Generalized linear models for longitudinal data with biased sampling designs

A sequential offsetted regression approach

L.S. McDaniel[1] J.S. Schildcrout[2] E.F. Schisterman[3]
P.J. Rathouz[4]

[1]Biostatistics Program, School of Public Health
LSU Health Sciences Center, New Orleans

[2]Department of Biostatistics, Department of Anesthesiology
Vanderbilt University School of Medicine

[3]*Eunice Kennedy Shriver* National Institute of Child Health and Human Development
National Institutes of Health

[4]Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Example - ADHD Study

- Goal: Identify risk and prognostic factors for ADHD in **early** childhood
- Sampling: 255 subjects, about half cases/controls
  - Cases: Referred by parent or teacher
  - Controls: Matched demographically
- Followed for 15 years (we have 8)
- Analyze: Time course of hyperactivity symptom count

Subject-level Sampling:

A subject is either in or out of the study

Observation-level Sampling:

A subject may be sampled at each time point

- Y_j: count or continuous outcome at times $t_j = 1, \ldots, T$
- x_j: p-vector of covariates at times t_j
- $X = (x_1, \ldots, x_T)'$ is a $T \times p$ matrix of covariates

Marginal population mean model for Y_j:

$$\mu_{P_j} = E(Y_j|X) = g^{-1}(x_j'\beta)$$

Finally,

- Z_j: subject was referred at time t_j
- S_j: subject was sampled at time t_j

1. No interference assumption

$$E(Y_j|X) = E(Y_j|x_j)$$

2. Known value for

$$\frac{\Pr(S_j = 1|Z_j = 1)}{\Pr(S_j = 1|Z_j = 0)} = \frac{\pi(1)}{\pi(0)}$$

3. Sampling only depends on Z_j, and possibly baseline covariates

Modeling Steps

The assumptions allow for three modeling steps:

1. Estimate $\Pr(Z_j = 1|Y_j, X)$ from sample, for each t_j
2. Compute $\Pr(S_j = 1|Y_j, X)$, for each t_j
3. Estimate $E(Y_j|X)$ from sample

Step 1: Estimate $\Pr(Z_j = 1|Y_j, X)$

Let w_j be a vector of covariates (possibly overlapping x_j)
In the population:

$$\begin{aligned} Pr(Z_j = 1|Y_j, X) =& \lambda_{P_j}(y, X) \\ =& \text{logit}^{-1}\left\{w_{1j}'\gamma_1 + h(y)w_{2j}'\gamma_2\right\} \end{aligned}$$

Then, in the sample:

$$\begin{aligned} Pr(Z_j = 1|Y_j, X, S_j) =& \lambda_{S_j}(y, X) \\ =& \text{logit}^{-1}\left\{w_{1j}'\gamma_1 + h(y)w_{2j}'\gamma_2 + \log \pi(1)/\pi(0)\right\} \end{aligned}$$

Step 2: Compute $\Pr(S_j|Y_j, X)$

$$\begin{aligned}\rho_j(y, X) =& \Pr(S_j = 1|y, X) \\ =& \pi(0)\left\{1 - \lambda_{P_j}(y, X)\right\} + \pi(1)\lambda_{P_j}(y, X)\end{aligned}$$

Gain stability by using

$$\frac{\rho_j(y, X)}{\rho_j(y_0, X)} = \frac{1 - \lambda_{P_j}(y, X) + \{\pi(1)/\pi(0)\}\,\lambda_{P_j}(y, X)}{1 - \lambda_{P_j}(y_0, X) + \{\pi(1)/\pi(0)\}\,\lambda_{P_j}(y_0, X)}$$

Step 3: Estimate $E(Y_j|X)$

In the population, conditional density is **exponential family**:

$$f_P(y|X) = \exp\left\{\frac{\theta_j y - b(\theta_j)}{\phi} + c(y;\phi)\right\}$$

Use canonical link:

$$g(\mu_{P_j}) = g(\mathrm{E}(Y_j|X)) = x_j'\beta = \theta_j$$

In sample:

$$f_S(y|X) \propto \exp\left\{\frac{\theta_j y - b(\theta_j)}{\phi} + c(y;\phi) + \log\rho_j(y,X)\right\}$$

View the solution as stacked estimating equations:

$$\sum_i \begin{pmatrix} \mathbf{T}_i(\gamma) \\ \mathbf{U}_i(\gamma, \beta) \end{pmatrix} = \mathbf{0}.$$

Use sandwich estimate for SE

Applicable Outcomes

Worked out for

- Binary data (binomial)
- Count data (Poisson)
- Continuous data (normal)

For continuous data, need to estimate variance

For each subject, i,

$$\log \mu_{ij} = \beta_0 + \beta_{x_1} x_{1i} + \beta_t t_j + \beta_{tx_1}(t_j \times x_{1i})$$

x_{1i} is time-invariant, binary covariate.

Oversample **subjects** with high values for Y_{i1}

Table includes % bias and coverage probability (target of 95%)

Estimation	$\beta_0 = -1.4$	$\beta_{x_1} = 0.4$	$\beta_t = -0.1$	$\beta_{tx_1} = 0.1$
Naive GEE	-42 (0)	-21 (88)	24 (84)	-11 (92)
IPW	0 (94)	-1 (95)	3 (93)	2 (94)
SOR	1 (95)	2 (95)	2 (94)	0 (95)

Efficiency Relative to Simple Random Sampling

Estimation	$\beta_0 = -1.4$	$\beta_{x_1} = 0.4$	$\beta_t = -0.1$	$\beta_{tx_1} = 0.1$
IPW	1.17	1.17	0.63	0.60
SOR	1.40	1.37	1.26	1.14

For each subject, i,

$$Y_{ij} = \beta_0 + \beta_{x_1} x_{1i} + \beta_{x_2} x_{2ij} + \beta_{x_3} x_{3ij} + \epsilon_{ij}$$

x_{1i} is time-invariant, x_{2ij} and x_{3ij} vary with time.

Oversample **observations** with high values for Y_{ij}

Estimation	$\beta_0 = 1$	$\beta_{x_1} = 1$	$\beta_{x_2} = 1$	$\beta_{x_3} = 1$
Naive GEE	443 (0)	29 (74)	29 (70)	28 (72)
IPW	-1 (94)	1 (95)	1 (94)	-1 (95)
SOR	-5 (90)	0 (95)	0 (95)	-1 (96)

Efficiency Relative to Simple Random Sampling

Estimation	$\beta_0 = 1$	$\beta_{x_1} = 1$	$\beta_{x_2} = 1$	$\beta_{x_3} = 1$
IPW	0.34	0.37	0.34	0.33
SOR	0.75	1.33	1.31	1.26

Repsonse is hyperactivity symptom count.
Coefficients are exponentiated

	Naive		SOR	
Intercept	4.01	$(3.32, 4.90)$	2.75	$(2.23, 3.39)$
t	0.98	$(0.90, 1.05)$	1.06	$(0.97, 1.15)$
$\mathbf{(t-2)_+}$	**0.95**	$\mathbf{(0.88, 1.03)}$	**0.89**	$\mathbf{(0.80, 0.97)}$
age	0.87	$(0.75, 1.00)$	0.87	$(0.75, 1.01)$
sex	**0.80**	$\mathbf{(0.53, 1.22)}$	**0.62**	$\mathbf{(0.41, 0.93)}$
afr	1.58	$(1.25, 2.03)$	1.67	$(1.30, 2.12)$
other	1.11	$(0.63, 1.95)$	1.03	$(0.58, 1.84)$
sex*t	1.04	$(0.84, 1.28)$	1.07	$(0.85, 1.35)$
sex*$(t-2)_+$	0.91	$(0.71, 1.16)$	0.89	$(0.68, 1.15)$