A Bayesian Approach to Combining Population and Survey Data for Male Fertility Estimation

Ryan Admiraal, Mark S. Handcock, Kara Joyner, H. Elizabeth Peters, Michael S. Rendall, and Felicia Yang

Short Abstract

Where survey data are used for quantitative assessment of the relationship of socio-economic variables to male fertility, data quality is a concern. Survey data on male fertility are known to be worse for men than for women. Population-level data, especially for non-marital births, are also available in less detail for men, and often with lower degrees of completeness. Nevertheless, given the known problems with men's fertility reports in survey data, it may be even more necessary to use what data are available from population sources. We do so in a study of the socio-economic associations with male fertility, using different statistical methods that successively relax assumptions about the accuracy of the population data from birth registrations and the survey data from the 2002 National Survey of Family Growth.

Extended Abstract

The close connection among family formation, childrearing, the acquisition of education, and labor market decisions for women has long been recognized. With the recent focus in the policy and research communities on the role of fathers in families, understanding the tradeoffs and complementarities between family life and market work takes on a central role in analyzing the transition to adulthood for men as well. Because both labor market opportunities (e.g., declining wages for low skilled men) and family processes (e.g., increases in non-marital births and absent fathering) have undergone major changes in recent decades, it is even more critical to understand the factors that affect the transition to fatherhood.

Among the questions that may be addressed with survey data are: What is the relationship between the transition to biological fatherhood and other transitions to adulthood, such as marriage, educational completion, and entry into the workforce? What are the social, economic, policy, relationship and individual factors associated with men fathering further children after they have already become fathers, and what factors lead men to have additional births with more than one partner?

While survey data are needed for quantitative assessment of these socio-economic questions, data quality is a concern. Not only are survey data on male fertility worse for men than for women (see Garfinkel et al 1998 for the NSFH, Mott et al., 2003, for NLSY79; Harris and Boisjoly, 2004, for AddHealth; Lindberg et al 1998 for NSAM, and Rendall et al, 1999, for the PSID). Population-level data, especially for non-marital births, are also available in less detail for men, and often for low degrees of completeness. This makes the benchmarking of survey estimates against population data a further challenge. Nevertheless, because of the known problems with men's fertility reports in survey data, it may be even more important in the case of men's fertility than women's to use what data are available from population sources. To do so, we will experiment with different statistical methods that successively relax assumptions about the accuracy of the population and survey data.

1. Exact and inexact population constrained estimates of male fertility with "representative" survey data

In the first two approaches, we will first use population-level data to either "exactly" or "inexactly" constrain estimates from a single survey data source, the 2002 NSFG, under the assumption that the data from the NSFG are missing fatherhood episodes "randomly by covariate" (Rubin 1977). We call this a "representative" survey dataset.

The population level data are birth registration data with imputed father's age where missing on the birth certificate, and Census Bureau annual population estimates by age and sex. These are used to calculate age-specific male fertility rates for the years 1975-present. The method of imputation when men's data are missing from the birth certificate is that described in the National Center for Health Statistics (NCHS) Vital Statistics Reports (2003). Specifically, we use the age of father reported on the birth certificate to identify fathers by single-year age. Father's age is missing from the birth certificate in substantial numbers of cases. In 2002, information on the age of the father was missing in 13% of all births, 24% of births to mothers aged less than 25, and 38% of non-marital births. In those cases we will allocate births to fathers of a given age in the same proportion as births where the father's age is not missing. As the National Center for Health Statistics report (2003) argues, "This procedure avoids the distortion in rates that would result if the relationship between age of mother and age of father were disregarded." The method is described in detail in the Technical Notes to the report. Note that this method assumes that for a given age of the mother, father's age is not missing in any systematic way.

Our estimated population age-specific male fertility rates, however, will be inaccurate to the degree that this assumption is violated. Earlier research by Elo et al. (1999) investigates this issue by comparing the distribution of fathers' ages from vital statistics data with more complete data on fathers' ages from the 1988 National Maternal and Infant Health Survey (NMIHS). Their results suggest that "there is no systematic bias in the reporting of fathers' ages… on the birth certificate."

In our first estimation, we assume that these population values are exact. This is the estimation assumption in the women's fertility examples addressed by Handcock, Huovilainen (2000) and Handcock, Rendall, and Cheadle (forthcoming). A computationally convenient way of addressing this type of estimation problem when there are multiple covariates and multiple constraints is provided by Chaudury, Handcock, and Rendall (2005).

In our second estimation, we assume that the need for imputation of men's age in some of the NCHS data will lead to population estimates that may not reasonably be characterized as exact, and that may even be biased. Our solution to adjust for this type of bias is to use a Bayesian statistical approach that brings in a third type of information: expert opinion. Under this approach, expert opinion is solicited to weigh the probability that a given value of the constraint (including the value estimated by the NCHS) equals the true value. This approach recognizes that the NCHS value may not be the true value, and that alternative values of the constraint may be. Formally, this is implemented by attaching a "prior" probability distribution over the alternate values of the constraints. The "best" point estimate will receive the highest weight, but the other estimates will receive some weight too in the process of arriving at the Bayesian point estimate.

We now provide specific details of Bayesian statistical inference about the parameter θ_0 given expert opinion about the constraint value ϕ. The Bayesian paradigm explicitly uses probability distributions to express information about θ_0, ϕ or unobserved data in terms of probability statements that are based on the combined information of *each* of the three types. Let $q^{[\phi]}(\phi)$ denote the probability from expert opinion about the constraint value ϕ. This also implies expert opinion about the value θ_0 through the constraint function $C(\theta_0) = \phi$. To combine this information with the survey data we apply Bayes theorem to this distribution and the model. This produces the distribution representing what we know about $\theta_{0.}$ "posterior" to the survey data:

$$p^{[\theta_0]}(\theta; y, x) = \alpha(y, x) \times q^{[\varphi]}(C(\theta)) \times L(\theta; y, x)$$

The first term α(y,x) does not depend on θ, so interest focuses the likelihood L(θ; y, x) and the implied prior distribution for θ_0, $q^{[\varphi]}(C(\theta))$. In this role, the latter term represents the population level information expressed in the constraints and the former expresses the information in the survey sample data about θ_0. Under the Bayesian paradigm, this posterior distribution represents our complete knowledge of θ_0 based on prior knowledge about the constraint value, potentially some prior knowledge about the parameter value for given values of the constraint, and finally the sample survey data itself.

The reduction in sampling error under a Bayesian prior about the constraint will be less than in the case that the constraint is exact. Reduction in sampling error will be achieved, however, in cases of an informative prior. By 'informative', we simply mean that the function $q^{[\phi_1]}(\phi)$ gives a stronger weighting to values of ϕ given by and near those given by the NCHS or similar population value than they give to some randomly chosen value of the constraint. The greater the confidence of expert opinion in the precision of the demographic analyses and data that went into forming the constraint estimation, the greater will be the reductions in both variance and bias. The Bayesian approach has its strength, and weakness, is the quality of the expert judgment. In particular, the degree of bias reduction is directly related to the expert's ability to capture the direction and range of possible values of the constraint. An additional strength of the Bayesian approach is the ability to easily combine expert opinion with additional empirical information from either sample or population data.

2. Bayesian priors that take into account biases in the survey dataset

The above approach fails to take into account the problem that men's births are unlikely to be missing at random by covariate, and are instead likely to induce elements of "non-representativeness" in the survey dataset. For example, non-marital fertility reporting may be expected to be much worse where the man is not living with the mother then when he is living with the mother, and may be worse among black men than among white men. This information (coresidence and race) is contributed by the survey data only in our example study, as we do not attempt to estimate race-specific fertility from the NCHS data.

This problem of non-random missingness on the covariates of interest may be addressed by forming Bayesian priors that are specific to the possible deficiencies of a given survey source ---- here the 2002 NSFG. For example, the retrospective nature of the NSFG may make it more likely that births in previous marriages are not reported as compared to in panel data that picks up the births while the marriage is still intact (Rendall et al 1999).

We sketch out in this section a Bayesian approach to dealing with non-representative survey data, in the more general case in which two non-representative survey data sources are combined with the population data. This allows for different representations of the uncertainty about survey data. We indicate some of the challenges for the implementation and further development of this approach.

Using the notation of the previous section, let $C_1(\theta_0) = \phi_1$ be the constraints on the first survey and $q^{[\phi_1]}(\phi)$ the prior distribution representing expert knowledge for the constraint value. This knowledge may be derived from prior demographic analysis. One such example is the comparing of total births reported by men with total births of a given type reported by men (Rendall et al 1999). Let $C_2(\theta_0) = \phi_2$ and $q^{[\phi_2]}(\phi)$ be the corresponding elements of the second survey. As in the section above, the effect of the constraints might have the result, for example, of pushing up the intercept parameter of men's fertility in previous marriages by a greater amount in the constraint applied to the first survey dataset than in the second. They might at the same time, however, decrease the coefficient parameter for black men's fertility relative to white men's fertility in the first survey but increase it in the second. If the expert opinion (and demographic analysis underlying it) is valid, the surveys augmented by this information will be approximately representative, at least with respect to the dimensions of overall and race-specific fertility.

Formally, under this approach the combined posterior distribution of θ_0 is:

$$p^{[\theta_0]}(\theta; y,x) = \alpha_2(y,x) \times q^{[\varphi_1]}(C_1(\theta)) \times L_1(\theta; y,x) \times q^{[\varphi_2]}(C_2(\theta)) \times L_2(\theta; y,x)$$

Here $L_1(\theta; y,x)$ and $L_1(\theta; y,x)$ are the likelihoods for the first and second surveys, respectively. If we had not conducted a demographic analysis of the surveys, the equation would be reduced to the product of the two likelihood functions. Bias will then be left uncorrected. Its level will be a weighted average of the bias present in each of the two surveys, where their respective sample sizes provide the weights. Reductions in sampling error, however, will be realized by the pooling of the two datasets. The reduction in sampling error will be less than in the case that a demographic analysis had been undertaken leading to an informative prior. By 'informative', we now mean that the functions $q^{[\phi_1]}(\phi)$ and $q^{[\phi_2]}(\phi)$ give a stronger weighting to values of ϕ given by and near those given by the demographic analysis than they give to some randomly chosen value of this constraint. In the extreme, where the expert believes the constraint values to be exact, we are close to the special case first presented above, in which exact population constraints were combined with the representative survey data. The difference in this general presentation is that now we realize reductions in bias as well as variance by using the constraint with non-representative survey data.

References

Chaudhury, S., M. S. Handcock, and M. S. Rendall Generalised linear models incorporating population information: An empirical likelihood based approach. Center for Statistics and the Social Sciences Working Paper, University of Washington, Seattle.

Elo, I. T., R. B. King, and F. F. Jr. Furstenberg. 1999. "Adolescent Females: Their Sexual Partners and the Fathers of Their Children." Journal of Marriage and the Family 61(1): 74-84.

Fitzgerald, J., P. Gottschalk, and R. Moffitt. 1998. "An Analysis of Sample Attrition in the Michigan Panel Study of Income Dynamics." Journal of Human Resources 33:251-99.

Garasky, S., and D.R. Meyer. (1996). Reconsidering the Increase in Father-only Families. Demography 33(3):385-393.

Garfinkel, I., S. S. McLanahan, and T. L. Hanson. (1998). A Patchwork Portrait of Nonresident Fathers. In Fathers Under Fire: The Revolution in Child Support Enforcement, edited by I. Garfinkel, S. McLanahan, D. Meyer, and J. Seltzer. New York: Russell Sage.

Gelman, A. Carlin, J. B., Stern, H. S., and Rubin D. B. (1995). Bayesian Data Analysis New York: Chapman Hall.

Goldstein H. (1991) Nonlinear Multilevel Models, with an Application to Discrete Response Data Biometrika 78(1):45-51.

Handcock, M. S., S.M. Huovilainen, and M. S. Rendall. (2000) Combining Registration-System and Survey Data to Estimate Birth Probabilities Demography.

Handcock, M. S., M. S. Rendall, and J. E. Cheadle. (Forthcoming). Improved regression estimation of a multivariate relationship with population data on the bivariate relationship Sociological Methodology.

Hansen, L. P. (1982) Large Sample Properties of Generalized Method of Moments Estimators. Econometrica 50:1029-1054.

Harris, K.M. and J. Boisjoly. 2004. "Measuring Male Fertility in Add Health." Working paper.

Hellerstein, J., and G.W. Imbens (1999) Imposing moment restrictions from auxiliary data by weighting Review of Economics and Statistics 81(1):1-14.

Hill, M.S. 1992. The Panel Study of Income Dynamics: A User's Guide. Newbury Park, California: Sage.

Holt, D. and Smith T. F. M. (1979) Post-Stratification. Journal of the Royal Statistical Society, Series A. 142:33-46.Imbens, G.W. and T. Lancaster (1994) Combining micro and macro data in microeconometric models. Review of Economic Studies 61: 655-680.

Imbens, G.W. and T. Lancaster (1994) Combining micro and macro data in microeconometric models. Review of Economic Studies 61: 655-680.

Ireland, C. T., and Kullback, S. (1968) Contingency tables with Given Marginals. Biometrika 55:179-188.

Juby, H., and C. LeBourdais. (1998). The changing context of fatherhood in Canada: A life course analysis. Population Studies 52:163-175.

Kish, L. (1965) Survey Sampling New York: Wiley.

Lancaster, T., and G.W. Imbens (1996) Case control studies with contaminated controls Journal of Econometrics 71:145-160.

Lindberg, L. D., F. L. Sonenstein, G. Martinez, and J. Marcotte. 1998. "Completeness of Young Fathers' Reports of Fertility.Journal of Economic and Social Measurement 24(1): 15-23.

Little, R. J. A. (1993) Post-stratification : A modeler's perspective Journal of the American Statistical Association 88(423):1001-1012.

Little, R. J. A. and Rubin D. B. (1987) Statistical Analysis of Missing Data. New York: John Wiley.

Little, R. J. A., and Wu, M. M. (1991) Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ Journal of the American Statistical Association 86(413):87-95.

Manski, C. (1988) Analog Estimation Methods in Econometrics New York: Chapman Hall.

Mott, F.L., Dawn S. Hurst, and Thomas Gryn. (2003). Center for Human Resource Research, The Ohio State University. Working paper.

Owen, A. B. (1988) Empirical Likelihood Ration Intervals for a Single Functional. Biometrika 75:237-249.

Poole, D., and Raftery, A. E. (1998) Inference for Deterministic Simulation Models: The Bayesian Melding Approach. Technical Report 436, Department of Statistics, University of Washington.

Raftery, A. E., Givens, G. H., and Zen, J. E. (1995) Inference from a Deterministic Population Dynamics Model for Bowhead Whales (with discussion). Journal of the American Statistical Association 90(423):402-430.

Rendall, M.S., L. Clarke, H. E. Peters, N. Ranjit, and G. Verropoulou (1999) Incomplete Reporting of Male Fertility in the United States and Britain: A Research Note. Demography 36(1):135-144.

Rice J. A. (1995) Mathematical Statistics and Data Analysis Pacific Grove:Wadsworth.

Rubin, D. B. (1976) Inference and Missing Data. Biometrika 63:581-592.

Rubin, D. B. (1977) Formalizing Subjective Notions About The Effect on Non-respondents in Sample Surveys. Journal of the American Statistical Association 72(405):538-543.

Wolpert, R. L. (1995) Comment: Borels paradox. Journal of the American Statistical Association 90(423):426-427.