International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2011), 79, 2, 221–230 doi:10.1111/j.1751-5823.2011.00144.x

Discussions

Anastasios A. Tsiatis and Marie Davidian

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA
E-mail: tsiatis@ncsu.edu

We congratulate the authors (henceforth LSD) on a long overdue, detailed review of the connection between the AIPW estimators derived in the incomplete data context via semiparametric theory by Robins, Rotnitzky, and colleagues and the survey calibration estimators used widely in survey sampling. Although this connection has been noted previously (e.g. Robins & Rotnitzky, 1998; Rotnitzky, 2009), the present article appears to be the first in the statistical literature to offer a more comprehensive account. We had only a passing familiarity with this connection, and we are grateful to the editors for the opportunity to offer this discussion, whose preparation required us to acquire a deeper understanding. In what follows, we hope to complement the presentation of LSD by highlighting some further relationships and differences between the two perspectives.

We adopt the notation used by the authors and consider estimation of the population total T. We focus on the regression estimator $\widehat{T}_{reg}$, which, as in Section 2.1 of LSD may be written equivalently as $\widehat{T}(\widehat{\beta})$, where

$$\widehat{T}(\beta) = \widehat{T} - \sum_{i=1}^{N} \left(\frac{R_i - \pi_i}{\pi_i} \right) x_i \beta, \tag{1}$$

a representation that may be more familiar to statisticians well-versed in the AIPW literature. We reiterate and expand upon some important differences between the missing data and survey sampling contexts noted by LSD.

In survey sampling, the realizations $(x_1, y_1), \ldots, (x_N, y_N)$ that comprise the population are regarded as fixed, and inference on $T = \sum_{i=1}^{N} y_i$, or, equivalently, the population mean $N^{-1}T$, is the goal. This is based on data $(x_i, R_i, R_i y_i)$, $i = 1, \ldots, N$, drawn from the population according to a fixed, known design, where $n = \sum_{i=1}^{N} R_i$, and $\Pr(R_i = 1) = \pi_i$ and $\Pr(R_i = 1, R_j = 1) = \pi_{ij}$ for π_i known and π_{ij} known or unknown, $i, j = 1, \ldots, N$. The (x_i, y_i) may be viewed as realizations of random variables (X_i, Y_i), $i = 1, \ldots, N$, representing an independent and identically distributed (iid) sample from some super-population; however interest focuses on the fixed quantity T (Särndal *et al.*, 2003).

In contrast, in the incomplete data context, interest is in estimation of $\mu = E(Y)$, a parameter associated with the super-population. Instead of observing a realization of i.i.d. $(X_i, Y_i), i = 1, \ldots, N$, we observe a realization of i.i.d. $(X_i, R_i, R_i Y_i), i = 1, \ldots, N$, where R_i is an indicator of whether or not the value of Y_i is observed or missing. Ordinarily, the probabilities of observing Y_i for each i are not fixed by design; rather, missingness arises according to some unknown mechanism about which some assumption is made. A common assumption is that R_i is conditionally independent of Y_i given X_i, the so-called "missing at random" (MAR) assumption, under which $\pi(X_i) = \Pr(R_i = 1|X_i, Y_i) = \Pr(R_i = 1|X_i)$. MAR cannot be verified from the observed data, so the

validity of inference on μ depends on its unknown relevance. Even if MAR is plausible, which we assume henceforth, the function $\pi(X)$ is not known and thus must be estimated based on the observed data, usually via maximum likelihood for a posited parametric model, yielding predicted values $\widehat{\pi}_i$.

From the survey sampling perspective, because the π_i are known, $\widehat{T}(\beta)$ in (1) is a consistent estimator for T for any fixed β, and the choice $\widehat{\beta}$ given in Section 2 of LSD leading to $\widehat{T}_{reg}$ is meant to yield an estimator that is more precise than $\widehat{T}$. In the incomplete data setting, $N^{-1}\widehat{T}(\beta)$ with the $\widehat{\pi}_i$ substituted for the π_i, need not be consistent for μ unless the model for $\pi(X_i)$ is correctly specified. If this model is correct, then it follows from semiparametric theory that all regular, asymptotically linear estimators for μ may be written in the form (Robins *et al.*, 1994)

$$N^{-1}\sum_{i=1}^{N}\left\{\frac{R_i Y_i}{\widehat{\pi}_i} - \left(\frac{R_i - \widehat{\pi}_i}{\widehat{\pi}_i}\right)\phi(X_i)\right\}, \tag{2}$$

where $\phi(X)$ is an arbitrary function of X. The choice of $\phi(X)$ leading to the most precise estimator within class (2) is $\phi(X) = E(Y|X)$. Accordingly, a (usually parametric) model for $E(Y|X)$ may be posited and fitted, and the predicted values substituted for $\phi(X_i)$ in (2).

Scharfstein *et al.* (1999) made the critical observation that such an estimator is "doubly robust" (DR), i.e. is consistent for μ as long as at least one of the posited models for $\pi(X)$ or $E(Y|X)$ is correct. Because of the protection afforded by this property, DR estimators have been advocated for routine use. In practice, posited models for $E(Y|X)$ might be linear, generalized linear, or arbitrarily nonlinear in a parameter β, depending on the nature of Y. Usually, the posited model is fitted using ordinary or iteratively reweighted least squares based on the pairs (x_i, y_i) for which $R_i = 1$, which, under MAR, would yield a consistent estimator for β in a correctly specified model $m(X, \beta)$, say, for $E(Y|X)$.

Kang & Schafer (2007) evaluated the performance of the usual DR estimator for μ in a missing data context under specific simulation scenarios with continuous Y, linear $m(x, \beta) = x\beta$, and β estimated by ordinary least squares. The estimator exhibited poor performance under scenarios where the models for $\pi(X)$ and $E(Y|X)$ were only slightly misspecified and/or when the relative magnitudes of the $\widehat{\pi}_i$ were extremely disparate, with $\widehat{\pi}_i$ relatively very small for some i, leading Kang & Schafer to issue a strong warning against its routine use. Because of the observational nature of the data, where missingness is by happenstance rather than design, such $\widehat{\pi}_i$ may be encountered in practice.

These results led us (Cao *et al.*, 2009; see also Tan, 2006, 2007 and Tsiatis & Davidian, 2007) to speculate that this poor performance may be partly a consequence of the method used to estimate β in the posited model $m(x, \beta)$. We proposed considering the class of DR estimators in (2), where $\phi(X_i)$ is replaced by $m(X_i, \beta)$, and, among such estimators indexed by β, found the value of β that minimizes the variance of estimators within the class when $\pi(X)$ is correctly specified regardless of whether or not $m(X_i, \beta)$ is correct, and a means of estimating this optimal β. The estimator for the optimal β is not ordinary or usual weighted least squares but, rather, involves, in the ideal case where the π_i were known, a weighted regression with weights $(1 - \pi_i)/\pi_i^2$, with modification when π_i are estimated by $\widehat{\pi}_i$ as above. Cao *et al.* (2009) reported simulations showing that the DR estimator incorporating this estimator for the optimal β demonstrated vastly improved performance in the Kang & Schafer and other scenarios. Tsiatis *et al.* (2011) extended this idea to DR estimators in the more complex setting of longitudinal studies with monotone dropout.

We were interested to learn that the same tactic of finding the optimal β minimizing the variance of estimators for T of the form $\widehat{T}(\beta)$ in (1) and an estimator for this optimal β using a weighting scheme similar in spirit to that in Cao *et al.* (2009) was proposed by Montanari

(1987); see also Berger *et al.* (2003). Given that in the survey sampling context, with the π_i determined by design, the issue of disparate π_i would not be as pronounced, we suspect that the gains in performance realized by such an approach may not be as dramatic.

We again compliment the authors on an insightful and useful article.

Acknowledgements

This work was supported by grants R37 AI031789, R01 CA051962, R01 CA085848, and P01 CA142538 from the National Institutes of Health.

References

Berger, Y.G., Tirari, M.E.H. & Tillé, Y. (2003). Towards optimal regression estimation regression in sample surveys. *Aust. & N. Zeal. J. Statist.*, **45**, 319–329.

Cao, W., Tsiatis, A.A. & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, **96**, 723–734.

Kang, J.D.Y. & Schafer, J.L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, **22**, 523–539.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *Int. Statist. Rev.*, **55**, 191–202.

Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, **89**, 846–866.

Rotnitzky, A. (2009). Inverse probability weighted methods. In *Longitudinal Data Analysis*, Eds. G. Fitzmaurice, M. Davidian, G. Vereke & G. Molenberghs, pp. 453–476. Boca Raton: Chapman & Hall/CRC.

Scharfstein D.O., Rotnitzky, A. & Robins, J.M. (1999). Rejoinder to "Adjusting for nonignorable drop-out using semiparametric nonresponse models". *J. Amer. Statist. Assoc.*, **94**, 1135–1146.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101**, 1619–37.

Tan, Z. (2007). Understanding OR, PS and DR. *Statist. Sci.,* **22**, 560–8.

Tsiatis, A.A. & Davidian, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, **22**, 569–573.

Tsiatis, A.A., Davidian, M. & Cao, W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics*, **67**, 536–545.

[*Received March 2011, accepted May 2011*]

Mark S. Handcock

Department of Statistics, University of California, Los Angeles, USA
E-mail: handcock@ucla.edu

Lumley, Shaw, and Dai (LSD) are to be congratulated on producing an insightful, clarifying, and accessible paper on the relationships between various methods of survey inference when auxiliary information is available.

As for many statistical areas with broad applicability, inference from sample survey data has accumulated a plethora of terminology due to the multidisciplinary sources of its development. LSD do well to re-focus on the core statistical principals rather than emphasize the incidental differences. The identification of influence functions as a means to connect the regression and calibration frameworks is particularly insightful.

The calibration methods have the advantage of typically expressing the effect of auxiliary information via weights. As the statistical agencies known, most users are familiar and comfortable with analysing survey data with weights and this aids the acceptance of the methods. However, most users think of the weights as exogenous and in-volatile, so that in circumstances were the weights are endogenous and change with the nature of the auxiliary information the usual interpretation can lead users astray. This will become a concern as the calibration methods become more sophisticated.

It is interesting to see if a broader framework can be developed to connect even more apparently disparate approaches. Likelihood frameworks for survey inference typically depend on postulating a super-population sampling process. These enable classical estimators to be reinterpreted as maximum likelihood estimators under the super-population process. For example, Chen & Qin (1993) and Chen & Sitter (1999) use an empirical likelihood based method to incorporate auxiliary information under simple random sampling and probability sampling, respectively. The auxiliary information is incorporated via constraints on the likelihood and can be interpreted as both a calibration and as a maximum likelihood estimator. Chaudhuri and colleagues have developed approaches for generalized linear models that are simple computationally (Chaudhuri *et al.*, 2008). Wu and colleagues (notably Chen *et al.*, 2002; Wu & Rao, 2006; Rao & Wu, 2008 among others) study the method of Chen & Sitter (1999) extensively and apply it to several design based surveys. Kim (2009) approximates the sampling process via Poisson sampling to develop alternative estimators, while Chaudhuri *et al.* (2010) incorporate the sampling design information through the conditional expectation of the sampling probabilities. I would be interested in the author's thoughts on the value of (parametric and non-parametric) likelihood framings in connecting the various survey estimators.

There has been much recent interest and developments in designs for hard-to-reach populations (see, for example, Gile, 2008 and the references therein). These populations are characterized by the difficulty in survey sampling from them using standard probability methods. Typically, a sampling frame for the target population is not available, and its members are rare or stigmatized in the larger population so that it is prohibitively expensive to contact them through the available frames (Gile & Handcock, 2010). Chain-referral and link-tracing designs exploit an underlying social network of ties between the population members to sample. For these designs the inclusion probabilities π_i are typically unknown (Gile, 2009; Handcock & Gile, 2010). LSD focus on designs where π_i is known. Do the authors see hope for calibration estimators in designs where the π_i are, at least partially, unknown?

Acknowledgements

This work was supported by grant number 1R21HD063000 from NICHD and grant number MMS-0851555 from NSF, and grant number N00014-08-1-1015 from ONR. Its contents are solely the responsibility of the author and do not necessarily represent the official views of the Demographic & Behavioral Sciences (DBS) Branch, the National Science Foundation, or the Office of Naval Research.

References

Chaudhuri, S., Handcock, M.S. & Rendall, M.S. (2008). Generalized linear models incorporating population level information: an empirical-likelihood-based approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **70**, 311–328.

Chaudhuri, S., Handcock, M.S. & Rendall, M.S. (2010). A conditional empirical likelihood based approach to incorporate sampling weights and population level information. Technical report, National University of Singapore.

Chen, J., Sitter, R.R. & Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, **89**(1), 230–237.
Chen, J.H. & Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, **80**(1), 107–116.
Chen, J.H. & Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statist. Sinica*, **9**(2), 385–406.
Gile, K.J. (2008). Inference from Partially-Observed Network Data. PhD in Statistics, University of Washington.
Gile, K.J. (2009). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J. Amer. Stat. Assoc.*, **106**(493), 135–146.
Gile, K.J. & Handcock, M.S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociol. Methodol.*, **40**, 285–327.
Handcock, M.S. & Gile, K.J. (2010). Modeling networks from sampled data. *Ann. Appl. Stat.*, **272**(2), 383–426.
Kim, J.K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Stat. Sin.*, **19**(1), 145–157.
Rao, J. & Wu, C. (2008). *Empirical Likelihood Methods*, Vol. 29B, pp. 189–208. Amsterdam: Elsevier.
Wu, C.B. & Rao, J.N.K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *Canad. J Stat.-Revue Canadienne De Statistique*, **34**(3), 359–375.

[*Received May 2011, accepted May 2011*]

Jerald F. Lawless[1] and John D. Kalbfleisch[2]

[1]*Department of Statistics and Actuarial Science, University of Waterloo, Canada*
E-mail: jlawless@uwaterloo.ca
[2]*Department of Biostatistics and Statistics, University of Michigan, USA*
E-mail: jdkalbfl@umich.edu

The paper by Lumley, Shaw, and Dai (LSD) provides insights into previous work on estimation in some problems with incomplete data. A variety of techniques exists in the survey sampling and general statistical literature for addressing missing data, or for "strengthening" estimation of a target parameter through the use of auxiliary data. These techniques include various methods of imputation, calibration, post-stratification, weighting, and modelling. A number of authors have previously discussed connections among methods. For example, Zhang (2000) considers calibration, post-stratification and regression estimation in survey contexts; Kang & Schafer (2007) and Robins *et al.* (2007) discuss doubly robust (DR) estimating functions, stratification and regression estimation.

These articles reveal a vast and sometimes bewildering array of approaches, the performance of which can depend on many factors including the (approximate) validity of modelling assumptions, the nature of the parameters of interest, the strengths of association among fully and incompletely observed variables, and the pattern and nature of the missingness. The main message we discern in LSD is that, in many applications, calibration is an effective, convenient, and unifying way to incorporate auxiliary information into estimation of a target parameter.

Despite its intuitive appeal, the calibration approach often seems less than transparent with respect to implicit assumptions, efficiency or sensitivity to departures from assumptions. Likelihood and pseudo-likelihood estimating functions, with weights incorporated if necessary to reflect sample design or other features of observation, seem clearer to us in general. Likelihood-related methods are convenient and efficient in many settings (e.g. Kalbfleisch & Lawless, 1988; Lawless *et al.*, 1999; Chen & Little, 1999; Chatterjee *et al.*, 2003; Zhang & Rockette,

2005; McLeish & Struthers, 2006; Zhao *et al.*, 2009), and more generally, a wide variety of "augmented" estimating functions (Robins *et al.*, 1995) exist.

For settings involving auxiliary variables that are not easily approached via maximum likelihood, extension of an estimating function approach of Chen & Chen (2000, hereafter CC) provides an alternative to calibration that is relatively transparent, and produces estimators similar to ones discussed in LSD. Suppose that a parametric model $f(y; \beta)$ involving variables Y_i observed in a sample of size n is the target of estimation and that auxiliary variables $X_i (i = 1, \ldots, N)$ are available for all individuals in a cohort (or population) from which the sample is selected. A partial model linking Y and β through an estimating function could alternatively be considered. In this presentation, we assume the cohort is itself a random sample from a conceptual super-population. Let R_i indicate that individual i is selected for the sample, and let $\pi_i = \Pr(R_i = 1 | Y_i, X_i)$, which is assumed positive for all $i = 1, \ldots, N$.

Let $U_i(Y_i; \beta) = \partial \log f(Y_i; \beta)/\partial \beta'$ and consider the weighted pseudo-likelihood estimating function

$$U(\beta) = \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_i(Y_i; \beta) \tag{1}$$

which is easily seen to be unbiased under usual regularity conditions. In conjunction with (1), consider an estimating function

$$V(\gamma) = \sum_{i=1}^{N} \frac{R_i}{\pi_i} V_i(X_i; \gamma), \tag{2}$$

where $V_i(X_i; \gamma)$ is a function of X_i and a vector of parameters γ. Consider also the estimating function

$$V_N(\gamma) = \sum_{i=1}^{N} V_i(X_i; \gamma) = 0 \tag{3}$$

which is assumed to be unbiased at γ^*. (That is, $E\{V_N(\gamma^*)\} = 0$). It is further assumed that the solution $\hat{\gamma}_N$ to (3) converges in probability to γ^* as $N \to \infty$. Note that we do not assume the $V_i(X_i; \gamma)$ are based on any "correct" model specification. CC consider the case of a simple random sample from the cohort but their arguments readily extend to the situation here. Define the matrices

$$A(\theta) = -E \begin{pmatrix} \partial U(\beta)/\partial \beta' & 0 \\ 0 & \partial V(\gamma)/\partial \gamma' \end{pmatrix} \qquad B(\theta) = \begin{pmatrix} \mathrm{Var(U)} & \mathrm{Cov}(U, V) \\ \mathrm{Cov}(V, U) & \mathrm{Var}(V) \end{pmatrix},$$

where $\theta = (\beta', \gamma')'$. Under mild additional conditions, $\sqrt{N}(\hat{\theta} - \theta^*)$ converges to a multivariate normal with mean 0 and covariance matrix $NA^{-1}(\theta^*)B(\theta^*)A^{-1}(\theta^*)$ for $N \to \infty$ and $n/N > 0$ fixed, where $\theta^* = (\beta_0', \gamma^{*\prime})'$ and β_0 is the true value of β. CC propose the adjusted estimator $\tilde{\beta}$ which is based on the mean of $\hat{\beta}$ given $\hat{\gamma}$ in the limiting distribution. Thus,

$$\tilde{\beta} = \hat{\beta} - \hat{A}_1^{-1} \hat{B}_{12} \hat{B}_{22}^{-1} \hat{A}_2 (\hat{\gamma} - \gamma^*), \tag{4}$$

where A_1 and A_2 are the diagonal blocks in $A(\theta)$, $B_{12} = \mathrm{Cov}(U, V)$ and $B_{22} = \mathrm{Var}(V)$. These matrices are estimated using $\hat{\theta} = (\hat{\beta}', \hat{\gamma}')'$. Since γ^* in (4) is unknown, we replace it with the estimator $\hat{\gamma}_N$ based on (3). CC show that when $U_i(Y_i; \beta_0)$ and $V_i(X_i; \gamma^*)$ are sufficiently highly correlated, $\tilde{\beta}$ can be substantially more efficient than $\hat{\beta}$.

An alternative and asymptotically equivalent approach would utilize an "adjusted" estimating function for β based on $E\{U(\beta)|V(\gamma^*)\}$ from the limiting normal approximation for $U(\beta)$,

$V(\gamma^*)$. This is similar to the augmented estimating function approach of Robins *et al.* (1995), and leads to the estimating function

$$\tilde{U}(\beta) = \sum_{i=1}^{N} \frac{R_i}{\pi_i} \left\{ U_i\left(Y_i; \beta\right) - \hat{B}_{12}\hat{B}_{22}^{-1} V_i\left(X_i; \gamma^*\right) \right\}.$$

Since $V_N(\gamma_N^*) = 0$, $\tilde{U}(\beta)$ can be rewritten as

$$\tilde{U}(\beta) = \sum_{i=1}^{N} \left\{ \frac{R_i}{\pi_i} U_i\left(Y_i; \beta\right) + \left(1 - \frac{R_i}{\pi_i}\right) \hat{B}_{12}\hat{B}_{22}^{-1} V\left(X_i; \gamma^*\right). \right\} \tag{5}$$

With γ^* replaced with $\hat{\gamma}_N$, this estimating function yields an estimate of β that is asymptotically equivalent to $\tilde{\beta}$. By way of comparison with Robins *et al.* (1995), note that $\hat{B}_{12}\hat{B}_{22}^{-1} V\left(X_i, \hat{\gamma}_N\right)$ is an estimate of $E\{U_i(Y_i; \beta) | V(X_i; \gamma^*)\}$.

Further extensions can be made for the case where weights π_i are estimated, but given space limitations, we omit this here. The following examples illustrate this approach in two simple settings.

Example 1. As in the initial example in LSD, suppose that Y is a scalar variable and $\mu_Y = E(Y)$ is the target for estimation. Let X_i be a scalar auxiliary variable with finite mean μ_X. Letting $U_i(Y_i; \mu_Y) = Y_i - \mu_Y$ and $V_i(X_i; \mu_X) = X_i - \mu_X$, we find after a little algebra that (4) produces the adjusted estimator

$$\tilde{\mu}_y = \bar{y} - \hat{\delta}\left(\bar{x}_n - \bar{x}_N\right) = \hat{\alpha} + \hat{\delta}\bar{x}_N, \tag{6}$$

where $\hat{\alpha}$ and $\hat{\delta}$ are the weighted least squares estimators of the intercept and slope from a regression of Y on X with data $\{(y_i, x_i), i = 1, \ldots, n\}$. This is the regression estimator for μ_Y corresponding to $\hat{T}_{reg}$ in LSD.

Example 2. Consider the same setting as in Example 1 but suppose in addition that X is a discrete variable taking on K values $x_1, \ldots, x_K$. For simplicity we suppose the Y_i constitute a random sample of size n. In this case associate γ with probabilities $\{g(x_1), \ldots, g(x_K)\}$ and consider the estimating function $V(\gamma)$ with components

$$V(\gamma)_j = \sum_{i=1}^{n} \left\{ I\left(X_i = x_j\right) - \gamma_j \right\} \quad j = 1, \ldots, K-1.$$

After some tedious algebra, we find that (4) produces the stratification estimator

$$\tilde{\mu}_Y = \sum_{j=1}^{K} \left(\frac{N_j}{N}\right) \bar{y}_j, \tag{7}$$

where $\bar{y}_j$ is the mean of the sampled Y_i's with corresponding $X_i = x_j$ and N_j is the number of individuals in the full cohort with $X_i = x_j$.

This approach can be applied to other examples in LSD, and it would be of interest to explore it further. The fact that it uses ordinary estimating function theory makes it adaptable to a range of problems and has some substantial appeal. We also note that the approach used to incorporate auxiliary data here is parametric, but γ and $V(\gamma)$ could be adapted to deal with semiparametric models, as is done with semiparametric maximum likelihood.

References

Chatterjee, N., Chen, Y. & Breslow, N.E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *J. Amer. Statist. Assoc.*, **98**, 158–168.

Chen, Y.-H. & Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *J. R. Stat. Soc. B*, **62**, 449–460.

Chen, H.Y. & Little, R.J.A. (1999). Proportional hazards with missing covariates. *J. Amer. Statist. Assoc.*, **94**, 896–908.

Kalbfleisch, J.D. & Lawless, J.F. (1988). Likelihood analysis of multistate models for disease incidence and mortality. *Statist. Med.*, **1**, 149–160.

Kang, J.D.Y. & Schafer, J.L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, **22**, 523–539.

Lawless, J.F., Kalbfleisch, J.D. & Wild, C.J. (1999). Semiparametric methods for response selective and missing data problems in regression. *J. R. Stat. Soc. B*, **61**, 413–438.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *J. Amer. Statist. Assoc.*, **99**, 546–556.

McLeish, D.L. & Struthers, C.A. (2006). Estimation of regression parameters in missing data problems. *Canad. J. Statist.*, **34**, 233–259.

Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. (2007). Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Statist. Sci.*, **22**, 544–559.

Samuelsen, S.O., Anestad, H. & Skrondal, A. (2007). Stratified case-cohort analysis of general cohort sampling designs. *Scand. J. Statist.*, **34**, 103–119.

Zhang, L.-C. (2000). Post-stratification and calibration - a synthesis. *Amer. Statist.*, **54**, 178–184.

Zhang, Z. & Rockette, H.E. (2005). On maximum likelihood estimation in parametric regression with missing covariates. *J. Statist. Plann. Inference*, **134**, 206–223.

Zhao, Y., Lawless J.F. & McLeish, D.L. (2009). Likelihood methods for regression models with expensive variables missing by design. *Biom. J.*, **51**, 1–14.

[*Received April 2011, accepted May 2011*]

Alastair J. Scott and Christopher J. Wild

Department of Statistics, The University of Auckland, Private Bag 92019, Auckland, New Zealand
E-mail: c.wild@auckland.ac.nz

Using population information on auxiliary variables to improve estimates of population means or totals has a long (and successful) history in survey sampling – ratio and regression estimators have been used for at least 75 years and probably more (see Watson, 1937, for example). This early work has been extended and formalized over the last 20 years through the work of Sarndal, Deville, and collaborators on calibration and GREG estimation. At the same time, Robins, Rotnitzky, and colleagues were developing rather similar techniques in the very different context of semiparametric methods for incomplete data. For most of this time, there was very little awareness in either field of the developments in the other. Recently, however, there has been a growing cross-fertilisation between the sample survey literature with the biostatistics literature with a Washington biostatistics group based loosely around Norm Breslow being important protagonists. This excellent and timely paper continues this Washington tradition. Not only do the authors convey deep insights into the close connections between survey-calibration and the AIPW estimators from the Harvard biostatistics group associated with Jamie Robins, they express them in very accessible ways. Hopefully this will encourage researchers in both fields to learn about, and borrow strength from, the work done in the other field.

We do have one slight quibble, however. The key result that $\text{var}[\widehat{T}_{\text{reg}}] \approx (1-\rho^2)\text{var}[\widehat{T}]$ (see the sentence immediately following equation (2)) is only true under fairly strong unstated conditions on the π_is. It is true for simple random sampling but certainly not true, for example, if the y_is are positive and $\pi_i \propto y_i$ or, more realistically, $\pi_i \propto x_i$ where x is some design variable highly correlated with y. The comparison between the decompositions in equations (2) and (6) is still interesting but somewhat less compelling as an explanation of the "estimated weights" paradox.

We too have been exploring the effect of using estimated weights in weighted estimating equations but using the approach of modelling the π_is outlined in Section 2.3, rather than calibration. Suppose that π_i can be modelled by some function, say $\pi_i = p_i(\alpha)$, of an unknown parameter α and we have enough information in the data to estimate α. The estimate of θ is then obtained by solving the estimating equation

$$S_0(\widehat{\alpha}, \theta) = \sum_{i:R_i=1} w_i(\widehat{\alpha}) U_i(\theta) = 0,$$

where $w_i(\alpha) = 1/p_i(\alpha)$. If $\widehat{\alpha}$ is obtained by solving the estimating equation $S_1(\widehat{\alpha}) = 0$, then we can estimate α and θ together by solving $S(\alpha, \theta) = 0$, where $S(\alpha, \theta) = \binom{S_0(\alpha, \theta)}{S_1(\alpha)}$, and hence get an explicit expression for $\text{ACov}\{\widehat{\theta}\}$ using standard methods for estimating equations. In the special case in which the R_is are independent, the observations are missing at random in the sense of Rubin (1974), and an efficient estimator is used for $\widehat{\alpha}$, this reduces to $\text{ACov}\{\widehat{\theta}\} = \text{V}_{00} - \text{V}_{01}\text{V}_{11}^{-1}\text{V}_{01}^T$, where $\text{V}_{00} = \text{Cov}\{Q_0\}$, $\text{V}_{01} = \text{Cov}\{Q_0, S_1\}$, and $\text{V}_{11} = \text{Cov}\{S_1\}$. Here $\text{I}_{00} = E\{-\partial S_0/\partial\theta^T\}$ and $Q_0 = \text{I}_{00}^{-1}S_0$, the sum of the influence functions $\text{I}_{00}^{-1}U_i$. (Details are given in Scott & Wild, 2011, and are an extension of results in Lawless *et al.*, 1999.) Note that V_{00} would be the value of $\text{ACov}\{\widehat{\theta}\}$ if the true π_is were used so that $\text{V}_{01}\text{V}_{11}^{-1}\text{V}_{01}^T$ represents the reduction in asymptotic variance from estimating them. Note also that $\text{V}_{00} - \text{V}_{01}\text{V}_{11}^{-1}\text{V}_{01}^T$ is the covariance matrix of the residual vector when Q_0 is regressed on S_1 so the reduction is maximized by including terms in $p_i(\alpha)$ whose scores are highly correlated with Q_0, i.e highly correlated with the influence functions, whether or not they actually affect the π_is. This is a very similar message to that given for choosing calibrating variables in Section 4 of this paper. We note that when we have a saturated model for π_i the approach discussed here and the calibration approach are identical.

Another point of intersection of our interests with those of this paper is in improved Horvitz-Thompson estimation of regression parameters in two-phase studies (Section 2.5) in which the response variable Y_i is observed together with some components of X, say X_{1i}, at Phase 1, while at Phase 2 the remaining components X, X_{2i}, are observed for units with $R_i = 1$ but are unobserved otherwise. For the two-phase problem $U_i(\theta) = U(Y_i, X_{1i}, X_{2i}; \theta)$. The regression-estimation version of the approach given in Section 2.5 is to solve the APIW equations $T(\theta) = 0$ where

$$T(\theta) = \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_i(\theta) + \left(1 - \frac{R_i}{\pi_i}\right)\phi_i \tag{1}$$

for some choice of ϕ_i. As the authors note, the efficient choice for ϕ_i is $\text{E}\{U_i(\theta_0)|Y_i, X_{1i}\}$ where θ_0 solves $\sum_{i=1}^{N} U_i(\theta) = 0$. In practice some strategy for estimating this quantity is required, and more particularly, one that does not rely on correctly specifying a parametric model for the distribution of X_2 given Y and X_1.

Breslow *et al.* (2009) deal with the situation in which there is good Phase-1 information available for forming predictions $\widehat{X}_{2i}$ of the missing X-variable(s) and, following Kulich & Lin

(2004), use $\widehat{\phi}$ with $\widehat{\phi}_i = U(Y_i, X_{1i}, \widehat{X}_{2i}; \widetilde{\theta})$ as their regression estimation or calibration variables. Here $\widetilde{\theta}$ solves $U(Y_i, X_{1i}, \widehat{X}_{2i}; \widetilde{\theta}) = 0$. In an unpublished MSc thesis, Fiona Grimson has explored this, together with iterative versions in which $\widehat{\phi}_i$ is updated when θ is updated, and also strategies employing direct estimation of U_i as a function of the phase 1 variables rather than indirect estimation via $\widehat{X}_{2i}$. She was working in cases where the model of interest was a linear or logistic regression model. Her results show clear benefits from iterative updating of θ in forming $\widehat{\phi}$.

References

Breslow, N.E., Lumley, T., Ballantyne, C.M., Chambless, L.E. & Kulich, M. (2009). Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Stat. Biosci.*, **1**, 32–49.

Grimson F.L. (2011). Methods for Utilising Partially Observed Data in Two-Phase Sampling. Unpublished MSc Thesis, University of Auckland.

Watson, D.J. (1937). The estimation of leaf areas. *J. Agric. Sci.*, **27**, 474.

Lawless, J.F., Kalbfleisch, J.D. & Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. R. Stat. Soc. Ser. B*, **61**, 413–438.

Kulich M. & Lin D.Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Amer. Statist. Assoc.*, **99**, 832–844.

Scott, A.J. & Wild, C.J. (2011). Fitting regression models with response-biased samples. *Canad. J. Statist*, **39**, in press.

[*Received May 2011, accepted May 2011*]

International Statistical Review (2011), 79, 2, 230–232 doi:10.1111/j.1751-5823.2011.00143.x

Rejoinder

Thomas Lumley

Department of Biostatistics, University of Washington, Seattle, WA and Fred Hutchinson Cancer Research Center, Seattle, WA
E-mail: tlumley@uw.edu

My co-authors and I would like to thank all the discussants for their thoughtful comments. Mark Handcock raises some important points about the use of auxiliary information. I would agree only in part with the concern that endogenous weights changing with the analysis could lead users astray. When calibration is used only to add population-level information to a genuine probability sample, as we have considered in this paper, the target of inference is not changed. Parameters estimated with calibrated weights mean the same thing as those estimated with uncalibrated weights, and two estimates of the same quantity as part of differently-calibrated analyses will not be identical, but should agree to well within the sampling uncertainty. Whether this level of agreement is sufficient depends on the context; it could well matter for major economic indicators, but is less likely to be an issue in research. On the other hand, the main practical role of calibration in large surveys is to correct for non-response and imperfections in the sampling

frame. The precise details of the calibration approach do then affect the interpretation of the results, as they determine which parts of the non-response bias are corrected and which parts remain.

When the manuscript was originally being written I was unaware of the related literature on empirical likelihood, stretching back to Hartley & Rao (1968) and now undergoing active development. The two-step empirical likelihood estimator of Chaudhari *et al.* (2008) is very close to being a calibration estimator in the case of independent sampling. The empirical likelihood estimator finds weights w_i for individual i in the sample that maximize

$$\sum_{i=1}^{n} \log n w_i,$$

subject to the population information constraints and the constraint that the weights sum to 1, and then solves the weighted score equations for the regression coefficients. A calibration estimator finds weights that minimize

$$\sum_{i=1}^{n} \frac{1}{\pi_i} G(g_i)$$

for some suitable function G, subject to the calibration constraints, where g_i are the calibration adjustments and $1/\pi_i$ the sampling weights. If we could take $G(x) = -\log(x)$, the calibration estimator would minimize

$$\sum_{i=1}^{n} \frac{-\log(g_i)}{\pi_i}$$

and be equivalent to the empirical likelihood estimator, which has all π_i equal and so also has $g_i \propto w_i$.

For computational reasons, however, the class of calibration estimators was defined to require G to be convex and increasing, with $G(1) = G'(1) = 0$ and $G''(1) = 1$, and $G(x) = -\log(x)$, being convex and decreasing, does not qualify. The impact of this largely technical difference seems to be that the estimation algorithms fail in different ways when the constraints are unattainable. Another source for relationships between the approaches is Chan (2010), who examines some "generalized empirical likelihood" estimators, and also discusses multiple-robustness properties when π_i are not known, as for missing data.

The major difficulty in extending empirical likelihood to general survey designs is going beyond the exchangeability present in simple or stratified random sampling. Rao & Wu (2010) manage this by rescaling the likelihoods using a design effect, an approach that requires a purely design-based estimator as a starting point and so limits the extensions to more general Bayesian models. I have not had time to fully digest the technical report of Chaudhari *et al.* (2010), but it seems to have a broader scope.

I do not hold out much hope for calibration approaches in respondent-driven sampling, not so much because the sampling probabilities are unknown, but because there is little precise auxiliary information available about hard-to-reach populations. Techniques from multiframe sampling seem more promising, where different strategies are used to sample the population, each one possibly incomplete but covering a different subset.

As Tsiatis and Davidian illustrate, there is a lot known now about the details of double-robust estimation, and these details can matter in practical application. My main concern about double robustness is that the statistical literature often overstates the likely benefit. The chance of getting one model out of two right is only much higher than the chance of getting a single model right

if both numbers are small, and the real benefit of double robustness seems to be the ability to use substantive knowledge of the problem in two qualitatively different ways.

Scott and Wild are completely correct in criticizing the claim following equation (2)

$$\text{var}\left[\hat{T}_{reg}\right] \approx (1-\rho^2)\ \text{var}\left[\hat{T}\right].$$

This claim does implicitly assume that none of the $1-\rho^2$ gain in precision available from the population-level information has already been absorbed by the design. The actual gain could be either greater or less than $1-\rho^2$, depending on how well the design is targeted at the parameter being estimated. However, I would argue that for regression parameters the approximation is likely to be good. Any use of population information in the design is likely to be limited to stratification on individual variables. I would expect stratification on individual predictor or outcome variables to add relatively little information about a regression parameter, for the same reasons that calibration on individual variables adds little information. It is certainly my experience that design effects for regression parameters typically vary less than design effects for the means or totals of the same variables.

The benefits of iteration found by Grimson are interesting. I had explored iteration for the Cox model, in the research that later became Breslow *et al.* (2009), and had not found any benefit. Further research is clearly needed to characterize the conditions where iteration is useful.

Lawless and Kalbfleisch describe the approach of Chen & Chen (2000), which in fact is yet another approach equivalent to calibration and to the regression estimator, using $V(X_i;\hat{\gamma})$ as the auxiliary variables. The question here is how to choose $V()$ so that it is correlated with the estimating functions $U()$. It is clear either from the regression approach or the estimating functions approach that any choice of $V()$ is valid; the difficult part is choosing $V()$ so that the correlation is high. The optimal choice,

$$V(X_i,\gamma) = E\{U(Y_i;\gamma)|X_i;\gamma\}$$

is infeasible, and as we showed in Section 4.1, there are obvious and plausible choices that perform very poorly when β is a regression parameter. Our strategy is to work by analogy with the convolution theorem and argue that a good way to make $V()$ correlated with $U()$ is to make it an estimating function for the same parameter, at least under a correctly-specified imputation model.

Lawless and Kalbfleisch say *By way of comparison with Robins et al. (1995), note that* $\hat{B}_{12}\hat{B}_{22}^{-1}V(X_i,\hat{\gamma})$ *is an estimate of* $E\{U(Y;\hat{\beta})|V(X;\gamma^*)\}$. This is true; furthermore, it is a weighted least squares regression estimator, and is the same as our $\hat{T}_{\text{reg}}$. Tastes will vary, so a diversity of explanations and constructions is useful, but I find the explanation of efficiency gains from least squares regression more transparent than those from estimating function theory.

References

Chan, K-W.G. (2010). Oracle and multiple robustness properties of survey calibration estimator in missing response problem (December 30, 2010). UW Biostatistics Working Paper Series. Working Paper 368. http://www.bepress.com/uwbiostat/paper368.

Hartley, H.O. & Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, **55**, 547–557.

Rao, J.N.K. & Wu, C. (2010). Bayesian pseudo empirical-likelihood intervals for complex surveys. *J. R. Stat. Soc. B*, **72**, 533–544.

[*Received May 2011, accepted May 2011*]