Problem Set 8: Logistic growth curves

S-043/Stat-151 (Fall 2023)

Again, don't panic :-)

This is a structured assignment, continuing last week's assignment. The structured assignment 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.

Goals

The goal of this assignment is to assess whether enhanced nutrition worked (we are going to ignore the second factor of responsive stimulation).[1]

Pedagogically, the goal is to learn about fitting logistic growth models. Last assignment, we looked at `ch_haz`. We now look at a binary indicator, `ch_stunted`, of stunted growth. Our research question is:

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

[1] 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.[2]

[2] 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.[3]
3. Review your work from Problem Set 7, and reread the description of the variables, etc., in that assignment.

[3] Assignment Guidelines

The Problems

In this problem set we will model impact on the probability of stunting. We will be doing this with logistic regression and growth curves (but see disclaimer at end of this assignment for comments about the field).

1: Fit a logistic regression model

Fit a random-effects linear growth model with `ch_stunted` as an outcome, allowing a time by EN interaction. Use logistic regression.[4] Include random intercepts and slopes for both child and LHW. Do not use the `time.f` factor trick we did in the prior assignment; this is a vanilla linear model.

[4] Here linear refers to the shape of the latent growth curve in logit space. We are modeling the probability of stunting as a linear trend, and then the data itself is the 0/1 binary outcome. The logistic **links** our linear growth to our binary outcome.

2. Interpret the results of the model (Answer RQ 2)

Answer the following:

a) Assess whether EN significantly impacts the initial odds of being stunted. Does this result make sense?

b) Assess whether EN significantly impacts how the odds of a child being stunted changes over time.

c) Interpret the fixed effect coefficients of the logistic linear growth model in terms of odds multipliers.

3. Make and interpret a median growth trajectory plot

Enter in and adapt the following code so it works with your script:

```
npd = expand.grid( ch_id = -1,
                   lhw_id = -1,
                   mtime = seq( 0, 24, by = 1 ),
                   EN = unique( dat$EN ) )
npd$ch_stunted = predict( MY_MODEL, newdata=npd,
                          allow.new.levels=TRUE,
                          type="response" )
ggplot( npd, aes( mtime, ch_stunted, col=EN ) ) +
    geom_line() + geom_point()
```

Interpret what this plot shows us in terms of individual or population average impacts. If individual, describe the individual we are seeing.

4. Plot the aggregate of the individual predictions as a model check

As a final step for assessing impacts we can estimate the growth curves for all our children and then average them to describe the model-predicted population trends with an overall *population average plot*. This plot is what our model says we should see for the population average trends, which we can compare to the actual data as a diagnostic check.

We start with aggregation:

```
alldat = expand.grid( ch_id = unique( dat$ch_id ),
                      EN = unique( dat$EN ),
                      mtime = unique( dat$mtime ) )

# Add in the child's LHW
children = dplyr::select( dat, ch_id, lhw_id ) %>% unique()
alldat = merge( alldat, children, by="ch_id" )

# Predict for all these points
alldat$pr_stunt = predict( MY_MODEL, newdata=alldat, type="response" )

aggdat = alldat %>% group_by( EN, mtime ) %>%
  summarise( pr_stunt = mean( pr_stunt ) )
```

Using the `aggdat` dataframe, make a plot just like you did in the first part of this assignment. How does this plot compare to your raw data? Is there evidence of model misfit?

5. A sensitivity check

People get excited about including additional covariates in RCTs to correct for chance imbalance and to improve power. Add in the child baseline characteristics to your model[5] and see if your results subsantively change.

To compare the models, people often use a regression table[6]. Make one and then briefly comment on the following:

a) Did the point estimate of impact change when including covariates?

b) Did the SE for the point estimate change?

c) Did anything else change?

d) Are any of the baseline covariates actually connected to the outcome?

[5] Traditionally you would add them as main effects, but if interested in growth you can also add them interacted with time.
[6] E.g., from `texreg()` or `stargazer()`

6. Write a formal wrap-up

Using your models and plots, write a few sentences giving your final results as if you were writing the results in a research paper.

A disclaimer and discussion

There are several general types of models that you can fit with binary outcome data. In the above we used logistic regression and worked with odds ratios (the odds multiplier for a binary covariate can be thought of as an odds ratios of

the treatment group relative to the control group). One can also work with risk ratios or linear probability models (LPMs). Luckily, if you plot things in terms of raw probabilities and aggregates of the predicted probabilities, then it doesn't matter what models you use, as long as the models fit reasonably well. We next discuss these alternative choices, in case you are interested.

Risk Ratios. Most students of epidemiology do not like using models that work with log-odds. They would instead fit the model for the binary outcome as a log-binomial, which gives you an estimate of the risk ratio (RR) rather than the odds ratio (OR).[7] With rare outcomes, the OR approximates the RR, so it doesn't matter how you fit the model and everyone's happy. With common outcomes, like here, the OR does not approximate the RRs.

[7] The risk ratio of two things is the probability of the first divided by the probability of the second. Compare to an odds ratio which is the ratio of the odds. See here for more.

RRs are arguably a better measure of difference: it's much easier to tell people they have a higher risk of a disease than a higher odds of a disease as odds are arguably harder to explain than risks.

In this assignment, however, we use the OR, and the treatment impact is a measure of how the odds of stunting change over time differently for different treatment groups. This would be a common interpretation in some worlds, but not the world where this data would actually be analyzed. To do everything in terms of RRs, one could use the Poisson model instead of the log-binomial.

Linear probability models. A final alternative is to fit an LPM, with your 0s and 1s as your Y value. You will end up reporting percentage-point differences (which is the same as the risk difference). This is a simple method that works well for population average descriptions. It doesn't work as cleanly for individual growth curves, in general.