

A new model to deal with design missingness in survey studies using auxiliary information

George Vamvakas

Biostatistics and Health Informatics, King's College London

2021 Stata Conference
September 9th-10th, 2021

Goals and contributions

Estimation of population parameters from two-stage surveys, where:

- first stage involves the use of an inexpensive and easy to use diagnostic tool to screen a large population (the target population).
- second stage proceeds with more intensive assessments of a sub-population.
 - Sub-population selection based on the diagnostic tool.

Traditionally, analyses of sub-population data deploy weights to make inferences for the target population.

We develop a model that incorporates the diagnostic tool as an auxiliary variable, to obtain unbiased and more efficient population parameters than a weighted model.

Backdrop - the motivational study

SCALES - a population based survey on language development

- **Screening phase:** all state schools in Surrey invited to take part - 61% participated (n=263 schools).
 - Background data on 7267 children who began reception class in 2011 (aged between 4 yrs 9 mths and 5 yrs 10 mths).
 - This included 'initial' diagnostic data on language ability: **CCC-S** scores.
- **Intensive phase:** a sub-sample of children, selected for detailed assessment.
 - 636 were selected.
 - Selection based (primarily) on the CCC-S questionnaire
 - So far, we have 3 measurements per child from (at least) six in-depth questionnaires for expressive and receptive language skills:
 - School year 1: 528/636 kids.
 - School year 3: 499/528 kids.
 - School year 6: 386/499 kids.

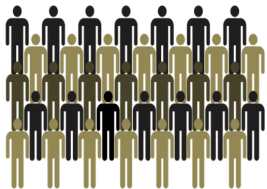
Backdrop - the selection criterion

The Children's Communication Checklist-Short (CCC-S) questionnaire

- Shown to be as effective as standardised assessment in identifying children at risk for clinically significant language impairment
- Scores range from 0-39 (discrete variable): 39 is indicative of greatest communication deficit.
- Used to identify low- and high-risk children.
- Low-risk children had a lower sampling fraction.

Backdrop - inference

Full (target) population
Screened for language ability
using primarily the CCC-S.

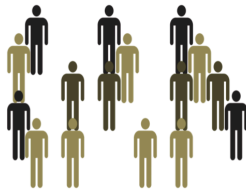


Selection using CCC-S as
the dominant criterion



**Sub-population (not
representative of the target
population, as it stands)**

Collected data on: vocabulary
(expressive/receptive), grammar
(expressive/receptive), narrative
recall, narrative comprehension.



Goal: inference



Use of weights

Typically used to control for the unequal probability of selection and produce estimates for the target population.

We've previously used weights to produce population charts using the LMS method (Vamvakas et al. (2019)).

$W = W(\text{design factors})$

- Design factors=risk of impairment (measured by CCC-S qaire), number of pupils screened
 - Also allowed for unit non-response.

Use of weights

Known disadvantages of weights:

- Probability-weighted estimators are generally inefficient compared to unweighted estimators.
 - Use complete data cases only.
- Many standard inferential procedures, such as the *likelihood ratio test*, are not applicable alongside probability weighting.

Use of auxiliary information

Auxiliary variable(s) as an alternative way to control for unequal probability of selection.

Borrowing one of Collins et al's (2001) definitions, an **auxiliary** variable is a variable that correlates with the partially observed measure of interest and with the determinants of the 'missingness' mechanism.

Use of auxiliary variables in statistical models

A lot of evidence in terms of unbiased and more efficient parameters compared with weighted estimators.

However,

- Most evidence comes from the multiple imputation setting.
- How do we put them in use under maximum likelihood?
 - Not aware of how non-Normally distributed auxiliary variables would behave in this context.

Incorporation of auxiliary information under ML

Graham (2003)

Extra DV model:

- AV predicted by X
- Residual AV associated with residual Y

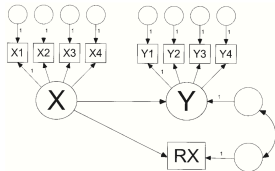


FIGURE 4 Model 2: "extra DV" model (latent variable version).

Saturated Correlates model

- AV associated with residual X and residual Y .
- If X and Y exogenous, AV associated directly.

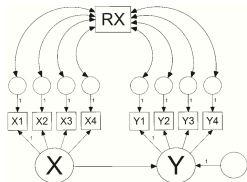
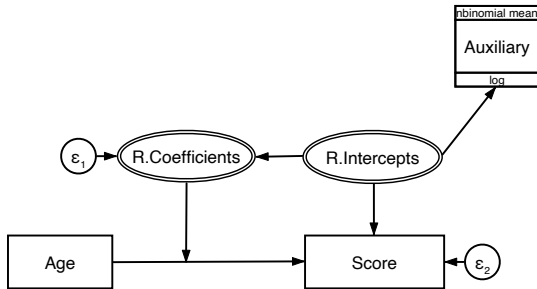


FIGURE 6 Model 3: "saturated correlates" model (latent variable version).

Our model

Vamvakas, Courtenay and Pickles (2021)



```
gsem (Score <- c.Age F1[id]@1 c.Age#F2[id]@1) ///  
(Auxiliary <- F1[id], fam(nbin) link(log)) ///  
(F2[id] <- F1[id]), iterate(50)
```

Performance

Tested in simulations and applied to real (SCALES) data.

Simulations:

- 1000 datasets were created, each containing 2000 subjects.
- Data distributions were based on parameters from SCALES.
 - Each subject had 3 repeated measurements.
- Approximately 50% of subjects were missing in each dataset.
 - If missing at baseline then missing from subsequent time points.
- Probability of being missing dependent on the auxiliary variable.

Performance - simulated data

	id	Score	Age	Auxiliary	
1	1	.	.	0	
2	1	.	.	.	
3	1	.	.	.	
4	2	52.86395	-34.87704	0	
5	2	70.15619	-11.27158	.	
6	2	102.3556	27.45946	.	
7	3	82.40202	-26.44148	2	
8	3	83.65213	-2.836022	.	
9	3	135.6448	35.89502	.	
10	4	.	.	1	
11	4	.	.	.	
12	4	.	.	.	
13	5	72.30627	-30.35404	2	
14	5	102.4044	-6.748582	.	
15	5	126.8928	31.98245	.	
16	6	.	.	2	
17	6	.	.	.	
18	6	.	.	.	
19	7	67.71524	-27.37874	0	
20	7	62.21595	-3.773278	.	
21	7	103.753	34.95776	.	

Performance - extract from simulation results

- The proposed models (PoAux and NbAux) were compared against a weighted growth model (and several other models).
 - PoAux: assumes the auxiliary variable is Poisson distributed
 - NbAux: assumes the auxiliary variable is Negative Binomial distributed
- Performance measures include bias, efficiency and coverage.
- Tested parameters:
 - Fixed intercept
 - Fixed effect of age
 - Random intercepts
 - Random coefficients
 - Covariance of random effects
- See Vamvakas, Courtenay and Pickles (2021) for details.

Performance - bias

Table 1 Bias (and Monte Carlo standard error) of parameters by model

	Fixed intercept	Fixed Age	R.Intercepts	R.Coefficients	R.Covariance
Complete	-0.0015 (0.0103)	-0.0000 (0.0002)	-0.0769 (0.2163)	0.0000 (0.0001)	-0.0013 (0.0021)
Naive	2.7674 (0.0152)	0.0056 (0.0002)	19.4979 (0.3273)	0.0001 (0.0001)	0.0349 (0.0034)
Weighted	-0.0175 (0.0136)	0.0003 (0.0002)	-0.1934 (0.3014)	0.0001 (0.0001)	-0.0042 (0.0032)
PoAux	-0.0062 (0.0132)	0.0003 (0.0003)	0.1243 (0.2835)	0.0000 (0.0001)	-0.0047 (0.0030)
TPoAux	0.1947 (0.0133)	0.0006 (0.0002)	14.8592 (0.3140)	0.0001 (0.0001)	0.0255 (0.0033)
PoMiss	-2.9330 (0.0148)	-0.0054 (0.0002)	42.4033 (0.3852)	0.0003 (0.0001)	0.0778 (0.0037)
TPoMiss	-2.2639 (0.0167)	-0.0042 (0.0002)	32.7086 (0.3604)	0.0002 (0.0001)	0.0603 (0.0036)
NbAux	0.0214 (0.0160)	0.0003 (0.0003)	0.2479 (0.3468)	0.0001 (0.0001)	-0.0070 (0.0037)

Values highlighted in red denote significant biases, as defined in the text. The models are abbreviated as follows: *Complete* the complete data model, *Naive* model 1 fitted to incomplete data, *Weighted* the Weighted model, *PoAux* the Poisson auxiliary model, *TPoAux* the transformed Poisson auxiliary model, *PoMiss* the Poisson/missingness model, *TPoMiss* the transformed Poisson/missingness model, *NbAux* the negative binomial auxiliary model

Performance - efficiency

Table 2 % increase (decrease) in precision (and Monte Carlo standard error) relative to the complete data model

	Fixed intercept	Fixed Age	R.Intercepts	R.Coefficients	R.Covariance
Complete
Naive	-53.9224 (2.0275)	-52.2847 (2.1722)	-56.3302 (1.8290)	-51.3789 (2.1773)	-59.2077 (1.7561)
Weighted	-42.8771 (2.3390)	-56.3795 (2.0384)	-48.4894 (2.2415)	-52.8240 (2.1576)	-55.2589 (2.0031)
PoAux	-39.1988 (2.3750)	-55.1849 (2.0828)	-41.7869 (2.3549)	-49.9521 (2.2659)	-49.7765 (2.1854)
TPoAux	-40.2721 (2.5402)	-55.1450 (2.0816)	-52.5498 (1.9508)	-51.0450 (2.2035)	-57.1591 (1.8617)
PoMiss	-51.7732 (2.7294)	-61.0153 (1.9040)	-68.4668 (1.4196)	-50.5739 (2.2350)	-66.2951 (1.4668)
TPoMiss	-61.8300 (1.8261)	-60.1179 (1.9233)	-63.9754 (1.5789)	-51.3581 (2.1897)	-63.8080 (1.5764)
NbAux	-39.0280 (2.9616)	-54.2624 (2.5905)	-42.5276 (2.8291)	-49.6582 (2.8349)	-49.4564 (2.7454)

Values highlighted in red indicate the largest difference from the complete data model by parameter. The models are abbreviated as follows: *Complete* the complete data model, *Naive* model 1 fitted to incomplete data, *Weighted* the Weighted model, *PoAux* the Poisson auxiliary model, *TPoAux* the transformed Poisson auxiliary model, *PoMiss* the Poisson/missingness model, *TPoMiss* the transformed Poisson/missingness model, *NbAux* the negative binomial auxiliary model

Performance - coverage

Table 4 Coverage of model (and Monte Carlo standard error) by parameter

	Fixed intercept	Fixed Age	R.Intercepts	R.Coefficients	R.Covariance
Complete	94.5 (0.7209)	94.3 (0.7332)	95.1 (0.6826)	95.0 (0.6892)	96.1 (0.6122)
Naive	0.0 (0.0000)	87.3 (1.0529)	57.2 (1.5647)	94.2 (0.7392)	94.6 (0.7147)
Weighted	96.7 (0.5649)	93.4 (0.7851)	94.9 (0.6957)	95.0 (0.6892)	95.1 (0.6826)
PoAux	95.0 (0.6892)	93.4 (0.7851)	95.3 (0.6693)	95.0 (0.6892)	94.8 (0.7021)
TPoAux	92.9 (0.8122)	93.3 (0.7906)	68.2 (1.4726)	94.2 (0.7392)	94.5 (0.7209)
PoMiss	0.0 (0.0000)	87.7 (1.0386)	4.4 (0.6486)	94.7 (0.7085)	90.8 (0.9140)
TPoMiss	0.6 (0.2442)	89.4 (0.9735)	12.4 (1.0422)	94.4 (0.7271)	92.4 (0.8380)
NbAux	94.7 (0.8624)	92.9 (0.9864)	95.3 (0.8156)	94.5 (0.8736)	94.5 (0.8736)

Values highlighted in red represent extremely low rates of coverage. The models are abbreviated as follows: *Complete* the complete data model, *Naive* model 1 fitted to incomplete data, *Weighted* the Weighted model, *PoAux* the Poisson auxiliary model, *TPoAux* the transformed Poisson auxiliary model, *PoMiss* the Poisson/missingness model, *TPoMiss* the transformed Poisson/missingness model, *NbAux* the negative binomial auxiliary model

Application to SCALES data

We use our model to construct language percentiles and growth charts for the **target** population.

For this, no data are needed - only parameters from the model:

$$C_{age} = \mathbf{X}_i \mathbf{b} + K \sqrt{\text{var}(y_{ji} | \mathbf{X}_i)}$$

where:

C_{age} : centile value at a specific age.

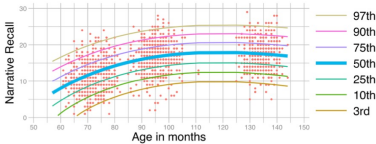
$\mathbf{X}_i \mathbf{b}$: fixed part of the model.

K : a value from the inverse cumulative Standard Normal distribution.

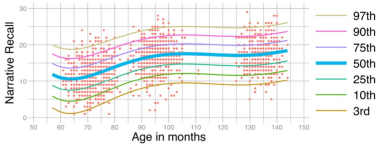
$\text{var}(y_{ji} | \mathbf{X}_i)$: the conditional variance of the total residual of the growth model.

Population growth charts

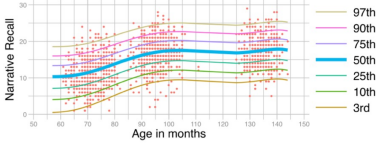
3rd degree polynomial in age



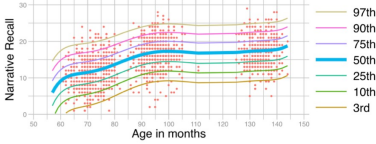
5th degree polynomial in age



6th degree polynomial in age



7th degree polynomial in age









Advantages of our auxiliary variable model

- 1 Inferences can be made for the full target population.
- 2 More efficient estimates than a weighted estimator.
- 3 Avoids the use of weights.
- 4 Flexible construction of norms, based only on model parameters.
 - Z-scores
 - Percentiles
 - Velocities
- 5 Main growth model can be treated as any other regression based model.
 - Use of polynomial terms.
 - Standard statistical tests apply, such as the likelihood ratio test.

Acknowledgements

- Prof. Andrew Pickles, King's College London
- Prof. Courtenay Norbury, University College London
- Participants and their families
- Wellcome
- NIHR Biomedical Research Centre (BRC) at South London and Maudsley NHS Foundation Trust and Kings College London

References

-  Vamvakas, G., Norbury, C., Pickles, A. (2021) *Two-stage sampling in the estimation of growth parameters and percentile norms: sample weights versus auxiliary variable estimation*, BMC Medical Research Methodology, 21(173)
-  Vamvakas, G., et al. (2019) *Standardizing test scores for a target population: The LMS method illustrated using language measures from the SCALES project*, PLoS ONE, 14(3)
-  Norbury, C., et al. (2016) *The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study*, J. Child Psychol. Psychiatry, 57(11), p. 124757.
-  Graham, J. W. (2003). *Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models*, Structural Equation Modeling, 10(1), p.80-100.
-  Collins, L. M., Schafer, J. L., & Kam, C.M. (2001). *A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures*, Psychological Methods, 6(4), p.330-351.
-  Holt, D., Smith, T., & Winter, P. (1980). *Regression Analysis of Data from Complex Surveys*, Journal of the Royal Statistical Society. Series A (General), 143(4), p.474-487.