HW8 (Due June 1)

1. Reproduce Tables 48, 50, 51 and 52.

Table 48: Descriptive stats

Treatment t0 mean t1 mean t0 var t1 var

1 Drug A 9.3 5.3 22.67778 21.56667

2 Drug B 10.0 6.1 27.55556 37.87778

3 Drug C 12.9 12.3 15.65556 51.12222

Table 50: Log-linear model parameter estimates

Coefficients:

Estimate Std.err Wald Pr(>|W|)

(Intercept) 2.373354 0.080138 877.101 <2e-16 \*\*\*

time -0.002877 0.157005 0.000 0.9854

timeA -0.562571 0.221983 6.423 0.0113 \*

timeB -0.495284 0.234201 4.472 0.0344 \*

Estimated Scale Parameters:

Estimate Std.err

(Intercept) 3.214 0.4998

Estimated Correlation Parameters:

Estimate Std.err

alpha 0.7384 0.08149

Parameter estimates are not quite the ones in the slide, allegedly is a different optimization algorithm/settings. I used website code, so I suppose this is the closest you can get to SAS output. Still not discernibly different.

Table 51: Log-linear model parameter estimates, where antibiotic treatments are merged into one variable

Coefficients:

Estimate Std.err Wald Pr(>|W|)

(Intercept) 2.37335 0.08014 877.10 <2e-16 \*\*\*

time -0.00286 0.15700 0.00 0.9855

timeAB -0.52783 0.19883 7.05 0.0079 \*\*

Estimated Scale Parameters:

Estimate Std.err

(Intercept) 3.23 0.52

Estimated Correlation Parameters:

Estimate Std.err

alpha 0.738 0.081

Results again are not quite the same, but difference is not discernible.

Table 52: parameter estimates of a logistic regression model for onycholysis data

Was not able to fit the model with unstructured correlation structure, as R crashed when I attempted it. Reading the manual of the package used (geepack), it says “Use "unstructured" correlation structure only with great care. (It may cause R to crash)”. Hence I ran a model with a first order autoregressive correlation structure instead. Here are the (not discernible different) results:

Coefficients:

Estimate Std.err Wald Pr(>|W|)

(Intercept) -0.5781 0.1215 22.66 1.9e-06 \*\*\*

month -0.1474 0.0255 33.42 7.4e-09 \*\*\*

I(trt \* month) -0.0867 0.0437 3.94 0.047 \*

1. Reproduce Table 56 and comment.

Table 56, using generalized linear mixed effect model with random intercepts

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.6971 0.3298 -5.15 2.7e-07 \*\*\*

month -0.3885 0.0433 -8.97 < 2e-16 \*\*\*

I(trt \* month) -0.1424 0.0649 -2.19 0.028 \*

Random effects:

Groups Name Variance Std.Dev.

id (Intercept) 16 4

In this model the fixed effect for the treatment by month interaction estimates the mean subject effect of treatment on the rate at which the risk of onycholysis changes. This is referred as subject-specific effect. This is different from the marginal model, where the interaction term estimates the mean population effect of treatment on the rate at which the risk of onycholysis changes. This is referred as population effect.

This difference is because of the non-linearity between subject-specific probabilities and odds, where a non-linear contrast of the averages (population effect, marginal models) is not equal to the average of the non-linear contrasts (subject-specific effect, GLMM).

Of note, for the subject-specific model the change in risk due to treatment (and for any time-invariant covariate) applies for hypothetical subject with the same underlying baseline propensity (same value of random intercept) to have onycholysis.

1. Reproduce Table 57 and comment.

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.0708 0.1403 7.63 2.3e-14 \*\*\*

time -0.0005 0.1091 0.00 0.996

trt 0.0512 0.1927 0.27 0.790

time:trt -0.3062 0.1504 -2.04 0.042 \*

Random effects:

Groups Name Variance Std.Dev. Corr

id (Intercept) 0.500 0.707

time 0.232 0.482 0.16

With corr = 0.16 the Cov (b1i, b2i) = 0.16\*sqrt(0.500\*0.232) = 0.0545

The R package lme4 is not able to fit ML with > 1 quadrature points for models with more than a single, scalar random-effects term. Hence, I had to use only one quadrature point. Still very close to SAS output with 50 quadrature points.

Interpretation: there is a significant subject-specific effect of treatment in the change of the rate of seizures, with the progabide group showing a greater reduction in the rate of seizures from baseline, when compared with the placebo group. For a patient receiving placebo there virtually no change in the expected rate of seizures (1- e^-0.0005 ≈ 0), while for a typical subject treated with progabide, the expected drop in the rate of seizures is 26% (1 – e^-0.0005-0.3062 ≈ 0.26).

Estimated random effects indicate that there’s substantial variability in the baseline seizure rate and in the subject to subject changes in the seizure rate in response to treatment. Also, the correlation between random effects is weak (0.16), indicating that expected change in the seizure rates is not directly related to the baseline rate of seizures.

1. We learned the two commonly used methods for analyzing longitudinal data. One is the 1) Marginal model approach, and the other is the 2) Conditional model approach. Explain the similarities and differences, and how to decide and implement/program in practice.

Similarities: they both fit models to correlated data using: 1)a linear model of the expectation, 2)a link function, 3) distributional assumptions

Differences: Marginal models don’t fully specify the joint probability distribution of the vector of multivariate responses, they only specify the distribution of the mean, variance, and covariance among repeated measures. On the other hand, GLMM fully specify the joint distribution of multivariate responses, which allows ML estimation (or numerical approximations) of parameter estimates.

Because of the latter GLMM are computationally more intensive, with no close form solutions for the integrals of the conditional distribution. As result, they might not fit sometimes (as it happened for some models in this HW).

Another important difference (as mentioned throughout the HW) is that the target of inference is different for these methods. While for marginal models the target is the population level effect of covariates on the expected value of the outcome, for the latter the target is the subject specific effect of covariates in the expected value of the outcome.

How to implement/program? Find the appropriate package and code your way to the results (not without some pain as you immerse yourself through the Internet forums).

1. Fit the same data with longitudinal or clustered binary outcome (of your choice) with population-averaged model and subject-specific model. Compare and comment on Odds ratio (point and interval estimate) and statistical significance (p-value) in these 2 approaches.

I’ll use the amenorrhea data from the book: a clinical on two different doses (high and low) of DMPA and their effect on amenorrhea (binary outcome variable). Here time represents 4 consecutive 90 intervals post randomization, and is fitted through a quadratic term. Due to randomization it is assumed that the baseline risk is the same for both dosage groups (i.e. no treatment main effect). This is the GLMM model

And the parameter estimates

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.80568 0.30497 -12.479 < 2e-16 \*\*\*

time 1.13319 0.26821 4.225 2.39e-05 \*\*\*

time2 -0.04192 0.05481 -0.765 0.44435

trt.time 0.56444 0.19224 2.936 0.00332 \*\*

trt.time2 -0.10955 0.04961 -2.208 0.02722 \*

Random effects:

Groups Name Variance Std.Dev.

id (Intercept) 5.065 2.25

Although the model did not converge to the convergence criterion (likelihood gradient was 0.002 and the algorithm tolerance was 0.001), results are virtually the same as in the book.

Interpretation of parameter estimates:

Model results provide evidence that the subject-specific log odds of amenorrhea increases over the 12 months of the trial, and that subject-specific changes in the risk of amenorrhea depend on the dose of DMPA. For a woman assigned to the low dose group the subject specific log odds of amenorrhea at 12 months (4th measurement occasion) is 3.86 (i.e. 4 x 1.1332 – 16 x 0.0419), which corresponds to an odds ratio of 47.6 when compared with baseline. On the other hand, for a woman assigned to the high dose group the subject specific log odds of amenorrhea at 12 months is 4.37 (i.e. 4 x (1.1332 + 0.5644)– 16 x( 0.0419 + 0.1096) ), which corresponds to an odds ratio of 78.8 when compared to baseline.

As for the previous model onycholysis, the interpretation of the dose by time interactions is subject specific, i.e. it compares the change in the log odds of amenorrhea for two different women who happen to have the same underlying risk of experiencing amenorrhea prior to randomization (at baseline).

The expected odds ratio of amenorrhea at 12 months for a woman assigned to the high dose, versus a woman with same underlying risk of amenorrhea prior to randomization who was assigned to the low dose is 1.66 (i.e. ), with a 95% confidence interval of 1.03 to 2.66 (please see below for CI estimation…a total pain, BTW).

Variance estimation:

95% CI estimation:

Lower bound =

Upper bound =

Now the results from the GEE model with unstructured pair wise log odds ratio pattern. Here’s the model specification

Here are the parameter estimates from the model

Coefficients:

Estimate Std.err Wald Pr(>|W|)

(Intercept) -2.2407 0.1765 161.18 <2e-16 \*\*\*

time 0.6979 0.1581 19.49 1e-05 \*\*\*

time2 -0.0314 0.0318 0.98 0.3233

trt.time 0.3379 0.1097 9.49 0.0021 \*\*

trt.time2 -0.0690 0.0284 5.90 0.0151 \*

Not quite, but very similar to the ones from the book (although the same for the website).

Here the interpretation is a population based one, as opposed to the subject-specific from the GLMM.

Here the interaction terms estimate the mean population effect of DMPA dose on the log odds of amenorrhea in the population of women treated with DMPA. This is referred to as a population effect.

This difference in estimates (in marginal models the effects are smaller in absolute value) is because of the non-linearity between subject-specific probabilities and odds, as mentioned in problem 2. The difference in p-values (here are smaller) is because of the smaller SE of the parameter estimates in the GEE model. Why? … I don’t know.

Odds ratio estimates and CI: for the population of women receiving a high dose of DMPA the expected odds ratio of amenorrhea at 12 months post randomization when compared to the population of women in the low dose group is , with 95% confidence interval of 0.97 to 1.69 (please see below for CI estimation).

Variance estimation:

95% CI estimation:

Lower bound =

Upper bound =

1. Instructor thinks there are some incorrect or misleading information in S443 and S672/673. What do you think? You can provide your thoughts.

Just saw this in class today. But, is it misleading? It is hierarchical data. Perhaps mentioning the weights could have made it clearer, although I don’t think it as the attempt to show how to analyze this data, but to show an example of multilevel data structure.

[Remark: You don’t really need nlmixed (in slides) as you have glimmix (programs on authors’ website) nowadays. But it is good to check capacities of these 2 approaches. I used proc nlin for broken-stick regression (that you learned in this course) and there is proc model, too!]