Small Area Estimation with R

Unit 3: Model-based estimators

V. Gómez-Rubio

Department of Mathematics
Universidad de Castilla-La Mancha, Spain

useR! 2016
27th June 2016, Stanford University

Model-based estimation

- Direst estimators cannot cope efficiently with estimates for areas that have not been included in the sample
- Model-based estimation relies on a parametric model
- The domains are assumed to be part of a *superpopulation* whose characteristics are estimated by the models

Types of models

Unit level models

- Use the survey data directly
- Area level covariates will be needed to provide small area estimates
- Sometimes access to individual data is difficult because of problems of confidentiality

Area level models

- Based on direct estimates and area level covariates
- No confidentiality issues involved
- Given that these are based on ecological data, the coefficients of the covariates in the model must be interpreted with care

Area level models

Fay-Herriott estimator

The Fay Herriott estimator combines direct estimations with linear regression:

$$\hat{\overline{Y}}_i = \mu_i + \epsilon_i;\ \epsilon_i \sim N(0, \hat{\sigma}_i^2)$$

$$\mu_i = \alpha + \beta \overline{X}_i$$

- $\hat{\overline{Y}}_i$ is a direct estimator
- $\hat{\sigma}_i^2$ is a design variance
- $\overline{X}_i$ is a vector of area level covariates (i.e., area level means in this case)
- α and β can be estimated by means of Generalised Least Squares

Standard linear regression

An alternative is to fit a standard linear regression model:

$$\hat{\overline{Y}}_i = \mu_i + \epsilon_i;\ \epsilon_i \sim N(0, \sigma^2)$$

$$\mu_i = \alpha + \beta \overline{X}_i$$

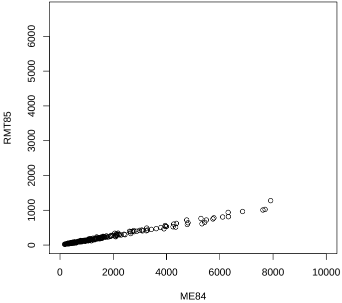
- This is useful when the design variances are not known
- No finite-population correction is implicitely done

Linear Regression

There are two approaches when using linear regression for SAE

- `lm` assumes that the sample comes from an *infinite* population
- `svyglm` accounts for the survey design and provides a correction for *finite population* in the estimation of the standard errors

We are trying to model the total tax revenues according to the number of municipal employees




```
> survlm <- lm(RMT85 ~ ME84, dsmp)
> survglm <- svyglm(RMT85 ~ ME84, svy)
> plot(MU284$ME84, MU284$RMT85, xlab = "ME84",
+   ylab = "RMT85", xlim = c(0, 10000))
```

Unit level models

Unit level models

$$y_{ij} = \mu_{ij} + \varepsilon_{ij};\ \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\mu_{ij} = \alpha + \beta x_{ij}$$

- y_{ij} unit level target variable
- x_{ij} unit level covariates

Small Area Estimation

- Additional area level covariates are required to provide small area estimates

$$\hat{\overline{Y}}_i = \hat{\alpha} + \hat{\beta}\overline{X}_i$$

- We ignore that we know a small proportion of the units in the domain

Unit level models

Unit level models with area level variances

$$y_{ij} = \mu_{ij} + \varepsilon_{ij};\ \varepsilon_{ij} \sim N(0, \sigma_i^2)$$

$$\mu_{ij} = \alpha + \beta x_{ij}$$

- σ_i^2 is the variance of the units in area i
- This model allows for internal variation to change between areas and it is likely that it will provide a better fit

Package **nlme**

- Function `gls` can be used to fit these models
- The variance structure of the data must be defined using `varFunc`

Variance and Correlation structures

Variance-Covariance structure

The family of functions `varFunc()` provide different ways of defining the covariance between the small areas

- These are passed as argument `weights=`
- `varIdent()`: Allows different variances per group
- `varFixed()`: Allows fixed variances depending on a covariate

Correlation structure

The family of functions `corClasses()` can be used to define correlation structures

- These are passed as argument `correlation=`
- `corGaus()`: Gaussian Correlation Structure
- `corExp()`: Exponential Correlation Structure

Example: MU284 data set

First of all, we fit a model with $\sigma_i^2 = \sigma^2$. This is equivalent to use `lm()`

```
> library(nlme)
> #Region level covariates
> REGCOV <- data.frame(ME84 = as.vector(by(MU284$ME84, MU284$REG,
+   mean)))
> #One variance
> gls1 <- gls(RMT85 ~ ME84, data = dsmp)
> synth1 <- predict(gls1, REGCOV, interval = "confidence")
```

Then, we fit model that considers a different variance per region. Note that we have 8 regions and we estimate the variances as compared to that of region 1.

```
> #Region-level variances
> l <- as.list(rep(1,7))
> names(l) <- as.character(2:8)
> vf1 <- varIdent(l, form = ~ 1 | REG)
> gls2 <- gls(RMT85 ~ ME84, data = dsmp, weights = vf1)
> synth2 <- predict(gls2, REGCOV, interval = "confidence")
```

Data 'missing' by design

- Usually, some areas are not covered in the survey design
- This means that some data will be 'missing'
- The missingness mechanism can be ignored because the data are 'missing' by design
- Synthetic estimation is used to provide estimates in these areas
- Information between areas is borrowed by means of the coefficients of the covariates
- The model that we will fit will only consider $\sigma_i^2 = \sigma^2$
- Prediction is done by means of *synthetic estimators*

Example: Synthetic estimation

First of all, we produce some estimates using the model with a common variance

```
> #One variance
> glsmiss1 <- gls(RMT85 ~ ME84, data = dsmp)
> synthmiss1 <- predict(glsmiss1, REGCOV, interval = "confidence")
```

The following example fits a model based on a two-stage sampling and allowing for region-level variances

```
> #Region-level variances
> regs <- unique(dsmpcl2$REG)
> l <- as.list(rep(1,length(regs) - 1))
> names(l) <- as.character(regs[-1])
> vfmiss1 <- varIdent(l, form = ~ 1 | REG)
> glsmiss2 <- gls(RMT85 ~ ME84, data = dsmpcl2, weights = vfmiss1)
> synthmiss2 <- predict(glsmiss2, REGCOV, interval = "confidence")
```

Composite estimator

The composite estimator aims at combining the good properties of direct and model based estimators:

- Direct estimators ($\hat{\bar{Y}}_{D,i}$) are design-unbiased
- Model-based estimators ($\hat{\bar{Y}}_{M,i}$) have a lower variance, because they combine information from different areas

$$\hat{\bar{Y}}_{C,i} = \gamma_i \hat{\bar{Y}}_{D,i} + (1 - \gamma_i)\hat{\bar{Y}}_{M,i}$$

- $0 \leq \gamma_i \leq 1$ is a shrinkage parameter to weight both estimators
- $\gamma_i = 0$ if $n_i = 0$ and the composite estimator reduces to the synthetic estimator
- Otherwise, γ_i can be estimated in different ways

Estimation of the shrinkage parameter

Different shrinkage parameters

γ_i is obtained by minimising $MSE(\hat{Y}_{C,i})$ when $Cov(\hat{Y}_{D,i}, \hat{Y}_{M,i}) \approx 0$

$$\hat{\gamma}_i = 1 - \frac{Var[\hat{Y}_{D,i}]}{(\hat{Y}_{M,i} - \hat{Y}_{D,i})^2}$$

Common shrinkage parameter

$\gamma_i = \gamma$ is obtained by minimising $\sum_i MSE(\hat{Y}_{C,i})$ when $Cov(\hat{Y}_{D,i}, \hat{Y}_{M,i}) \approx 0$

$$\hat{\gamma}_i = \hat{\gamma} = 1 - \frac{\sum_i Var[\hat{Y}_{D,i}]}{\sum_i (\hat{Y}_{M,i} - \hat{Y}_{D,i})^2}$$

Example: Composite estimator of regional values

Computation of individual weights

```
> gammaw1 <- 1 - (destdom$se^2)/((synth1 - destdom$RMT85)^2)
> gammaw1
```

```
[1]  -0.1031647   0.3586167   0.6764069   0.6744215   0.9805117   0.9857506
[7]  -1.1085866 -11.4097047
attr(,"label")
[1] "Predicted values"
```

```
> gammaw1[gammaw1 < 0] <- 0
> gammaw1[gammaw1 > 1] <- 1
> comp1 <- gammaw1 * destdom[, 2] + (1 - gammaw1) * synth1
> comp1
```

```
[1] 531.43796 287.96903 263.83818 193.89251 110.11218  80.67756 206.22962
[8] 145.13694
attr(,"label")
[1] "Predicted values"
```

Computation of a common weight

```
> gammaw2 <- 1 - sum(destdom$se^2) / sum((synth2 - destdom$RMT85)^2)
> gammaw2
```

```
[1] 0.3572664
```

```
> comp2 <- gammaw2 * destdom[,2] + (1 - gammaw2) * synth2
> comp2
```

```
[1] 487.8543 290.7588 225.8291 227.6599 208.7756 127.0239 234.6232 157.6470
attr(,"label")
[1] "Predicted values"
```

Example: Comparison of different types of estimators

```
              AEMSE
DIRECT        198763.46148
SYNTH 1          119.49474
SYNTH 2           50.43277
COMP GAMMA_i    6132.50711
COMP GAMMA      2242.32048
```

Other models

- Regression models can be extended to account for different types of effects
- When the relationship between a covariate and the target variable is not linear, splines can be used (see package **mgcv**)
- Temporal models can fitted by modelling areas as longitudinal data
- Spatial effects are more difficult to model, and they are usually considered as random effects