Biometrics-by-the-Harbour, Hobart, 2015

Longitudinal data with outcome-related sampling

Alastair Scott & Chris Wild

Department of Statistics
University of Auckland

(work with John Neuhaus & Ross Boylan, UCSF, Yannan Jiang, UoA)

ADHD (Attention Deficit Hyperactivity Disorder) study - Hartung et al. (2002)

- Interested in time course of Y = ADHD in children & association with X-variables
- Y = ADHD relatively rare so used outcome-related design
- Case-control sampling on Z = "ADHD suspected":
  - Suspicion-cases: referred because ADHD suspected by teacher or parent
  - Suspicion-controls: sample of children not so suspected
- 138 cases, 117 controls followed annually for up to 8 visits
  - Y = "actual" ADHD status determined at each visit (Am. Psych. Assoc. criteria)
  - Z = "ADHD suspected" – strongly related to ADHD at 1st visit but not a perfect predictor
- Call sampling on variable(s) Z related to Y (& maybe X) **"outcome-related"**.

Osteoarthritis Initiative (OAI))

- A multi-centre longitudinal study of knee osteoarthritis (OA).
  - objective is to understand risk factors for OA and OA progression.

- Data gathered from 4796 men and women aged 45-79 years. Includes:
  - clinical evaluation data;
  - a biospecimen repository;
  - radiological images (x-ray and MRI)

- MRIs yield both binary and continuous measurements of OA status and are more accurate, but more expensive, than X-rays.
  - To reduce cost, investigators want to select a subset of longitudinal MRIs to evaluate based on X-ray and clinical data (e.g. pain).

- Sets of selected longitudinal MRIs form an outcome-related cluster sample.

Sacramento Area Latino Study on Aging (SALSA))

- A study of elderly Mexican-American living near Sacramento
  - objective is to understand risk factors for changes in cognitive functioning and dementia.
- Data gathered every 12 months for up to 7 years from 1735 men and women. Includes:
  - cognitive function (3MSE);
  - physical and quality-of-life measurements;
  - stored blood samples
- Interested in levels of beta-amyloid (a protein fragment that may cause Alzheimer's disease) in the blood samples
  - expensive so only measure for a targeted subset.

Notation and set up

- We have longitudinal or clustered responses Y_{ij} and covariates $\boldsymbol{x}_{ij}$:
  - i indexes subjects (clusters) ($i = 1, \ldots, m$)
  - j indexes units within subjects (clusters) ($j = 1, \ldots, n_i$),
- Also have auxiliary design variables $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ik_i})$ associated with $\boldsymbol{Y}_i$ and possibly $\boldsymbol{x}_i$
  - Choose i^{th} subject (cluster) for study with probability based on $\boldsymbol{Z}_i$
- Objective: to assess the **individual-specific** (within-subject or within-cluster) assoc. of $\boldsymbol{Y}$ with $\boldsymbol{X}$ and also to examine within-subject aggregation.
  - We will use random effects models (generalized linear mixed models)
  - Schildcrout and Rathouz (2010) estimate **population-averaged** effects using estimating equation methods

Possible approaches

- Semi-parametric likelihood. This leads to efficient estimators, but
  - needs $pr(selection \mid \boldsymbol{Y}, \boldsymbol{X})$, and this needs a model for the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{Y}$ and $\boldsymbol{X}$
  - outside the linear case, can only handle a small number of random effects because of computational constraints

- Use sample survey methods.
  - e.g. in the ADHD study, we have a simple stratified cluster sample with strata based on the design variable $\boldsymbol{Z}$
  - could use
    - weighted pseudo-likelihood
    - weighted composite likelihood

Semiparametric Approach

- We have a cohort of N clusters with values $(\boldsymbol{z}_i, \boldsymbol{y}_i, \boldsymbol{x}_i)$, drawn from some joint distribution
- The model of interest is $f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$ (in the process that generated the cohort)
- Have information on all $\boldsymbol{z}_i$s (which could include elements of $\boldsymbol{y}_i$ and/or $\boldsymbol{x}_i$). Based upon $\boldsymbol{z}_i$, we either
  - observe remaining components of $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ (set $R_i = 1$)
  - or do not (set $R_i = 0$)
    - the marginal distribution of R_i contains no information about the parameters of interest

Conditional Maximum Likelihood

- The conditional MLE is obtained by solving

$$\mathbf{S}_0(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_i^N R_i \frac{\partial \log f(\boldsymbol{y}_i | \boldsymbol{x}_i, R_i = 1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0,$$

where $f(\boldsymbol{y}_i | \boldsymbol{x}_i, R_i = 1; \boldsymbol{\theta})$ is the conditional density of $\boldsymbol{y}_i$ **in the sample**:

$$f(\boldsymbol{y}_i | \boldsymbol{x}_i, R_i = 1; \boldsymbol{\theta}) = \pi(\boldsymbol{x}_i, \boldsymbol{y}_i) f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\theta}) / \pi(\boldsymbol{x}_i; \boldsymbol{\theta}).$$

Here

- $\pi(\boldsymbol{x}_i, \boldsymbol{y}_i) = \Pr(R_i = 1 | \boldsymbol{x}_i, \boldsymbol{y}_i)$,
- $\pi(\boldsymbol{x}_i; \boldsymbol{\theta}) = \Pr(R_i = 1 | \boldsymbol{x}_i) = \int \pi(\boldsymbol{x}_i, \boldsymbol{y}) f(\boldsymbol{y} | \boldsymbol{x}_i; \boldsymbol{\theta}) d\boldsymbol{y}$.

Conditional Maximum Likelihood

- The conditional MLE is obtained by solving

$$\mathbf{S}_0(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_i^N R_i \frac{\partial \log f(\boldsymbol{y}_i | \boldsymbol{x}_i, R_i = 1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0,$$

where $f(\boldsymbol{y}_i | \boldsymbol{x}_i, R_i = 1; \boldsymbol{\theta})$ is the conditional density of $\boldsymbol{y}_i$ **in the sample**:

$$f(\boldsymbol{y}_i | \boldsymbol{x}_i, R_i = 1; \boldsymbol{\theta}) = \pi(\boldsymbol{x}_i, \boldsymbol{y}_i) f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\theta}) / \pi(\boldsymbol{x}_i; \boldsymbol{\theta}).$$

Here

  - $\pi(\boldsymbol{x}_i, \boldsymbol{y}_i) = \Pr(R_i = 1 | \boldsymbol{x}_i, \boldsymbol{y}_i)$,
  - $\pi(\boldsymbol{x}_i; \boldsymbol{\theta}) = \Pr(R_i = 1 | \boldsymbol{x}_i) = \int \pi(\boldsymbol{x}_i, \boldsymbol{y}) f(\boldsymbol{y} | \boldsymbol{x}_i; \boldsymbol{\theta}) d\boldsymbol{y}$.

- Note that calculating $\pi(\boldsymbol{x}_i, \boldsymbol{y}_i)$ may need a model for $f(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$.

Known selection probabilities

- If the $\pi(\boldsymbol{x}_i, \boldsymbol{y}_i)$s are known for all the sampled units, we can use standard likelihood theory to show that, under suitable conditions, $\widehat{\boldsymbol{\theta}}$, the solution to $\mathbf{S}_0(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$:
  - is consistent and asymptotically normal
  - has asymptotic covariance matrix $\boldsymbol{\mathcal{I}}_{00}^{-1}$ where $\boldsymbol{\mathcal{I}}_{00} = E\left\{-\partial \mathbf{S}_0 / \partial \boldsymbol{\theta}^T\right\}$
- Note that $\mathbf{S}_0$ only involves the sampled units
  - ignores any information that we have for the whole cohort

Unknown π_is

- If π_i is not known, we have to estimate it from the full cohort data.
- Suppose we fit a binary regression model, $\pi(\boldsymbol{z}_i; \boldsymbol{\alpha})$ say, for $Pr(R_i = 1)$
  - Let $\mathbf{S}_1(\boldsymbol{\alpha})$ be the corresponding score function
  - Estimating $\boldsymbol{\theta}$ (and $\boldsymbol{\alpha}$) now equivalent to solving

$$\mathbf{S}(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \begin{pmatrix} \mathbf{S}_0(\boldsymbol{\theta}, \boldsymbol{\alpha}) \\ \mathbf{S}_1(\boldsymbol{\alpha}) \end{pmatrix} = 0$$

Unknown π_is

- If π_i is not known, we have to estimate it from the full cohort data.
- Suppose we fit a binary regression model, $\pi(\boldsymbol{z}_i; \boldsymbol{\alpha})$ say, for $Pr(R_i = 1)$
  - Let $\mathbf{S}_1(\boldsymbol{\alpha})$ be the corresponding score function
  - Estimating $\boldsymbol{\theta}$ (and $\boldsymbol{\alpha}$) now equivalent to solving

$$\mathbf{S}(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \begin{pmatrix} \mathbf{S}_0(\boldsymbol{\theta}, \boldsymbol{\alpha}) \\ \mathbf{S}_1(\boldsymbol{\alpha}) \end{pmatrix} = 0$$

- Some notation: let

$$\boldsymbol{\mathcal{I}} = E\left\{ \begin{pmatrix} -\partial \mathbf{S}_0 / \partial \boldsymbol{\theta}^T & -\partial \mathbf{S}_0 / \partial \boldsymbol{\alpha}^T \\ 0 & -\partial \mathbf{S}_1 / \partial \boldsymbol{\alpha}^T \end{pmatrix} \right\} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{00} & \boldsymbol{\mathcal{I}}_{01} \\ 0 & \boldsymbol{\mathcal{I}}_{11} \end{pmatrix}$$

Unknown π

Then we can show that

$$ACov\{\widehat{\boldsymbol{\theta}}\} = \boldsymbol{\mathcal{I}}_{00}^{-1} - \boldsymbol{\mathcal{I}}_{00}^{-1}\boldsymbol{\mathcal{I}}_{01}\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{01}^{T}\boldsymbol{\mathcal{I}}_{00}^{-1}.$$

Unknown π

Then we can show that

$$ACov\{\widehat{\boldsymbol{\theta}}\} = \boxed{\boldsymbol{\mathcal{I}}_{00}^{-1}} - \boldsymbol{\mathcal{I}}_{00}^{-1}\boldsymbol{\mathcal{I}}_{01}\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{01}^{T}\boldsymbol{\mathcal{I}}_{00}^{-1}.$$

Using the true values of π_i

Unknown π

Then we can show that

$$ACov\{\widehat{\boldsymbol{\theta}}\} = \boldsymbol{\mathcal{I}}_{00}^{-1} - \boldsymbol{\mathcal{I}}_{00}^{-1}\boldsymbol{\mathcal{I}}_{01}\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{01}^{T}\boldsymbol{\mathcal{I}}_{00}^{-1}$$

Using the true values of π_i

Effect of estimating π_i

Unknown π

Then we can show that

$$ACov\{\widehat{\boldsymbol{\theta}}\} = \boldsymbol{\mathcal{I}}_{00}^{-1} - \boldsymbol{\mathcal{I}}_{00}^{-1}\boldsymbol{\mathcal{I}}_{01}\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{01}^{T}\boldsymbol{\mathcal{I}}_{00}^{-1}$$

Using the true values of π_i

Effect of estimating π_i

- The value of $ACov\{\widehat{\boldsymbol{\theta}}\}$ is **smaller** when we use estimated selection probabilities than when we use the true values!
- Although it may look paradoxical at first, by using the **estimated** π**s** we are actually using more information thus **better estimates**:
  - essentially equivalent to calibration on $\boldsymbol{z}$ component totals in survey sampling

Unknown π

Note that

$$ACov\{\widehat{\boldsymbol{\theta}}\} = \boldsymbol{\mathcal{I}}_{00}^{-1}\left(\boldsymbol{\mathcal{I}}_{00} - \boldsymbol{\mathcal{I}}_{01}\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{01}^{T}\right)\boldsymbol{\mathcal{I}}_{00}^{-1} = \boldsymbol{\mathcal{I}}_{00}^{-1}\boldsymbol{\mathcal{C}}_{R}\boldsymbol{\mathcal{I}}_{00}^{-1}$$

- $\boldsymbol{\mathcal{C}}_R$ is the cov. matrix of resid. vector when $\mathbf{S}_0$ is regressed on $\mathbf{S}_1$, i.e. $\boldsymbol{\mathcal{C}}_R = \inf_B \mathrm{Cov}\{\mathbf{S}_0 - B\mathbf{S}_1\}$;
- Adding any extra variables to inclusion model never increases, & may decrease, $\boldsymbol{\mathcal{C}}_R$ even if the π_is do not actually depend on them;
- Size of reduction depends on the relationship between the score for the added variable and $\mathbf{S}_0(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0)$ and not at all on the strength of its effect on the inclusion probabilities;
- If $\boldsymbol{z}$ has finite support then, most efficient to fit a saturated model for $\boldsymbol{\pi}$.

Improving estimates

We can do even better. The conditional log-likelihood has the form

$$\ell_c(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_i^N R_i \log f(\boldsymbol{y}_i | \boldsymbol{x}_i, R_i = 1; \boldsymbol{\theta}, \boldsymbol{\pi})$$

- If we replace $\boldsymbol{\pi}$ by the modelled value $\boldsymbol{\pi}(\boldsymbol{\alpha})$, this conditional log-likelihood becomes a function of $\boldsymbol{\alpha}$ as well as $\boldsymbol{\theta}$
  - the corresponding score function, $\tilde{\mathbf{S}}(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \dfrac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$, carries extra information
  - Replace $\mathbf{S}_1(\boldsymbol{\alpha})$ by $\mathbf{S}_1(\boldsymbol{\alpha}) + \lambda \tilde{\mathbf{S}}(\boldsymbol{\theta}, \boldsymbol{\alpha})$ (optimal $\lambda = -1$).
- This gives the fully efficient semi-parametric estimator when $\boldsymbol{z}$ can take only a finite number of values
  - Can give useful improvements more generally.

Generalized linear mixed models

- Given vector of random cluster-level parameters $\boldsymbol{b}_i$, conditional density of Y_{ij} (for the jth unit in i^{th} cluster) is of the form

$$f_Y(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{x}_{ij}) = \exp[\{y_{ij}\Delta_{ij} - c(\Delta_{ij})\}\phi + d(y_{ij}, \phi)],$$

  where c and d are known functions, ϕ is scale parameter and Δ_{ij} is a function of $\mu_{ij} = E(Y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{w}_{ij}, \boldsymbol{x}_{ij})$ depending on the x_{ij}s through

$$\mu_{ij} = g^{-1}(\boldsymbol{w}_{ij}^T \boldsymbol{b}_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}) \equiv g^{-1}(\eta_{ij}).$$

  - $\boldsymbol{x}_{ij}^T$ and $\boldsymbol{w}_{ij}^T$ are covariate row vectors relating the fixed and random effects, resp

- Assume that $Y_{i1}, \ldots, Y_{in_i}$ are independent, given the random cluster effects $\boldsymbol{b}_i$.

- Simulations focus on models with random intercepts and slopes.

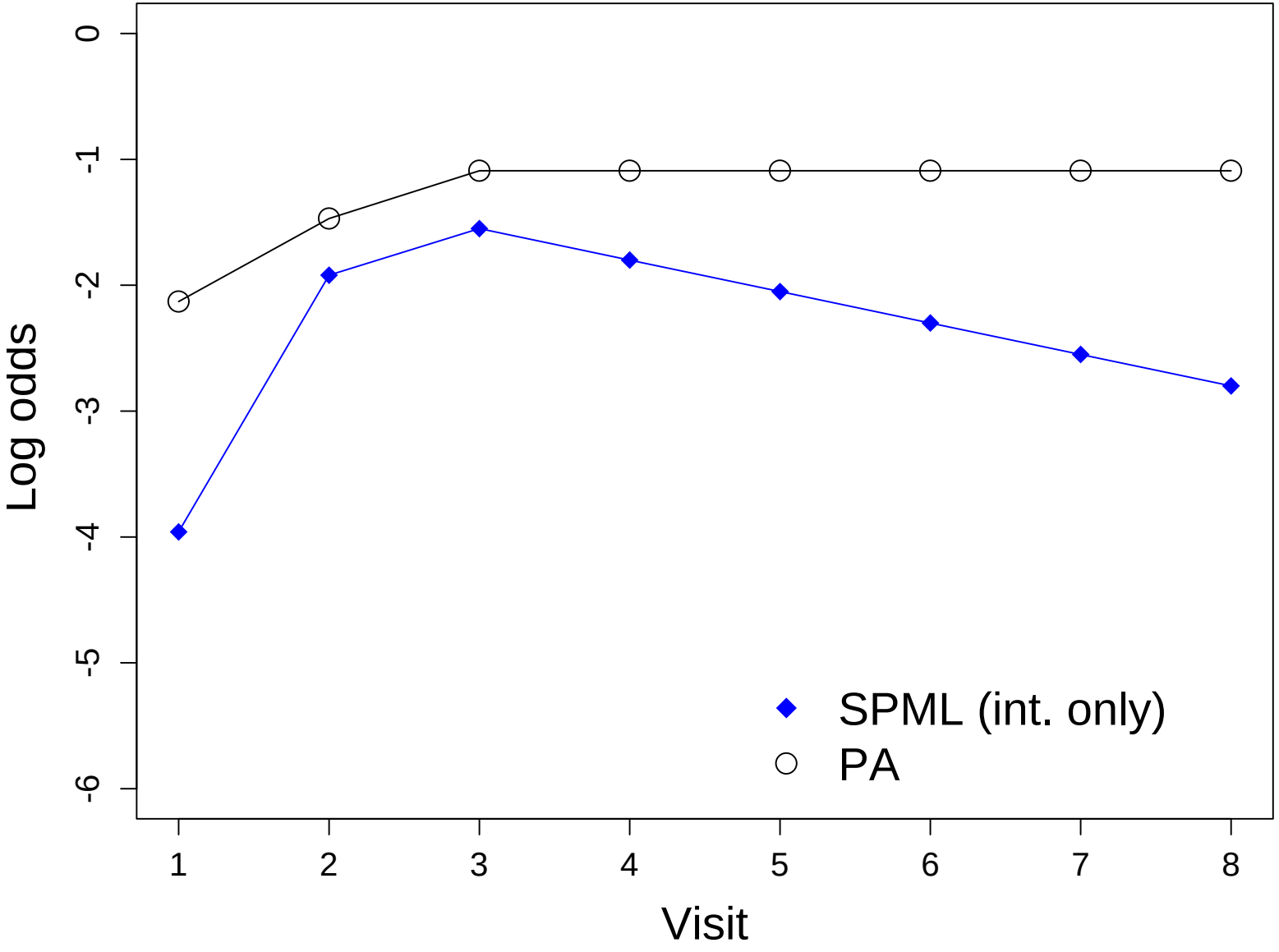
Simulations

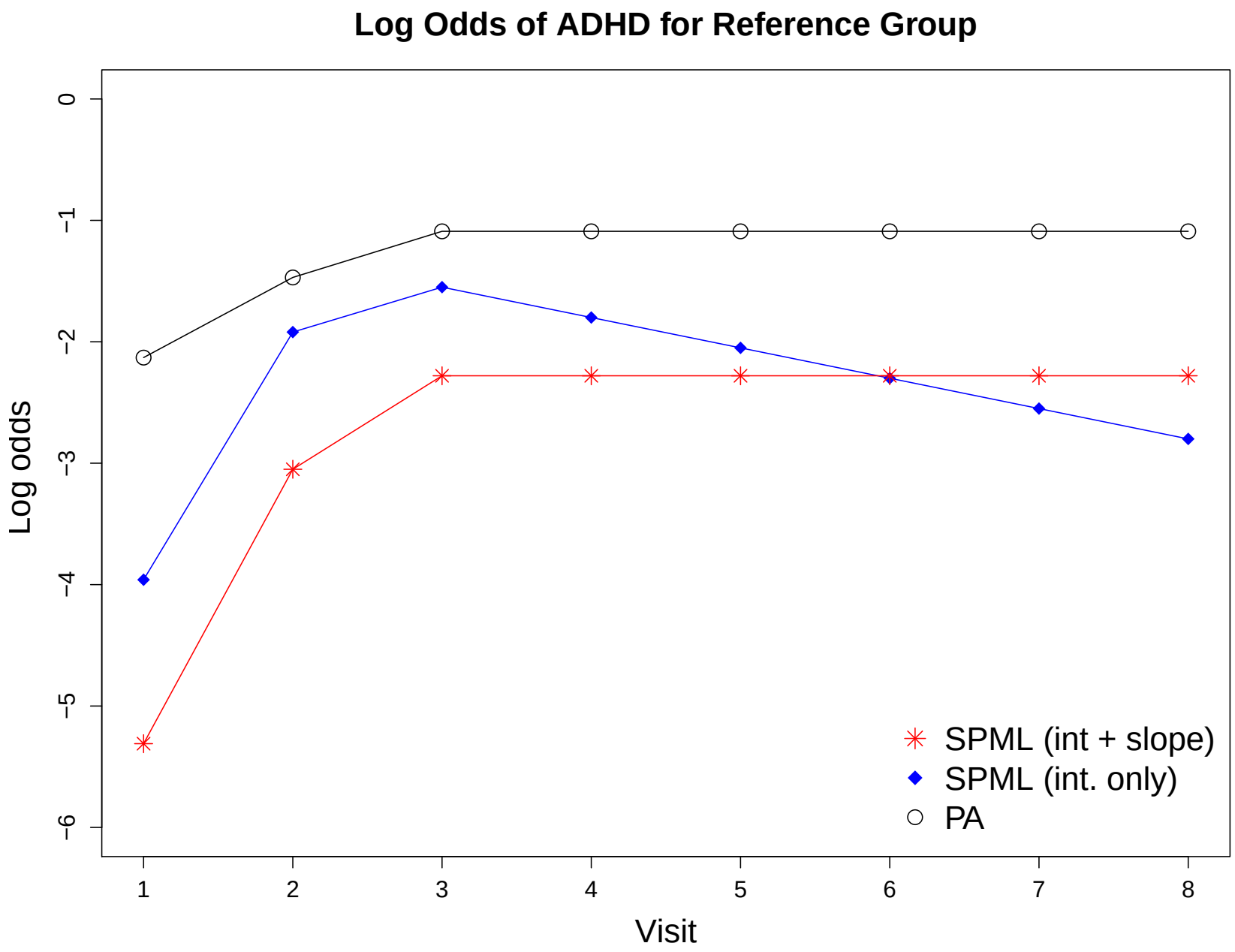
- Simulations based on the data from ADHD study
  - longitudinal binary data generated from logistic regression models
    - Time (visit) variable x_t over 8 (or 4) time points
    - a continuous x-variable and a binary x-variable unrelated to time
    - an auxiliary variable $Z = 1$ or $Z = 0$
  - simulations generated populations of 5000 subjects with 500 sampled on value of Z
- 2 main sets of simulations:
  - random intercepts only
  - using both random intercepts and slopes
- Results:
  - semi-parametric maximum likelihood performance as hoped (very low bias, excellent coverage)
  - standard mixed effects logistic regression uncorrected for biased sampling led to large biases for all parameters

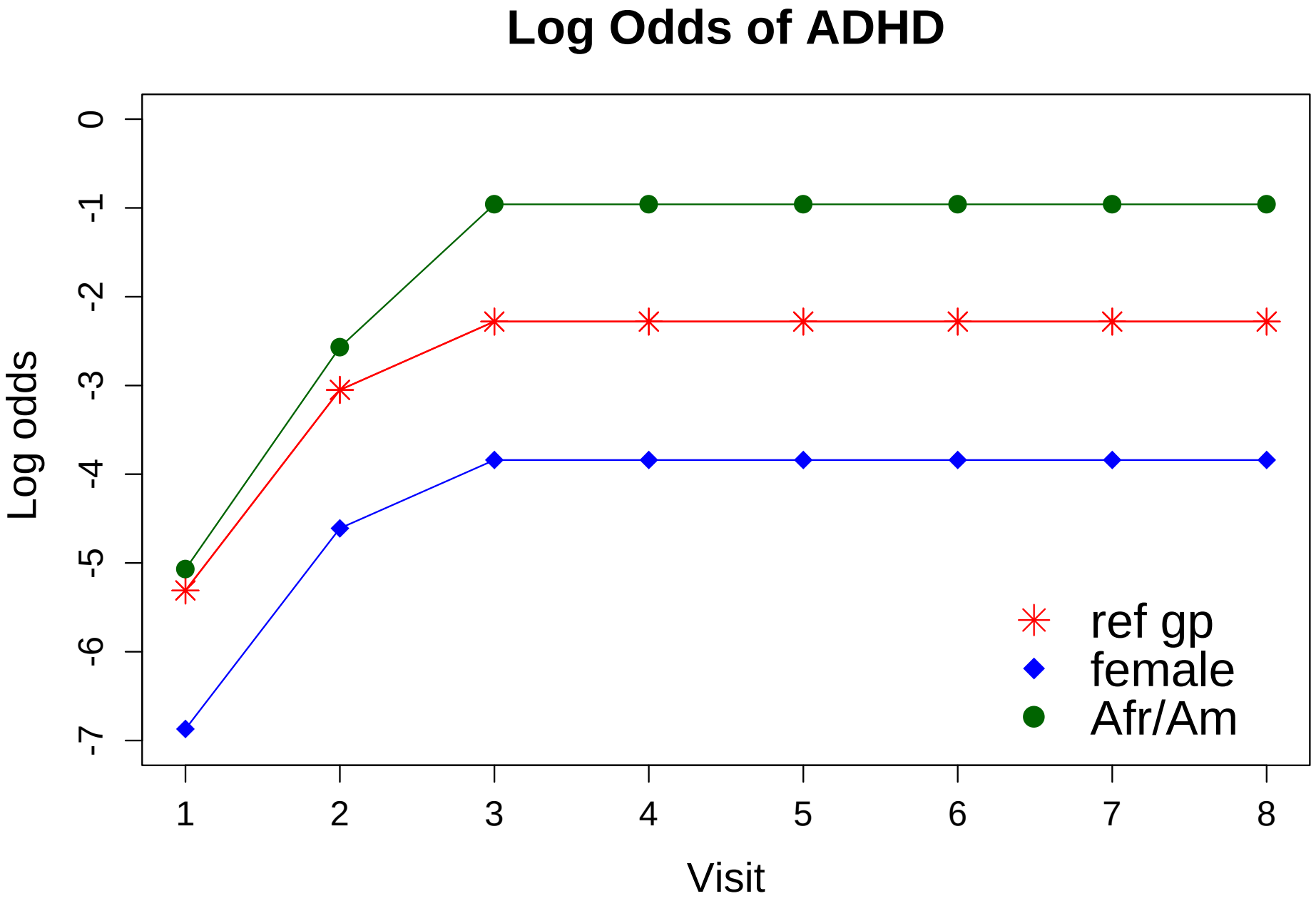
ADHD example

- Fitted three models
  - a marginal model with no random effects
  - random intercepts only
  - using both random intercepts and slopes
- All models included fixed effects for
  - visit number,
  - sex,
  - ethnicity,
  - + interactions

Log Odds of ADHD for Reference Group




Log Odds of ADHD for Reference Group
Log odds
0
−1
−2
−3
−4
−5
−6
1
2
3
4
5
6
7
8
Visit
SPML (int + slope)
SPML (int. only)
PA


Log Odds of ADHD
Log odds
0
-1
-2
-3
-4
-5
-6
-7
1
2
3
4
5
6
7
8
Visit
ref gp
female
Afr/Am

Some questions

- How can we manage the computation with more random effects in nonlinear models?
  - better numerical techniques (adaptive Gaussian quadrature, etc)?
  - Could a weighted survey approach do better?
- How do we choose a good design?
  - How should we use X-rays, pain, mobility etc to decide which MRIs to look at in the Osteoarthritis Initiative?
  - How should we use measures of cognitive function, quality of life etc to decide when to measure beta-amyloid levels in SALSA?