Fay-Herriot Model for a Income and Living Condition Survey: Application with R

Tobias Schoch (February 15, 2022)

Morales et al. (2021, Chapter 1.2) have prepared data on the basis of a living conditions survey (LCS).[1] The goal is to estimate average income for 26 small areas.

First, we introduce the LCS survey data. Then, we study the direct estimator of average income by area. Finally, we estimate the Fay-Herriot model. Along our discussion, you will be asked to solve tasks.

- **Task 1.** Compute the direct estimator of average income;
- **Task 2.** Compute the generalized variance function of the direct estimator's variance (this will be discussed below);
- **Task 3.** Compute estimates of average income by area with the Fay-Herriot model (incl. model diagnostics, and estimation of mean square error).

We will come back to the tasks as we go along.

1 Survey Data

1.1 LCS Survey

The survey data `datLCS` contains 6 variables, which are measured for individuals living in private households. The households are identified by variable `house`. In total, the dataset contains data on 2512 individuals living in 962 households. The sampling weight `w` refers to the households. There are 26 small areas (identified by the domain indicator `dom`). The variables of `datLCS` are described below.

Variable	Description
`sex`	man=1; woman=2
`house`	household identifier
`income`	net equivalent income in euros
`lab`	labor status (0=child, i.e., age < 16 years; 1=employed; 2=unemployed; 3=inactive)
`dom`	domain indicator (small area)
`w`	sampling weight (level: household; calibrated)

The data are stored as `datLCS.txt` file. We can load the data by

```
datLCS <- read.table("datLCS.txt", header = TRUE, sep = "\t", dec = ",")
```

The first three lines of `datLCS` are printed below.

```
dom  sex  house          w    income  lab
 27    1     68   3022.840   6262.40    3
 27    2     68   3022.840   6262.40    3
 27    1     68   3022.840   6262.40    3
```

1.2 Direct Estimator

Average income in the areas $i = 1, \ldots, 26$ is computed with the Hajek estimator, which is defined as

$$\bar{y}_i = \frac{1}{\widehat{N}_i} \sum_{j \in s_i} w_j \cdot \text{income}_j, \qquad \text{with} \quad \widehat{N}_i = \sum_{j \in s_i} w_j,$$

where s_i is the part of the sample s that falls into the i-th area, and w_i denotes the sampling weight.

There are several equivalent ways to compute the Hajek estimator for the small areas (defined by `dom`) with the R software (R Core Team, 2022).[2] We stick to the functions of the R `base` package. First, we split the `datLCS` data into a list by `dom` .

```
datLCS_dom <- split(datLCS, datLCS$dom)
```

The object `datLCS_dom` is a list with 26 list entries (one for each area). The list entries consist of the area-specific part of the `datLCS` data. In the next step, we use `sapply()` to compute the Hajek estimator (i.e., weighted mean) by area.

```
sapply(datLCS_dom, function(u) weighted.mean(u$income, u$w))
```

A more complete function of the Hajek estimator (which is also capable of computing an approximate variance of the estimator) is given by

```
hajek <- function(x, w)
{
    avg <- weighted.mean(x, w)                        # Hajek estimator
    ni <- length(w)                                   # sample size
    vi <- sum(w * (w - 1) * (x - avg)^2) / sum(w)^2   # variance
    c(avg = avg, vi = vi, ni = ni)
}
```

Note that function `hajek()` also retrieves the sample size n_i. The return value of the function is a vector of size 3.

Task 1. Use the `hajek()` function to compute the Hajek estimator and its variance, $v_i = \text{var}(\hat{\theta}_\text{i})$, for all areas $i = 1, \ldots, 26$. Also, compute the coefficient of variation (in %), defined as

$$cv_i = 100 \cdot \frac{v_i}{\bar{y}_i},$$

for all $i = 1, \ldots, n$ areas. Finally, we want to combine the Hajek estimator (`avg`), its variance (`vi`), sample size in the i-th area (`ni`), and the coefficient of variation (`cv`) to one data.frame called `direct` using the `as.data.frame()` function; maybe you must transpose the result using `t()` in order to obtain a rectangular representation of data (with area-specific observations on the rows and variables in the columns).

2 Auxiliary Data

The file `auxLCS.txt` contains aggregated auxiliary data for all $i = 1, \ldots, 26$ areas. We can load the data by

```
auxLCS <- read.table("auxLCS.txt", header = TRUE, sep = "\t", dec = ",")
```

The first three observations of `auxLCS` are printed below

```
dom     TOT     Mwork     Mnowork     Minact         ss
  3   82001 0.3632226 0.12764258 0.3574135 0.5695195
  5  251866 0.3564652 0.15503770 0.3192122 0.4323160
  6  190653 0.3405221 0.15860230 0.3158246 0.5998553
```

where

- `TOT`: total number of individuals in area,
- `Mwork`: domain mean of `lab=1`,
- `Mnowork`: domain mean of `lab=2`,
- `Minact`: domain mean of `lab=3`.

We will utilize the auxiliary information as explanatory variables in the Fay-Herriot model. The dataset `auxLCS` has been processed by Morales et al. (2021) and is ready to use. In practice, we have to take care of this process ourselves.

3 Fay-Herriot Model

In theory, we can take the estimated variances v_i of the direct estimators $\hat{\theta}_i$ and estimate the Fay-Herriot model without further ado—in practice, we usually cannot because some of the v_i's are too unstable.

In Section 3.1, we introduce the notion of *generalized variance functions* (GVF). The variances computed using GVF's tend to be much more stable than the variances of the direct estimator. Therefore, this approach is preferred in practice. Readers who are not interested in the details of GVF can skip Section 3.1 and go directly to Section 3.2. In Section 3.2, we provide a recipe for computing a GVF (without having to know details).

3.1 Generalized variance function (theory)

For very small areas, the estimated variances v_i are typically very unstable and thus unreliable. Some additional stability can be gained by using *generalized variance functions* for estimating the area-specific variances. This approach has a long history in survey sampling.

A generalized variance function (GVF) is a *model* or method that attempts to compute the variances of an estimator by exploiting a simple mathematical relationship connecting the variance to the expectation of the estimator (Wolter, 2007, p. 272). To be explicit, let us consider estimating a population characteristic θ (e.g., mean). Let $\hat{\theta}$ be an unbiased survey estimator of the population parameter θ with variance $v(\hat{\theta})$. The form of the estimator and the sampling design are left unspecified. We define the *relative variance* of $\hat{\theta}$ by

$$V^2 = \frac{v(\hat{\theta})}{\theta^2}. \tag{A}$$

A GVF attempts to model the relationship in (A). Most of the GVF's are based on the *premise* that V^2 is a decreasing function of θ—that is, V^2 becomes smaller as θ increases. A simple GVF (or model) which has this property is

$$V^2 = \alpha + \frac{\beta}{\theta}, \tag{B}$$

where α and $\beta > 0$ are unknown parameters to be estimated. Observe the similarity of (A) and (B)—the denominator of the second term on the r.h.s. in (B) is θ not θ^2. Clearly, the parameters α and β depend upon the population, the sampling design, the estimator, etc. This model has been used in the US Current Population Survey since 1947 (Wolter, 2007, p. 274).

Remarks.

- The GVF in (B) is one possible model; other commonly used models are, e.g., $V^2 = (\alpha + \beta\theta)^{-1}$ or $\log\left(V^2\right) = \alpha - \beta \log\left(\theta\right)$.
- With empirical data, we substitute the v_i's for V and replace the $\hat{\theta}_i$'s for θ in (B) for all $i = 1, \ldots, 26$ areas. Estimates of α and β can be obtained by—for instance—ordinary least squares.
- It is helpful to plot v_i against $\hat{\theta}_i$ (scatter plot) to learn more about the functional form of the V^2 vs. θ relationship (in order to select an appropriate model).
- We seek to achieve a good empirical fit (model selection).
- It can be useful to estimate the parameters (i.e., α and β in Equation B) only on a subset of the data or use some kind or grouping. For instance, suppose that 5 out of 26 small areas have extremely unreliable v_i's; thus, we exclude the 5 areas and fit the model to the data of the remaining areas.
- One danger to be avoided is the possibility of negative variance estimates. This can be avoided by using some kind of restricted estimating method (e.g., restricted least squares such that α is constrained to be positive).
- GVF's proved to be very useful in practice. Unfortunately, there is very little theoretical justification for any of the models (Wolter, 2007, p. 274).

Suppose we have fitted model (B). The estimated parameters are denoted by $\widehat{\alpha}$ and $\widehat{\beta}$. Next, we can predict the variances under the GVF model in (B) as $v_i^* = \widehat{\alpha} + \widehat{\beta}/\hat{\theta}_i$, for $i = 1 \ldots, 26$. This variance should be much more stable than the v_i's.

3.2 Application of GVF for LCS data

According to Morales et al. (2021, p. 454), we fit the following GVF model by ordinary least squares to the data of all $i = 1, \ldots, 26$ small areas

$$\log\left(v_i\right) = b_0 + b_1\hat{\theta}_i + b_2 n_i + b_3 n_i \hat{\theta}_i + e_i, \tag{C}$$

where $\hat{\theta}_i$ is the Hajek (direct) estimator of average income, v_i its variance, n_i is the sample size in the i-th area, and e_i is a random error with $e_i \sim N(0, \sigma_e^2)$.

Task 2. This task includes three steps.

- Fit the GVF model in (C) by ordinary least squares and assign the fitted model to object `est`, i.e., call `est <- lm(...)`. [Note: In Task 1, we have generated all data that is required for this task]
- Given the fitted GVF model (which we have assigned to `est`), compute the residual variance (i.e., an estimate of σ_e^2), which is defined as

  ```
  sigma_e2 <- sum(residuals(est)^2) / df.residual(est)
  ```

- Obtain the predicted values of the estimated model `est` using the `predict()` command. Assign the predicted values to the object `p`. Now, we are ready to compute the variances of the GVF model, v_i^*, as `exp(p + sigma_e2 / 2)` for all $i = 1, \ldots, 26$ small areas.[3] Assign the so computed v_i^*'s to the data.frame `direct` and call the new variable `vi_gvf` (where the suffix `_gvf` reminds us that these variances have been computed by the GVF approach).

3.3 Estimation of the Fay-Herriot model

Now, we are almost ready to fit the Fay-Herriot model to our data. In the last step of preparation, we add the auxiliary information `auxLCS` to the data.frame `direct`. To be on the safe side, we sort the data `auxLCS` by `dom` using `auxLCS[order(auxLCS$dom), ]`; and we do the same for the `direct` data.frame. Then, we can safely add the two data.frames with the `cbind()` command (without worrying to have generated mismatch).

Task 3.

- Load the `sae` R package.
- Estimate the Fay-Herriot model for average income (direct Hajek estimator `avg` with GVF variance `vi_gvf`) using the auxiliary variables `Mwork`, `Mnowork`, and `Minact`. We are interested in the REML estimate of variance.
- Estimate same model but this time without variable `Mwork`. Why do we drop variable `Mwork`?
- Use the model from the last step and compute the second-order approximation to the mean square error (MSE) using the function `mseFH()`. Assign the estimated MSE to object `m`.
    - Extract the EBLUP by `m$est$eblup[, 1]`. Make a scatter plot the EBLUP against the Hajek estimator (`avg`). Add the 45-degree line to the plot.
    - The estimated MSE can be extracted by `m$mse`. Make a line plot of the estimated MSE for the $i = 1, \ldots, 26$ areas. Add a line for the GVF variance of the Hajek estimator (`vi_gvf`).

Notes

[1] The LCS data synthetically generated data that imitate the structure of an income an living condition survey; see Morales et al. (2021, p. 4).

[2] Instead of the functions in the R `base` package, we may use R packages `data.table` or `tidyr` to compute the area-specific estimates.

[3] The naive predictor under model (C) is $\exp\left(\mathbf{x}_i^T\widehat{\mathbf{b}}\right)$, where $\widehat{\mathbf{b}} = (\hat{b}_0, \ldots, \hat{b}_3)^T$ is the least squares estimate and $\mathbf{x}_i$ denotes the vector of explanatory variables. This predictor can be heavily biased—in particular if the estimate of σ_e^2 is large. If we assume that the v_i's have a lognormal distribution, the predictor is $\exp\left(\mathbf{x}_i^T\widehat{\mathbf{b}} + \widehat{\sigma}_e^2/2\right)$.

References

Morales, Esteban, Pérez & Hobza (2021). *A Course in Small Area Estimation and Mixed Models: Methods, Theory and Applications in R*, Cham: Springer Nature.

Wolter (2007). *Introduction to Variance Estimation*, New York: Springer.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.