نموذج مستوى الوحدة

المعهد العربي للتدريب والبحوث الاحصائية ـعمان

دورة حول

تقديرات المجالات او المناطق الصغيرة

Model-Based Small Area Estimation

14-6 to 8-7, 2021

نموذج مستوى الوحدة

Area level model

Fay and Herriot, 1979

اليوم الثالث ـ الخميس 2021-06-17

الدكتور عبدالحكيم عبدالحي عيده

استاذ مشارك في احصاءات ومنهجيات المسوح

جامعة القدس ـابو ديس ـفلسطين

a.eideh.3s@gmail.com

00970599662851

نموذج مستوى الوحدة

تقديرات المناطق الصغيرة المعتمدة على النموذج

نموذج مستوى الوحدة

Area level model

Fay and Herriot, 1979

النموذج الإحصائي: مجموعة من الافتراضات التي تحدد السلوك العشوائي لخصائص المجتمع قيد الاهتمام.

النموذج الاحصائي يسمح بمعالجة البيانات الثنائية والعدد والمتصلة واستخدام مع المتغيرات المساعدة الفردية ، واستعارة القوة من مناطق أخرى والمسوحات السابقة ، وبناء فترات ثقة حتى مع أحجام عينات صغيرة

رموز اساسية:

θ_i :الخاصية الحقيقية للمنطقة محل الاهتمام (على سبيل المثال ، معدل المنطقة)

$\tilde{\theta}_i$ - المقدر المباشر ل θ_i

m - عدد المناطق مع بيانات العينة

n_i - حجم عينة المنطقة i

N_i - الحجم الحقيقي للمنطقة

نموذج مستوى الوحدة

نموذج مستوى الوحدة الاساسي (Fay and Herriot, 1979)

نموذج المعاينة-sampling model

$$\tilde{\theta}_i = \theta_i + e_i$$

نموذج الربط المناطق-linking model

$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

x_i متجه عامودي ـمستوى المنطقة

e_i and u_i متغيرات عشوائية مستقلة (اخطاء عشوائية)

σ_{Di}^2 and σ_u^2 $e_i \sim (0,)$ و $u_i \sim (0,)$

لاحظ ايضا:

$\tilde{\theta}_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i + e_i \rightarrow$ نموذج خطي مختلط

e_i-خطا المعاينة بالنسبة للتقدير المباشر $\tilde{\theta}_i$

u_i: الفرق بين قيمة المنطقة الحقيقية θ_i والانحدار $\mathbf{x}_i'\boldsymbol{\beta}$

تشمل التأثير المشتركة لخصائص المنطقة والتي لا يتم احتسابها من قبل المتغيرات المشتركة المعلومة على مستوى المنطقة المدرجة في $\mathbf{x}_i$

نموذج مستوى الوحدة

افضل متنبىء خطي غير متحيز

Best Linear Unbiased Predictor (BLUP)

$$\hat{\theta}_i = \gamma_i \tilde{\theta}_i + (1-\gamma_i)\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS}$$

$$\gamma_i = \frac{\sigma_u^2}{\sigma_{Di}^2 + \sigma_u^2}$$

$$\hat{\boldsymbol{\beta}}_{GLS} = (\sum_{i=1}^{m} \frac{1}{\nu_i}\mathbf{x}_i\mathbf{x}_i')^{-1} \sum_{i=1}^{m} \frac{1}{\nu_i}\mathbf{x}_i\tilde{\theta}_i$$

$$\nu_i = \sigma_{Di}^2 + \sigma_u^2.$$

اذا كان $n_i = 0 \Rightarrow \tilde{\theta}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS}$ --التقدير التركيبي

يسمى $\gamma_i = \sigma_u^2/(\sigma_{Di}^2 + \sigma_u^2)$

- عامل الانكماش - shrinkage factor - لانه ينكمش المقدر المباشر نحو المقدر التركيبي

نموذج مستوى الوحدة

تابع ـ نموذج مستوى الوحدة الاساسي

$$\begin{aligned}\hat{\theta}_i &= \gamma_i\tilde{\theta}_i + (1-\gamma_i)\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS} \\ &= \tilde{\theta}_i - (1-\gamma_i)(\tilde{\theta}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS}) \\ &= \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS} + \gamma_i(\tilde{\theta}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS}) = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS} + \hat{u}_i\end{aligned}$$

σ_{Di}^2 يعتمد على حجم عينة المنطقة n_i، لكن على عكس الطريقة المعتمدة على التصميم، γ_i يعتمد ايضا على σ_u^2 (تباين θ_i حول x_i)

اذا كان – $n_i = 0 \Rightarrow \tilde{\theta}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS}$ لا يوجد نظرية تعتمد على تصمصيم المعاينة !!! تدعم تقدير المناطق الصغيرة في حالة $n_i = 0$

يمكن برهنة ان:

- $E_M(\hat{\theta}_i - \theta_i) = 0$ $\hat{\theta}_i$ تقدير غير متحيز ل θ_i

$$\gamma_i = Corr_M^2(\tilde{\theta}_i, \theta_i)$$

تابع- افضل متنبىء خطي غير متحيز

افرض عدم وجود معلومات او متغيرات مساعدة، اي ان $x_i = 1$ لكل i

$\tilde{\theta}_i = \bar{y}_i$ - معدل عينة المنطقة i

يمكن برهنة ان

$$\hat{\theta}_i = \gamma_i \bar{y}_i + (1-\gamma_i)\bar{y} = \bar{y}_i - (1-\gamma_i)(\bar{y}_i - \bar{y}) = \bar{y} + \gamma_i(\bar{y}_i - \bar{y})$$

$$\bar{y} = (\sum_{i=1}^{m} \frac{1}{v_i})^{-1} \sum_{i=1}^{m} \frac{1}{v_i} \bar{y}_i \; ; \; v_i = \sigma_{Di}^2 + \sigma_u^2 \; .$$

حالة خاصة:

اذا كان

$$\sigma_{Di}^2 = \sigma_D^2 \text{ for all } i, \; \gamma_i = \sigma_u^2/(\sigma_D^2 + \sigma_u^2) = \gamma \text{ and } \bar{y} = \sum_{i=1}^{m} \bar{y}_i / m$$

فان هذا يفسر فكرة الانكماش: مقدرات العينة المباشرة (معدلات العينة البسيطة $\bar{y}_i$) تقلص او تنكمش الى إلى المعدل العام $\bar{y}$

تباين (خطا التنبؤ) افضل متنبىء خطي غير متحيز

$$Var_M(\hat{\theta}_i - \theta_i) = E_M(\hat{\theta}_i - \theta_i)^2 \quad [\text{ since } E_M(\hat{\theta}_i - \theta_i) = 0 \text{ }]$$
$$= \gamma_i \sigma_{Di}^2 + (1-\gamma_i)^2 \mathbf{x}_i' Var(\hat{\boldsymbol{\beta}}_{GLS}) \mathbf{x}_i = g_{1i} + g_{2i}$$

حيث

$$Var(\hat{\boldsymbol{\beta}}_{GLS}) = \sigma_u^2 [\sum_{i=1}^{m} \gamma_i \mathbf{x}_i \mathbf{x}_i']^{-1}$$

وعليه

$$Var_M(\hat{\theta}_i) = \gamma_i \sigma_{Di}^2 + (1-\gamma_i)^2 \mathbf{x}_i' Var(\hat{\boldsymbol{\beta}}_{GLS}) \mathbf{x}_i = g_{1i} + g_{2i}$$

اذا كان m كبيرة فان

$$Var(\hat{\theta}_i) \approx g_{1i}$$

ملاحظة هامة:

$$Var(\hat{\theta}_i) \approx g_{1i} = \gamma_i \sigma_{Di}^2 \text{ اقل من } \sigma_{Di}^2$$

هذه العلاقة بين التباين في حالة النموذج وتباين المقدر المباشر توضح تفوق استخدام النموذج على استخدام المقدر المباشر

الحالة: σ_u^2 التباين غير معلوم
EBLUP

يمكن تقدير σ_u^2 باستخدام طرق الارجحية العظمى (ML) او

الارجحية العظمى المقيدة (REML) او طريقة العزوم (MM)

من المألوف اعتبار ان تباين المعاينة معلوم $\sigma_{Di}^2 \approx \sigma_D^2 / n_i$

طريقة العزوم1 – تقدير σ_u^2

$$\tilde{\sigma}_u^2 = \frac{1}{(m-p)}[\sum_{i=1}^{m}(\tilde{\theta}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{OLS})^2 - \sum_{i=1}^{m}\sigma_{Di}^2 h_i]$$

$$h_i = (1 - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i)$$

$$\hat{\sigma}_u^2 = \max(0, \tilde{\sigma}_u^2)$$

نموذج مستوى الوحدة

$$\begin{aligned}\hat{\theta}_i &= \gamma_i\tilde{\theta}_i + (1-\gamma_i)\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS} \\ &= \tilde{\theta}_i - (1-\gamma_i)(\tilde{\theta}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS}) \\ &= \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS} + \gamma_i(\tilde{\theta}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS}) = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GLS} + \hat{u}_i\end{aligned}$$

نحصل على افضل متنبىء تجريبي خطي غير متحيز (EBLUP)

اذا كان $\hat{\sigma}_u^2 = 0$

$$\gamma_i = \frac{\sigma_u^2}{\sigma_{Di}^2 + \sigma_u^2}$$

فاننا نحصل على المقدر التركيبي اي ان:

$$\mathbf{EBLUP} = \mathbf{x}_i'\hat{\boldsymbol{\beta}} \rightarrow \mathbf{Synthetic\ Estimator.}$$

Maximum and Restricted Maximum Likelihood Estimation Methods

Maximum Likelihood Estimation

5.2.4 ML and REML Estimators

We now provide formulas for the ML and REML estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ under the general linear mixed model (5.2.1) and the associated asymptotic covariance matrices, assuming normality (Cressie 1992). Under normality, the log-likelihood function is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\delta}) = c - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \qquad (5.2.16)$$

where c denotes a generic constant. The partial derivative of $l(\boldsymbol{\beta}, \boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is given by $\mathbf{s}(\boldsymbol{\beta}, \boldsymbol{\delta})$, with the jth element

$$s_j(\boldsymbol{\beta}, \boldsymbol{\delta}) = \partial l(\boldsymbol{\beta}, \boldsymbol{\delta})/\partial\delta_j = -\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{V}_{(j)}) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{(j)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{V}_{(j)} = \partial\mathbf{V}/\partial\delta_j$ and $\mathbf{V}^{(j)} = \partial\mathbf{V}^{-1}/\partial\delta_j = -\mathbf{V}^{-1}\mathbf{V}_{(j)}\mathbf{V}^{-1}$, noting that $\mathbf{V} = \mathbf{V}(\boldsymbol{\delta})$. Also, the matrix of expected second-order derivatives of $-l(\boldsymbol{\beta}, \boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is given by $\mathcal{I}(\boldsymbol{\delta})$ with (j, k)th element

$$\mathcal{I}_{jk}(\boldsymbol{\delta}) = \frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{V}_{(j)}\mathbf{V}^{-1}\mathbf{V}_{(k)}). \qquad (5.2.17)$$

The ML estimator of $\boldsymbol{\delta}$ is obtained iteratively using the Fisher-scoring algorithm, with updating equation

$$\boldsymbol{\delta}^{(a+1)} = \boldsymbol{\delta}^{(a)} + [\mathcal{I}(\boldsymbol{\delta}^{(a)})]^{-1}\mathbf{s}[\tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}^{(a)}), \boldsymbol{\delta}^{(a)}], \qquad (5.2.18)$$

Restricted Maximum Likelihood Estimation

A drawback of the ML estimator of $\boldsymbol{\delta}$ is that it does not take account of the loss in degrees of freedom (df) due to estimating $\boldsymbol{\beta}$. For example, when $y_1, \ldots, y_n$ are iid $N(\mu, \sigma^2)$, the ML estimator $\hat{\sigma}^2 = [(n-1)/n]s^2$ is not equal to the customary unbiased estimator of $\sigma^2, s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$. The REML method takes account of the loss in df by using the transformed data $\mathbf{y}^* = \mathbf{A}^T\mathbf{y}$, where $\mathbf{A}$ is any $n \times (n-p)$ full-rank matrix that is orthogonal to the $n \times p$ matrix $\mathbf{X}$. It follows that $\mathbf{y}^* = \mathbf{A}^T\mathbf{y}$ is distributed as a $(n-p)$-variate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{A}^T\mathbf{V}\mathbf{A}$. The logarithm of the joint density of $\mathbf{y}^*$, expressed as function of $\boldsymbol{\delta}$, is called restricted log-likelihood function and is given by

$$l_R(\boldsymbol{\delta}) = c - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{y}, \tag{5.2.20}$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}. \tag{5.2.21}$$

Note that $\mathbf{P}\mathbf{y} = \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$. The partial derivative of $l_R(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is given by $\mathbf{s}_R(\boldsymbol{\delta})$, with the jth element

$$s_{Rj}(\boldsymbol{\delta}) = \partial l_R(\boldsymbol{\delta})/\partial\delta_j = -\frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{V}_{(j)}) + \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{V}_{(j)}\mathbf{P}\mathbf{y}.$$

نموذج مستوى الوحدة

تطبيقات

Software SAE R Package

- **The R package sae contains two specific functions for the FH area level model.**

- **Function eblupFH gives the EBLUP estimates of θi, i = 1, … , m, based on the FH model**

- **Function mseFH returns the same EBLUP estimates together with MSE estimates, depending on the model fitting method specified in argument method**

- **Both functions admit ML, REML (default), and moments fitting methods. The calls to these functions are as follows**

- **eblupFH(formula, vardir, method = "REML", MAXITER = 100, PRECISION = 0.0001, data)**
-
- **mseFH(formula, vardir, method = "REML", MAXITER = 100, PRECISION = 0.0001, data)**

نموذج مستوى الوحدة

Formula

- **the regression equation for the fixed (nonrandom) part of the model as a usual R regression formula. The left-hand side of the regression formula is the vector of direct estimates $\hat{\theta}$ = ($\hat{\theta}$ 1, … , $\hat{\theta}$ m) T , and the right-hand side contains the auxiliary variables included in zi, separated by + and with an intercept included by default**

vardir

- **The vector of sampling variances of direct estimators, ψi, must also be specified in the argument vardir**

MAXITER and PRECISION

- **The optional arguments MAXITER and PRECISION might be used, respectively, to specify the maximum number of iterations of the Fisher-scoring fitting algorithm and the relative difference between the model parameter estimates in two consecutive iterations below which the Fisher-scoring algorithm stops**

- **The function eblupFH returns a list of two objects: the vector eblup with the EBLUP estimates $\hat{\theta}$H i for the domains, and the list fit, containing the results of the fitting method.**

نموذج مستوى الوحدة

- **The function mseFH returns a list with two objects:**

- **The first170 EBLUP: BASIC AREA LEVEL MODEL object is the list est containing the results of the model fit (in another list called fit) and the vector of EBLUP estimates (eblup)**

- **The second object is the vector mse, which contains the second-order unbiased MSE estimates.**

- **The list fit contains the results of the fitting procedure, namely**

- **the fitting method (method),**

- **a logical value indicating if convergence is achieved in the specified maximum number of iterations (convergence), the number of iterations (iterations),**

- **the estimated regression coefficients (estcoef),**

- **the estimate of the random effects variance $\sigma 2\ v$ (refvar),**

- **and the usual goodness-of-fit criteria mAIC, BIC, and loglikelihood (AIC, BIC, and loglikelihood)**

نموذج مستوى الوحدة

Application

Milk Expenditure

إنفاق الأسرة على الحليب كامل الدسم

- **We illustrate the use of the mseFH function (the function eblupFH is used similarly) with the predefined data set milk on fresh milk expenditure.**

- **This data set was used originally by Arora and Lahiri (1997) and later by You and Chapman (2006)**

- **It contains m = 43 observations on the following**
- **six variables:**

- **SmallArea containing the areas of inferential interest,**
- **ni with the area sample sizes**

- **yi with the average expenditure on fresh milk for the year 1989 (direct estimates)**

- **SD with the estimated standard deviations of direct estimators**

- **CV with the estimated coefficients of variation of direct estimators**

نموذج مستوى الوحدة

- **MajorArea containing major areas created by You and Chapman (2006)**

- **We will obtain EBLUP estimates of average area expenditure on fresh milk for 1989 based on the FH model with fixed effects for MajorArea categories**

- **We will also calculate second-order unbiased MSE estimates and coefficients of variation (CVs)**

- **The gain in efficiency of the EBLUP estimators in comparison with direct estimators will be analyzed.**
- **The following R code loads the sae package and the milk data set: (see below)**
- **With the previous sentence, the output of the function mseFH has been placed in the object called FH.**
- **Then, FHesteblup gives the vector of EBLUP estimates and FH$mse gives the MSE estimates.**
- **We calculate coefficients of variation (CVs) as square root of MSE estimates divided by the point estimates, in percentage.**

نموذج مستوى الوحدة

Area Level Model – SAE R Package

library(sae)
data("milk")
attach(milk)

```
FH <- mseFH(yi ~ as.factor(MajorArea), SD^2)
cv.FH <- 100 * sqrt(FH$mse) / FH$est$eblup
results <- data.frame(Area = SmallArea, SampleSize = ni, DIR = yi,  cv.DIR = 100 * CV, eblup.FH = FH$est$eblup, cv.FH)
detach(milk)
results <- results[order(results$SampleSize, decreasing = TRUE), ]

plot(results$DIR, type = "n", ylab = "Estimate", ylim = c(0.4, 1.6), xlab = "area (sorted by decreasing sample size)", cex.axis = 1.5,  cex.lab = 1.5)
points(results$DIR, type = "b", col = 1, lwd = 2, pch = 1, lty = 1)
points(results$eblup.FH, type = "b", col = 4, lwd = 2, pch = 4, lty = 2)
legend("top", legend = c("Direct", "EBLUP FH"), ncol = 2, col = c(1, 4), lwd = 2, pch = c(1, 4), lty = c(1, 2), cex = 1.3)
plot(results$cv.DIR, type = "n", ylab = "CV", ylim = c(5, 40),  xlab = "area (sorted by decreasing sample size)", cex.axis = 1.5,  cex.lab = 1.5)
points(results$cv.DIR, type = "b", col = 1, lwd = 2, pch = 1, lty = 1)
points(results$cv.FH, type = "b", col = 4, lwd = 2, pch = 4, lty = 2)
legend("top", legend = c("Direct", "EBLUP FH"), ncol = 2, col = c(1, 4), lwd = 2,  pch = c(1, 4), lty = c(1, 2), cex = 1.3)
```

****

Conclusion:

- **EBLUP and direct area estimates of average expenditure are plotted for each small area in Figure a.**
- **CVs of these estimates are plotted in Figure b.**
- **In both plots, small areas have been sorted by decreasing sample size.**
- **Observe in Figure a that EBLUP estimates track direct estimates but seem to be more stable.**
- **See also in Figure b that CVs of EBLUP estimates are smaller than those of direct estimates for all the small areas.**
- **In fact, direct estimates have CVs over 20% for several areas, whereas the CVs of the EBLUP estimates do not exceed this limit for any of the areas.**
- **Moreover, the gains in efficiency of the EBLUP estimators tend to be larger for areas with smaller sample sizes (those in Figure b).**
- **Thus, in this example, EBLUP estimates based on FH model seem to be more reliable than direct estimates**