Index of Lecture 8–L

— in addition to overheads by Davis (2002):
538–43, 554–57, 560, 590–596

Subject-Specific vs. Population-Averaged

As usual, a 2-level data structure (herds — animals).

Subject-specific (SS) and population-averaged (PA) generalized linear models may be written as:

$$\begin{aligned} \text{subject-specific:} \quad g\left[E(Y|u)\right] &= X'\beta^{\text{SS}} + u, \\ \text{population-averaged:} \quad g\left[E(Y)\right] &= X'\beta^{\text{PA}}, \end{aligned}$$

where g is the link function and u is the herd random effect (with variance σ_h^2).

- $\beta^{\text{SS}} = \beta^{\text{PA}}$ in the following cases:
  - ∗ no link function, (linear mixed models!),
  - ∗ no clustering, ($\sigma_h = 0$),
  - ∗ log link (except for $\beta_0^{\text{SS}} = \beta_0^{\text{PA}} - \sigma_h^2/2$),
- logistic regression: β^{PA} is closer to the null than β^{SS},[1]

$$\beta^{\text{PA}} \approx \beta^{\text{SS}} / \sqrt{1 + 0.346\sigma_h^2}\,.$$

Choice of inference based on predictor types:

- herd-level predictors *inherent* in the herd:
  PA inference most appropriate,
- within-herd predictors:
  individual (SS) or average across herds (PA) effect?

[1] The approximation formula holds also for >2 levels when replacing the single variance component by the sum of all variance components.

Transitional Models

Result: A linear, normal model with autocorrelated errors (1st order) corresponds to a transitional model, as follows:

- model: $Y_{it} = (X\beta)_{it} + \varepsilon_{it}$,
- $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in}) \sim \mathrm{N}(0, \Sigma)$ with Σ of `ar(1)` form
  $\Rightarrow$ the errors are related by the equation,
  $$\varepsilon_{it} = \rho\varepsilon_{i,t-1} + Z_{it}, \tag{1}$$
  where the Z_{it} are i.i.d. and $\sim \mathrm{N}(0, (1-\rho^2)\sigma^2)$,
- inserting the model for time $(t-1)$ into equation (1) and substituting back into the model yields,
  $$Y_{it} = (X\beta)_{it} + \rho(Y_{i,t-1} - (X\beta)_{i,t-1}) + Z_{it}, \tag{2}$$
  — a transition model with independent errors.

Comments:

- higher order `ar` models have similar interpretation,
- β's of (2) have a marginal interpretation, but that is no longer true without the centering of $Y_{i,t-1}$,
- for generalized linear models, a simple relationship only holds in simple cases,
- recommendations about transitional models:
  - use when focus is on within-cluster predictors,
  - do not use when focus is on between-cluster (time-independent) predictors, due to loss of power/efficiency.

GEE for Binary Repeated Measures

Peeking into the Statistician's Laboratory

The scc_40 data — a hard and realistic example (pretty long repeated measures series, plus an additional hierarchical level (herd)),

- create a binary outcome by thresholding the t_lnscc, e.g.

$$\texttt{t\_highscc} = \begin{cases} 1 \text{ if } \texttt{t\_lnscc} \geq \ln(200), \\ 0 \text{ if } \texttt{t\_lnscc} < \ln(200), \end{cases}$$

- 14357 binary outcomes (40 herds, 2178 cows), overall 39.4% positive tests,
- get some idea of the fixed effects:
run ordinary logistic regression (not correct, of course),
- get some idea of the amount of clustering/correlation:
    - could run a 3-level hierarchical model, but dataset too large for SAS glimmix (Stata gllamm ?),
    - run some GEEs to see the working correlations (within cows and within herds),
    - run logistic regression with Williams' correction for overdispersion at either cow or herd level (gives a value for the ICC),
    - could also run a beta-binomial model for cow-aggregated data (skipping test level predictors),

⇒ quite strong clustering within cows (0.45), minor clustering only (0.05) within herds,

- get some idea of the correlation structure for the repeated measures:
  - ∗ run some GEEs with repeated measures working correlation structures (still not correct, because ignoring the herds),

— moderate decay in correlations (roughly 0.48 → 0.20),

- run a 3-level ALR (still not correct but maybe a reasonable approximation); table of estimates next page, some dramatic changes in fixed effects (`t_season`, `h_size`),
- run a 2-level GEE with herd fixed effects (⇒ need to drop `h_size`) and a flexible correlation structure (e.g., Toeplitz structure); next page, some (strange) changes in fixed effects estimates (`t_season`),
- run a 3-level hierarchical model using other software (MLwiN, iterative weighted least squares (IWLS) method),
  - ∗ with binomial and extra-binomial dispersion (to see if underdispersion is observed),
- possible next steps:
  - ∗ try different software (candidates are S-Plus/R or MLwiN macros for GLMM-type models),
  - ∗ try another approach (transitional models to follow).

Summary of Analyses for `scc_40` Data

Regression coefficients and clustering parameters:

Statistic	stand. LR	ALR 3-level	GEE herd fixed	IWLS (MLwiN) $\phi = 1$	IWLS (MLwiN) $\hat{\phi} = 0.75$
$\hat{\beta}$ for `h_size`[1]	1.15 (.10)	0.75 (.48)	–	1.24 (.76)	1.27 (.78)
$\hat{\beta}$ for `c_heifer`	-1.10 (.04)	-1.16 (.07)	-1.29 (.07)	-1.90 (.12)	-1.96 (.13)
$\hat{\beta}$ for winter	-0.08 (.05)	-0.02 (.05)	-0.02 (.05)	-0.03 (.08)	-0.03 (.07)
$\hat{\beta}$ for spring	0.08 (.05)	0.01 (.05)	0.05 (.05)	-0.01 (.08)	-0.03 (.07)
$\hat{\beta}$ for summer	0.11 (.05)	0.03 (.04)	0.18 (.05)	0.04 (.08)	0.03 (.07)
$\hat{\beta}$ for `t_dim`[1]	0.44 (.02)	0.48 (.03)	0.47 (.03)	0.86 (.04)	0.91 (.03)
$\hat{\sigma}_h^2$	–	0.22* (.06)	–	0.57 (.16)	0.60 (.17)
$\hat{\sigma}_c^2$	–	2.23* (.09)	–	4.68 (.21)	5.35 (.23)

[1] estimates multiplied by 100; * value of log OR

Observations/Interpretations:

- very good agreement between IWLS and ALR after having done the approximative correction between SS and PA estimates; e.g. for `c_heifer`,

$$-1.90/\sqrt{1 + 0.346 \cdot (4.68 + 0.57)} = -1.13,$$

  compared to -1.16 by ALR estimation,
- with $\hat{\phi} = 0.75$, there seems to be some indication of an underdispersion caused by ignoring the autoregressive structure.

TRANSITIONAL LOGISTIC REGRESSION

Idea: include at time t the outcome at $t-1$ as a predictor (fixed effect) in the model; in a formula,

$$\text{logit}[P(Y_{ti}=1|Y_{t-1,i})] = (X'\beta)_i + (Zu)_i + \gamma Y_{t-1,i}.$$

The probability is *conditional* on $Y_{t-1,i}$.

Consequences:

- changed interpretation of model effects because conditional and marginal probabilities[2] are not the same,

$$\text{logit}[P(Y_{ti}=1|Y_{t-1,i})] \neq \text{logit}[P(Y_{ti}=1)],$$

- model focuses on transitions $0 \to 1$ and $1 \to 1$,
  - $*$ $0 \to 1 \sim$ new "success" (new cases of disease),
  - $*$ $1 \to 1 \sim$ maintained "success" (non-cured disease),
- regression coefficients for x's not necessarily the same for the two types of transitions $\Rightarrow$ interactions with previous state needed (natural) in the model,
- regression coefficients for first time point are marginal, and may need to be estimated separately in model,
- changed correlations in model, in particular one would expect the short-range correlations (decaying over time) to be taken care of by the previous state
$\Rightarrow$ constant correlations maybe more reasonable (?).

[2] Note that the terms conditional and marginal here do not refer to subject-specific and population-averaged modelling, respectively.

Transitional Models for `scc_40` Data

For demonstration purposes, restrict the dataset to

- series without "gaps" (like Y_2, Y_3, Y_4, Y_6, Y_7),
- observations Y_t for which Y_{t-1} (on the same cow) available for use as a fixed predictor in the model.

Regression coefficients and clustering parameters:

Statistic	ALR (SAS) margin.	ALR (SAS) transit.	IWLS[2] (MLwiN) margin.	IWLS[2] (MLwiN) transit.
$\hat{\beta}$ for `h_size`[1]	0.77 (.50)	0.72 (.39)	1.35 (.79)	0.96 (.54)
$\hat{\beta}$ for `c_heifer`	-1.17 (.08)	-1.03 (.07)	-1.94 (.13)	-1.31 (.09)
$\hat{\beta}$ for winter	-0.02 (.06)	-0.06 (.06)	-0.02 (.09)	-0.09 (.08)
$\hat{\beta}$ for spring	0.02 (.06)	0.00 (.05)	0.03 (.09)	-0.01 (.09)
$\hat{\beta}$ for summer	0.01 (.05)	0.01 (.05)	0.03 (.09)	0.03 (.08)
$\hat{\beta}$ for `t_dim`[1]	0.52 (.05)	0.54 (.04)	0.97 (.05)	0.70 (.05)
$\hat{\beta}$ for `prev`	–	1.55 (.15)	–	2.12 (.15)
$\hat{\beta}$ for `prev*t_dim`[1]	–	-0.15 (.07)	–	-0.16 (.08)
$\hat{\sigma}_h^2$	0.22* (.06)	0.19* (.05)	0.60 (.17)	0.28 (.08)
$\hat{\sigma}_c^2$	2.38* (.10)	1.53* (.08)	5.11 (.25)	1.86 (.11)

[1] estimates multiplied by 100; [2] with $\phi = 1$; * value of log OR

Observations/Interpretations:

- fair agreement between marginal and transitional fixed effects (ALR[3]): one additional (weakly) significant effect,
- reduction in variance parameters for transitional model,
- extra-dispersion parameter slightly less extreme for transitional model ($\hat{\phi} = 0.70$ vs. 0.73, not shown in table).

[3] The disagreement between IWLS estimates is due to the random effects.