Discussion of “Data Defect Correlation: A Unified Quality Metric for Probabilistic Sample and Non-Probabilistic Sample” by XL Meng

Rod Little

Department of Biostatistics and Institute for Social Research

University of Michigan
Large vs Little

In the crimson corner:
Representing Harvard, AKA Michigan of the East):
Distinguished JPSM Speaker “Xtra Large” Meng

In the maize and blue corner:
Representing Michigan, home of survey gods Leslie Kish and Bob Groves:
Proposer of the Vote of Thanks (ha ha!)
“Ol’ Fogey” Little, a dyed in the wool survey statistician
The topic


See also XL’s discussion of “Perils and potentials of self-selected entry to epidemiological studies and surveys” by Keiding and Louis (2016 JRSSB)

The (very topical) topic: Bias/variance tradeoffs in selecting a sample from a population

Random (“scientific”) sampling is somewhat under siege from the forces of “big data.”
Centrality of the random sampling assumption in statistics

From my comment on “Models as Approximations 1: Consequences Illustrated with Linear Regression” by A. Buja et. al. (2019 Statist. Sci.):

“Buja et al. state a key random sampling assumption (italics mine):

‘In fact, it may rely on no more than the assumption that the rows \((y_i, x_i)\) of the data matrix … are iid samples from a joint multivariate distribution subject to some technical conditions.’

This assumption, routine in much of mathematical statistics, is both crucial – in some respects it trumps all the other assumptions in the statistical model – and often very questionable. If, as is usual, units are not selected by simple random sampling, this is an assumption, and if violated then estimates under model-trusting or model-robust paradigms -- are subject to unknown biases.”
Covid 19

• Keiding and Louis (2016 JRSSA with discussion) argue for the importance of probability sampling in epidemiologic modeling
• There are no estimates of prevalence of Covid 19 in the US based on scientific (probability) sampling
• Applications of SIR models make the questionable assumption that reported prevalences are unbiased for true prevalences
  – But reported prevalences are vulnerable to selection bias
  – Testing is not carried out on a random sample of target population
• This assumption is often buried relative to other assumptions like homogeneous transmission, measurement error, etc.
• UM proposed a national probability survey but not funded
The broad question of the Meng paper

What’s the trade-off between probability sampling (scientific but hard and expensive) and “big data” (large sample, but high potential for selection bias)?
Meng’s Law of Large Populations (LLP)

\[ \bar{G}_N, \sigma_G = \text{target population mean, sd of survey variable } G; \]
\[ \bar{G}_n = \text{sample estimate} \]
\[ R = \text{sample indicator; } \rho_{R,G} = \text{corr}(R,G); \]
\[ \text{LLP: } \bar{G}_n - \bar{G}_N = \rho_{R,G} \times \sqrt{\frac{1-f}{f}} \times \sigma_G \]

Various equivalent identities can be constructed from the components – the question is whether LLP is the most meaningful way of doing it.

The more important practical issue is that key quantities in comparing prob samples and big data are generally not known.
Terminology matters

LLP: \( \bar{G}_n - \bar{G}_N = \rho_{R,G} \times \sqrt{\frac{1-f}{f}} \times \sigma_G \)

Error = Data quality \( \times \) Data Quantity \( \times \) Problem Difficulty

• I’d say \( \rho_{R,G} \) is a measure of selection bias, not data quality -- “Data quality” in surveys generally refers to measurement error

• A more natural measure of data quantity is simply \( n \)

• A more standard term for \( \) is heterogeneity – “problem difficulty” could mean the conceptual difficulty of the quantity under study (e.g. Annual Wages vs Wealth)

• See e.g. familiar decompositions of “total survey error”
Points of Agreement with XL

• It’s an identity – not disputing the math

• (I) “Probabilistic sampling ensures high data quality by controlling $\rho_{R,G}$ at the level of $N^{-1/2}$”
  – (true, though a simpler statement is that probability sampling eliminates *selection bias*)

• (IV) “When combining data sources for population inferences, those relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.”

• RMSE from a large data set subject to selection bias can be equivalent to RMSE for a much smaller random sample
Points of disagreement

• XL elevates $\rho_{R,G}$ to the status of a universal quantity -- I am not convinced
  – I don’t find the correlation particularly interpretable. It’s bounded and dimensionless, but unconnected with the substance of the survey – can you distinguish $\rho_{R,G} = 0.01$ and $\rho_{R,G} = 0.001$?
  – When applied to sampling, this correlation depends on sampling fraction $f = n/N$. Fixing $\rho_{R,G}$ implies fixing $f$, and then $n$ is tied to $N$.
  – E.g. nonresponse with response rate $r/n$: correlation based on $r/n$ is different from correlation based on $r/N$, but size of $n/N$ is not important provided it’s small.
  – XL’s two stage application somewhat addresses this.
Points of disagreement

• XL draws the implication that the key quantity in the trade-off between random sampling and big data is $N$, rather than the conventional view that the key quantity is $n$. For example, two XL quotes:

“For population inferences, a key “policy proposal” of the current paper is to shift from our traditional focus on assessing probabilistic uncertainty (e.g., in a sample mean) in the familiar form of Standard Error $\propto \sigma / \sqrt{n}$ to the practice of ascertaining systematic error in non-probabilistic Big Data captured by Relative Bias $\propto \rho^* \sqrt{N}$.”

“[LLP] implies that once we lose control of probabilistic sampling, then the driving force behind the estimation error is no longer the sample size $n$, but rather the population size $N$.”
Alternative well-known expressions

$$RMSE = \sqrt{\left(1 - \frac{n}{N}\right)\left(b^2 + \frac{\sigma^2}{n^*}\right)}$$

$b = \text{bias (often somewhat independent of } n, N)\n
n^* = \text{effective sample size (} n^* = n \text{ for srs)}$

When $n^*$ is large, bias predominates

When $n^*$ is small, precision $\sigma / \sqrt{n^*}$ matters

Note: there is no $N$ here, providing $f = n / N$ is not substantial
Random survey with nonresponse vs big data

For probability survey with nonresponse, \( n / N \) small:

\[
\bar{G}_R = \text{mean of } r \text{ respondents} \\
\text{bias}(\bar{G}_R) = b = (1 - r / n)(\bar{G}_R - \bar{G}_{NR}) \\
\text{RMSE}(\bar{G}_R) = \sqrt{b^2 + \sigma^2 / n^*}
\]

For big data, \( N_S \) out of \( N \) selected, \( N_S \) large

\[
\bar{G}_S = \text{mean of } N_S \text{ selected cases} \\
\text{RMSE}(\bar{G}_S) = \left| \text{bias}(\bar{G}_S) \right| = (1 - N_S / N)\left| \bar{G}_S - \bar{G}_{NS} \right|
\]

\( N \) is not a key quantity in either of these expressions

Can adjust survey means for variables measured for whole sample

With survey nonresponse, key issue is relative sizes of the survey bias and the big data bias
Nonignorable Nonresponse models

- Pattern-mixture models: factor G|R,X
- Selection models: R|G,X
- XL’s formula is similar to a selection model – the correlation is analogous to the correlation between R and G in the Heckman model for selectivity bias
- My formula for bias is analogous to a pattern-mixture model – compares the mean of G in respondent and nonrespondent populations
- For sensitivity analysis for nonresponse that is missing not at random, pattern-mixture models are easier to implement and (in my view) easier to interpret for non-statisticians.
- See for example, Proxy Pattern-Mixture Analysis (Andridge and Little 2011), chapter 15 of Little and Rubin (2019)
Measures of departure from ignorable selection

• Popular existing measures based on propensity, like the R-indicator, assume selection at random and do not reflect the association between selection and the survey variable in XL’s formula.

• Little et al. (2020) develop indices and corrections for selection bias of non-probability samples based on proxy pattern-mixture models. Extensions have been developed for binary outcomes (Andridge et al. 2019) and for linear and probit regression (West et al. 2021)
Conclusion

• XL’s examples have helped to highlight the dangers of small selection biases in large data sets, but…

• His “law of large populations” is treats $\rho_{R,G}$ as a form of “universal constant” – I think there are simpler existing approaches to the problem

• I disagree with the implication that the population size $N$ is the important quantity in bias/variance trade-offs, rather than the sample size $n$ and fraction of the population covered by the data.
References


