Annotated Bibliography

Merkouris et al. (2023)

Title: Combining National Surveys with Composite Calibration to Improve the Precision of Estimates from the United Kingdom's Living Costs and Food Supply

Authors: Takis Merkouris, Paul A. Smith, Andy Fallows

Journal (Year): JSSAM (2023)

Summary: In some sense this is the application of ideas from Merkouris (2004) and Merkouris (2010) to combine information from the Living Costs and Food Survey (LCF) and the Labour Force Survey (LFS) in the United Kingdom. The overall approach is to use a linear combination of regression estimators from each survey:

$$\hat{Z}^{CR} = B\hat{Z}_1^R + (I - B)Z_2^R$$

where B is a matrix, Z_1^R is the regression estimator from the first survey and Z_2^R is the regression estimator from the second survey. Like Merkouris (2004), they choose the optimal B that minimizes the variance of the estimator while still having the estimation of $\hat{Z}$ be the same for each survey. To incorporate results from Merkouris (2010), they also control totals at the regional level (a small-area estimation problem). (More details regarding the derivation of the optimal B is left to the summary of Merkouris (2004) and Merkouris (2010).)

Methodologically, the biggest innovation is the jackknife variance estimation that they use. However, even this seems relatively straightforward. I think that the main contribution of this paper is to show a dramatic reduction in the variance of regression estimates due to data integration.

Limitations:

- They only integrate two surveys.
- They do not consider informative sampling.

Extensions:

- I think that it would be interesting if the supports of several control variables were not the same. This would be evidence of selection bias and we would need to account for it.

Merkouris (2004)

Title: Combining Independent Regression Estimators from Multiple Surveys

Author: Takis Merkouris

Journal (Year): JASA (2004)

Summary: This paper develops the methodology to combine regression estimators from multiple surveys. Given two surveys the paper suggests combining the regression estimators as,

$$\hat{Z}^{CR} = B\hat{Z}_1^R + (I - B)\hat{Z}_2^R$$

where $B = \phi Z_2' L_2 Z_2((1-\phi)Z_1' L_1 Z_1 + \phi Z_2' L_2 Z_2)^{-1}$, $\phi = \frac{n_1/d_1}{n_1/d_1 + n_2/d_2}$, n_i is the sample size from survey i, d_i is the design effects from survey i of z, Z_i is a matrix of common columns between the two surveys, $L_i = I - X_i(X_i'X_i)^{-1}X_i'$, and X_i are additional covariates that we have information to calibrate. In this way, the combined regression estimator can have minimal variance while having the estimates of Z from each survey match. The weights also adjust for sample size and the strength of the correlation between Z and X. This is the main method used in Merkouris (2023) except that Merkouris (2023) also adjusts for regional totals as developed in Merkouris (2010).

Merkouris (2010)

Title: Combining information from multiple surveys by using regression for efficient small domain estimation

Author: Takis Merkouris

Journal (Year): JRSSB (2010)

Summary: The bulk of this paper compares three estimators: the population control $\hat{Y}_d^R \equiv \hat{Y}_d + \hat{\beta}_d'(t_x - \hat{X})$ where $\hat{\beta} = (X_s'\Lambda_s X_s)^{-1}X_s'\Lambda_s Y_{s_d}$, the domain control $\check{Y}_d^R \equiv \hat{Y}_d + \check{\beta}_d'(t_{x_d} - \hat{X}_d)$ where $\check{\beta}_d = (X_{s_d}'\Lambda_{s_d}X_{s_d})^{-1}X_{s_d}'\Lambda_{s_d}Y_{s_d}$, and the domain size control $\tilde{Y}_d^R \equiv \hat{Y}_d^R + \tilde{\beta}_d'(N_d - \hat{N}_d^R)$ where $\tilde{\beta}_d = (1_{s_d}'L_s 1_{s_d})^{-1}1_{s_d}'L_s Y_{s_d}$. Notice that the last estimator is based off of the population control $\hat{Y}_d^R$ instead of the HT-estimator $\hat{Y}_d$. Since this model incorporates both the population variables and the domain indicators, the corresponding regression estimator can be thought of as an application of the augmented regression in Merkouris (2004).

Chen et al. (2022)

Title: Nonparametric Mass Imputation for Data Integration

Author: Sixia Chen, Shu Yang, Jae Kwang Kim

Journal (Year): JSSM (2022)

Summary: This paper aims to combine data from a probability sample A and a non-probability sample B using non-parametric methods. We assume the X is observed in A and B but Y is only available in B. It achieves this goal by using a kernel estimator for the outcome model $m(x)$ on sample B and then using this model to predict missing values of Y for A. Since the sample weights in A are valid, these are used to adjust the predicted values $\hat{m}(x)$. Variance estimates are also given.

Limitations:

- The model require MAR for the response model in B.

Extensions:

- Can we extend this to multiple non-probability samples?

Hidiroglou (2001)

Title: Double Sampling

Author: M. A. Hidiroglou

Journal (Year): Survey Methodology (2001)

Summary: This paper produces the optimal regression estimator for non-nested probability samples. This occurs when sample A observes X while B observes X and Y but unlike two-phase sampling there is no overlap between the observed values for X in A and B.

This paper also explores the connection between non-nested samples and calibration.

Limitations:

- Unclear to me which class of estimators we are optimizing within.

Extensions:

- How to extend to a third category?

Park and Kim (2019)

Title: Mass Imputation for Two-Phase Sampling

Author: Seho Park and Jae Kwang Kim

Journal (Year): arXiv (2019)

Summary: This paper proposes a mass imputation estimator for estimating the total from a two-phase sample. Instead of using an estimate from phase 1 data and combining it with phase 2 data, this approach uses a regression model to predict the missing data in phase 1. The paper shows how this result is consistent with previous estimates and they provide a variance estimator.

Limitations:

- What if the regression model is incorrect?
- Does this account for enough uncertainty if just the predicted values are used?

Extensions:

- How can this be extended to another phase.