PROYECTO DE ESTIMACIÓN DE ÁREAS PEQUEÑAS: DIFUSIÓN DE RESULTADOS Y ESTADO ACTUAL DEL PROYECTO

Iosune Azula, Patxi Garrido y Haritz Olaeta

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

PROYECTO DE ESTIMACIÓN DE ÁREAS PEQUEÑAS: DIFUSIÓN DE RESULTADOS Y ESTADO ACTUAL DEL PROYECTO

Iosune Azula, Patxi Garrido y Haritz Olaeta

Indice

Capítulo
1

Introducción

Eustat, consciente de la creciente necesidad de estadísticas de calidad, constituyó hace dos año un equipo de investigación compuesto por miembros de diferentes departamentos de Eustat para trabajar en la mejora de las técnicas de estimación en diferentes operaciones estadísticas e introducir técnicas de estimación en áreas pequeñas basadas en modelos. Este equipo está supervisado por las profesoras Ana Fernández Militino y Lola Ugarte de la Universidad Pública de Navarra.

Este proyecto se puede dividir en dos subproyectos, complementarios en cierta medida:

1. **Encuesta Industrial**. Las fases del proyecto han sido:

   a. Análisis del sistema de estimación utilizado en la Encuesta Industrial desde su puesta en marcha y derivación matemática de los estimadores de los coeficientes de variación asociados.
   b. Estudio de estimadores alternativos.
   c. Selección, propuesta y decisión de cambio en el sistema de estimación (a partir de la Encuesta Industrial del 2005, siguiente cambio de año base).
   d. Estudio de diferentes estimadores de áreas pequeñas para la Encuesta Industrial.
   e. Selección, propuesta y adopción de un sistema de estimación en áreas pequeñas.

   Este subproyecto se dará por finalizado en diciembre de 2005. La publicación de las primeras estimaciones, correspondientes a la Encuesta Industrial del año 2003, está prevista para el último trimestre del año 2005.

2. **Población en Relación a la Actividad. Las fases del proyecto son:**

   a. Análisis del sistema de estimación utilizado en la encuesta de Población en Relación a la Actividad.
   b. Estudio de estimadores alternativos.
   c. Selección del estimador a utilizar.
   d. Estudio de diferentes estimadores de áreas pequeñas para la encuesta de Población en Relación a la Actividad.
   e. Selección, propuesta y adopción, si procede, de un sistema de estimación en áreas pequeñas.

   Las fases a), b) y c) están actualmente básicamente ejecutadas y se está trabajando en la fase d).

Este trabajo se centra en el subproyecto de la Encuesta Industrial y se divide en cuatro partes. Comienza con una breve introducción a la Encuesta Industrial de la Comunidad Autónoma de Euskadi. A continuación, se describen los estimadores tanto de los totales como de los coeficientes de variación asociados utilizados además de la propuesta de cambio en los mismos aprobada en Eustat. Se pasa a continuación a describir el plan de estimación en áreas pequeñas adoptado por Eustat. Finalmente se describe el plan de difusión que seguirá Eustat con las estimaciones comarcales de la Encuesta Industrial del año 2003, que serán difundidas a lo largo del último trimestre del año 2005.

Capítulo
2

Encuesta industrial de la Comunidad Autónoma de Euskadi

Antecedentes

Esta operación se puso en marcha en 1981, teniendo desde su creación como objetivo fundamental el conocimiento pormenorizado del entramado industrial vasco, dada su importancia tanto en términos de valor añadido como de empleo. La información básica para ello se obtiene a partir de las principales partidas de la cuenta de pérdidas y ganancias, y la consiguiente estimación, a partir de ellas, de las principales macromagnitudes.

Esta operación estadística se realiza en colaboración con el Servicio de Estadística y Análisis Sectorial del Departamento de Agricultura y Pesca, Organo Estadístico específico de dicho Departamento.

Características Técnicas

Ambitos

Universo: El ámbito poblacional se circunscribe a aquellos establecimientos cuya actividad principal, medida en términos de valor añadido generado, sea industrial.

Incluye, según la Clasificación Nacional de Actividades Económicas de 1993 (en adelante CNAE-93), las siguientes secciones:

Sección C: Industrias extractivas

Sección D: Industria manufacturera

Sección E: Producción y distribución de energía eléctrica, gas y agua

Geográfico. Las unidades estadísticas que estén ubicadas en el ámbito geográfico de la C.A. de Euskadi, aun cuando su sede social o gerencia se encuentre fuera de ella.

Temporal. El período de referencia es el ejercicio económico del año natural. Excepcionalmente, de presentarse establecimientos cuya contabilidad vaya referida a períodos de tiempo que no correspondan al año natural, se referirá la información a los ejercicios que finalizan dentro de los años correspondientes.

Marco de la encuesta

El marco de la encuesta es el Directorio de Actividades Económicas de Eustat. Su utilización permite la elaboración de un muestreo probabilístico que acote los errores muestrales.

Unidad Estadística

La unidad estadística es el establecimiento definido como una unidad que ejerce, exclusiva o principalmente, una o varias actividades situada en un mismo emplazamiento geográfico.

Diseño muestral y extrapolación

Se realiza un muestreo probabilístico en dos fases: una primera en la que se seleccionan con probabilidad "uno" todas las unidades que tengan más de 19 empleados; en la segunda fase, se realiza un muestreo aleatorio estratificado donde las variables de estratificación son:

1. Territorio Histórico: Araba, Bizkaia y Gipuzkoa.

b) Actividad: Clasificación Nacional de Actividades Económicas (CNAE-93) a nivel de subclase, es decir, a 5 dígitos. Posteriormente para su difusión se utiliza la clasificación normalizada de EUSTAT A84. La clasificación A84 es una desagregación de la A60 (CNAE-93 a 2 dígitos) en función de la estructura económica de la C.A. de Euskadi.

El tamaño de la muestra seleccionada es de 3.000 unidades estadísticas, aproximadamente.

Previamente a la extrapolación, se post-estratifican los establecimientos muestrales, según los tres Territorios Históricos (Araba, Bizkaia, Gipuzkoa) , subclase de la CNAE-93 y 5 tamaños de establecimientos, que son:

1. Entre 1 y 19 empleados
2. Entre 20 y 49 empleados
3. Entre 50 y 99 empleados
4. Entre 100 y 499 empleados
5. Mayores o iguales a 500 empleados.

El paso de datos muestrales a los poblacionales se realiza a través de una matriz de elevadores por cada estrato. La variable utilizada para la obtención de los elevadores ha sido el número de ocupados de los establecimientos industriales. El uso de esta variable está justificado en que es la más correlacionada con las principales variables económicas que intenta medir la encuesta.

Capítulo
3

Estimadores utilizados en la Encuesta Industrial

Estimadores utilizados actualmente en Eustat

Todas las unidades estadísticas con más de 19 empleados son autoponderadas, por lo que el interés radica principalmente en las estimaciones dentro del estrato de empleo de 1-19 empleados. En lo que sigue se describe la estimación dentro de un sector de actividad el total de una variable cualquiera y así como de su correspondiente coeficiente de variación.

Actualmente, para la estimación del total de la variable y se utiliza el estimador indirecto de razón o estimador sintético utilizando como información auxiliar el número de empleados de los establecimientos.

El estimador indirecto de razón de una variable de interés cualquiera, y, cuando se dispone de una variable auxiliar x está, en el caso de la Encuesta Industrial, asistido por el modelo de regresión lineal simple heterocedástico del tipo:

$$y_{hj} = x_{hj}\beta + \varepsilon_{hj} \qquad \text{con} \quad \text{var}(\varepsilon_{hj}) = \sigma^2 x_{hj}\,, \tag{1}$$

donde h hace referencia al estrato, j a la unidad estadística y el resto de la notación es la habitual.

En el caso de la Encuesta Industrial la variable auxiliar utilizada, tras comprobar el poder explicativo que ésta presenta para la mayoría de las variables más relevantes de la Encuesta Industrial, es la variable empleo y los estratos son los Territorios Históricos (Araba, Bizkaia y Gipuzkoa) dado que en todo lo que sigue se supone el interés radica en un único sector de actividad (CNAE a 5 dígitos).

El estimador del total de la variable y en un sector dado en el Territorio Histórico h viene dado por:

$$\hat{t}_{yh.SYN} = X_h \hat{\beta} = X_h \frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj} y_{hj}}{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj} x_{hj}},$$

donde $X_h = \sum_{j=1}^{N_h} x_{hj}$, w_{hj} es el peso de muestreo de la unidad j en el Territorio Histórico h, x_{hj} recoge el empleo del establecimiento j del Territorio Histórico h y n_h es el tamaño de la muestra en el Territorio Histórico h.

El estimador de la varianza del estimador indirecto de razón se puede aproximar con:

$$\hat{\text{var}}(\hat{t}_{yh.SYN}) \approx N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\left(\frac{X_h}{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj}x_{hj}}\right)^2 \hat{\text{var}}(\varepsilon)$$

donde $\hat{\text{var}}(\varepsilon)$ es la varianza muestral de los residuos del modelo heterocedástico (1) con todos los datos muestrales (es decir, se calculan los residuos en toda la CA de Euskadi, no sólo en el Territorio Histórico h) y el resto de la notación es la habitual.

Särndal and Hidiroglou (1989) proporcionan una aproximación del sesgo del estimador sintético

según la cual $E\left(\hat{t}_{yh.SYN}\right)-\hat{t}_{yh.SYN} \approx -\sum_{j=1}^{N}\varepsilon_j$ donde $\varepsilon_j = y_j - x_j\hat{\beta}$. Luego, el estimador será aproximadamente insesgado si se verifica que $\sum_{j=1}^{N}\varepsilon_k = 0$. Esta condición no se satisface normalmente. Si el modelo no ajusta bien en el dominio de interés, la suma de residuales puede estar lejos de cero, indicando un sesgo considerable. En caso contrario, podemos esperar un sesgo limitado. Por ello, es deseable estimar el error cuadrático medio como medida de precisión del estimador. Viene dado por

$$MSE\left(\hat{t}_{yh.SYN}\right) = \text{var}\left(\hat{t}_{yh.SYN}\right) + \left(sesgo_{h.SYN}\right)^2,$$

y se estima mediante la expresión:

$$\hat{MSE}\left(\hat{t}_{yh.SYN}\right) = \hat{\text{var}}\left(\hat{t}_{yh.SYN}\right) + \left(\sum_{j=1}^{n_h}\hat{\varepsilon}_j\right)^2,$$

donde $\hat{\varepsilon}_j \quad j=1,\ldots,n$ son los residuos obtenidos a partir del modelo estimado (1) con todos los datos muestrales, aunque en cada Territorio Histórico solamente se suman los específicos de ese Territorio Histórico. El coeficiente de variación se define como

$$\hat{cv}\left(\hat{t}_{yh.SYN}\right) = \frac{\hat{rmse}\left(\hat{t}_{yh.SYN}\right)}{\hat{t}_{yh.SYN}},$$

donde $\hat{rmse}\left(\hat{t}_{yh.SYN}\right) = \sqrt{\hat{MSE}}\left(\hat{t}_{yh.SYN}\right)$.

Estimadores aprobados a partir de la Estadística Industrial 2005

Tras analizar los errores cuadráticos medios proporcionados por el estimador sintético en la Encuesta Industrial en diferentes años, se optó por proponer una batería de estimadores alternativos y comparar sus errores cuadráticos medios con el del sintético. Entre los estimadores analizados destacar, además de diferentes estimadores directos e indirectos, una batería de estimadores compuestos. Tras comparar los errores cuadráticos medios de los distintos estimadores, se escogió (siguiendo un criterio de minimización de errores cuadráticos medios pero, además, penalizando la introducción de sesgos importantes) un estimador compuesto que se describe a continuación. La introducción de dicho estimador se ha programado para la Encuesta Industrial del año 2005, siguiente cambio de año base previsto.

El estimador compuesto propuesto para la Encuesta Industrial es una combinación del estimador directo de razón1 con el estimador indirecto de razón que se utiliza en la actualidad propuesto por Pfefferman (2002). Este estimador se construye para compensar el posible sesgo del estimador indirecto de razón con la insesgadez de un estimador directo y la imprecisión del estimador directo con la precisión del estimador indirecto de razón. Viene dado por:

$$\hat{t}_{yh.C} = \phi_h \hat{t}_{yh.D} + (1-\phi_h)\hat{t}_{yh.SYN} \qquad \text{con} \quad \phi_h = \frac{n_h}{N_h}.$$

Esta elección de ϕ_h es especialmente adecuada para poblaciones de tamaño pequeño, ya que en otro caso el cociente $\frac{n_h}{N_h}$ no favorecería necesariamente al estimador directo cuando n_h crece. Con estos pesos, el peso del estimador directo o indirecto es mayor según sea su representación muestral. Esto es, a mayor fracción muestral, mayor contribución del estimador directo. Cuando la población está poco representada en la muestra, es el estimador indirecto el que tiene más peso en el estimador compuesto. Puede ocurrir también que $n_h = 1$ y $N_h = 1$, en cuyo caso el estimador compuesto sería igual al directo.

El error cuadrático medio de este estimador compuesto se puede aproximar mediante

$$MSE(\hat{t}_{yh.C}) \approx \phi_h^2 MSE(\hat{t}_{yh.D}) + (1-\phi_h)^2 MSE(\hat{t}_{yh.SYN}) + 2\phi_h(1-\phi_h)E\left[(\hat{t}_{yh.D} - Y_h)(\hat{t}_{yh.SYN} - Y_h)\right].$$

Su estimación no es fácil, ya que puede ocurrir que el tercer término de este sumatorio, es decir el de la covarianza, no sea pequeño. En Eustat se ha aproximado del siguiente modo:

[1] Tanto el estimador del total como del correspondiente coeficiente de variación se puede encontrar en cualquier libro de muestreo básico.

$$E\left[\left(\hat{t}_{yh.D} - Y_h\right)\left(\hat{t}_{yh.SYN} - Y_h\right)\right] \approx MSE\left(\hat{t}_{yh.D}\right) - Y_h\left(sesgo_{h.SYN}\right).$$

Por lo tanto, el estimador del error cuadrático medio utilizado en Eustat viene dado por

$$\hat{MSE}\left(\hat{t}_{yh.C}\right) \approx \phi_h^2\ \hat{MSE}\left(\hat{t}_{yh.D}\right) + \left(1-\phi_h\right)^2\ \hat{MSE}\left(\hat{t}_{yh.SYN}\right) + 2\phi_h\left(1-\phi_h\right)\left[\hat{MSE}\left(\hat{t}_{yh.D}\right) - \hat{t}_{yh.SYN}\left(sesgo_{h.SYN}\right)\right].$$

El estimador del coeficiente de variación viene, por consiguiente, dado por

$$\hat{cv}\left(\hat{t}_{yh.C}\right) = \frac{\hat{rmse}\left(\hat{t}_{yh.C}\right)}{\hat{t}_{yh.C}},$$

donde $\hat{rmse}\left(\hat{t}_{yh.C}\right) = \sqrt{\hat{MSE}\left(\hat{t}_{yh.C}\right)}$.

Capítulo
4

Sistema de estimación de áreas pequeñas en la Encuesta Industrial

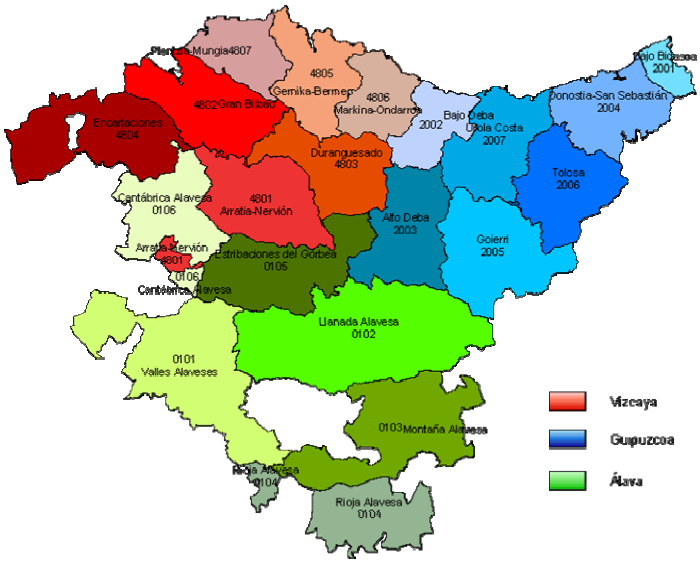
Introducción

La C.A .de Euskadi se divide en las siguientes 20 comarcas administrativas:

- Araba: Valles Alaveses, Llanada Alavesa, Montaña Alavesa, Rioja Alavesa, Estribaciones del Gorbea y Cantábrica Alavesa.

- Bizkaia: Arratia-Nervión, Gran Bilbao, Duranguesado, Encartaciones, Gernika-Bermeo, Markina-Ondarroa y Plentzia-Mungia.

- Gipuzkoa: Bajo Bidasoa, Bajo Deba, Alto Deba, Donosita-San Sebastián, Goierri, Tolosa y Urola Costa.



La actividad industrial de la CA de Euskadi no está uniformemente repartida en las 20 comarcas administrativas y tanto la importancia del sector industrial como su tamaño varía enormemente entre comarcas. De esta forma, el número de establecimientos industriales (no de construcción) que aparecen en el Directorio de Actividades Industriales del año 2003 es:

Valles Alaveses: 63

- Llanada Alavesa: 1344
- Montaña Alavesa: 28
- Rioja Alavesa: 605
- Estribaciones del Gorbea: 171
- Cantábrica Alavesa: 229
- Arratia-Nervión: 210
- Gran Bilbao: 4451
- Duranguesado: 1065
- Encartaciones: 205
- Gernika-Bermeo: 241
- Markina-Ondarroa: 196
- Plentzia-Mungia: 308
- Bajo Bidasoa: 504
- Bajo Deba: 757
- Alto Deba: 603
- Donosita-San Sebastián: 2222
- Goierri: 575
- Tolosa: 567
- Urola Costa: 769

Como se ve, no sólo se trata de un entramado industrial heterogéneo si no que realmente hay comarcas en las que la actividad industrial es realmente pequeña por lo que la tarea de estimación comarcal requiere ciertamente de técnicas de estimación en áreas pequeñas.

Los modelos de áreas pequeñas suponen la existencia de un modelo subyacente que siguen todos los datos de la población, pero que se estima con los datos de la muestra (Rao, 2003). Eustat utiliza para la obtención de estimaciones comarcales en la Encuesta Industrial dos tipos de modelos: el modelo lineal de efectos fijos y el modelo de regresión lineal con efectos fijos y aleatorios, llamado también modelo mixto. En el modelo mixto el predictor consta de un término común de efectos fijos y otro diferenciado para los elementos de cada comarca d ($d = 1,...,t$). Este término diferenciado está formado por los efectos aleatorios (v_d), de modo que todos los datos de la misma comarca comparten el mismo efecto aleatorio. En el caso del modelo de efectos fijos no existen términos diferenciados para cada comarca ya que la parte sistemática ($X\beta$) es común para todas las comarcas. Sin embargo, la especificidad se consigue al proyectar el coeficiente común (β) a la información auxiliar específica (X_d) de cada comarca.

Modelo lineal mixto

Se parte de una población formada por los N establecimientos de una CNAE concreta. En cada comarca d ($d=1,...,t$) hay N_d establecimientos, de modo que $N=\sum_d N_d$. En dicha CNAE se han muestreado n establecimientos de los que n_d pertenecen a la comarca d. Se propone el siguiente modelo lineal mixto heterocedástico

$$y_{dj}=\beta_0+\beta_1 x_{dj}+v_d+e_{dj}, \qquad d=1,...,t \quad j=1,...,n_d,$$

donde para el establecimiento j de la comarca d, y_{dj} es la variable de interés y x_{dj} es el número de empleados del establecimientos. El número total de establecimientos muestreados en la comarca d es n_d. Los efectos fijos del modelo son β_0 y β_1. El efecto aleatorio común para todos los establecimientos de la comarca d es v_d y e_{dj} son los errores aleatorios específicos de cada establecimiento. Además, se supone que $v_d \subset N\left(0,\sigma_v^2\right)$ y $e_{dj} \subset N\left(0,\sigma_e^2 c_{dj}^{-1}\right)$ son independientes. Para corregir la heterocedasticidad presente en los datos se utilizan los pesos $c_{dj}=1/x_{dj}$. Cuando $c_{dj}=1$, $\forall d,j$, este modelo es similar al propuesto por Battese et al (1988).

Cuando la fracción de muestreo por comarca $f_d=n_d/N_d$ no es despreciable, la literatura recomienda utilizar la versión predictiva para obtener la predicción del total de la comarca d en lugar de la versión proyectiva. Esta versión consiste en diferenciar la parte muestreada de la no muestreada. Así, la predicción de la parte muestreada es la misma muestra, mientras que la no muestreada se predice con el predictor de tipo proyectivo.

Para obtener la versión predictiva se descompone el total $\sum_{j\in N_d} y_{dj}=\sum_{j\in d_r} y_{dj}+\sum_{j\in d_s} y_{dj}$, donde d_s indica la muestra en la comarca d y d_r el resto de los establecimientos no pertenecientes a la muestra de la comarca d. Se deriva sin excesivas dificultades el estimador del total como:

$$\hat{t}_d=X_d^{'}\hat{\beta}+(N_d-n_d)\hat{\gamma}_{dc}\left(\bar{y}_{dc}-\bar{x}_{dc}\hat{\beta}\right)+\sum_{j=1}^{n_d} y_{dj}, \qquad d=1,...,t$$

donde $X_d^{'} = (N_d, X_d)$ $\hat{\gamma}_{dc} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / c_{d.}}$, $c_{d.} = \sum_{j=1}^{n_d} c_{dj}$,

$\bar{y}_{dc} = \dfrac{1}{c_{d.}} \sum_{j=1}^{n_d} c_{dj} y_{dj}$, $\bar{x}_{dc} = \dfrac{1}{c_{d.}} \sum_{j=1}^{n_d} c_{dj} x_{dj}^{'}$, $x_{dj}^{'} = (1, x_{dj})$ y donde $\hat{\beta} = \hat{\beta}_c(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$ ha sido evaluada con las estimaciones de los componentes de varianza.

Partiendo de Prasad y Rao (1990), la estimación del error cuadrático medio del total estimado para comarca utilizado en Eustat viene dado por:

$$\hat{MSE}[\hat{t}_d] = N_d^2 [g_{1d}(\hat{\sigma}^2) + g_{2d}(\hat{\sigma}^2) + 2g_{3d}(\hat{\sigma}^2),]$$

donde los términos de la varianza vienen dados por:

$$g_{1d}(\hat{\sigma}^2) = (1 - f_d)^2 (1 - \hat{\gamma}_{dc}) \hat{\sigma}_v^2,$$

$$g_{2d}(\hat{\sigma}^2) = (1 - f_d)^2 \left[(\bar{X}_d - \hat{\gamma}_{dc} \bar{x}_{dc})^{'} \hat{\Phi}_s (\bar{X}_d - \hat{\gamma}_{dc} \bar{x}_{dc}) \right],$$

$$g_{3d}(\hat{\sigma}^2) = (1 - f_d)^2 c_d^{-1} (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / c_{d.})^{-3} [\hat{\sigma}_e^4 \, \hat{\text{var}}(\hat{\sigma}_v^2) + \hat{\sigma}_v^4 \, \hat{\text{var}}(\hat{\sigma}_e^2) - 2\hat{\sigma}_e^2 \hat{\sigma}_v^2 \, \hat{\text{cov}}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)]$$

con $\hat{\Phi}_s = \text{var}(\hat{\beta})$.

El estimador del coeficiente de variación viene, por consiguiente, dado por

$$\hat{cv}\left(\hat{t}_d\right) = \frac{\hat{rmse}\left(\hat{t}_d\right)}{\hat{t}_d},$$

donde $\hat{rmse}\left(\hat{t}_d\right) = \sqrt{\hat{MSE}}\left(\hat{t}_d\right)$.

Modelo lineal de efectos fijos

El modelo lineal de efectos fijos propuesto para cada establecimiento viene dado por

$$y_{dj} = \beta x_{dj} + e_{dj}, \qquad d = 1,...,t \qquad j = 1,...,n_d,$$

donde para el establecimiento j de la comarca d, y_{dj} es el valor que toma la variable de interés y x_{dj} es el número de empleados del establecimiento. El número total de establecimientos muestreados en la comarca d es n_d, β es el único efecto fijo del modelo y $e_{dj} \subset N\left(0, \sigma_e^2 c_{dj}^{-1}\right)$ son los errores aleatorios. Para corregir la heterocedasticidad presente en los datos se utilizan los pesos $c_{dj} = 1/x_{dj}$.

El estimador del total para la comarca d se obtiene como

$$\hat{t}_d^F = \sum_{j=1}^{n_d} y_{dj} + X_d \hat{\beta}$$

donde la notación ha sido previamente introducida.

El error cuadrático medio se estima mediante

$$\hat{MSE}(\hat{t}_d^F) = N_d (1 - f_d)^2 \left[\overline{X}_d \, \hat{\text{var}}\left(\hat{\beta}\right) \overline{X}_d\right] + \sigma^2 \sum_{j=n_d}^{N_d} x_{dj} \,,$$

por lo que el estimador del coeficiente de variación viene, por consiguiente, dado por

$$\hat{cv}\left(\hat{t}_d^F\right) = \frac{\hat{rmse}\left(\hat{t}_d^F\right)}{\hat{t}_d^F},$$

donde $\hat{rmse}\left(\hat{t}_d^F\right) = \sqrt{\hat{MSE}}\left(\hat{t}_d^F\right)$.

Plan de estimación en la Encuesta Industrial

Dentro del grupo de investigación se ha programado una aplicación informática ad hoc en SAS para la introducción del cálculo de estimaciones de áreas pequeñas en la producción estadística de Eustat. Se trata de un programa específico para la Encuesta Industrial pero que fácilmente puede adaptarse a otro tipo de encuestas económicas. Son diversas las decisiones que se han tomado:

Se ha considerado necesario establecer un número mínimo de establecimientos para proceder al cálculo de los modelos mixtos o fijos. Si no se dispone de este número mínimo de establecimientos, se procede a hacer agregaciones de CNAEs con un dígito menos. Primero se estima el modelo mixto a ese nivel de agregación y si $\sigma_v^2 = 0$ ó

$\sigma_e^2 = 0$ entonces se estima el modelo de efectos fijos. Este número mínimo se ha fijado (puede ser variado) en la actualidad en 5 establecimientos.

Se ha tomado la decisión de utilizar el modelo de efectos fijos cuando el modelo mixto no es válido debido a que se considera prioritaria la decisión de no agrupar CNAES. Es por ello, que se prefiere un modelo de efectos fijos con 5 dígitos por ejemplo, a un modelo mixto de 3 dígitos.

Cuando se realiza una agregación, ésta permite estimar los coeficientes del modelo pero las predicciones se hacen de forma particularizada a la CNAE considerada.

Si la predicción hecha en una CNAE para alguna comarca es negativa se sustituye por la suma de los valores muestrales (sólo se considera esta parte en la predicción). Si no hay muestra se sustituye por 0.

El uso de la variable auxiliar "número de empleados" introduce heterocedasticidad en los modelos, ya que habitualmente la variable respuesta y tiene mayor variabilidad a medida que aumenta el número de empleados. Por ello, todos los modelos de efectos fijos y mixtos consideran que la varianza del error es proporcional al número de empleados.

En cada CNAE los totales por Territorio Histórico y por CA de Euskadi se obtienen de manera agregada a partir de las estimaciones por comarcas. Los totales por sector A84 se obtienen agregando las predicciones obtenidas a nivel de CNAE. Lo mismo sucede para los totales por TH y CAE. Se procede de igual modo para otros tipos de agregaciones.

En cada CNAE, para calcular las ráices cuadradas de los errores cuadráticos medios de las predicciones a nivel de Territorio Histórico, se aplican fórmulas específicas, ya que no se obtiene como raíz cuadrada de la suma de los errores cuadráticos medios de las predicciones por comarcas. Ello es debido a que en cada CNAE las estimaciones por comarcas no son independientes en ninguno de los modelos.

Sin embargo, una vez hechas las estimaciones de los RMSE por CNAEs para cada Territorio Histórico, el cálculo de las estimaciones por CAE es directo, a que ahora se cumple la hipótesis de independencia. A partir de los cálculos por CNAEs los RMSE de la variable A84 o de cualquier otra agrupación se obtienen de forma directa, es decir, calculando la raíz cuadrada de la suma de los MSE de las CNAEs que forman cada sector.

Capítulo
5

Difusión de las estimaciones comarcales de la Encuesta Industrial del año 2003

La difusión de los primeros resultados comarcales, correspondientes a la Encuesta Industrial del año 2003, está prevista para el último trimestre del año 2005. El objetivo para este año es ofrecer a los usuarios estimaciones del total del Valor Añadido Bruto a coste de factores por comarcas y A31 (clasificación propia de Eustat). Estas estimaciones irán acompañadas de las correspondientes estimaciones de los coeficientes de variación.

En cuanto a nivel de precisión mínimo exigido a estas estimaciones, se ha optado por no ofrecer aquellos totales cuyo coeficiente de variación superen el 0.20. Tampoco se ofrecerán aquellas estimaciones susceptibles de violar el secreto estadístico.

Debido a la relativa dificultad metodológica de los estimadores de áreas pequeñas, se considera necesario acompañar las estimaciones de un documento metodológico exhaustivo en el que se formulen de forma explícita tanto los modelos utilizados como los supuestos realizados.

Capítulo
6

Conclusiones

El subproyecto de estimación en áreas pequeñas en la Encuesta Industrial se ha ejecutado dentro de los plazos programados. La publicación de estimaciones comarcales referentes al año 2003 incrementará considerablemente la calidad de los productos que Eustat ofrece a sus usuarios.

El proyecto de estimación en áreas pequeñas de Eustat sigue su curso programado y actualmente se está realizando un estudio comparativo de una batería de estimadores específicos para datos discretos. El interés se centra en la obtención de datos comarcales para le encuesta de Población en Relación a la Actividad.

Capítulo
6

Bibliografía

[1] BATTESE, G.E. HARTER, R.M. and FULLER, W.A. (1988)

An Error-Components Model for Prediction of Country Crop Areas Using Survey and Satellite Data. Journal of the American Statistical Association, 83, 28-36 .

[2] PFEFFERMAN, D. (2002).

Small Area Estimation – New Developments and Directions. International Statistical Review, 70, 125-143.

[3] PRASAD, N.G.N. and RAO, J.N.K. (1990)

The Estimation of Mean Squared Error of Small Area Estimators. Journal of the American Statistical Association, 85, 163-171

[4] RAO, J.N.K. (2003)

Small Area Estimation. Wiley Series in Survey Methodology

[5] SÄRNDAL, C.E. HIDIROGLOY, M.A. (1989).

Small Domain Estimation: A Conditional Analysis. Journal of the American Statistical Association, 84, 266-275