CNSTAT Workshop (January 18-19, 2018):

Improving Health Research for Small Populations

Bayesian Methods for Small Population Analysis

Thomas A. Louis, PhD
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
www.biostat.jhsph.edu/~tlouis/
tlouis@jhu.edu

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Pre-Summary

- Seldom can inferences from small populations stand on their own, because estimates are unstable (have low precision)
  - Also, large-sample assumptions/conveniences may not apply
- Therefore, modeling or other stabilization/(information enhancement) is necessary; strategies include:
  - Aggregation
  - Regression both within and across populations
  - Hierarchical (Bayesian/EB) modeling to 'borrow information' within and between data sources
  - Trimming survey weights

Pre-Summary

- Seldom can inferences from small populations stand on their own, because estimates are unstable (have low precision)
  - Also, large-sample assumptions/conveniences may not apply
- Therefore, modeling or other stabilization/(information enhancement) is necessary; strategies include:
  - Aggregation
  - Regression both within and across populations
  - Hierarchical (Bayesian/EB) modeling to 'borrow information' within and between data sources
  - Trimming survey weights
- Stabilization/enrichment targets include,
  - Estimated regression slopes and residual variances
  - A control group, using historical data
  - Clinical trial subgroup estimates (Henderson et al., 2016)
  - Transporting, e.g., adults $\longrightarrow$ children
  - Small Area (Domain) estimates (SAEs)
  - Estimated SMRs and the challenges of low information
  - Survey weights (Gelman, 2007)
- The Bayesian formalism is effective in meeting these goals

Preview

The Bayesian formalism

- Modern Contraceptive Rates in Uganda
- Estimating rates of bone loss
- Stabilizing variance estimates

Combine, don't pool

- Historical controls in carcinogenicity testing

Making use of Big Data

- Embed a high-resolution study in a larger, lower-resolution one

Design-based inference loosens its grip on the survey world

- Combine survey estimates: SIPP aided by the ACS
- Small Area Income and Poverty Estimates (SAIPE)
- Alternative language determinations as required by Section 203 of the voting rights act

Health Provider Profiling

- Shrinkage/stabilization can be controversial
- The challenges of low information

Closing

Trading off Variance and Bias (for the linear model)

- K units (individuals, clusters, institutions, studies, regions, domains, ...)
- Each with an underlying feature of interest (θ_k):
  - Poverty Rate, Relative Risk, treatment effect, residual variance, ...
- A direct (unbiased) estimate of it (Y_k), with estimated variance $(\hat{\sigma}_k^2)$
- Unit-specific attributes $\mathbf{X}_k$ (tax data, age, exposure) produce,

$$\begin{aligned} \text{regression prediction} &= \hat{\boldsymbol{\beta}} X_k \quad (\text{e.g., } \hat{\beta}_0 + \hat{\beta}_1 X_k) \\ \text{residual} &= Y_k - \hat{\boldsymbol{\beta}} X_k \end{aligned}$$

- Inviting three choices for estimating the θ_k:

Direct:	Use the Y_k (unbiased, but possibly unstable)
Regression:	Use the regression (stable, but possibly biased)
Middle ground:	A weighted average of Regression and Direct

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Trading off Variance and Bias (for the linear model)

- K units (individuals, clusters, institutions, studies, regions, domains, ...)
- Each with an underlying feature of interest (θ_k):
  - Poverty Rate, Relative Risk, treatment effect, residual variance, ...
- A direct (unbiased) estimate of it (Y_k), with estimated variance $(\hat{\sigma}_k^2)$
- Unit-specific attributes $\mathbf{X}_k$ (tax data, age, exposure) produce,

$$\begin{aligned} \text{regression prediction} &= \hat{\boldsymbol{\beta}} X_k \quad (\text{e.g., } \hat{\beta}_0 + \hat{\beta}_1 X_k) \\ \text{residual} &= Y_k - \hat{\boldsymbol{\beta}} X_k \end{aligned}$$

- Inviting three choices for estimating the θ_k:

  Direct: Use the Y_k (unbiased, but possibly unstable)
  Regression: Use the regression (stable, but possibly biased)
  Middle ground: A weighted average of Regression and Direct

$$\begin{aligned} \hat{\theta}_k &= \text{regression prediction} + (1 - \hat{B}_k) \times \text{residual} \\ &= \hat{\boldsymbol{\beta}} \mathbf{X}_k + (1 - \hat{B}_k) \cdot (Y_k - \hat{\boldsymbol{\beta}} \mathbf{X}_k) \\ \hat{B}_k &= \hat{\sigma}_k^2 / (\hat{\sigma}_k^2 + \hat{\tau}^2) \\ \hat{\tau}^2 &= \text{residual/unexplained variance, model lack of fit} \end{aligned}$$

- For general models use the Bayesian formalism (Carlin and Louis, 2009; Gelman et al., 2013; Kadane, 2015)

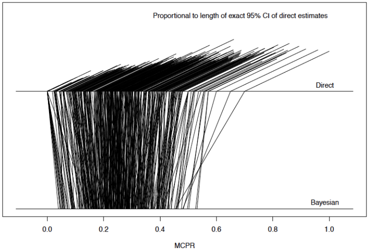
Small Area Estimates

Modern Contraceptive Prevalence Rate (MCPR) in Uganda

- Performance Monitoring and Accountability 2020 (PMA2020) survey data
- Woman-specific information: $\approx$ 13,100 inputs
  (109 areas) $\times$(4 rounds) $\times$($\approx$ 30 women per round)
- Logistic regression with covariates and an area-specific random effect

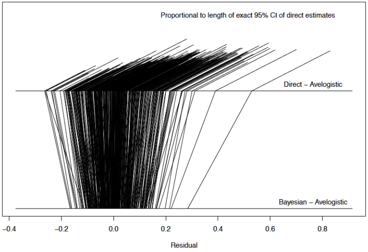
Change in Estimates

Direct $\rightarrow$ Bayes



Amount of shrinkage

(Direct - Regression) $\rightarrow$ (Bayes - Regression)



JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Age-specific rate of bone loss Hui and Berger (1983)

- Woman/age–specific, locally linear slope estimates
  - Positive values are 'loss'
  - Positive trend indicates increasing rate of loss with age
- Short follow-up, so slope and residual variance estimates are imprecise
- Use empirical Bayes to calm the variation

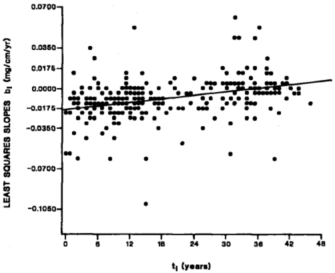


Figure 2. Individual least squares estimates of rate of bone loss b_i vs. t_i, where the t_i are suitably chosen points in the follow-up intervals.

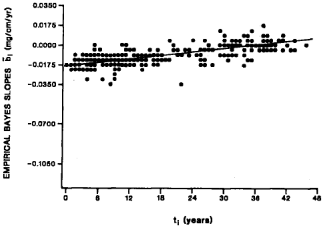


Figure 4. Individual empirical Bayes estimates of rate of bone loss $\bar{b}_i$ versus t_i.

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Stabilizing Variance Estimates

(Less controversial than stabilizing attributes of primary interest)

- The woman/age-specific, estimated residual variance is $\hat{\sigma}_k^2$
- With degrees of freedom, $d_k = \#\{\text{measurements}\} - 2$
- The σ_k^2 come from a (Gamma) prior with,
  - Estimated mean $\hat{m}$
  - Estimated effective sample size $\hat{M}$
- The empirical Bayes estimates are,

$$
\begin{aligned}
\tilde{\sigma}_k^2 &= \hat{m} + (1 - B_k)(\hat{\sigma}_k^2 - \hat{m}) \\
B_k &= \hat{M}/(\hat{M} + d_k) \\
\widetilde{d}_k &\approx B_k d_+ + (1 - B_k) d_k
\end{aligned}
$$

Stabilizing Variance Estimates

(Less controversial than stabilizing attributes of primary interest)

- The woman/age-specific, estimated residual variance is $\hat{\sigma}_k^2$
- With degrees of freedom, $d_k = \#\{\text{measurements}\} - 2$
- The σ_k^2 come from a (Gamma) prior with,
  - Estimated mean $\hat{m}$
  - Estimated effective sample size $\hat{M}$
- The empirical Bayes estimates are,

$$
\begin{aligned}
\tilde{\sigma}_k^2 &= \hat{m} + (1 - B_k)(\hat{\sigma}_k^2 - \hat{m}) \\
B_k &= \hat{M}/(\hat{M} + d_k) \\
\widetilde{d}_k &\approx B_k d_+ + (1 - B_k) d_k
\end{aligned}
$$

- The distribution of $\tilde{\sigma}_k^2$ isn't chi-square, but a fully Bayesian analysis (possibly via MCMC) produces the joint posterior distribution of the slopes and variances, and supports valid intervals and other inferences
- This 'full probability processing' is one of the principal advantages of the Bayesian formalism

Historical Controls (combine, don't 'pool')

	C	E	Total
Tumor	0	3	3
No Tumor	50	47	97
	50	50	100

- Fisher's exact one-sided P $= 0.121$
- But, pathologists get excited:
  - The 3 tumors are 'Biologically Significant'
- Statisticians protest:
  - But, they aren't 'Statistically Significant'

Historical Controls (combine, don't 'pool')

	C	E	Total
Tumor	0	3	3
No Tumor	50	47	97
	50	50	100

- Fisher's exact one-sided $P = 0.121$
- But, pathologists get excited:
  - The 3 tumors are 'Biologically Significant'
- Statisticians protest:
  - But, they aren't 'Statistically Significant'

We need to stop using these terms!

Include Historical Data

- There may be historical information for the same species/strain, same Lab, recent time period with 0 tumors in 450 control rodents
- Pooling gives,

Pooled Analysis

	C	E	Total
Tumor	**0**	3	3
No Tumor	500	47	547
	500	50	550

- Fisher's exact one-sided P $\doteq$ .0075
- Convergence between biological and statistical significance
- The Bayesian formalism should be be used to bring in history, in general, giving it only partial credit

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Bringing in history

Identify 'relevant' experiments, and use the Bayesian formalism

- Control rates come from a Beta distribution with

$$\begin{aligned} \text{mean} &= \mu \\ \text{Variance} &= \frac{\mu(1-\mu)}{M+1} \end{aligned}$$

- Use all the data to produce $\hat{\mu}$ and $\widehat{M}$
- Augment concurrent control group by pseudo-data with mean $\hat{\mu}$ and sample size $\widehat{M}$ (adaptive down-weighting of history)
- Female, Fisher F344 Male Rats, 70 historical experiments (Tarone, 1982)

Tumor	N	$\widehat{M}$	$\hat{\mu}$	$\frac{\widehat{M}}{N}$
Lung	1805	513	.022	28.4%
Stromal Polyp	1725	16	.147	0.9%

See Ibrahim et al. (2014) for a clinical trials example

Big Data and Data Synthesis

Chatterjee et al. (2016)

- Have a fine-grained study, with internally valid estimates; and stable, reduced dimension, external information
  - e.g, a joint distribution of a subset of the within-study variables
- Constrain the within-study estimates to be compatible with the externally determined (marginal) distributions in the spirit of,
  - Stabilizing estimates in a contingency table by ‘benchmarking’ to marginal distributions estimated from other data
  - Using external prevalence data so that a case-control study can estimate relative risk (RR) or a risk difference

Big Data and Data Synthesis

Chatterjee et al. (2016)

- Have a fine-grained study, with internally valid estimates; and stable, reduced dimension, external information
  - e.g, a joint distribution of a subset of the within-study variables
- Constrain the within-study estimates to be compatible with the externally determined (marginal) distributions in the spirit of,
  - Stabilizing estimates in a contingency table by 'benchmarking' to marginal distributions estimated from other data
  - Using external prevalence data so that a case-control study can estimate relative risk (RR) or a risk difference
- The key issue is whether stochastic features of the external data are sufficiently similar to those for the internal data so that in the end MSE is reduced
- Resonates with external validity, representativity of a sample, transporting within-sample estimates to a reference population, . . .

See, Keiding and Louis (2016); Keiding and Louis (2018)
Pearl and Bareinboim (2014); National Academies (2017)

Combining Surveys

With other data, see Lohr and Raghunathan (2017)

Combining Estimates from Related Surveys
via Bivariate Models

(Application: using ACS estimates to improve estimates
from smaller U.S. surveys)

William R. Bell and Carolina Franco, U.S. Census Bureau

2016 Ross-Royall Symposium

February 26, 2016

Application I: 2010 Disability Rates for U.S. States: SIPP borrowing from ACS

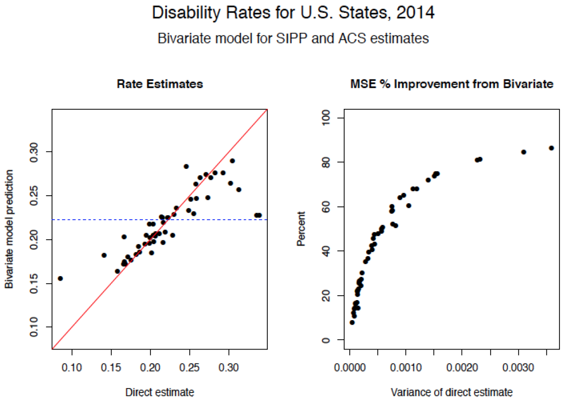
y_{1i} = SIPP disability estimate, y_{2i} = ACS disability estimate

Smoothing of SIPP direct sampling variance estimates is applied.

$\hat{\rho}$ = **.82**

- Univariate shrinkage yields an MSE decrease of $2\% - 67\%$ from direct, with a median of 19%
- The MSE decrease from bivariate vs. univariate model is $6\% - 59\%$ with a median of 29%
- The MSE decrease from bivariate vs. direct is **8 – 86%, with a median decrease of 43%**

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH


Disability Rates for U.S. States, 2014
Bivariate model for SIPP and ACS estimates
Rate Estimates
Bivariate model prediction
Direct estimate
0.10
0.15
0.20
0.25
0.30
MSE % Improvement from Bivariate
Percent
Variance of direct estimate
0
20
40
60
80
100
0.0000
0.0010
0.0020
0.0030

SAIPE and Section203

(Bayesian) hierarchical modeling is essential

SAIPE: Small Area income and Poverty Estimates (Bell et al., 2016)

- Allocate \$12+ billion a year
- 'Direct' Data are from the ACS and other surveys
- Xs are tax rates, etc.

SAIPE and Section203

(Bayesian) hierarchical modeling is essential

SAIPE: Small Area income and Poverty Estimates (Bell et al., 2016)

- Allocate \$12+ billion a year
- 'Direct' Data are from the ACS and other surveys
- Xs are tax rates, etc.

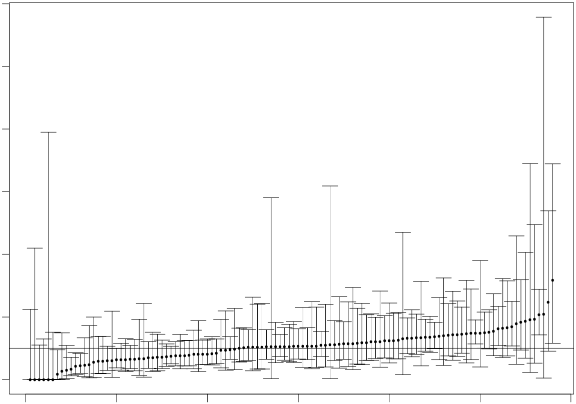
Section 203 of the Voting Rights Act (Slud and Ashmead, 2017)

- In order to make the determinations, it is necessary to estimate the total population of voting age persons who are citizens, of citizens who have limited English proficiency, and of citizens with limited English proficiency who are illiterate in approximately 8000 jurisdictions, 570 American Indian and Alaska Native Areas (AIA/ANAs), and 12 Alaska Native Regional Corporations (ANRCs), separately for 68 Language Minority Groups
- Potential estimation domains $\approx$ 560,000 $= 70 \times 8000$

USRDS, SMRs: MLEs and exact CIs

(1, 41, 81, . . . ordered MLEs)

- SMR = Standardized Mortality Ratio = observed/expected deaths

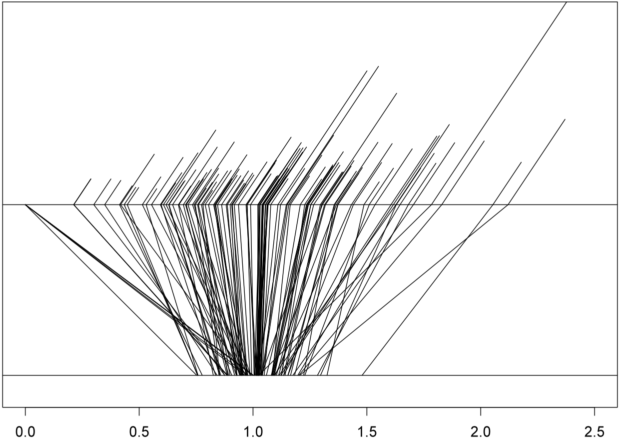


- Sampling variability has a wide range over units

Bayesian analysis, $\rho = \text{SMR}$ (Lin et al., 2009)

$\hat{\rho}^{mle}$, $\hat{\rho}^{pm}$, SE($\hat{\rho}^{mle}$) using USRDS dialysis data

middle = MLE :: whisker = SE :: bottom = Posterior Mean



JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Shrinkage can be controversial (Normand et al., 2016)

- Direct estimates with greatest uncertainty are shrunken closest to the regression surface, potentially conferring undue benefits or punishments
- Especially troublesome when the model is mis-specified (always true!) and sample size is informative so that the degree of shrinkage is ‘connected at the hip’ to the underlying truth
- Standard model fitting gives more weight to the stable units, consequently the units that ‘care about’ the regression model have less influence on it
- Recent approaches increase the weights for the relatively unstable units, paying some variance, but improving estimation performance for mis-specified models (Chen et al., 2015; Jiang et al., 2011)

Closing

- Statistics has always been about combining information; think $\bar{X}$
- Careful development and assessment is necessary, and the Bayesian formalism is an effective aid to navigation and inferential framework
- Advances in data science (annotation, harmonization, storage and retrieval), computing (hardware & software), and statistical methods; make evermore relevant,

> *All of statistics involves combining evidence over basic units to make inferences for a population. The current challenge involves broadening the scope of inputs and inferences in a scientifically valid and credible manner. Development and application of these meta-modeling strategies will challenge and inform in the next and subsequent decades.* *(Louis, 1989)*

Closing

- Statistics has always been about combining information; think $\bar{X}$
- Careful development and assessment is necessary, and the Bayesian formalism is an effective aid to navigation and inferential framework
- Advances in data science (annotation, harmonization, storage and retrieval), computing (hardware & software), and statistical methods; make evermore relevant,

> *All of statistics involves combining evidence over basic units to make inferences for a population. The current challenge involves broadening the scope of inputs and inferences in a scientifically valid and credible manner. Development and application of these meta-modeling strategies will challenge and inform in the next and subsequent decades. (Louis, 1989)*

- However,

  Space-age procedures will not rescue stone-age data

#thank you

JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

Bibliography

Bell, W., Basel, W., and Maples, J. *Analysis of Poverty Data by Small Area Estimation*, chapter An Overview of the U. S. Census Bureau's Small Area Income and Poverty Estimates Program. John Wiley & Sons (2016).

Carlin, B. P. and Louis, T. A. *Bayesian Methods for Data Analysis, 3rd edition*. Boca Raton, FL: Chapman and Hall/CRC Press, 3^{nd} edition (2009).

Chatterjee, N., Chen, Y. H., Maas, P., and Carroll, R. J. "Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-level information from External Big Data Sources (with discussion)." *J Am Stat Assoc*, 111(513):107–131 (2016).

Chen, S., Jiang, J., and Nguyen, T. "Observed Best prediction for small area counts." *Journal of Survey Statistics and Methodology*, 3:136–161 (2015).

Gelman, A. "Struggles with survey weighting and regression modeling." *Statistical Science*, 22(2):153–164 (2007).

Gelman, A., Carlin, J., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. *Bayesian Data Analysis, 3rd edition*. Boca Raton, FL: Chapman and Hall/CRC Press (2013).

Henderson, N., Louis, T., Wang, C., and Varadhan, R. "Bayesian Analysis of Heterogeneous Treatment Effects for Patient-Centered Outcomes Research." *Health Services and Outcomes Research Methodology*, 16:213–233 (2016).

Hui, S. L. and Berger, J. O. "Empirical Bayes estimation of rates in longitudinal studies." *Journal of the American Statistical Association*, 78:753–759 (1983).

Ibrahim, J., Chen, M., Lakshminarayanan, M., Liu, G., and Heyse, J. "Bayesian probability of success for clinical trials using historical data." *Statistics in Medicine*, 34:249–264 (2014).

Jiang, J., Nguyen, T., and Rao, J. S. "Best Predictive Small Area Estimation." *Journal of the American Statistical Association*, 106:732–745 (2011).

Kadane, J. "Bayesian Methods for Prevention Research." *Prevention Science*, 16:1017–1025 (2015).

Keiding, N. and Louis, T. "Web-based Enrollment and other types of Self-selection in Surveys and Studies: Consequences for Generalizability." *Annual Review of Statistics and Its Application*, 5:to appear (2018).

Keiding, N. and Louis, T. A. "Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussion and response)." *J Roy Statist Soc, Ser A*, 179:319–376 (2016).

Lin, R., Louis, T., Paddock, S., and Ridgeway, G. "Ranking of USRDS, provider-specific SMRs from 1998-2001." *Health Services and Outcomes Reseearch Methodology*, 9:22–38 (2009).

Lohr, S. and Raghunathan, T. "Combining Survey Data with Other Data Sources." *Statistical Science*, 32:293–312 (2017).

Louis, T. "Meta-modeling." In *Challenges for the 90's, ASA Sesquicentennial visioning* (1989).

National Academies. *Federal Statistics, Multiple Data Sources and Privacy Protection: Next Steps*. National Academies Press (2017).

Normand, S.-L., Ash, A. S., Fienberg, S. E., Stukel, T., Utts, J., and Louis, T. A. "League Tables for Hospital Comparisons." *Annual Review of Statistics and Its Application*, 3:21–50 (2016).

Pearl, J. and Bareinboim, E. "External Validity: From do-calculus to Transportability across Populations." *Statistical Science*, 29:579–595 (2014).

Slud, E. and Ashmead, R. "VRA Section 203 Determinatinos: Statistical Methodology Summary." Technical report, U. S. Census Bureau (2017).

Tarone, R. "The use of historical control information in testing for a trend in proportions." *Biometrics*, 38:215–220 (1982).