Optimal Sampling Design is Sensitive to Model Misspecification

Thomas Lumley
University of Auckland, New Zealand
and Tong Chen
MCRI, Melbourne, Australia

2023–11–30

Two-phase studies

We have some data (‘phase I’)

We can get new variables or better measures of old variables on a subsample of the same people.

We want to estimate some regression parameters and try to get the same answer as if we measured everyone

Problem

In two-phase studies without non-response we have model-based and design-based estimators.

- Model-based estimators are optimal if the outcome model is correct
- Design-based ('raking') estimators are optimal if the outcome model is modestly misspecified

They sometimes imply very different optimal designs

Designs sometimes similar

Logistic regression in case–control sampling, both design-based and model-based

- 1:1 case–control ratio is optimal for small β
- more controls is optimal for large β

(probably not identical, but qualitatively similar)

Designs sometimes differ

For linear regression:

- model-based estimator optimality: sampling **extremes**
- design-based estimator optimality: sampling **everywhere**

We know the transition happens over quite small amounts of model misspecification ($O_p(n^{-1/2})$).

What does it look like?

Example: big difference for MLE at truth

Fitted outcome model:

$$Y = \beta_0 + \beta_1 X + N(0,1)$$
$$A = X + N(0,1)$$

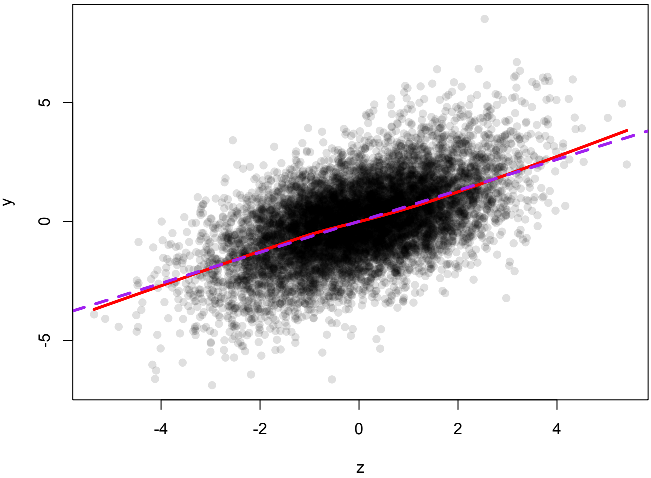
True generative model: Y is linear spline in X with knots at ± 1.

Sampling model: sampling from 10 strata at deciles of A, total 10% (and extreme tail sampling for MLE only)

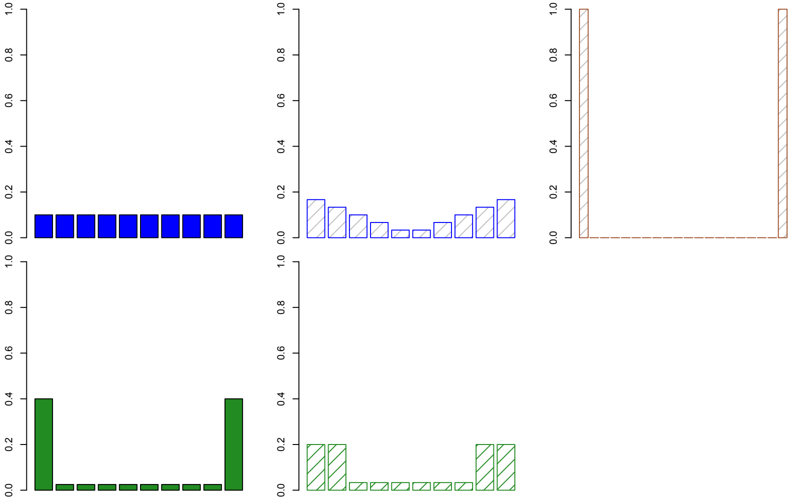
Target of inference: $\hat{\beta}_1$ estimated in full cohort

Estimators: IPW, parametric MLE based on subsample

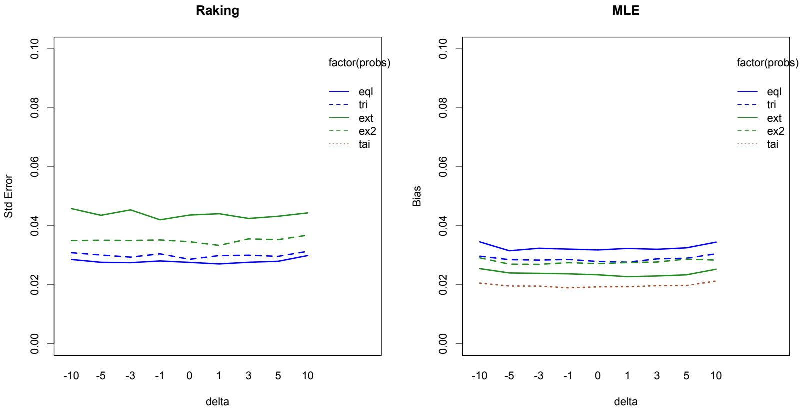
The data: largest misspecification



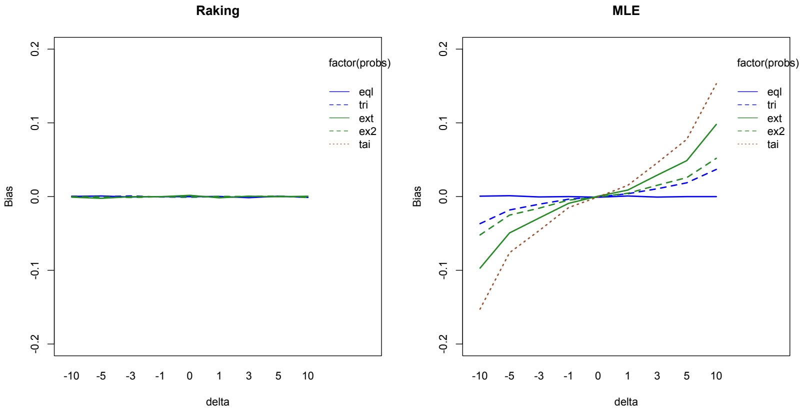
Sampling patterns



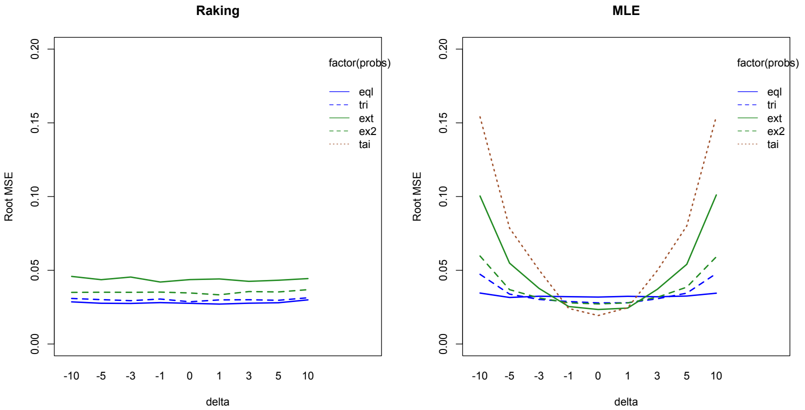
Standard error

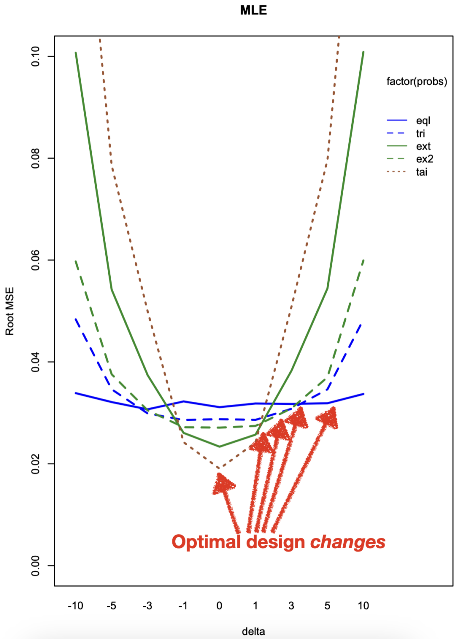


Bias


Raking
MLE
factor(probs)
eql
tri
ext
ex2
tai
Bias
0.2
0.1
0.0
-0.1
-0.2
-10
-5
-3
-1
0
1
3
5
10
delta

RMSE


Raking
MLE
Root MSE
0.00
0.05
0.10
0.15
0.20
-10
-5
-3
-1
0
1
3
5
10
delta
factor(probs)
eql
tri
ext
ex2
tai


MLE
Root MSE
0.00
0.02
0.04
0.06
0.08
0.10
delta
-10
-5
-3
-1
0
1
3
5
10
factor(probs)
eql
tri
ext
ex2
tai
Optimal design *changes*

Summary

- Optimal design for model-based estimator becomes less extreme with even slight misspecification
- Optimal design for design-based estimator stays roughly the same
- Extreme-sampling design for model-based estimator is quite sensitive to model specification

Example: small difference for MLE at truth

Fitted outcome model:

$$Y = \beta_0 + \beta_1 X + N(0, 1)$$

$$A = X + N(0, 1)$$

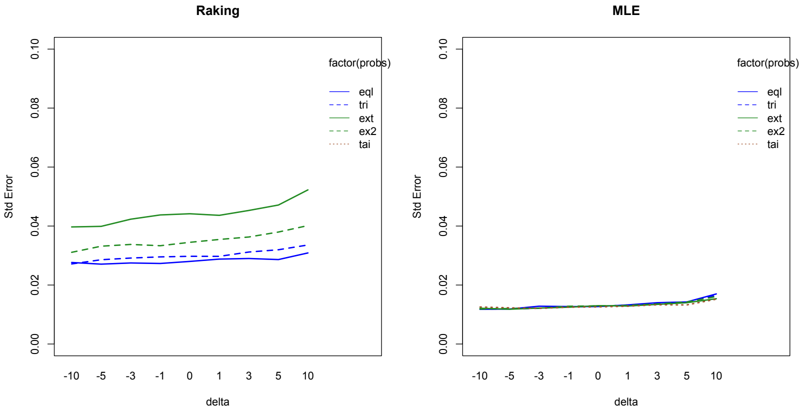
True generative model: Y is linear spline in X with knots at ± 1, biased measurement error in A: $E[A|X = x] = (1 - \gamma)x$

Sampling model: sampling from 10 strata at deciles of A, total 10% (and tail sampling and extreme residual sampling for MLE only)

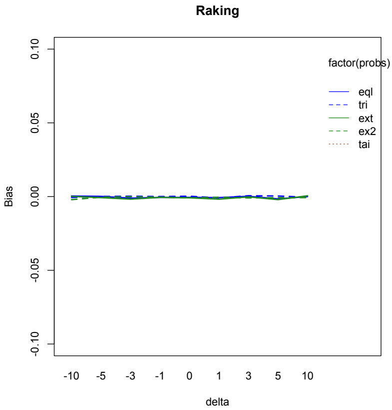
Target of inference: $\hat{\beta}_1$ estimated in full cohort

Estimators: Raking/AIPW, parametric MLE based on subsample

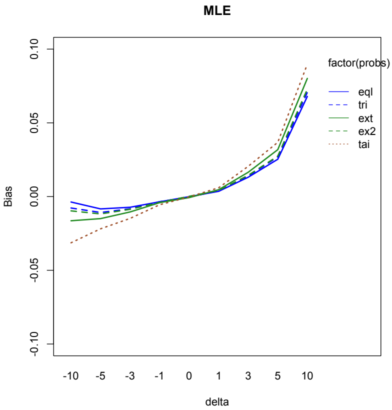
Standard error



Bias

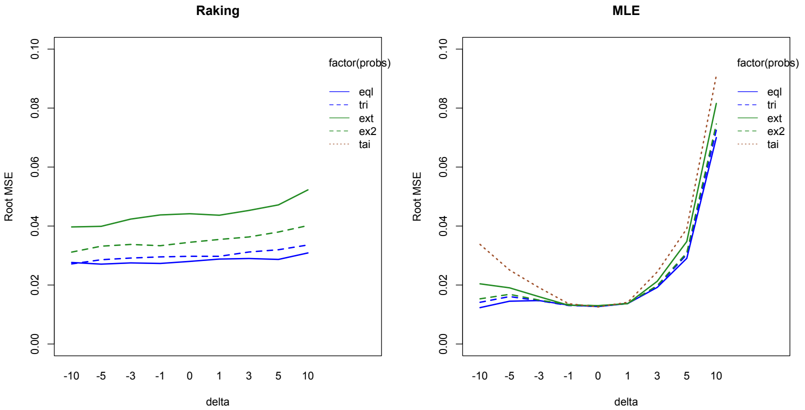


Raking
Bias
0.10
0.05
0.00
-0.05
-0.10
factor(probs)
eql
tri
ext
ex2
tai
-10
-5
-3
-1
0
1
3
5
10
delta



MLE
Bias
0.10
0.05
0.00
-0.05
-0.10
factor(probs)
eql
tri
ext
ex2
tai
-10
-5
-3
-1
0
1
3
5
10
delta

RMSE


Raking
MLE
Root MSE
0.00
0.02
0.04
0.06
0.08
0.10
factor(probs)
eql
tri
ext
ex2
tai
-10
-5
-3
-1
0
1
3
5
10
delta

Summary

- As model misspecification increases, designs become more different for MLE
- Optimal design for design-based estimator stays roughly the same
- Efficient design for raking is also more robust for MLE (less dramatically)

Conclusions

- Design optimality can be quite sensitive to model specification
- Designs that are good for the raking estimator seem to be more robust to model misspecification
- That's how raking and MLE-optimal designs converge under model misspecification
- If you're going to optimise, it's worth checking under misspecification

Conjecture: something like this is true more generally for the worst-case misspecification direction and the best raking estimator (tricky to prove for designs with zero sampling probabilities)

Questions?

Weka, by Giselle Clarkson