Longitudinal Homework 6

Tim Vigers

15 November 2019

1. Model planning

a. Software

i. Medication use

In order to adapt a generalized linear model (GzLM) for serially correlated data, I would use a generalized estimating equation (GEE). This could be fit using PROC GENMOD in SAS, with a Poisson distribution since this is count data. There is also a package called "geepack" in R that can fit GEEs, but I've never used it and don't know much about it, so PROC GENMOD is probably safer.

ii. FEV1

I think you can probably get away with a normal theory model for FEV1 data, unless it's really skewed. If so, I would use either PROC MIXED in SAS or gls() in R. If normal theory models won't work, then I would use PROC GENMOD or geeglm() like above, to model the outcome with a non-normal distribution.

b. Data

i. Medication use

Because we need a GEE for this outcome, I would set up the data so that each subject has a row for every day during the relevant timeframe. On days without an albuterol count the outcome would be filled in as missing (NA in R), but the temporal spacing would be equal between rows within subject.

The SAS code would look something like:

```
proc genmod data=albuterol;
class id friday;
model albuterol_use =
  friday ln_mmax_pm25 temperature pressure humidity / solution dist=poisson;
repeated subject=id / TYPE = AR(1);
run;
```

ii. FEV1

Again, assuming that a normal mixed model will work, the PROC MIXED code would be something like:

```
proc mixed data = albuterol;
class id friday;
model fev1 = date friday ln_mmax_pm25 temperature pressure humidity / solution;
repeated / type = AR(1) subject = id;
run;
```

I'd use the same data structure as above for this outcome as well.

c. Binary outcome for medication

In order to fit a GzLMM model with a random intercept for subject, I would use PROC GLIMMIX in SAS (pretty much the same code as above, but with distribution = binary). PROC NLMIXED would also work, but I find it a little more confusing. The default in PROC GLIMMIX is to approximate the true likelihood

using Laplace's method, but you can specify method = quad to use adaptive Gaussian quadrature. This can also be done using glmer() in R.

One drawback of quadrature is that it only approximates the true likelihood and there's a bias/variance tradeoff. Also, page 19 of the s13 says that "the GzLMM quadrature approach overestimates SE's since it does not account for the underdispersion," although there was some debate about this in office hours. Finally, you may end up with a different number of quadtrature points using PROC NLMIXED vs. PROC GLIMMIX. This likely wouldn't make a big difference, but you do have to be a little careful.

d. Random intercept and serial correlation

Correlated count data like this probably requires a generalized linear mixed model (GzLMM) where the outcome is modeled as Poisson-distributed, a random intercept for subject, and with an AR(1) or spatial power correlation for repeated measures. I think the best way to do this in SAS is to use PROC GLIMMIX, and in R glmmPQL() should work.

There isn't a REPEATED statement in PROC GLIMMIX, so you need to include another random effect with the "*residual*" keyword and the correlation structure. Something like:

```
proc glimmix data=albuterol;
model albuterol_use =
  friday ln_mmax_pm25 temperature pressure humidity / solution distribution=poisson;
random intercept / subject=id;
random _residual_ / subject=id type=ar(1);
run;
```

The R code would be something like:

```
glmmPQL(fixed = albuterol_use ~ friday + ln_mmax_pm25 + temperature + pressure + humidity,
        random = ~1|id, family = "poisson",correlation = corAR1(),
        data = df)
```

Estimation using pseudolikelihoods can be biased, although this goes away with large sample sizes. Also, you cannot specify non-simple R matrices or have random effects at multiple levels.

2. Albuterol data

a. GEE

The results of PROC GENMOD are below, with the scale parameter in the red box:

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-8.3451	3.4749	-15.1559	-1.5344	-2.40	0.0163
date		0.0004	0.0002	0.0000	0.0008	2.09	0.0368
friday	0	1.2167	0.0810	1.0579	1.3754	15.02	<.0001
friday	1	0.0000	0.0000	0.0000	0.0000	.	.
ln_mmax_pm25		0.0532	0.0151	0.0237	0.0828	3.53	0.0004
temperature		-0.0063	0.0013	-0.0089	-0.0037	-4.69	<.0001
pressure		0.0013	0.0019	-0.0024	0.0051	0.69	0.4886
humidity		-0.0039	0.0006	-0.0052	-0.0027	-6.17	<.0001

Analysis Of GEE Parameter Estimates

Model-Based Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-8.3451	4.5588	-17.2801	0.5899	-1.83	0.0672
date		0.0004	0.0003	-0.0001	0.0009	1.75	0.0802
friday	0	1.2167	0.0376	1.1429	1.2904	32.34	<.0001
friday	1	0.0000	0.0000	0.0000	0.0000	.	.
ln_mmax_pm25		0.0532	0.0193	0.0155	0.0910	2.76	0.0057
temperature		-0.0063	0.0015	-0.0093	-0.0033	-4.16	<.0001
pressure		0.0013	0.0026	-0.0038	0.0065	0.51	0.6132
humidity		-0.0039	0.0009	-0.0056	-0.0022	-4.58	<.0001
Scale		0.8573	.	.	.	.	.

Adding the scale parameter generally increases the standard errors (except for friday = 0 for some reason which I can't figure out).

b. GzLMM

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
Intercept	id	0.5452	0.1134
SP(EXP)	id	0.6045	0.03411
Residual		0.6619	0.01270

Solutions for Fixed Effects						
Effect	friday	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		-5.5776	4.0478	56	-1.38	0.1737
date		0.000281	0.000216	5914	1.30	0.1942
friday	0	1.2008	0.03845	5914	31.23	<.0001
friday	1	0	.	.	.	.
ln_mmax_pm25		0.04600	0.02000	5914	2.30	0.0215
temperature		-0.00561	0.001332	5914	-4.21	<.0001
pressure		0.000509	0.002396	5914	0.21	0.8318
humidity		-0.00300	0.000789	5914	-3.80	0.0001

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
date	1	5914	1.69	0.1942
friday	1	5914	975.51	<.0001
ln_mmax_pm25	1	5914	5.29	0.0215
temperature	1	5914	17.74	<.0001
pressure	1	5914	0.05	0.8318
humidity	1	5914	14.40	0.0001

The residual estimate is equivalent to the scale parameter $\hat{\phi}$, while GEE reports $\sqrt{\hat{\phi}}$. So, to compare the two, $\sqrt{0.6619} = 0.814$ which is a little bit lower than the scale parameter from the GEE (0.857), but I'd say they're pretty close. In order to get this model to converge I had to use sp(exp)(date) rather than the spatial power structure and lower the convergence criteria, so I think that could be partially responsible for the difference.

c. Dispersion

These scale parameter estimates suggest underdispersion because they are less than 1.

d. Slopes and SEs

GzLMM:

ln_mmax_pm25		0.04600	0.02000	5914	2.30	0.0215

GEE MBSE:

ln_mmax_pm25		0.0532	0.0193	0.0155	0.0910	2.76	0.0057

GEE Empirical:

ln_mmax_pm25		0.0532	0.0151	0.0237	0.0828	3.53	0.0004

The estimate from the GzLMM is about 14% smaller, but the standard error from the GzLMM is very close to the SE from the model-based GEE approach. The SE from the empirical GEE approach is smaller than both of the others.

e. Interpretation

The SD of the pollution variable is 0.592, so for each 1 SD increase in pollution, the rate of children's albuterol use increases by about 2.76% (95% CI: 0.52% - 5.06%, p = 0.0215).

3. Exacerbation data

a. Parameter estimates

For each 1 unit increase in day, odds of an exacerbation are 1.009 times higher (95% CI: 1.004 - 1.014, $p < 0.0001$). Or, for each week odds of an exacerbation are 1.066 times higher (95% CI: 1.034 - 1.099, $p < 0.0001$) and for each month odds of an exacerbation are 1.316 times higher (95% CI: 1.156 - 1.497, $p < 0.0001$).

The odds of an exacerbation are 0.812 times lower (95% CI: 0.695 - 0.948, p = 0.0084) on the weekend compared to weekdays.

I don't think the two approaches differ much, but it does sort of depend on what counts as a meaningful change in exacerbation odds. The approach with no repeated measures says that odds of exacerbation are 1.01 times higher per increase in day. which is very close to 1.009. However, the approach without AR(1) says that the odds of an exacerbation are 0.794 times lower on the weekend, which is a little lower than 0.812. It seems close enough to me, but I suppose it's possible that the difference is clincally meaningful.

b. SS vs. PA effects

The slope estimates have subject-specific interpretations, because MSPL was used to approximate the likelihood. This option means that instead of averaging the function over subjects, it determines the function for the average subject. The pseudodata is expanded around subjects as opposed to the population.

c. Estimates

A GzLM/GEE will have population-averaged interpretations, which will generally be lower than the subject-specific beta estimates. The slope of the marginal mean is more attenuated (i.e. a flatter curve) than the conditional mean in a logistic model. This makes some intuitive sense, because if you're averaging the function across multiple subjects, it's impossible for the average to be steeper than the subject specific curves.