outcomes that you can then plot.

Make a plot of both these predicted lines and the lines from the raw plot from the EDA code we provided.[9] How well does your model capture the overall population trends? Is there evidence of model misfit?

At this point you might decide to allow different treatment impacts at each time point by interacting `time.f` with treatment instead of interacting `mtime` with treatment.

[9] The prior EDA plotting code can essentially be reused on your new aggregated data frame. You can add in the new plot points via something like `+ geom_line( data=groups, lty=2 )`

## 7. Use predicted values to create plots for 16 random children

Give a further assessment of model fit by using your model to predict HAZ for all children using, e.g., `dat$haz_pred = predict( MY_MODEL )`. Now make a plot of 16 randomly selected children along with their raw data to get 16 small multiple plots.[10] For your 16 children, do the plots appear to follow the actual data in a reasonable way?

[10] Use `facet_wrap( ~ ch_id )`

## 8. Interpret the results of the model (Answer RQ 1)

Given your previous work (and possibly some extra code to calculate point estimates, conduct likelihood ratio tests, or calculate confidence intervals) answer RQ1:

a) Assess whether EN significantly impacts baseline HAZ. Does this result make sense?

b) Assess whether EN significantly impacts child growth over time.

c) Does the estimated size of the impact seem notable? E.g., you might consider how much extra does a child grow over the first 24 months?

Be sure to identify what model(s) you are using to answer these questions. If some models do not seem appropriate, do not use them. Answer these questions as best you can given the work you did above, and provide any caveats you like to your findings, again based on the work you did above.