Some issues of efficiency and robustness

Chris Wild
The University of Auckland
New Zealand

c.wild@auckland.ac.nz

(Joint work with Yannan Jiang and Alastair Scott)

ABSTRACT

We investigate some issues of efficiency, robustness and study design affecting semiparametric maximum likelihood and survey-weighted analyses for linear regression under two-phase sampling and bivariate binary regressions, especially those occurring in secondary analyses of case-control data.

Motivation: ABC STUDY

(Auckland Birthweight Cooperative Study)

- Designed as a hospital-based Case-control study for SGA
  - Small for Gestational Age $\approx$ lowest 10% for gestational age and sex
- Approximately 900 cases (SGA) and 900 controls (non-SGA)
- Large number of covariates

HOWEVER

- approx 20,000 births in the same hospital over the same time period
- For "all" of them, we have
  - Birthweight; gestational age; sex; mother's "dimensions", smoking habits and a number of other largely demographic variables.

ANALYSES

- Primary planned analyses were logistic regressions.
- Although sampling is stratified on SGA, also want analyses with Birthweight itself as the response.
- Be nice to include all-babies data in the analysis

SOME NOTATION

- $\boldsymbol{y}$ is response
- $\boldsymbol{z}$ contains the covariates that are not always available
- $\boldsymbol{v}$ contains covariates and correlated variables that are always available
- Interest centers on $f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$ where $\boldsymbol{x}$ contains $\boldsymbol{z}$ and $\boldsymbol{v}_1$, a subset of $\boldsymbol{v}$ — impose no further parametric assumptions

Applications

- Structured missing data (obv.)
- Measurement error problems (cf. Carroll, Ruppert & Stefanski 1995, Ch 9)
  - Surrogates in $\boldsymbol{v}$
- Intermediate (for efficiency gains)

Likelihood

$$L[\boldsymbol{\theta}, \{g(\cdot|\boldsymbol{v}_s)\}) = \prod_{s=1}^{S} \prod_{i:\boldsymbol{v}_i=\widetilde{\boldsymbol{v}}_s} f(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \widetilde{\boldsymbol{v}}_{1s}; \boldsymbol{\theta})^{\Delta_{1i}} g(\boldsymbol{z}_i \mid \widetilde{\boldsymbol{v}}_s)^{\Delta_{2i}} \mathrm{pr}(\boldsymbol{y}_i \mid \widetilde{\boldsymbol{v}}_s; \boldsymbol{\theta})^{\Delta_{3i}},$$

[Here $\mathrm{pr}(\boldsymbol{y} \mid \boldsymbol{v}; \boldsymbol{\theta}) = \int f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{v}_1; \boldsymbol{\theta}) dG(\boldsymbol{z}|\boldsymbol{v})$.]

Permits inclusion of observations on $\boldsymbol{v}$ and any or all of $(\boldsymbol{y}, \boldsymbol{v})$, $(\boldsymbol{y} \mid \boldsymbol{v})$, $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, $(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{v})$, $(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{v})$, $(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{v})$, $(\boldsymbol{z}, \boldsymbol{v})$ and $(\boldsymbol{z} \mid \boldsymbol{v})$

Semiparametric maximum likelihood

- Replace $g(\boldsymbol{z}|\boldsymbol{v}_s)$ by a discrete distribution with probability mass concentrated on the $\boldsymbol{z}$ values observed in the $\boldsymbol{v} = \boldsymbol{v}_s$ stratum.
- This gives us a likelihood of the form $L(\boldsymbol{\theta}, \boldsymbol{\delta})$
- Profile out $\boldsymbol{\delta}$ to get profile likelihood $L_P(\boldsymbol{\theta}) = L[\boldsymbol{\theta}, \widehat{\boldsymbol{\delta}}(\boldsymbol{\theta})]$
- Use for inference — standard errors from the inf matrix of $L_P(\boldsymbol{\theta})$

Actually, the engine of the program handles...

any likelihood of the form

$$\prod_{s=1}^{S} \prod_{i:\boldsymbol{v}_i=\boldsymbol{v}_S} f(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \boldsymbol{v}_{1s}; \boldsymbol{\theta})^{\Delta_{1i}} g_1(\boldsymbol{z}_i \mid \boldsymbol{v}_s)^{\Delta_{2i}} \mathrm{pr}\{\boldsymbol{h}_{\boldsymbol{v}_S}(\boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{h}^{[i]} \mid \boldsymbol{v}_s; \boldsymbol{\theta}\}^{\Delta_{3i}}$$

where Δ_{1i} and Δ_{2i} are 0/1 indicators and Δ_{3i} can take values -1, 0 and 1.

$\boldsymbol{h}_{vS}(\boldsymbol{y}, \boldsymbol{z})$ allows for

- sampling on a subset/pattern in $\boldsymbol{y}$ (response-selective)
  - Proband example $\boldsymbol{y} = (y_1, y_2)$ sample on $y_1 = 1$ (yes/no), or on $(y_1, y_2) = (0, 0)$ (yes/no)
- Failure time example

LOOK AT THE LINEAR MODEL

$$Y = \boldsymbol{\beta}^T \boldsymbol{x} + \sigma U$$

And at issues of ...

- Sampling Design
- Efficiency
- Robustness
- Note that in semiparametric errors-in-variables
  - Tend to obtain gold standard measure for random subsample
- Following based on some extensive simulation studies by our PhD student Yannan Jiang

Simulations of Linear Model $Y = \boldsymbol{\beta}^T \boldsymbol{x} + \sigma U$

- Y Response observed for all
- X_1 Observed for subsample
  - "expensive" or "gold standard"
- V Discrete surrogate for X_1
- X_j's Other variables observed for subsample
- Simulations use $N = 1,000$ and a 20% subsample

- Will use survey-weighted analyses as well as semiparametric maximum likelihood

Weighted estimation with stratified sampling

- Idea: solve $\sum_{i=1}^{N} w_i \frac{\partial}{\partial \boldsymbol{\theta}} \log \{ f(y_i \mid \boldsymbol{z}_i, \boldsymbol{v}_{1i}; \boldsymbol{\theta}) \} = \mathbf{0}$,
  - where $w_i = \Delta_i / \text{pr}(\Delta_i = 1)$.

With S strata $\mathcal{S}_1, \ldots, \mathcal{S}_S$ and units in $\mathcal{S}_s$ are sampled with probability π_s, then this is equivalent to solving

$$\sum_{s=1}^{S} w_s \sum_{\text{subsampled from } \mathcal{S}_s} \frac{\partial}{\partial \boldsymbol{\theta}} \log \{ f(y \mid \boldsymbol{z}, \boldsymbol{v}_1; \boldsymbol{\theta}) \} = \mathbf{0}, \tag{1}$$

- More efficient to use estimated weights
  - $\widehat{\pi}_s = n_s / N_s$, giving weights $w_s = N_s / n_s$
- More efficient still to use stratification based on $Y \times V$
  - May be poststratification. Tradeoff fineness (efficiency) versus sparseness
  - This is crude. More efficient weighted methods are possible – see papers by Robins and Rotnitzky in the mid 1990s.

With a 20% subsample

- Extremes
  - Take random 20% subsample – analyse subsample data only
  - Analyze complete data
- Relative efficiency (Full data/Subsample) 500%
- How close to that can we get with ...?
  - Different subsampling designs
  - Complete information on Y and V

Warning: Rel. efficiencies comparing sampling designs of round 500% based on 1,000 simulations have std err $\approx$ 30%

How to subsample?

- Purely randomly
- Stratify on V
- Stratify on Y
- Stratify on both V and Y
  - Ignore this last as have never managed to do significantly better than sampling on Y or V alone
  - All our subsampling oversamples extremes

(cf. usual experimental design)

Single Covariate

Here, $Y = \beta_0 + \beta_1 X_1 + \sigma U$

- Y Response observable for all
- X_1 N(0,1), observed for 20% subsample
- V 8-point discretization of X_1, observable for all

*Effect of Sampling Designs with a Single Covariate (**Mle**)*

standard error of $\widehat{\boldsymbol{\theta}}$

Sampling Design	Analysis uses V	X-Effect $\theta = 2$	$\theta = 1$	$\theta = 0.5$	$\theta = 0.3$	$\theta = 0$
All Observed	-	0.0251	0.0244	0.0246	0.0247	0.0240
YV-Conditional	Yes	0.0270	0.0253	0.0251	0.0252	0.0246
V-Conditional	Yes	0.0272	0.0255	0.0254	0.0252	0.0246
Y-Conditional	Yes	0.0278	0.0257	0.0258	0.0253	0.0244
Unconditional	Yes	0.0296	0.0262	0.0256	0.0255	0.0245
Y-Conditional	No	0.0461	0.0454	0.0435	0.0413	0.0424
Unconditional	No	0.0444	0.0457	0.0511	0.0542	0.0543
MCAR	-	0.0551	0.0555	0.0545	0.0558	0.0529

*Effect of Sampling Designs with a Single Covariate (**Mle**)*

$$\frac{VEst_{ss}}{VEst_{mcar}} * 100\%.$$

Sampling Design	Analysis uses V	X-Effect $\theta = 2$	$\theta = 1$	$\theta = 0.5$	$\theta = 0.3$	$\theta = 0$
All Observed		483	520	490	510	489
YV-Conditional	Yes	416	482	470	490	464
V-Conditional	Yes	410	473	460	490	464
Y-Conditional	Yes	392	466	447	486	470
Unconditional	Yes	345	449	454	480	466
Y-Conditional	No	143	150	157	183	156
Unconditional	No	154	148	114	106	95
MCAR	-	100	100	100	100	100

Rel. Efficiencies of Wtd compared with Mle with a Single Covariate

Sampling Design	Analysis uses V	X-Effect $\theta = 2$	$\theta = 1$	$\theta = 0.5$	$\theta = 0.3$	$\theta = 0$
All Observed	-	100	100	100	100	100
YV-Conditional	Yes	62	72	67	69	99
V-Conditional	Yes	83	81	85	80	78
Y-Conditional	Yes	43	39	45	53	48
Unconditional	Yes	64	64	61	61	63
Y-Conditional	No	79	67	64	60	64
Unconditional	No	94	90	96	97	102
MCAR	-	100	100	100	100	100

Here, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma U$

- Y Response observable for all
- X_1 N(0,1), observed for subsample
- V Discretization of X_1, observable for all
- X_2 Indep N(0,1) observed for subsample
- X_3 Indep Binary observed for subsample

X-Effects	Analysis . Sampling	Percentage possible gain (Mle)			(Wtd)			Rel. Eff. (Wtd/Mle)		
		X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
Strong	yvStrat . vCond	69	28	19	66	13	18	95	86	99
	yvStrat . yCond	28	24	-12	-26	-24	-76	65	68	68
	yStrat . yCond	10	17	-10	-45	-29	-74	68	70	68
	yvStrat . UnCond	29	34	4	25	20	4	97	87	99
	yStrat . UnCond	14	18	1	14	16	4	100	98	103
...	.....	..	..	..	..					

X-Effects	Analysis . Sampling	Percentage possible gain (Mle)			(Wtd)			R.E. (Wtd-Mle)		
		X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
Strong	yvStrat . vCond	69	28	19	66	13	18	95	86	99
	yvStrat . yCond	28	24	-12	-26	-24	-76	65	68	68
	yStrat . yCond	10	17	-10	-45	-29	-74	68	70	68
	yvStrat . UnCond	29	34	4	25	20	4	97	87	99
	yStrat . UnCond	14	18	1	14	16	4	100	98	103
Weak	yvStrat . vCond	99	-1	1	92	3	4	80	103	102
	yvStrat . yCond	99	45	51	73	-10	9	50	60	64
	yStrat . yCond	54	43	50	17	-7	12	66	63	66
	yvStrat . UnCond	99	-3	-4	82	-1	-1	60	102	102
	yStrat . UnCond	-3	-4	-7	-1	-1	-2	101	103	104
Weak . X_1	yvStrat . vCond	73	26	-1	71	18	-2	95	93	100
	yvStrat . yCond	8	30	-1	-53	-12	-44	65	69	75
	yStrat . yCond	-11	31	0	-73	-7	-44	69	71	74
	yvStrat . UnCond	31	26	2	27	17	1	96	91	99
	yStrat . UnCond	-4	28	6	-1	23	4	103	95	99
Weak . other	yvStrat . vCond	96	8	1	89	10	1	78	101	100
	yvStrat . yCond	93	15	7	53	-35	-51	42	68	67
	yStrat . yCond	38	19	8	8	-31	-49	74	67	67
	yvStrat . UnCond	92	-3	-3	70	1	0	58	103	103
	yStrat . UnCond	42	-1	-1	36	0	-1	92	101	100
Medium	yvStrat . vCond	93	11	-2	88	2	-2	85	93	100
	yvStrat . yCond	88	30	44	48	-11	-1	47	71	64
	yStrat . yCond	42	31	45	-12	-12	2	60	69	65
	yvStrat . UnCond	89	8	-1	71	6	-1	66	98	100
	yStrat . UnCond	9	5	-1	7	5	0	98	100	101

Lessons

- Most info. on X_1-coef can be recovered with a 20% subsample
  - together with a surprising amount of information on the unrelated variables in some situations
    - Largely (but not exclusively) limited to SPML and Y-conditional sampling
- More information recovered with V-conditional sampling than Y-conditional or unconditional sampling
  - Y-conditional not competitive when X_2 and X_3 effects strong
- Weighted analyses are bad news with Y-conditional sampling
- but have good efficiency with V-conditional sampling - never lost more than 20%

But what about robustness to error model?

- Generate data using one distribution
- Analyse data using a different one
- Used
  - Normal
  - Logistic
  - t_6 (heavy tails)
  - χ^2_{10} (skewed)
- Look at coverage of CIs of regression coefficients

Generate as t_6, solve Logistic

X-Effects	Sampling	% Coverage (Mle)			% Coverage (Wtd)		
		X_1	X_2	X_3	X_1	X_2	X_3
Strong	yvStrat.vCon	96	92	95	96	95	96
	yvStrat.yCon	89	94	92	94	95	95
	yStrat.yCon	95	95	94	95	95	94
	yvStrat.UnCon	94	93	95	95	94	96
	yStrat.UnCon	92	94	95	95	95	95
Weak	yvStrat.vCon	94	96	95	93	96	94
	yvStrat.yCon	94	94	95	94	94	95
	yStrat.yCon	95	95	95	95	93	96
	yvStrat.UnCon	94	94	95	94	94	95
	yStrat.UnCon	94	95	94	94	94	95
Weak.X_1	yvStrat.vCon	96	93	95	96	94	95
	yvStrat.yCon	95	95	94	96	95	95
	yStrat.yCon	96	95	93	95	95	95
	yvStrat.UnCon	95	93	94	95	94	95
	yStrat.UnCon	95	94	94	96	96	94
Weak.other	yvStrat.vCon	95	96	94	96	95	95
	yvStrat.yCon	95	94	96	95	94	95
	yStrat.yCon	95	93	95	96	94	95
	yvStrat.UnCon	95	96	95	95	96	96
	yStrat.UnCon	93	96	94	95	96	95
Medium	yvStrat.vCon	94	93	96	95	95	96
	yvStrat.yCon	95	92	94	95	95	94
	yStrat.yCon	94	94	94	95	94	94
	yvStrat.UnCon	96	94	95	96	95	96
	yStrat.UnCon	95	95	96	96	95	96

Generate Normal, solve Logistic

X-Effects	Sampling	% Coverage (Mle) X_1	X_2	X_3	(Wtd) X_1	X_2	X_3
Strong	yvStrat.vCon	94	92	94	95	94	96
	yvStrat.yCon	88	91	91	94	95	95
	yStrat.yCon	93	94	93	95	96	95
	yvStrat.UnCon	93	93	94	95	95	95
	yStrat.UnCon	91	90	93	95	94	95
Weak	yvStrat.vCon	95	90	90	96	94	94
	yvStrat.yCon	95	92	94	94	96	95
	yStrat.yCon	92	93	94	93	95	96
	yvStrat.UnCon	96	88	90	95	94	95
	yStrat.UnCon	92	88	90	96	94	94
Weak.X_1	yvStrat.vCon	95	92	94	96	94	95
	yvStrat.yCon	93	91	93	94	96	95
	yStrat.yCon	92	91	93	94	96	94
	yvStrat.UnCon	94	92	94	95	95	96
	yStrat.UnCon	94	94	94	96	96	96
Weak.other	yvStrat.vCon	93	92	93	94	95	96
	yvStrat.yCon	93	92	91	95	95	93
	yStrat.yCon	90	94	93	96	95	93
	yvStrat.UnCon	96	92	91	96	94	95
	yStrat.UnCon	94	94	94	95	94	95

But ... Generate Normal, solve Logistic cont.

X-Effects	Sampling	Coverage (%))					
		(Mle)			(Wtd)		
		X_1	X_2	X_3	X_1	X_2	X_3
Medium	yvStrat.vCon	94	84	89	95	95	93
	yvStrat.yCon	82	49	79	95	95	94
	yStrat.yCon	74	72	87	95	95	94
	yvStrat.UnCon	94	83	90	93	95	95
	yStrat.UnCon	91	91	94	94	95	95

Generate χ^2_{10}, solve Logistic

X-Effects	Sampling	% Coverage (Mle)			(Wtd)		
		X_1	X_2	X_3	X_1	X_2	X_3
Strong	yvStrat.vCon	95	77	91	95	95	96
	yvStrat.yCon	84	56	82	95	94	95
	yStrat.yCon	69	70	87	94	94	96
	yvStrat.UnCon	94	86	90	95	96	93
	yStrat.UnCon	90	92	91	95	96	93
Weak	yvStrat.vCon	95	93	91	96	95	95
	yvStrat.yCon	95	94	93	95	96	94
	yStrat.yCon	94	94	92	94	96	94
	yvStrat.UnCon	95	94	92	95	96	95
	yStrat.UnCon	93	93	92	95	96	95
Weak.X_1	yvStrat.vCon	95	78	90	95	94	95
	yvStrat.yCon	95	37	77	94	94	94
	yStrat.yCon	93	38	77	93	94	94
	yvStrat.UnCon	94	84	91	95	94	95
	yStrat.UnCon	92	84	91	94	95	95
Weak.other	yvStrat.vCon	95	94	92	95	97	96
	yvStrat.yCon	94	94	93	94	95	95
	yStrat.yCon	28	94	93	95	95	95
	yvStrat.UnCon	94	93	92	95	95	94
	yStrat.UnCon	84	93	92	95	94	94
Medium	yvStrat.vCon	93	90	93	93	94	95
	yvStrat.yCon	93	81	90	93	96	94
	yStrat.yCon	81	82	90	95	96	94
	yvStrat.UnCon	93	91	91	94	95	95
	yStrat.UnCon	91	91	93	95	96	95

Lessons

- Weighted method coverage always fine
  - BUT critical that standard errors allow for randomness in the weights
- The robustness probs with SemiPar ML occurred with Y-conditional sampling
  - Which is where it is much more efficient than weighting

Tentative conclusions

- Programme already allows for regression modelling of scale factor
- Could allow parameters in error model
- But in the absence of that ...
- Sample on a V-variable and use a weighted analysis
- Use a V-variable related to a variable of particular interest even if the sample was obtained unconditionally
- If sampling has been on Y,
  - do both analyses and treat conclusions that can only be obtained from SemiPar ML as hypothesis-generation rather than confirmatory

Important qualification

- Work less use for observational studies than designed studies
- Why?
  - Lack of comparability of population/1st phase information and sub-sample information
  - e.g. birthweight study. Detailed exclusion criteria applied to study individuals excluded significant numbers. No idea how non-study births would have fared