The Stata Journal (2019)
19, Number 3, pp. 626–644 DOI: 10.1177/1536867X19874238

The fayherriot command for estimating small-area indicators

Christoph Halbmeier
Freie Universität Berlin, DIW Berlin
Berlin, Germany
chalbmeier@diw.de

Ann-Kristin Kreutzmann
Freie Universität Berlin
Berlin, Germany
ann-kristin.kreutzmann@fu-berlin.de

Timo Schmid
Freie Universität Berlin
Berlin, Germany
timo.schmid@fu-berlin.de

Carsten Schröder
Freie Universität Berlin, DIW Berlin
Berlin, Germany
cschroeder@diw.de

Abstract. We introduce a command, `fayherriot`, that implements the Fay–Herriot model (Fay and Herriot, 1979, *Journal of the American Statistical Association* 74: 269–277), which is a small-area estimation technique (Rao and Molina, 2015, *Small Area Estimation*), in Stata. The Fay–Herriot model improves the precision of area-level direct estimates using area-level covariates. It belongs to the class of linear mixed models with normally distributed error terms. The `fayherriot` command encompasses options to a) produce out-of-sample predictions, b) adjust nonpositive random-effects variance estimates, and c) deal with the violation of model assumptions.

Keywords: st0570, fayherriot, disaggregated indicators, small-area estimation, (log-transformed) Fay–Herriot model, empirical best linear unbiased predictor

1 Introduction

Various national and international institutions, including the United Nations (Leadership Council of the Sustainable Development Solutions Network 2015) and the Organisation for Economic Co-operation and Development (Piacentini 2014), collect comprehensive indicator sets for monitoring purposes. Many indicators refer to subnational areas or domains: federal states, economic sectors, societal groups, etc.

In the socioeconomic context, domain-level indicators are usually derived from population surveys by direct estimation. Direct estimates are based only on the survey data, so small sample sizes can limit their precision. Thus, institutions that provide these indicators usually require a minimum number of observations per domain or impose limits on the variability of the estimates (Eurostat 2013; Tzavidis et al. 2018). Furthermore, direct estimates cannot be obtained for out-of-sample domains, that is, domains without any observation in the sample.

Small-area estimation techniques use auxiliary data from additional data sources to improve the precision of survey-based direct estimates. Two basic model types can be distinguished: unit- and area-level models. Unit-level models require survey and

auxiliary data at the unit level, that is, individual- or household-level information in each domain. Examples are the model proposed by Battese, Harter, and Fuller (1988) and the empirical best predictor by Molina and Rao (2010). In comparison, area-level models, such as the Fay–Herriot (FH) model (1979),[1] require only domain-level auxiliary data, hence their popularity in applied research.

`fayherriot` provides empirical best linear unbiased predictors (EBLUP), which are linear combinations of the domain-level direct estimator and a regression-synthetic component based on a linear model. The underlying model can also be expressed as a special linear mixed model. In contrast to a standard linear mixed model [encompassed in `mixed` (Rabe-Hesketh and Skrondal 2012) or `gllamm` (see [R] **gllamm**)], the FH model builds on two error terms on the domain level, with domain-specific variances of one error term and a common variance of the other error term. The model assumes linearity and normality of its two error terms. Corral et al. (2018) implement a standard version.

`fayherriot` extends the existing possibilities in Stata and performs the following:

- estimation of the FH model as described in Rao and Molina (2015, 123–129) with restricted maximum likelihood (REML) and maximum likelihood estimation (MLE) of the variance of the random effects,
- estimation of the mean squared error (MSE) as proposed in Datta and Lahiri (2000) and Prasad and Rao (1990),
- prediction and MSE estimation for out-of-sample domains (Rao and Molina 2015, 126 and 139),
- estimation with adjusted methods as proposed in Li and Lahiri (2010) and Yoshimori and Lahiri (2014) to deal with nonpositive estimates of the variance of the random effects,
- estimation of the log-transformed FH model including a bias correction by Slud and Maiti (2006) to deal with violations of model assumptions, for example, nonnormality of the error terms, and
- estimation of the FH model for proportions defined on the $[0, 1]$ interval, that is, with the dependent variable transformed by the arcsine square root transformation. The back-transformation and the corresponding boundaries of a bootstrap confidence interval following Casas-Cordero Valencia, Encina, and Lahiri (2016, 394–397) and Schmid et al. (2017, 1173–1177) are provided.

1. Applications include, for example, the estimation of income and poverty rates (Powers, Basel, and O'Hara 2008; Huang and Bell 2012) and educational indicators (Schmid et al. 2017).

2 The FH model

2.1 Modeling

The FH model (Fay and Herriot 1979) combines domain-level direct estimates (based on survey data) with aggregated domain-level covariates (for example, from register or administrative data). The direct estimator should be a linear statistic such as an arithmetic mean, total, or share.

The FH model builds on a sampling and a linking model. According to the sampling model,

$$\widehat{\theta}_d = \theta_d + e_d \quad \text{for} \quad d = 1, \ldots, D$$

the observed direct estimator for domains $d = 1, \ldots, D$, $\widehat{\theta}_d$, is composed of the true value, θ_d, and a sampling error, e_d, with mean zero and variance $\sigma^2_{e_d}$. The model assumes that the sampling error variance of each domain is known. In practice, the variance of the direct estimator is used frequently as an estimate for $\sigma^2_{e_d}$ (You and Chapman 2006). To consider sampling weights in the FH model, one can use the weighted direct estimator and its corresponding variance. For example, one can use the Horvitz–Thompson estimator for the mean (Horvitz and Thompson 1952). According to the linking model,

$$\theta_d = \mathbf{x}_d^\top \boldsymbol{\beta} + u_d \quad \text{for} \quad d = 1, \ldots, D$$

the true value, θ_d, is explained by the domain-specific covariates, $\mathbf{x}_d$; a random effect, u_d; and the regression parameters, $\boldsymbol{\beta}$. The random effect is independently, identically, and normally distributed with mean zero and variance σ^2_u. The model assumes interdomain correlations to be zero.

Combining the sampling and the linking models gives the FH model, which is a linear mixed model of the form

$$\widehat{\theta}_d = \mathbf{x}_d^\top \boldsymbol{\beta} + u_d + e_d \quad \text{for} \quad d = 1, \ldots, D \tag{1}$$

The FH estimator (EBLUP) is given by $\widehat{\theta}_d^{\text{FH}} = \mathbf{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{u}_d$. It can also be expressed more intuitively as a weighted average of the direct and a regression-synthetic estimator,

$$\widehat{\theta}_d^{\text{FH}} = \widehat{\gamma}_d \widehat{\theta}_d + (1 - \widehat{\gamma}_d) \mathbf{x}_d^\top \widehat{\boldsymbol{\beta}}$$

The estimate $\widehat{\gamma}_d = \widehat{\sigma}^2_u / (\widehat{\sigma}^2_u + \sigma^2_{e_d})$, or the "shrinkage factor", weights the direct estimate and the regression-synthetic part. The weight on the direct estimate decreases with the sampling error variance.

For out-of-sample domains, $\widehat{\gamma}_d$ is not defined, and the regression-synthetic estimate $\mathbf{x}_d^\top \widehat{\boldsymbol{\beta}}$ is used. A domain is treated as out-of-sample if either the direct estimate or the sampling error variance is missing. Missing values in the domain-specific covariates (usually obtained from register or administrative data) are not allowed; that is, each explanatory variable needs to have a value for each domain.

2.2 Estimating the variance of the random error

The FH model requires an estimation of the variance of the random error, σ_u^2, and of the regression parameters, $\boldsymbol{\beta}$. Standard estimation techniques for σ_u^2 are, among others, REML and MLE. These methods do not guarantee positive variance estimates (Yoshimori and Lahiri 2014; Li and Lahiri 2010). Especially if there are few domains, the variance estimates can be negatively biased or even below zero. In the latter case, the variance estimate is set to zero. Underestimating the variance component could lead to a significant overshrinkage of the direct estimate to the regression-synthetic part; that is, too much weight is put on the regression-synthetic part.

Adjusted estimation methods, such as the adjusted maximum residual-likelihood approach (ARYL) following Yoshimori and Lahiri (2014) and the adjusted maximum-profile likelihood (AMPL) following Li and Lahiri (2010), ensure strictly positive variance estimates. `fayherriot` allows the estimation of σ_u^2 with REML (as the default), MLE, ARYL, and AMPL.[2] The method can be specified in the option `sigmamethod()`. The vector of regression parameters, $\boldsymbol{\beta}$, is estimated by the empirical best linear unbiased estimator $\widehat{\boldsymbol{\beta}}$ (Rao and Molina 2015, 124).

2.3 Evaluating the precision

The precision of the EBLUP is evaluated by means of the MSE, defined as

$$\text{MSE}\left(\widehat{\theta}_d^{\text{FH}}\right) = E\left\{\left(\widehat{\theta}_d^{\text{FH}} - \theta_d\right)^2\right\}$$

Because the true value θ_d is unobserved, $\text{MSE}(\widehat{\theta}_d^{\text{FH}})$ must be estimated. For in-sample domains, MSE estimators have been proposed for estimates of σ_u^2 relying on REML (Prasad and Rao 1990, 167), MLE (Datta and Lahiri 2000, 619), ARYL (Yoshimori and Lahiri 2014), and AMPL (Li and Lahiri 2010, 886). For out-of-sample domains, MSE estimators have been proposed only for REML and MLE (Rao and Molina 2015, 139). `fayherriot` automatically selects the appropriate MSE estimator.

2.4 Dealing with model-assumption violations

The FH model assumes linearity and normality of its two error terms. If there is a violation of these assumptions, a log-transformation of the direct estimator might be an option (Slud and Maiti 2006). Choosing this option requires an appropriate transformation of the variance of the original direct estimator.[3] Neves, Silva, and Correa (2013) suggest the transformation,

2. See Yoshimori and Lahiri (2014) for a general discussion of the comparative advantages of each method.
3. It is not appropriate to take the logarithm of the variance, because the variance of a log-transformed variable is different from the log-transformed variance of the original variable.

$$\widehat{\theta}_d^* = \log\left(\widehat{\theta}_d\right)$$
$$\text{var}\left(\widehat{\theta}_d^*\right) = \left(\widehat{\theta}_d\right)^{-2} \text{var}\left(\widehat{\theta}_d\right)$$

with * indicating the transformed scale.

Equation (1) is estimated using $\widehat{\theta}_d^*$ as the direct estimate and $\text{var}(\widehat{\theta}_d^*)$ as the estimate for the sampling error variance. To bring the estimated EBLUP and MSE back from the transformed to the original scale, we advise a bias correction (Slud and Maiti 2006; Sugawasa and Kubokawa 2017). `fayherriot` includes two back-transformation methods: the "crude" method, shown in Neves, Silva, and Correa (2013) and Rao and Molina (2015), and (as the default) the bias correction proposed by Slud and Maiti (2006). For the point estimates, these methods are defined as

$$\widehat{\theta}_d^{\text{FH, crude}} = \exp\left\{\widehat{\theta}_d^{\text{FH}*} + 0.5\text{MSE}\left(\widehat{\theta}_d^{FH*}\right)\right\}$$
$$\widehat{\theta}_d^{\text{FH, Slud–Maiti}} = \exp\left\{\widehat{\theta}_d^{\text{FH}*} + 0.5\widehat{\sigma}_u^2\left(1 - \widehat{\gamma}_d\right)\right\}$$

with * indicating the transformed scale.

The Slud–Maiti back-transformation relies on MLE for the estimation of σ_u^2. Because it requires an estimate of $\widehat{\gamma}_d$, it is only applicable for in-sample domains. The crude back-transformation can be used for in- and out-of-sample predictions.

For estimating the precision of the back-transformed EBLUPs, Slud and Maiti (2006, 248) developed an MSE estimator when using the log-transformation. The crude method uses the estimates in the transformed scale and the following back-transformation:

$$\text{MSE}\left(\widehat{\theta}_d^{\text{FH, crude}}\right) = \exp\left(\widehat{\theta}_d^{\text{FH}*}\right)^2 \text{MSE}\left(\widehat{\theta}_d^{\text{FH}*}\right)$$

2.5 Overview of functionalities

Figure 1 gives an overview of the functionalities of the `fayherriot` command. Additionally, the arcsine square root transformation can be used for proportions, and `fayherriot` returns back-transformed EBLUPs and corresponding boundaries of bootstrap confidence intervals. For a detailed description, we refer to Casas-Cordero Valencia, Encina, and Lahiri (2016, 394–397) and Schmid et al. (2017, 1173–1177).

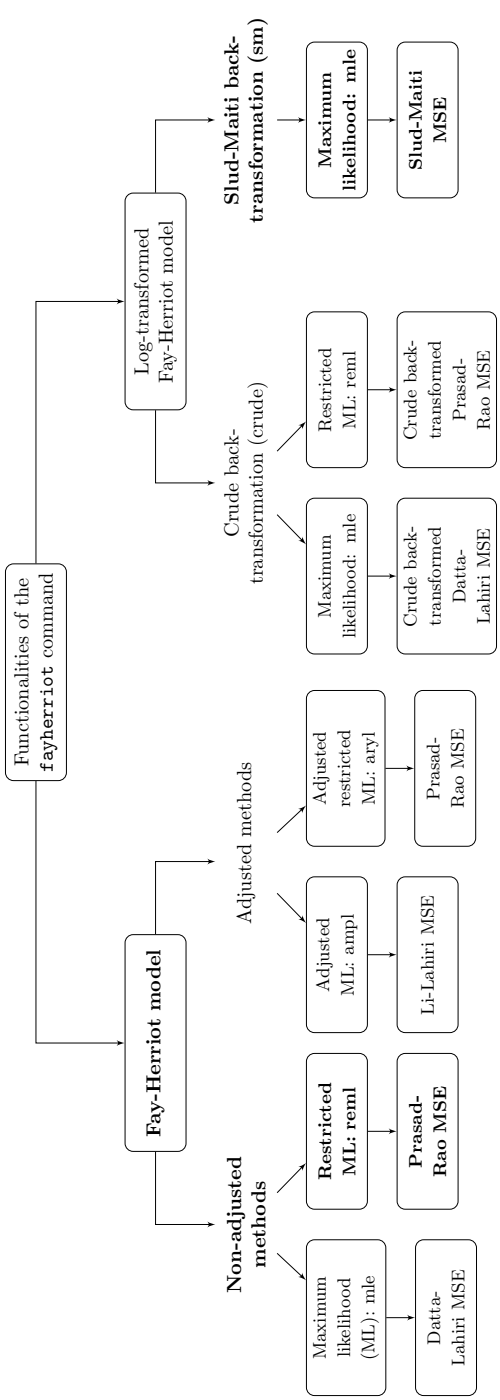

Figure 1. Functionalities of the `fayherriot` command. The two lowest levels describe the estimation methods of σ_u^2 and the corresponding MSE estimators, respectively. The default options are written in bold.

3 The fayherriot command

3.1 Syntax

`fayherriot` runs in Stata 12 and later versions. The syntax is

`fayherriot` *depvar* [*varlist*] [*if*] [*in*], `variance(`*varname*`)`
[`sigmamethod(`*method*`)` `logarithm` `arcsin` `biascorrection(`*method*`)`
`initialvalue(#)` `reps(#)` `level(#)` `eblup(`*name*`)` `mse(`*name*`)` `gamma`
`nolog`]

The command runs on datasets on the domain level with one observation per domain. *depvar* is the direct estimate, $\widehat{\theta}_d$ (in the documentation, theta), and *varlist* corresponds to the auxiliary explanatory variables, $\mathbf{x}_d$ (in the documentation, X).

3.2 Options for fayherriot

`variance(`*varname*`)` determines the variable containing the sampling error variances, $\sigma^2_{e_d}$. This variance is assumed to be known in the model. However, it often needs to be estimated from the data. One possibility is to use the estimated variance of the direct estimator specified in *depvar* for each domain. Whenever the direct estimator needs to be logarithmized with logdepvar = log(*depvar*), the estimated variance can be modified as logvar = var/(*depvar*^{2}) (Neves, Silva, and Correa 2013). In case the estimate is transformed by the arcsine transformation theta_arcsine = asin($\sqrt{\widehat{\theta}_d}$), the estimated variance can be approximated by sigma2_e_arcsine = $1/(4 \times$ effsample) with effsample being the effective sample size (Jiang et al. 2001). The effective sample size is an estimate of the sample size that a survey based on simple random sampling would have to have the same sampling error as the currently used survey with the corresponding sampling design. It can be estimated by the division of the sample size and the design effect (Lohr 2010, 239). `variance()` is required.

`sigmamethod(`*method*`)` specifies the method for the estimation of the variance of the random effect σ^2_u: `reml`, `mle`, `ampl`, or `aryl`. The default is `sigmamethod(reml)`. If a zero estimate is received for the variance—which is more likely when the number of domains is small—the adjusted maximum-likelihood methods `ampl` (Li and Lahiri 2010) and `aryl` (Yoshimori and Lahiri 2014) may help to estimate strictly positive variances.

`logarithm` indicates that the dependent variable in *depvar* is the log-transformed direct estimate. A log-transformed FH model is suitable when the linearity or normality assumption of the error terms is not fulfilled. `logarithm` automatically back-transforms EBLUP and MSE to the original scale.

`arcsin` indicates that the dependent variable in *depvar* is the direct estimate transformed by the arcsine square root transformation. This transformation is especially

suitable when the indicator of interest is a proportion confined to the $[0, 1]$ interval. `arcsin` automatically back-transforms EBLUP and the boundaries of the bootstrap confidence interval to the original scale.

`biascorrection(`*method*`)` determines the method for the back-transformation of EBLUP and MSE in a log-transformed FH model. The EBLUPs and MSEs in the transformed scale can be back-transformed using the bias correction proposed by Slud and Maiti (2006), which is set as a default, and a crude bias correction (Neves, Silva, and Correa 2013; Rao and Molina 2015). When the arcsine transformation is used, the EBLUP and the boundaries of the confidence interval are, by default, back-transformed by the inverse transformation as proposed in Casas-Cordero Valencia, Encina, and Lahiri (2016), and thus no method needs to be specified.

`initialvalue(#)` sets the initial value of the optimization algorithm for estimating the variance of the random effect σ_u^2 to #. The default is `initialvalue(0.0)`.

`reps(#)` sets the number of bootstrap repetitions for the confidence intervals to #. The default is `reps(100)`. The confidence intervals are returned if `arcsin` is specified.

`level(#)` sets the confidence level of the bootstrap confidence intervals to #. The default is `level(95)`, which corresponds to a 95% confidence level.

`eblup(`*name*`)` stores the EBLUP estimates in the variable *name*. For in-sample domains, the EBLUPs are defined as $\texttt{eblup()} = \mathbf{x}_d^\top \widehat{\boldsymbol{\beta}} + \widehat{u}_d$, where $\mathbf{x}_d^\top \widehat{\boldsymbol{\beta}}$ are the estimated fixed effects and $\widehat{u}_d$ is the estimated random effect. The EBLUP can also be expressed as the weighted average of the direct estimator and a synthetic part, $\texttt{eblup()} = \widehat{\gamma}_d \times \widehat{\theta}_d + (1 - \widehat{\gamma}_d)\mathbf{x}_d^\top \widehat{\boldsymbol{\beta}}$. For out-of-sample domains, the EBLUP shrinks to the synthetic part, $\texttt{eblup()} = \mathbf{x}_d^\top \widehat{\boldsymbol{\beta}}$.

`mse(`*name*`)` stores the MSE estimates in the variable *name*. The MSE depends on the estimation procedure of sigma2_u. For `sigmamethod(reml)`, the MSE estimator relies on Prasad and Rao (1990, 167); for `sigmamethod(mle)`, the MSE estimator relies on Datta and Lahiri (2000, 619); for `sigmamethod(ampl)`, the MSE estimator relies on Li and Lahiri (2010, 886); and for `sigmamethod(aryl)`, the MSE estimator relies on Yoshimori and Lahiri (2014). For the log-transformed FH model under the Slud–Maiti bias correction, the MSE is defined as in Slud and Maiti (2006, 248). It is only applicable to in-sample domains. Under the crude bias correction, for in- and out-of-sample domains, `mse(`eblup_backtransformed`)` $= \exp(\text{EBLUP})^2 \times$ `mse(`EBLUP`)` (Neves, Silva, and Correa 2013). In case `arcsin` is chosen, upper and lower bounds of bootstrap confidence intervals (Casas-Cordero Valencia, Encina, and Lahiri 2016, 394–397; Schmid et al. 2017, 1173–1177) are returned.

`gamma` reports summary statistics of the shrinkage factor, $\widehat{\gamma}_d = \widehat{\sigma}_u^2 / (\widehat{\sigma}_u^2 + \sigma_{e_d}^2)$, where $\widehat{\sigma}_u^2$ is the estimated variance of the random effect and $\sigma_{e_d}^2$ is the sampling error variance of each domain provided in `variance(`*varname*`)`.

`nolog` suppresses the display of the iteration log of the optimization algorithm.

3.3 predict after fayherriot

Syntax

The syntax for `predict` following `fayherriot` is

`predict` [*type*] *newvar* [*if*] [*in*] [`, eblup mse reps(#) level(#) ehat estandard uhat gamma cvdirect cvfh`]

Options

`eblup` generates the EBLUPs as defined above; this is the default.

`mse` generates estimates for the MSE or the boundaries of the confidence interval as defined above.

`reps(#)` sets the number of bootstrap repetitions for the confidence intervals to #. The default is `reps(100)`.

`level(#)` sets the confidence level of the bootstrap confidence intervals to #. The default is `level(95)`, which corresponds to a 95% confidence level.

`ehat` calculates the residuals. The residuals are defined as $\widehat{e}_d = (1 - \widehat{\gamma}_d) \times (\widehat{\theta}_d - \mathbf{x}_d^\top \widehat{\boldsymbol{\beta}})$, where $\widehat{\theta}_d$ corresponds to *depvar*.

`estandard` calculates the standardized residuals defined as $\widehat{e}_d / \sqrt{\sigma^2_{e_d}}$.

`uhat` calculates the random effects. The random effects are defined as $\widehat{u}_d = \widehat{\gamma}_d \times (\widehat{\theta}_d - \mathbf{x}_d^\top \widehat{\boldsymbol{\beta}})$.

`gamma` generates the shrinkage factor as defined above.

`cvdirect` calculates the coefficient of variation (CV) of direct estimates. `cvdirect` $= 100 \times \sqrt{\sigma^2_{e_d}} / \widehat{\theta}_d$, where $\widehat{\theta}_d$ corresponds to *depvar* and $\sigma^2_{e_d}$ is the sampling error variance provided in `variance()`. In case `logarithm` is specified, `cvdirect` $= 100 \times \sqrt{\sigma^{2'}_{e_d}} / \theta'$ with $\widehat{\theta}' = \exp(\widehat{\theta}_{\log})$, and $\sigma^{2'}_{e_d} = \mathrm{var}(\widehat{\theta}_{\log}) \times (\widehat{\theta}')^2$. In case `arcsin` is chosen, the CV for the direct estimate cannot be returned because the direct variance in the original scale is unknown within the `fayherriot` command.

`cvfh` calculates the CV based on EBLUPs: `cvfh` $= 100 \times \sqrt{\texttt{mse}} / \texttt{eblup}$. In case `arcsin` is chosen, the CV for the EBLUP cannot be returned because no MSE estimation is provided.

3.4 Stored results

Scalars

`e(N_in)`	number of observations used for estimation of `e(b)` and `e(sigma2_u)`
`e(N_out)`	number of out-of-sample observations for which EBLUP is calculated
`e(sigma2_u)`	estimated sigma2_u
`e(r2_a)`	adjusted R^2 of unweighted ordinary least squares
`e(r2_fh)`	R^2 according to Lahiri and Suntornchost (2015)
`e(p_e)`	p-value of Shapiro–Wilk test for normality of residuals
`e(V_e)`	test statistic of Shapiro–Wilk test of normality of residuals
`e(p_u)`	p-value of Shapiro–Wilk test for normality of the random effect
`e(V_u)`	test statistic of Shapiro–Wilk test of normality of the random effect

Macros

`e(cmd)`	`fayherriot`
`e(title)`	`Fay Herriot estimation`
`e(depvar)`	name of dependent variable
`e(variance)`	name of variance variable
`e(sigma_method)`	`sigmamethod()` estimation method
`e(bias_correction)`	bias-correction method for the back-transformation of transformed EBLUPs
`e(logarithm)`	logarithm true or false
`e(arcsin)`	arcsine true or false
`e(properties)`	`b V`
`e(predict)`	program to implement `predict`
`e(marginsok)`	predictions allowed by `margins`
`e(marginsnotok)`	predictions disallowed by `margins`

Matrices

`e(b)`	coefficient vector
`e(V)`	variance–covariance matrix of coefficients
`e(gamma)`	summary of values of shrinkage factor gamma

Functions

`e(sample)`	marks estimation sample

4 Example

We use the FH model to estimate households' material well being in 2015 in Germany: at the level of federal states (16 divisions), planning regions (96 divisions), and districts (402 divisions). Material well being is defined as region-specific average equivalent income, that is, household disposable income divided by the Organisation for Economic Co-operation and Development modified scale (Hagenaars, de Vos, and Zaidi 1994).

Following the policies used by several statistical agencies to evaluate the precision of the regional estimates, we rely on the CV, which is the standard error of the estimate divided by the estimate (in percent). For instance, Statistics Canada releases data without warning about low precision if the CV is below 16.5% (Statistics Canada 2013; Eurostat 2013).

4.1 Data description and direct estimates

We derive the direct estimates from the German Socio-Economic Panel (SOEP), which is a household survey covering about 15,000 households per year (Goebel et al. 2019).

Table 1 provides the division-specific numbers of SOEP households. Sample sizes by federal states are large (median: 624), ranging from 114 to 3,159 observations. Sample sizes by planning regions are considerably smaller (median: 132), ranging from 32 to 665 observations. Sample sizes by districts range from 10 to 648 observations (median: 32).[4] Because of small sample sizes, we expect that many direct estimates for planning regions and districts are measured with high imprecision.

Table 1. Number of regions and sample sizes

		Sample-size distribution				
Regional division	Number of regions	Minimum	p10	p25	p50	Maximum
Federal states	16	114	144	444	624	3159
Planning regions	96	32	61	88	132	665
Districts	357	10	14	20	32	648

NOTE: Data are from SOEP v33.1. Computations are our own.

For each regional level, table 2 provides direct estimates of mean equivalent income and coefficients of variation, our precision indicator.[5] The table suggests considerable regional heterogeneities in material well being. Across federal states, mean equivalent income ranges from €1,362 to €1,863; across planning regions, from €1,298 to €2,101; and across districts, from €1,023 to €2,976. As expected, coefficients of variation increase as we move to smaller regional levels. In line with the policy of Statistics Canada, not all estimates could be reported for the planning regions and the districts without warning of low precision. In the following, we show how this can be achieved using the FH model. In particular, we can a) improve the precision of all estimates and b) receive estimates for the districts without a direct estimator.

4. For confidentiality issues, we discarded areas with fewer than 10 observations. This left us with 357 out of 402 districts.
5. We estimated standard errors using the random group estimator to account for the survey sampling design (Rendtel 1995).

Table 2. Summary of mean equivalent household income and coefficients of variation by regional level

Regional division	Minimum	p10	p25	p50	p75	p90	Maximum
(A) Mean equivalized household income							
Federal states	1362	1398	1492	1683	1777	1841	1863
Planning regions	1298	1400	1495	1664	1780	1898	2101
Districts	1023	1311	1463	1641	1847	2049	2976
(B) Coefficient of variation							
Federal states	0.6	0.8	1.4	2.2	3.8	6.4	8.0
Planning regions	1.5	3.4	4.1	5.3	7.2	9.0	18.2
Districts	2.2	5.8	7.6	10.2	13.6	16.7	42.5

NOTE: Data are from SOEP v33.1. Computations are our own.

4.2 Estimation using fayherriot

For fitting the FH model, we rely on the direct estimates of average equivalent incomes (table 2); their sampling error variances, $\sigma^2_{e_d}$; and region-specific explanatory variables. The set of explanatory variables in this example includes the unemployment rate, the share of the population older than 65 years, and per-capita income tax revenue.[6]

FH model for the planning regions

In the following, we detail the application of `fayherriot` at the level of planning regions. In this example, all regions are sampled and the model assumptions are fulfilled. The underlying dataset includes 96 observations (one observation per region):

```
. use dataror.dta
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
income	96	1658.387	188.7142	1297.915	2100.683
directvari~e	96	11448.52	12856.35	612.4922	96107
unemployment	96	6.259375	2.579212	2.1	12.8
incometax	96	399.2719	105.6913	211.6	705
share65	96	56.48438	.8259385	54.9	58.2
N	96	162.3854	125.7412	32	665

6. The explanatory variables are obtained from INKAR (Bundesinstitut für Bau-, Stadt-, und Raumforschung 2017), a database of regional indicators derived from high-quality and large-scale national census and register data.

To fit the FH model, we type

```
. fayherriot income unemployment incometax share65,
> variance(directvariance) gamma nolog
Sigma2_u estimation method:     reml            N in sample      =         96
Transformation of depvar:       none            N out of sample  =          0
EBLUP and MSE bias correction:  none            Sigma2_u         =  4683.7208
                                                Adj R-squared    =     0.5769
                                                FH R-squared     =     0.7808
                    Gamma
  Min       5%      Median      95%       Max
0.0465    0.1464    0.3726    0.7307    0.8844

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
unemployment |    5.956309   6.664692     0.89   0.371    -7.106248    19.01887
   incometax |    1.278903   .1365014     9.37   0.000     1.011365    1.546441
     share65 |   -38.88107   18.04845    -2.15   0.031    -74.25537   -3.506762
       _cons |    3301.427   1013.564     3.26   0.001     1314.877    5287.976
------------------------------------------------------------------------------

Shapiro-Wilk test for normality:
 Residuals e (standardized)  V =    0.837  p-value = 0.653
 Random effects u            V =    0.392  p-value = 0.981
```

The syntax of the command is inline with the familiar Stata regression syntax: `income` contains the direct estimates of mean equivalent income and is regressed on the regional explanatory variables `unemployment`, `incometax`, and `share65`. `variance()` specifies the variable containing the sampling error variances, `directvariance`. We specify the `gamma` option to display summary statistics of shrinkage factors $\widehat{\gamma}_d$. `nolog` suppresses the iteration log of the optimization algorithm.

`N in sample` indicates that the full set of 96 planning regions was used in the estimation. `FH R-squared` is an indicator for the goodness of fit of the FH model, proposed by Lahiri and Suntornchost (2015, 317, AdjR^2_h). Similarly to the standard R^2, it expresses the explained variation of `income` in relation to the total variation, while taking into account that some variation in `income` is due to the sampling error. In this example, about 78% of the variation is explained.

The variance of the random effects, $\widehat{\sigma}^2_u = 4683.72$, is estimated using the REML approach (the default). Together with the sampling error variances $\sigma^2_{e_d}$, it determines the shrinkage factor $\widehat{\gamma}_d$. The shrinkage factor shows how direct estimates and model predictions are weighted when calculating the EBLUP. Large values of $\widehat{\gamma}_d$ mean that a large weight is given to the direct estimate $\widehat{\theta}_d$. In our example, the distribution of $\widehat{\gamma}_d$ ranges from 0.0465 to 0.8844 with its median being 0.3726. So for some regions, the EBLUP relies strongly on the model predictions (small value of $\widehat{\gamma}_d$) and strongly on the direct estimator for others (large value of $\widehat{\gamma}_d$). The Shapiro–Wilk test for normality shows that neither normality of the realized residuals, $\widehat{e}_d$, nor of the random effects, $\widehat{u}_d$, is rejected. Hence, the model assumptions are not violated.

Log-transformed FH model for the districts

In the district-level analysis, not all regions are sampled, and the normality assumption of the model is violated. Hence, we log-transform equivalent incomes and the variances of the sampling error,

```
. use datadistricts
. generate logincome = log(income)
(45 missing values generated)
. generate directlogvariance = directvariance/income^2
(45 missing values generated)
```

and fit the log-transformed FH model:

```
. fayherriot logincome unemployment incometax share65,
> variance(directlogvariance) nolog logarithm
Sigma2_u estimation method:     mle             N in sample      =        357
Transformation of depvar:       logarithm       N out of sample  =         45
EBLUP and MSE bias correction:  sm              Sigma2_u         =     0.0089
                                                Adj R-squared    =     0.2891
                                                FH R-squared     =     0.4745
------------------------------------------------------------------------------
   logincome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
unemployment |  -.0004102    .003304    -0.12   0.901    -.0068858    .0060655
   incometax |   .0007471   .0000904     8.26   0.000     .0005698    .0009243
     share65 |  -.0063528    .003548    -1.79   0.073    -.0133067    .0006011
       _cons |   7.241288   .1051244    68.88   0.000     7.035248    7.447328
------------------------------------------------------------------------------
Shapiro-Wilk test for normality:
 Residuals e (standardized)  V =    1.614  p-value = 0.128
 Random effects u            V =    0.830  p-value = 0.670
------------------------------------------------------------------------------
```

By specifying the `logarithm` option, `fayherriot` transforms the estimated EBLUP and MSE back to the original scale. Because we did not specify the bias-correction method, the estimation method is MLE and the bias correction follows Slud and Maiti (2006) (see figure 1). In this default setting, only estimates for the 357 in-sample districts are calculated. `biascorrection(crude)` could be specified to obtain in- and out-of-sample estimates.

4.3 Comparison of direct and FH estimates

Next we compare the direct with the FH point estimates (EBLUP) and assess their precision. There are two equivalent ways to obtain the EBLUPs and their level of precision (MSE). The first is specifying the `eblup(`*varname*`)` and `mse(`*varname*`)` options (here done for the planning regions):

```
. fayherriot income unemployment incometax share65,
> variance(directvariance) nolog eblup(eblupROR) mse(mseROR)
```

The second is using the postestimation `predict` routine directly after the `fayherriot` command:

```
. predict eblupROR, eblup
. predict mseROR, mse
```

An additional feature of `predict` is that it provides the CV for the direct and FH estimates.

```
. predict cvROR_FH, cvfh
. predict cvROR_direct, cvdirect
```

To assess the magnitude of adjustments, figure 2 presents the ratios of EBLUPs and direct estimates against region-specific sample sizes.[7] For federal states, the ratios are all close to 1, suggesting small adjustments of the direct estimator. For planning regions and districts, adjustments are larger, which is an expected result given smaller sample sizes of these domains.

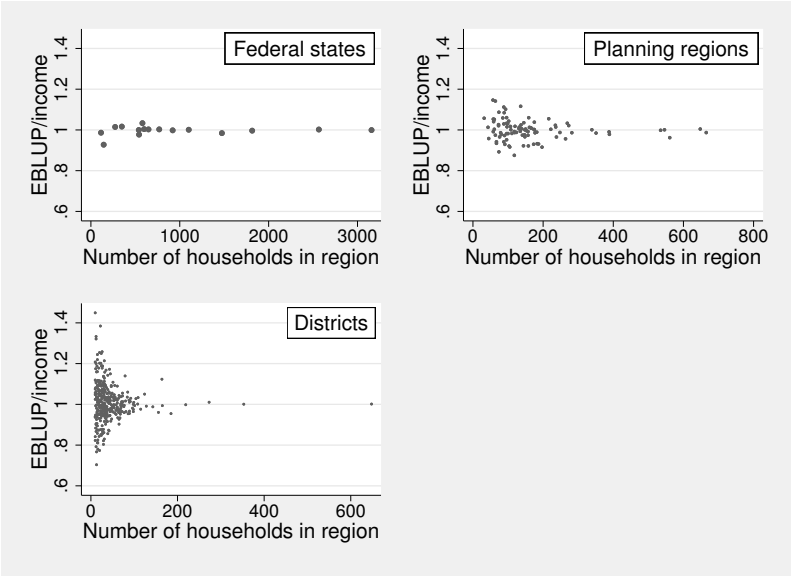

Figure 2. Ratio of the EBLUP to the direct (income) estimates plotted against regional sample sizes for all three regional divisions—federal states, planning regions, and districts. Only in-sample domains are plotted. Data are from SOEP v33.1. Computations are our own.

To assess the gain in precision, figure 3 provides box plots of coefficients of variation for the direct and FH estimates. The horizontal line indicates the threshold of 16.5 suggested by Statistics Canada. For the direct estimates, several CVs at the district and

7. For further comparison methods, see Brown et al. (2001).

planning region levels exceed the threshold. For the FH estimates, in contrast, CVs for all regional levels are under the threshold.

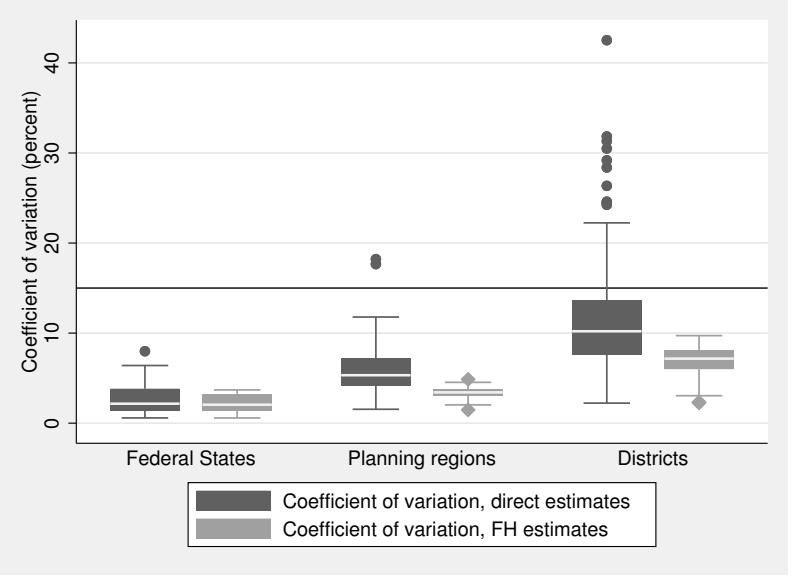

Figure 3. Box plots of the distribution of the coefficients of variation for the federal states, the planning regions, and the districts. The horizontal line indicates the precision threshold of 16.5%. Only in-sample domains are plotted. Data are from SOEP v33.1. Computations are our own.

5 Conclusion

We implemented the FH model in Stata. It is a small-area estimation technique and aims at improving the precision of direct estimators from a survey by using additional domain-level covariate information. We introduced the `fayherriot` command and provided an application to regional heterogeneities in material well being in Germany. The results showed that the precision of the FH model estimates is markedly higher than that of the direct estimates.

6 Acknowledgments

Halbmeier and Schröder thank Johannes König for his valuable remarks and comments, as well as Paul Brockmann, Deborah Anne Bowen, and Fabian Nemeczek for excellent research assistance. Kreutzmann and Schmid gratefully acknowledge support by the German Research Foundation within the project QUESSAMI (281573942) and the MIUR-DAAD Joint Mobility Program (57265468). We thank the editor and the referee for their constructive comments that helped to improve the article.

7 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 19-3
. net install st0570     (to install program files, if available)
. net get st0570         (to install ancillary files, if available)
```

8 References

Battese, G. E., R. M. Harter, and W. A. Fuller. 1988. An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* 83: 28–36.

Brown, G., R. Chambers, P. Heady, and D. Heasman. 2001. Evaluation of small area estimation methods—An application to unemployment estimates from the UK LFS. In *Symposium 2001—Achieving data quality in a statistical agency: A methodological perspective*. Ottawa: Statistics Canada.

Bundesinstitut für Bau-, Stadt-, und Raumforschung. 2017. Indikatoren und Karten zur Raum- und Stadtentwicklung. Datenlizenz Deutschland - Namensnennung - Version 2.0. http://www.inkar.de/.

Casas-Cordero Valencia, C., J. Encina, and P. Lahiri. 2016. Poverty mapping for the Chilean comunas. In *Analysis of Poverty Data by Small Area Estimation*, ed. M. Pratesi, 379–403. Hoboken, NJ: Wiley.

Corral, P., W. Seitz, J. P. Azevedo, and M. C. Nguyen. 2018. fhsae: Stata module to fit an area level Fay–Herriot model. Statistical Software Components, S458495, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458495.html.

Datta, G. S., and P. Lahiri. 2000. A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* 10: 613–627.

Eurostat. 2013. Handbook on precision requirements and variance estimation for ESS households survey. Methodologies and working papers, European Union.

Fay, R. E., III, and R. A. Herriot. 1979. Estimates of income for small places: An application of James–Stein procedures to census data. *Journal of the American Statistical Association* 74: 269–277.

Goebel, J., M. M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp. 2019. The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics* 239: 345–360.

Hagenaars, A. J. M., K. de Vos, and M. A. Zaidi. 1994. *Poverty Statistics in the Late 1980s: Research Based on Micro-data*. Luxembourg: Office for the Official Publications of the European Communities.

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47: 663–685.

Huang, E. T., and W. R. Bell. 2012. An empirical study on using previous American Community Survey data versus Census 2000 data in SAIPE models for poverty estimates. Research Report Series Statistics #2012-04, U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2012/adrm/rrs2012-04.pdf.

Jiang, J., P. Lahiri, S.-M. Wan, and C.-H. Wu. 2001. Jackknifing in the Fay–Herriot model with an example. In *Proc. Sem. Funding Opportunity in Survey Research*, 75–97. Washington, DC: Bureau of Labor Statistics.

Lahiri, P., and J. Suntornchost. 2015. Variable selection for linear mixed models with applications in small area estimation. *Indian Journal of Statistics* 77: 312–320.

Leadership Council of the Sustainable Development Solutions Network. 2015. Indicators and a monitoring framework for the Sustainable Development Goals. Report to the Secretary General of the UN, United Nations.

Li, H., and P. Lahiri. 2010. An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis* 101: 882–892.

Lohr, S. L. 2010. *Sampling: Design and Analysis*. 2nd ed. Boston: Cengage Learning.

Molina, I., and J. N. K. Rao. 2010. Small area estimation of poverty indicators. *Canadian Journal of Statistics* 38: 369–385.

Neves, A., D. Silva, and S. Correa. 2013. Small domain estimation for the Brazilian service sector survey. *Estadística* 65: 13–37.

Piacentini, M. 2014. Measuring income inequality and poverty at the regional level in OECD countries. OECD Statistics Working Papers 2014/03, OECD Publishing. https://doi.org/10.1787/5jxzf5khtg9t-en.

Powers, D., W. Basel, and B. O'Hara. 2008. SAIPE county poverty models using data from the American Community Survey. https://www.census.gov/content/dam/Census/library/working-papers/2008/demo/powersbaselohara2008asa.pdf.

Prasad, N. G. N., and J. N. K. Rao. 1990. The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* 85: 163–171.

Rabe-Hesketh, S., and A. Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*. 3rd ed. College Station, TX: Stata Press.

Rao, J. N. K., and I. Molina. 2015. *Small Area Estimation*. 2nd ed. Hoboken, NJ: Wiley.

Rendtel, U. 1995. *Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität*. Frankfurt: Campus.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski. 2017. Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in Senegal. *Journal of the Royal Statistical Society, Series A* 180: 1163–1190.

Slud, E. V., and T. Maiti. 2006. Mean-squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society, Series B* 68: 239–257.

Statistics Canada. 2013. 2013 National Graduate Survey (Class of 2009–2010). Microdata user guide, Statistics Canada.

Sugawasa, S., and T. Kubokawa. 2017. Transforming response values in small area prediction. *Computational Statistics & Data Analysis* 114: 47–60.

Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla. 2018. From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society, Series A* 181: 927–979.

Yoshimori, M., and P. Lahiri. 2014. A new adjusted maximum likelihood method for the Fay–Herriot small area model. *Journal of Multivariate Analysis* 124: 281–294.

You, Y., and B. Chapman. 2006. Small area estimation using area level models and estimated sampling variances. *Survey Methodology* 32: 97–103.

About the authors

Christoph Halbmeier is a research assistant at the SOEP at the German Institute for Economic Research and a PhD candidate of Prof. Carsten Schröder, Chair of Public Finance and Social Policy at the Freie Universität Berlin.

Ann-Kristin Kreutzmann is a research assistant at the Institute for Statistics and Econometrics at the Freie Universität Berlin and a PhD candidate of Prof. Timo Schmid, Professor of Applied Statistics at the Freie Universität Berlin.