$$fs_{1ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{\left| x_{ij} - \tilde{x}_{ij} \right|}{\tilde{x}_{ij}}$$

$$fs_{3ij} = \frac{w_i \left(\frac{\tilde{x}_{ij}}{y_{ij}} \right)}{\left(\frac{\tilde{X}_j}{Y_j} \right)} \times \frac{\left| \frac{x_{ij}}{y_{ij}} - \left(\frac{\tilde{x}_{ij}}{y_{ij}} \right) \right|}{\left(\frac{\tilde{x}_{ij}}{y_{ij}} \right)}$$

DEPURACIÓN SELECTIVA DE DATOS

2014

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA
www.eustat.eus

Organismo Autónomo del
EUSKO JAURLARITZA
GOBIERNO VASCO

Elaboración:
EUSTAT
Euskal Estatistika Erakundea
Instituto Vasco de Estadística

Edición:
EUSTAT
Euskal Estatistila Erakundea
Instituto Vasco de Estadística
Donostia-San Sebastián 1
01010 Vitoria-Gasteiz

Primera Edición
I/2015

Impresión y Encuadernación:
Servicio de Imprenta y Reprografía del GobiernoVasco-Eusko Jaurlaritza

ISBN: 978-84-7749-480-5

Depósito Legal: VI-792/2014

DEPURACIÓN SELECTIVA

Imanol Montoya Arroniz

imanolmontoya@gmail.com

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.eus
www.eustat.eus

Presentación

Disponer de métodos eficientes de depuración es fundamental para los organismos estadísticos ya que la depuración de los datos es una de las partes que más tiempo lleva y que resulta más cara en el proceso de mejorar la calidad de los datos.

El objetivo es el estudio y aplicación de las diferentes técnicas de depuración selectiva de bases de datos. La depuración selectiva sirve para depurar aquellos errores cuya corrección tiene una influencia significativa en los resultados a publicar, reduciendo por tanto costes y plazos de entrega.

Este documento consta de varios capítulos en los que, en primer lugar, se desarrolla la metodología, se realiza un estudio de simulación para evaluar dicha metodología y se describen las macros SAS que se han creado para realizar depuración selectiva de bases de datos. Posteriormente se presenta un ejemplo con datos reales, concretamente la nueva operación Estadística de Servicios de Eustat, donde se ha aplicado la metodología propuesta para la depuración selectiva de bases de datos. Finalmente, se muestran algunas conclusiones acerca de la eficacia y utilidad de esta metodología.

Vitoria-Gasteiz, Diciembre 2014

JOSU IRADI ARRIETA

Director General

Índice

. . . / . . .

Capítulo

1

1. Introducción

El contenido recogido en este Cuaderno Técnico, es fruto del trabajo realizado durante el disfrute de la beca de formación e investigación en metodologías estadístico-matemáticas, para el tema de depuración selectiva de bases de datos, concedida en el año 2012 por el Instituto Vasco de Estadística / Euskal Estatistika Erakundea.

El presente documento está dividido en los siguientes capítulos:

En el segundo capítulo se realiza una introducción y se mencionan los objetivos que han marcado la elaboración de este cuaderno técnico.

En el tercer capítulo se desarrolla la metodología de microselección, definiendo la función "score", se muestran diferentes tipos de funciones, qué estrategias existen para construirlas, cómo se pueden combinar en una función "score" global y por último se fija un umbral.

En el cuarto capítulo se desarrolla la metodología de macroselección, diferenciando el método del agregado y el método de la distribución.

En el quinto capítulo se muestra un estudio de simulación para estudiar la metodología previamente desarrollada. Se explica cómo se ha creado la base de datos simulada, cómo se han calculado las funciones "score" y se muestran los resultados obtenidos.

El sexto capítulo contiene una breve explicación acerca de las macros SAS que se han preparado para la depuración selectiva de bases de datos.

En el séptimo capítulo, se presenta un ejemplo real, concretamente la nueva operación Estadística de Servicios del Instituto Vasco de Estadística, donde se ha aplicado la metodología propuesta para la depuración selectiva de bases de datos.

Finalmente, se muestran algunas conclusiones acerca de la eficacia y utilidad de esta metodología.

Quiero agradecer el apoyo a todos los componentes del Área de Metodología, Innovación e I+D y, en general, la amabilidad de todo el personal de Eustat.

PALABRAS CLAVE: Depuración selectiva, función "score" local, función "score" global, microselección, macroselección.

Capítulo

2

2. Introducción a la depuración selectiva

Una de las partes que más tiempo lleva y que resulta más cara en el proceso de mejorar la calidad de los datos es la depuración manual o interactiva de los datos. En el pasado se solían depurar todos los registros manualmente, con el consiguiente coste en personal y de tiempo. En las últimas décadas se ha investigado el efecto que tiene esta depuración manual de los datos y se ha mostrado que el número de registros a depurar manualmente puede reducirse en gran medida ya que, para muchos registros, la depuración manual tiene una influencia insignificante en los estimadores de los principales parámetros de interés.

Estudios como los de (Granquist, 1995), (Granquist and Kovar, 1997) y (Hoogland, 2000), entre otros, mostraron que generalmente no es necesario corregir todos los errores para obtener una cifra fiable del parámetro de interés. Basta con corregir aquellos errores con mayor influencia.

El siguiente gráfico muestra el descenso en la influencia que tiene el corregir sucesivamente los errores menos importantes sobre la estimación del parámetro de interés, en este caso el número de empleados (Hoogland, 2000).

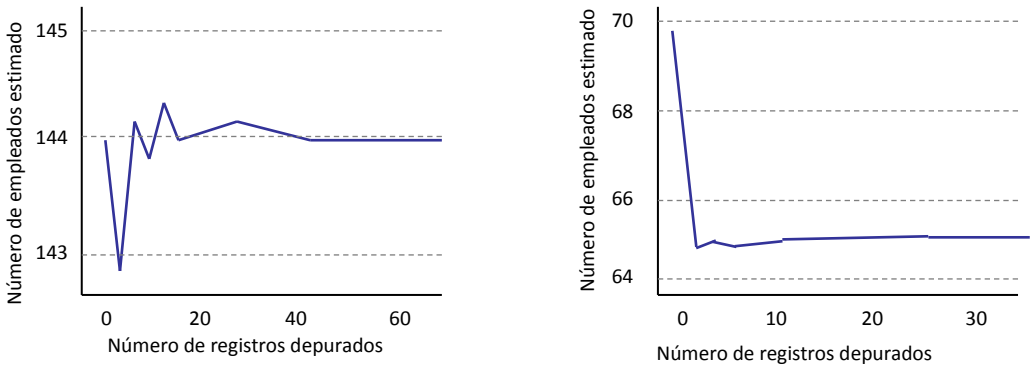


Figura 1. Número de empleados estimados en función del número de registros depurados, los cuales han sido ordenados dependiendo de su influencia en el estimador final. El gráfico de la izquierda se basa en registros de empresas de la construcción, mientras que el de la derecha está basado en empresas de ingeniería civil. (Hoogland, 2000).

La depuración selectiva es aquella estrategia en la que sólo se depuran aquellos errores cuya corrección tiene una influencia significativa en los resultados a publicar, reduciendo por tanto costes y plazos de entrega.

Existen diferentes métodos que permiten seleccionar los registros a depurar en una base de datos. Cuando se aplican en las primeras etapas de la recogida de datos, sin que sea preciso que se haya completado la recogida de datos, se les conoce como métodos de microselección. Estos métodos se aplican en general individualmente a cada registro y se basan en datos de periodos anteriores o en estimaciones de subgrupos homogéneos. Por otro lado, los métodos de macroselección están pensados para ser usados cuando se dispone de prácticamente todos los datos. Estos métodos usan la información de todos los datos disponibles para detectar valores influyentes.

En este cuaderno técnico se describen diferentes métodos que permiten seleccionar los registros más influyentes para depurar. Se ha basado principalmente en el capítulo sexto "*Selective Editing*" del libro "*Handbook of Statistical Data Editing and Imputation"* (de Wall, Pannekoek and Scholtus, 2011), en los cuadernos técnicos publicados por el Instituto Nacional de Estadística de Holanda (Hoogland, van der Loo, Pannekoek and Scholtus, 2011) y (de Wall, 2008) y en las recomendaciones de los proyectos europeos (EUREDIT Project, 2004) y (EDIMBUS, 2007).

Capítulo

3

3. Microselección

La idea principal de la microselección es poder seleccionar aquellos registros a depurar sin que sea preciso haber terminado con la recogida de datos.

En este capítulo se mostrará qué es la función “score", las formas más frecuentes de construirla, otras estrategias para construir la función, como combinarla a la hora de calcular un valor global y cómo determinar el valor umbral que permitirá seleccionar los registros a depurar.

La función “score"

La función “score” es el principal instrumento que se utiliza en el proceso de microselección a la hora de depurar los registros. Esta función asigna una puntación, un score, a cada registro para cada variable analizada. Dicha puntuación da una indicación del efecto esperado sobre el parámetro a estimar en caso de ser depurado. Registros con una puntuación alta serán los que primero sean seleccionados para depurar.

Se conoce como función “score” local a aquella función que mide la influencia de depurar una variable concreta de un registro. Esta función “score” local suele tener dos componentes: el riesgo y la influencia. El riesgo recoge el tamaño y la probabilidad de un error potencial, mientras que la influencia recoge la aportación de ese registro en la estimación del parámetro de estudio. Las puntuaciones locales se definen como el producto de estos dos componentes,

$$s_{ij} = F_{ij} \times R_{ij} = influencia_{ij} \times riesgo_{ij}$$

donde s_{ij} es la función “score” para el registro *i* en la variable *j*. El componente de influencia se mide en general como la contribución relativa del valor anticipado o esperado sobre el estimador total. El componente de riesgo se mide generalmente comparando el valor crudo con respecto a un valor anticipado o esperado. Desviaciones pequeñas entre ambos valores implican que no hay razón para suponer que haya un error, mientras que desviaciones grandes son una indicación de puede que haya un error.

La puntuación global es una función que combina las puntuaciones locales para crear una medida para todo el registro.

$$S_i = f(s_{i1},...,s_{iJ})$$

Los métodos de microselección se aplican sin que haya necesidad de que se haya terminado la recogida de datos. Una vez que un registro está disponible se obtiene su

puntuación global y se compara a un valor umbral, previamente determinado. Si la puntuación sobrepasa dicho umbral, entonces el registro será designado como no plausible. Estos registros serán los que entren por la rama de registros que tienen que ser depurados,

Formalmente esta selección se basa en el indicador de plausibilidad definido como:

$$Indicador\ \ plausibilidad_i = \begin{cases} 1 \quad (plausible) & si\ S_i \le C, \\ 0\ (no\ \ plausible) & en\ otro\ caso \end{cases}$$

siendo C el valor umbral.

La estrategia de microselección se puede resumir en los siguientes tres pasos:

- Calcular las funciones "score" locales para las variables principales de interés, usando como referencia valores anticipados o esperados basados en datos de periodos anteriores o en subgrupos homogéneos.
- Determinar una función que combine estas puntuaciones locales en una global.
- Determinar el valor umbral para los valores globales que seleccione los registros a depurar.

Tipos de funciones "score"

o **Funciones "score" básicas para totales**

Una función "score" deberá cuantificar el efecto de depurar el registro en el estimador de interés. Sea entonces x_{ij} el valor de la variable x_j en el registro i. Si nuestro estimador de interés es el total, éste puede definirse como:

$$\hat{X}_j = \sum_{i \in D} w_i \hat{x}_{ij}$$

donde D son el conjunto de datos y i los registros. Los pesos w_i corrigen por la probabilidad desigual de inclusión y/o la no respuesta. Los $\hat{x}_{ij}$ son los datos una vez depurados. Esto implica que ciertos registros de los datos crudos, x_{ij}, han pasado por el proceso de depuración y han sido corregidos. Para la mayoría de los registros, x_{ij}, se considerará correcta e igual a $\hat{x}_{ij}$. Por lo tanto, el efecto (aditivo) sobre el total de depurar un único registro puede definirse como la diferencia entre el total estimado

con o sin el registro depurado i. El total estimado sin depurar el registro i es $\hat{X}_j - w_i(\hat{x}_{ij} - x_{ij}) = \hat{X}_j^{(-i)}$, y por tanto la diferencia se puede expresar como:

$$d_i(\hat{X}_j) = \hat{X}_j^{(-i)} - \hat{X}_j = w_i(\hat{x}_{ij} - x_{ij})$$

La diferencia $d_i(\hat{X}_j)$ depende de un valor corregido desconocido $\hat{x}_{ij}$ y por lo tanto no puede ser calculada. Una función "score" se basa en una aproximación a este valor desconocido $\hat{x}_{ij}$, $\tilde{x}_{ij}$, conocido como valor esperado. Normalmente estos valores esperados son:

- Valores depurados del mismo registro de periodos anteriores en la misma encuesta, multiplicado por una estimación de la evolución entre los dos periodos de tiempo.
- El valor de una variable similar del mismo registro pero obtenido de otra fuente de datos diferente.
- La media o mediana de la variable de interés de un subgrupo homogéneo de registros similares de un periodo anterior.

La diferencia $d_i(\hat{X}_j)$ depende también de unos pesos w_i. Esto es así ya que estos pesos corrigen por la probabilidad desigual de inclusión, pero también la no respuesta, algo que se desconoce hasta la recogida final de los datos. La aproximación que se utiliza en estos casos es usa los "pesos del diseño", v_i que corrigen únicamente por la probabilidad desigual de inclusión.

Usando estas aproximaciones, el efecto de depurar el registro i puede cuantificarse por la función "score":

$$s_{ij} = v_i\left|x_{ij} - \tilde{x}_{ij}\right| = v_i\tilde{x}_{ij} \times \frac{\left|x_{ij} - \tilde{x}_{ij}\right|}{\tilde{x}_{ij}} = F_{ij} \times R_{ij} = influencia_{ij} \times riesgo_{ij}$$

Esta función "score", por lo tanto, puede entenderse como el producto entre un factor de influencia y otro de riesgo. El factor de riesgo es una medida relativa de la diferencia entre el valor crudo y el esperado $R_{ij} = \left|x_{ij} - \tilde{x}_{ij}\right| / \tilde{x}_{ij}$. Grandes diferencias indican que el valor puede ser erróneo. El factor de influencia, $v_i\tilde{x}_{ij}$, es la contribución del registro al total estimado.

Si se multiplica el riesgo por la influencia da una medida del efecto que tendría depurar el registro sobre el total estimado. Valores grandes indicarán que el registro puede contener un error influyente y que podría merecer la pena revisarlo. Valores pequeños por el contrario, indican que los registros podrían no contener errores influyentes y que, por lo tanto, no es del todo necesario depurarlos minuciosamente.

Para variables no negativas, como la mayoría de las encuestas económicas, el factor de riesgo se puede basar también en el ratio entre el valor crudo y el valor esperado, en vez de la diferencia absoluta entre estos valores.

$$(x_{ij} - \tilde{x}_{ij}) / \tilde{x}_{ij} = \frac{x_{ij}}{\tilde{x}_{ij}} - 1$$

De esta forma el riesgo se expresa como un ratio entre el valor crudo y el valor esperado, y se le añade -1 para asegurar que el riesgo es cero cuando los dos valores son iguales. De todas formas, esta expresión todavía no recoge que diferencias grandes y pequeñas en el ratio indican desviaciones con el valor esperado. Para corregir por esto, se define la siguiente función de riesgo basada en el ratio:

$$R_{ij} = \max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1$$

Esta función asegura que incrementos multiplicativos de igual cuantía, bien hacia arriba o hacia abajo, darán la misma puntuación. Multiplicando este riesgo por el factor de influencia da una función "score" alternativa a la aditiva:

$$s_{ij} = v_i \tilde{x}_{ij} \times (\max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1)$$

Por último, normalmente se utiliza una versión escalada de la función "score" reemplazando el factor de influencia F_{ij} por la influencia relativa $F_{ij} / \sum_i F_{ij}$, donde

$$\sum_i F_{ij} = \sum_i v_i \tilde{x}_{ij} = \tilde{X}_j$$

Por lo que el valor escalado obtenido es el valor original dividido por una estimación del total basado en valores esperados. El escalar el valor permite que dicho valor sea independiente del tamaño y unidad de la variable estudiada. Esto resulta útil cuando se van a combinar varias funciones "scores" para generar un valor global.

Resumiendo, dos funciones "scores" locales escaladas para totales, una aditiva y otra multiplicativa, son, respectivamente:

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{\left|x_{ij} - \tilde{x}_{ij}\right|}{\tilde{x}_{ij}} \quad \text{y} \quad s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times (\max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1)$$

- **Modelos para valores esperados**

En general, un valor esperado es una función de variables auxiliares y coeficientes:

$$\tilde{x}_{ij} = f(\hat{\mu}_1,...\hat{\mu}_K, z_{i1},...z_{iK})$$

Estas variables auxiliares se suelen obtener de la base de datos actual o, lo más habitual, de registros o encuestas anteriores que ya han sido depuradas.

Un valor esperado que se basa en variables auxiliares puede ser el valor estimado de la media o mediana de la variable de interés en un subgrupo en concreto. Por ejemplo, en registros económicos un subgrupo lo podría definir el tipo de industria y su tamaño.

Cuando hay alguna variable auxiliar muy correlacionada con la variable de interés, se suele dividir la variable de interés entre la variable auxiliar y se compara este ratio con el valor esperado para este ratio. Por ejemplo suponiendo que el número de empleados es nuestra variable auxiliar y la facturación la variable de interés, el ratio sería la facturación por empleado. La facturación puede ser muy variable entre diferentes establecimientos, incluso en el mismo tipo de industria, mientras que el ratio por empleado suele tener mucha menos variabilidad.

Cuando se usa el ratio entre dos variables en las funciones "score", se sustituye x_{ij} y $\tilde{x}_{ij}$ en el factor de riesgo por el valor crudo del ratio y por su esperado, $\frac{x_{ij}}{y_{ij}}$ y $\left(\frac{\tilde{x}_{ij}}{y_{ij}}\right)$.
De nuevo, el valor esperado para el ratio puede ser la media o mediana en un periodo anterior, a poder ser perteneciente a un subgrupo homogéneo.

En general, los modelos que se usan en la práctica para los valores esperados no suelen ser muy sofisticados. Por lo que sus predicciones no suelen ser del todo precisas. Aún así, estas predicciones suelen servir ya que el objetivo de la microselección es seleccionar correctamente aquellos registros que enviar a depurar, y no obtener una predicción precisa de los registros (Lawrence and McKenzie, 2000).

- **Funciones "score" con datos longitudinales**

En encuestas o registros que son recogidos cada cierto periodo de tiempo, es habitual utilizar los valores de los periodos anteriores como variables auxiliares. La siguiente fórmula muestra el componente de riesgo que usa como valores esperados valores de periodos anteriores, basado en el ratio, y fue propuesto por (Hidiroglou and Berthelot, 1986):

$$R_{ij} = \max\left(\frac{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}}\right)}, \frac{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}}\right)}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}\right) - 1$$

Donde $x_{ij,t}$ es el valor de la variable de interés x_j para el registro *i* en el periodo actual *t* y $\hat{x}_{ij,t-1}$ el valor correspondiente a la misma unidad en el periodo anterior una vez depurado. Como valor esperado para el cambio proponen la mediana de los cambios de todos los registros, aunque para ello es necesario tener recogidos todos los datos. Una alternativa para este caso es usar la mediana de los cambios de periodos anteriores, por ejemplo entre *t* -2 y *t* – 1, pero sólo cuando se asume que este cambio será similar al de *t* – 1 y *t*.

A la hora de calcular la influencia, también se puede tener en cuenta la información del periodo anterior:

$$F_{ij,t} = \left[\max\left(x_{ij,t}, \hat{x}_{ij,t-1}\right)\right]^c$$

Con $0 \le c \le 1$. Esta formulación hace que *c* sirva para controlar la importancia del parámetro de influencia. Por ejemplo, un estudio (Latouche and Berthelot, 1992) estimó un valor de *c* de 0,5 como razonable en sus datos. Por otro lado, la función máximo asegura que un error en $x_{ij,t}$, aun cuando se trate de una infraestimación, tendrá una influencia como mínimo igual al de su versión depurada del periodo anterior $\hat{x}_{ij,t-1}$.

Otras estrategias para construir la función "score"

Las diferentes maneras que se han visto para calcular las funciones "score" se basan en la desviación entre el valor crudo y el valor esperado. Este tipo de funciones son las más utilizadas en los institutos nacionales de estadística. Aún así, se han propuesto otro tipo de estrategias, aunque todavía no tengan tanta aceptación como la estrategia "tradicional".

- **Modelos paramétricos para datos con errores**

Una alternativa es especificar un modelo paramétrico que tiene en cuenta los posibles errores en los datos. Este modelo asume que los datos con errores y los datos sin errores provienen de distribuciones diferentes. Esta estrategia ha sido propuesta por (Ghosh-Dastidar and Schafer, 2006), (Di Zio, Guarnera and Luzi, 2008) y (Bellisai et al., 2009). Estos autores asumen que los datos correctos provienen de una distribución normal con media μ y varianza σ^2 y que los datos incorrectos provienen de una

distribución normal con la misma media pero con una varianza inflada por un factor *c* > 1. Estos supuestos da un modelo normal contaminado, el cual tiene como función de densidad:

$$f_x = \pi N(\mu, \sigma^2) + (1-\pi) N(\mu, c\sigma^2)$$

siendo la probabilidad π la proporción de datos sin errores. Usando este modelo, se puede estimar la probabilidad condicional, $\hat{\pi}_i$, de que un registro esté libre de error, dado su valor observado: $\Pr(x_i = x_i^* \mid x_i)$. Esta probabilidad, condicional en los datos observados, se le conoce como probabilidad a posteriori. Valores inferiores a un punto de corte apropiado, se consideraran como valores atípicos y serán enviados a depurar.

El modelo anterior se puede extender permitiendo la presencia de valores perdidos, transformando logarítmicamente la variable para hacerla más simétrica y que el supuesto de normalidad sea más realista. También se puede extender usando covariables que permiten que el valor de la media μ varíe.

- **Estrategia asociada a los "edits" de validación**

Otra estrategia propuesta por (Hedlin, 2003) es valorar hasta qué punto un registro ha fallado los "edits" de validación, esto es, cuántos "edits" no cumple y por cuánto no los cumple. La idea de esta estrategia es que los errores influyentes violarán varios de los "edits" o que el mismo fallo será de una cuantía considerable. Al final del estudio, (Hedlin, 2003) mostró que la estrategia de usar la función "score" daba mejores resultados que la estrategia asociada a los "edits" de validación.

- **Estrategia del modelo de predicción**

Esta estrategia propone construir un modelo que relaciona la presencia y tamaño de errores influyentes en la variable de análisis con otras variables predictoras del mismo registro. Esta estrategia precisa de unos datos de entrenamiento, que contienen tanto los datos crudos originales como los datos depurados. Usando estos datos de entrenamiento, se puede calcular la influencia de depurar cada registro en la estimación total.

Una vez se tenga la influencia de cada registro, se puede predecir la "probabilidad de error" π clasificando cada registro en una variable de, por ejemplo, tal y como hizo (Van Lancen, 2002) 6 categorías: la primera categoría es aquella en la que los registros no contenían error, las otras conteniendo un 20% de los registros con error, siendo la última clase aquella con los errores más influyentes. A cada clase se le asigna una probabilidad π : 0; 0,2; 0,4; 0,6; 0,8; y 1. Para predecir esta probabilidad se puede usar un modelo de regresión logística incluyendo variables predictoras y usando los datos de entrenamiento. Una vez calculados los parámetros del modelo, éstos se utilizan en los datos actuales y se estima para cada registro su probabilidad de contener un error influyente.

Esta estrategia no se ha mostrado superior a la estrategia basada en el cálculo de funciones "score" (de Wall, Pannekoek and Scholtus, 2011).

Función "score" global

Para poder seleccionar un registro entero y así depurarlo, se precisa de un valor que combine la información de las diferentes funciones "score". Este valor se conoce como puntuación global o "score" global. Esta puntuación tiene que reflejar la importancia de depurar el registro por completo. Para poder combinar las diferentes puntuaciones locales, es importante que las funciones "scores" locales estén medidas en escalas comparables. Para ello, lo más habitual es escalar estos valores locales dividiéndolos por su total o su total esperado.

Las opciones más comunes de combinar las funciones "score" locales, previamente escaladas, son:

- La suma de las funciones "score" locales (Latouche and Berthelot, 1992):

$$S_i = \sum_{j=1}^{J} s_{ij}$$

- El máximo de las funciones "score" locales (Lawrence and McKenzie, 2000):

$$S_i = \max\left(s_{ij}\right)$$

- Una propuesta que abarca las dos anteriores (Hedlin, 2008):

$$S_i^{(\alpha)} = \left(\sum_{j=1}^{J} s_{ij} \right)^{1/\alpha}$$

Donde $S_i^{(\alpha)}$ es el valor global en función del parámetro α, s_{ij} es el valor de la función "score" local *j*-ésima, *J* es el número de valores locales.

La primera manera de combinar las funciones locales tiene como desventaja que registros con muchas desviaciones pero moderadas, tendrán prioridad sobre registros con pocas pero importantes. En el segundo caso, tiene como ventaja con respecto al caso anterior que aquellos registros donde se hayan dado desviaciones importantes serán priorizados. Aún así, esta opción no será capaz de discriminar registros con una única puntuación local grande respecto a otros con muchas puntuaciones locales grandes. En la última opción, es α el que determina la influencia de los valores locales en el valor global. Un valor $\alpha = 1$ implica la primera opción, la suma de funciones "score". Un valor de α próximo a ∞ da como resultado la segunda opción, el máximo de las funciones "score" locales.

Otra opción es seleccionar el peso específico que tendrá cada variable en el valor global dependiendo de la importancia que se quiera dar. Este peso podrá ser asignado por expertos y variar, por ejemplo, entre 0, 1, 10 y 100.

Fijar un umbral

El objetivo final de una función "score" global es seleccionar los registros que posteriormente hay que depurar. Si la depuración puede esperar hasta que todos los datos han sido recogidos, se podría dejar de depurar cuando los parámetros de interés ya no cambien substancialmente. Normalmente esta forma de depurar conlleva unos plazos de tiempo muy largos si la cantidad de datos y variables es grande. Para poder empezar la depuración de datos en la fase de recogida, es necesario tomar una decisión, basado en el "score" de cada registro, sin necesidad de compararlo con el de otros registros. Por ello se fija un umbral por el cual si la función "score" global de un registro supera dicho valor, este registro deberá ser enviado a depurar.

Lo habitual para determinar dicho umbral es hacer un estudio de simulación en el que se investiga el efecto de diferentes valores umbrales, esto es, el efecto que tiene depurar más o menos registros sobre los parámetros de interés. Este estudio de simulación utiliza datos originales crudos y estos mismos datos depurados manualmente.

Para el estudio de simulación, los registros son ordenados en función de su valor en la función "score" global. Entonces se seleccionan los primeros *p* % registros para ser depurados manualmente. Esto se hace reemplazando estos registros por los datos de la base depurada. El subconjunto de los *p* % registros depurados se le conoce como E_p. Estos pasos se repiten para un rango de valores de *p*. Entonces, se estima el parámetro de interés basado en esta nueva base de datos con *p* % registros depurados y se compara con el parámetro estimado con la base de datos completamente depurada. El valor absoluto de la diferencia relativa entre estos estimadores se le denomina pseudo-sesgo absoluto:

$$AB_j(p) = \frac{1}{\hat{X}_j}\left|\sum_{i \notin E_p} w_i\left(x_{ij} - \hat{x}_{ij}\right)\right|$$

A este valor se le conoce como pseudo-sesgo absoluto ya que si se depuráse sólo errores, se trataría del verdadero sesgo que quedaría por no depurar todos los registros. Pero como no es seguro que los datos depurados sean los datos correctos, este sesgo es una aproximación al verdadero, por lo tanto un pseudo-sesgo.

El pseudo-sesgo de depurar el *p* % de registros, se puede interpretar también como un estimador de la ganancia en la precisión del estimador si se depuráse el 1 – *p* % restante de registros. Por lo que, si se calcula el pseudo-sesgo para un rango de valores de *p*, se podrá obtener una idea de la mejora en la precisión en función de *p*. En cierto punto de *p* se podría decidir que no merece la pena seguir depurando registros ya que no hay apenas mejoría en la precisión del parámetro de interés.

Por último, el estudio de simulación es una manera de poder comprobar la efectividad del proceso de depuración selectiva. Se puede comprobar si aquellos registros con

valores elevados en la función “score” global tenían en verdad errores influyentes o no, y a la inversa, registros con valores globales pequeños, contenían errores no influyentes.

Capítulo

4

4. Macroselección

La idea principal detrás de la macroselección es seleccionar aquellos registros a depurar una vez que la recogida de datos ha terminado o está prácticamente terminada.

Método del agregado

Una vez que se han recogido todos los datos se calculan los principales agregados para las variables de interés. Si estos agregados se diferencian mucho de lo esperado, basado por ejemplo en datos de periodos anteriores, entonces será necesario revisarlos.

Las razones por las que los agregados pueden diferenciarse de lo esperado son múltiples:

- Puede que haya errores influyentes en los datos.
- Puede que haya habido problemas con los pesos usados en el diseño.
- Puede que haya habido variaciones inesperadas que son reales.

Un ejemplo de una función "score" a nivel macro es:

$$S_j = X_j - \tilde{X}_j$$

donde X_j es el estimador agregado para la variable x_j basado en los datos sin depurar y $\tilde{X}_j$ un valor esperado para este estimador agregado.

También se podría calcular la diferencia relativa entre el agregado y su esperado:

$$S_j = \frac{X_j - \tilde{X}_j}{\tilde{X}_j}$$

En algunos casos, suele ser más eficiente utilizar ratios entre agregados que los propios agregados por separado:

$$S_j = \frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{X_k} \right)$$

O poniéndolo en términos relativos:

$$S_j = \frac{\frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{X_k}\right)}{\left(\frac{\tilde{X}_j}{X_k}\right)}$$

Un ejemplo de este caso puede ser el ratio entre el total de facturación y el total de costes para un tipo de industria, o el total salarial entre el número total de empleados para un tipo de industria.

Una manera por la cual se puede controlar por la varianza del agregado es dividir la diferencia en agregados o en los ratios por su desviación estándar relativa:

$$S_j = \frac{X_j - \tilde{X}_j}{d.e.\left(X_j - \tilde{X}_j\right)} \quad \text{y} \quad S_j = \frac{\frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{X_k}\right)}{d.e.\left(\frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{X_k}\right)\right)}$$

De igual manera que ocurría en la microselección, las diferencias entre los agregados o los ratios se pueden expresar de forma aditiva o multiplicativa.

Un vez que se ha detectado un agregado sospechoso para los datos, será preciso investigar los agregados a un nivel inferior, por ejemplo el agregado para un determinado tipo de industria. Al final de este proceso, habrá que mirar a los propios registros para buscar posibles errores influyentes.

Las principales diferencias que hay entre la microselección y la aplicación de estas técnicas una vez que se tiene los datos recogidos son:

1. Se puede utilizar los datos actuales como fuente de información para los valores esperados. Por ejemplo, se puede utilizar la mediana de un grupo homogéneo de los datos actuales en vez de la de un periodo anterior. Como la base de datos actual todavía no ha sido depurada es importante usar las medianas porque son estimaciones robustas en presencia de valores atípicos.

2. No es necesario usar una aproximación de los pesos con los pesos del diseño v_i ya que se cuenta con los pesos finales w_i a la hora de calcular los estimadores.

3. No es necesario calcular un umbral a priori, ya que las propias puntuaciones de las funciones “score” da un orden por el cual los registros serán depurados, pudiendo controlar cuánto cambia el estimador final. La depuración puede detenerse en el momento en que se considere insignificante la mejoría en la precisión del estimador.

Método de la distribución

El método de la distribución intenta identificar aquellos valores que parecen no ajustarse bien con la distribución observada. Esto se hace a través de herramientas gráficas y de medidas estadísticas. Los valores atípicos, aquellos registros más sospechosos, se comprueban. Si un valor atípico se demuestra que es un valor incorrecto e influyente, se corrige.

El método se aplica a variables cuantitativas y suele asumir cierta normalidad y en su defecto asimetría en los datos. En caso de que esto no se cumpla, lo habitual es aplicar algún tipo de transformación en los datos.

Una medida robusta habitualmente utilizada para detectar valores atípicos se basa en la mediana $x_{ij} - mediana(x_{ij})$. Esto es similar a utilizar una función "score" en la que el valor esperado es la mediana. Para poder comparar desviaciones respecto a la mediana en diferentes grupos se suele estandarizar esta diferencia por la mediana de sus valores absolutos, para los registros de cada grupo.

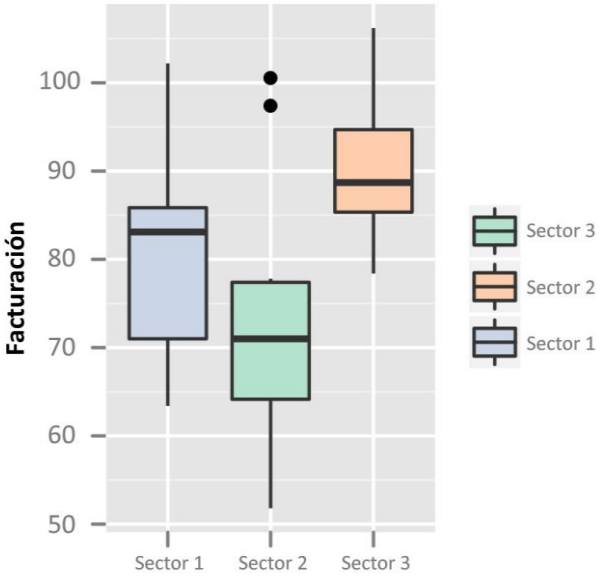
$$o_{ij,c} = \left| x_{ij} - med(x_{ij,c}) \right| / (1{,}4826 \times DAM_c(x_{ij,c}))$$

donde $med(x_{ij,c})$ es la mediana para los registros en el grupo *c* y $DAM_c(x_{ij,c})$es la desviación absoluta de la mediana para estos registros dada por

$$DAM_c(x_{ij,c}) = med(|x_{ij} - med(x_{ij,c})|)$$

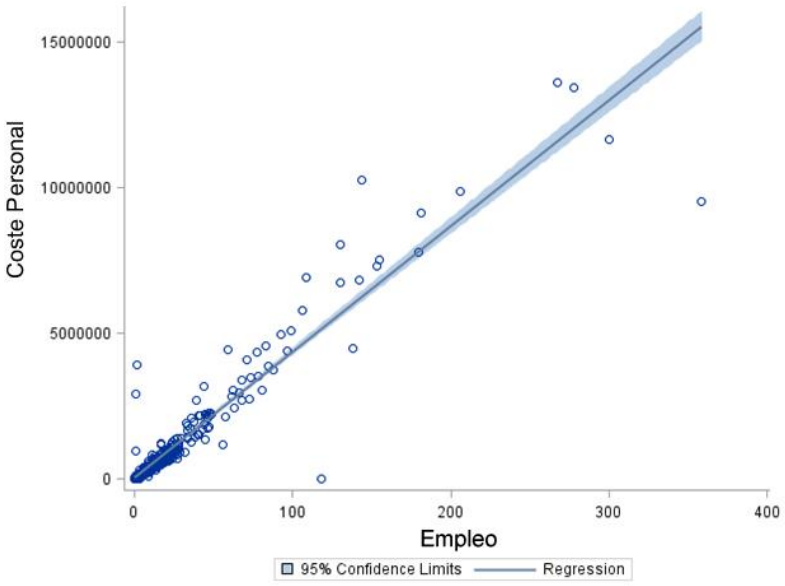
Para una distribución normal $1{,}4826 \times DAM$ es un estimador consistente de la desviación estándar. Otras medidas robustas que se utilizan para la detección de outliers son las medias Winsors o las medias truncadas. También se pueden usar medidas no robustas de dispersión como la varianza o la desviación estándar.

Es habitual usar gráficos como los diagramas de cajas o boxplots para representar las desviaciones respecto a la mediana. Estos gráficos muestran por un lado una caja donde se encuentra el 50% de los registros, unas líneas que normalmente delimitan 1,5 veces el rango intercuartílico respecto al primer y tercer cuartel. Valores más allá de estas líneas se consideran valores atípicos.



En este gráfico representa, para tres sectores, la distribución de la facturación de sus establecimientos en miles de euros. Como se puede ver, en el *Sector 2* existen 2 valores atípicos.

Otra técnica gráfica que se suele utilizar para detectar valores atípicos es el diagrama de puntos, *scatterplot*. A diferencia de los diagramas de cajas, el diagrama de puntos se suele usar cuando se está comparando la distribución de dos variables continuas.



En este gráfico se representa la relación existente entre el empleo y el coste de personal en el sector "maquinaria y equipo".

Capítulo

5

5. Simulación

El objetivo principal de este capítulo es evaluar cómo se comportan diferentes funciones “score” generando varios tipos de errores con una base de datos simulada.

Base de datos simulada

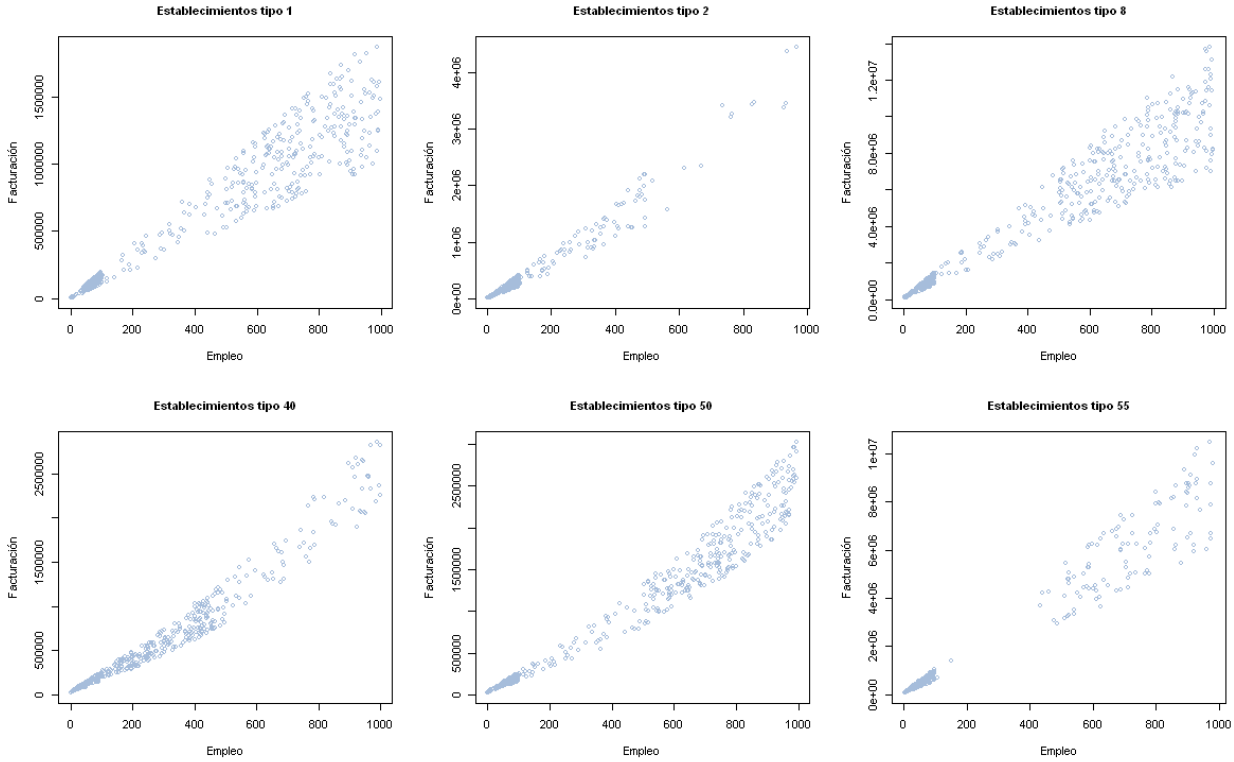
Se ha simulado una base de datos emulando los datos del Registro Mercantil. A la hora de generar la base de datos se ha tenido en cuenta diferentes características para hacerla lo más real posible.

En primer lugar, los datos provienen de empresas que han sido clasificadas según su tipo de actividad CNAE-2009 y según su tamaño en función del número de empleados. Cada empresa ha sido asignada a un territorio histórico con una probabilidad aproximada de estar en Bizkaia de 0,50, 0,33 de estar en Gipuzkoa y 0,17 de estar en Araba/Álava. El número de empresas en la base de datos simulada fue de 36.719.

La facturación de cada empresa se ha generado teniendo en cuenta el tipo de actividad, y el número de empleados. A mayor número de empleados la facturación será mayor, usando funciones lineales y cuadráticas. Se les ha añadido una cierta variabilidad o ruido. Esta variabilidad en la facturación será mayor a medida que aumente el número de empleados del establecimiento.

Como para el cálculo de algunas funciones “score" es preciso conocer el valor del periodo anterior, se han generado también valores previos para la variable facturación y número de empleados. De igual manera a cuando se ha generado la variable facturación, en este caso se le ha añadido un ruido o variabilidad, por lo que el valor del periodo anterior tiene relación con el actual más un ruido aleatorio.

En el siguiente gráfico se muestra la relación entre la facturación y el empleo en empresas de 6 tipos de actividad.



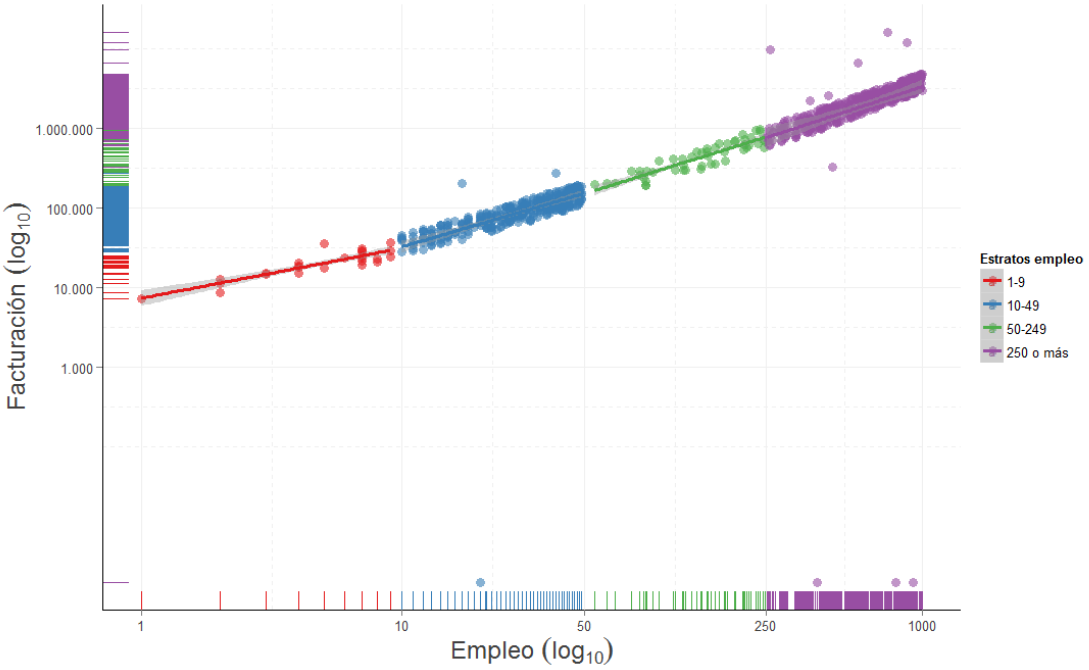
Generación de diferentes tipos de errores

Una vez generada la base de datos original, el siguiente paso fue añadirle errores a la variable facturación. Estos errores deberán ser lo más realistas posibles si se quiere comprobar cómo funcionan las funciones “score" a la hora de detectarlos.

En primer lugar se seleccionaron 500 empresas aleatoriamente de las 36.719 de la base de datos y se les añadió un error aleatorio con una distribución normal con media 0 y con una desviación estándar igual a dos veces su facturación original. En caso de dar un resultado negativo la facturación se igualaba a cero. Esto origina unos errores que pueden multiplicar la facturación hasta 4 o más veces o que pueden hacer que se convierta en 0.

Además, se han añadido errores de unidad. Se han seleccionado 50 empresas aleatoriamente cuya facturación ha sido multiplicada por 1.000 y otras 50 empresas cuya facturación se ha dividido por 1.000.

El siguiente gráfico muestra la relación entre el empleo y la facturación para todas las empresas pertenecientes a un tipo de actividad. Se ha usado la escala logarítmica para poder separar bien cada estrato de empleo y se muestra la línea de regresión lineal por tramos de empleo. Se pueden observar observaciones “anómalas”.



Funciones "score"

Éstas son las diferentes funciones "score" que se han calculado con la base de datos simulada. Para cada empresa se ha calculado su puntuación con cada una de las funciones.

Caso 1: Facturación y esperado de un grupo homogéneo

Aditiva

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{\left| x_{ij} - \tilde{x}_{ij} \right|}{\tilde{x}_{ij}}$$

Multiplicativa

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \left(\max \left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}} \right) - 1 \right)$$

Donde,

x_{ij} → Facturación

$\tilde{x}_{ij}$ → Facturación en un grupo homogéneo (mismo tipo de actividad y tamaño de empleo)

Caso 2: Ratio con el número de empleados y esperado de un grupo homogéneo

Aditiva

$$s_{ij} = \frac{w_i \left(\frac{\tilde{x}_{ij}}{y_{ij}} \right)}{\left(\frac{\tilde{X}_j}{Y_j} \right)} \times \frac{\left| \frac{x_{ij}}{y_{ij}} - \left(\frac{\tilde{x}_{ij}}{y_{ij}} \right) \right|}{\left(\frac{\tilde{x}_{ij}}{y_{ij}} \right)}$$

Multiplicativa

$$s_{ij} = \frac{w_i \left(\frac{\tilde{x}_{ij}}{y_{ij}} \right)}{\left(\frac{\tilde{X}_j}{Y_j} \right)} \times \left(\max \left(\frac{\frac{x_{ij}}{y_{ij}}}{\left(\frac{\tilde{x}_{ij}}{y_{ij}} \right)}, \frac{\left(\frac{\tilde{x}_{ij}}{y_{ij}} \right)}{\frac{x_{ij}}{y_{ij}}} \right) - 1 \right)$$

Donde,

$\frac{x_{ij}}{y_{ij}}$ → Facturación/Número de empleados

$\left(\frac{\tilde{x}_{ij}}{y_{ij}} \right)$ → Facturación/Número de empleados en un grupo homogéneo (mismo tipo de actividad y tamaño de empleo)

Caso 3: Ratio con el periodo anterior y esperado de un grupo homogéneo

Aditiva

$$s_{ij} = \frac{w_i \left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}{\left(\frac{\tilde{X}_{j,t}}{\widehat{X}_{j,t-1}} \right)} \times \frac{\left| \left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}} \right) - \left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right) \right|}{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}$$

Multiplicativa

$$s_{ij} = \frac{w_i \left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}{\left(\frac{\tilde{X}_{j,t}}{\widehat{X}_{j,t-1}} \right)} \times \left(\max \left(\frac{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}} \right)}{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}, \frac{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}} \right)} \right) - 1 \right)$$

Donde,

$\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}$ → Facturación en el periodo t / Facturación depurada en el periodo t - 1

$\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}}\right)$ → Facturación en el periodo t / Facturación depurada en el periodo t - 1 en un grupo homogéneo (mismo tipo de actividad y tamaño de empleo)

Caso 4: Influencia usando la facturación y riesgo con el ratio con el número de empleados.

Aditiva

$$s_{ij} = \frac{w_i \tilde{x}_{ij,t-1}}{\tilde{X}_{j,t-1}} \times \frac{\left| \frac{x_{ij,t}}{y_{ij,t}} - \left(\frac{\tilde{x}_{ij,t-1}}{y_{ij,t-1}} \right) \right|}{\left(\frac{\tilde{x}_{ij,t-1}}{y_{ij,t-1}} \right)}$$

Multiplicativa

$$s_{ij} = \frac{w_i \tilde{x}_{ij,t-1}}{\tilde{X}_{j,t-1}} \times \left(\max \left(\frac{\frac{x_{ij,t}}{y_{ij,t}}}{\left(\frac{\tilde{x}_{ij,t-1}}{y_{ij,t-1}} \right)}, \frac{\left(\frac{\tilde{x}_{ij,t-1}}{y_{ij,t-1}} \right)}{\frac{x_{ij,t}}{y_{ij,t}}} \right) - 1 \right)$$

Donde,

$x_{ij,t}$ → Facturación en el periodo t

$\tilde{x}_{ij,t-1}$ → Facturación en el periodo t-1 depurada

$\frac{x_{ij,t}}{y_{ij,t}}$ → Facturación/Número de empleados en el periodo t

$\left(\frac{\tilde{x}_{ij,t-1}}{y_{ij,t-1}}\right)$ → Facturación/Número de empleados en el periodo t-1 depurada

Caso 5: Facturación y estimado proveniente de la regresión robusta

Aditiva

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{\left| x_{ij} - \tilde{x}_{ij} \right|}{\tilde{x}_{ij}}$$

Multiplicativa

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \left(\max \left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}} \right) - 1 \right)$$

Donde,

x_{ij} → Facturación

$\tilde{x}_{ij}$ → Estimación de la facturación en función del número de empleados en la empresa. Se ha utilizado métodos de regresión robusta para cada estrato.

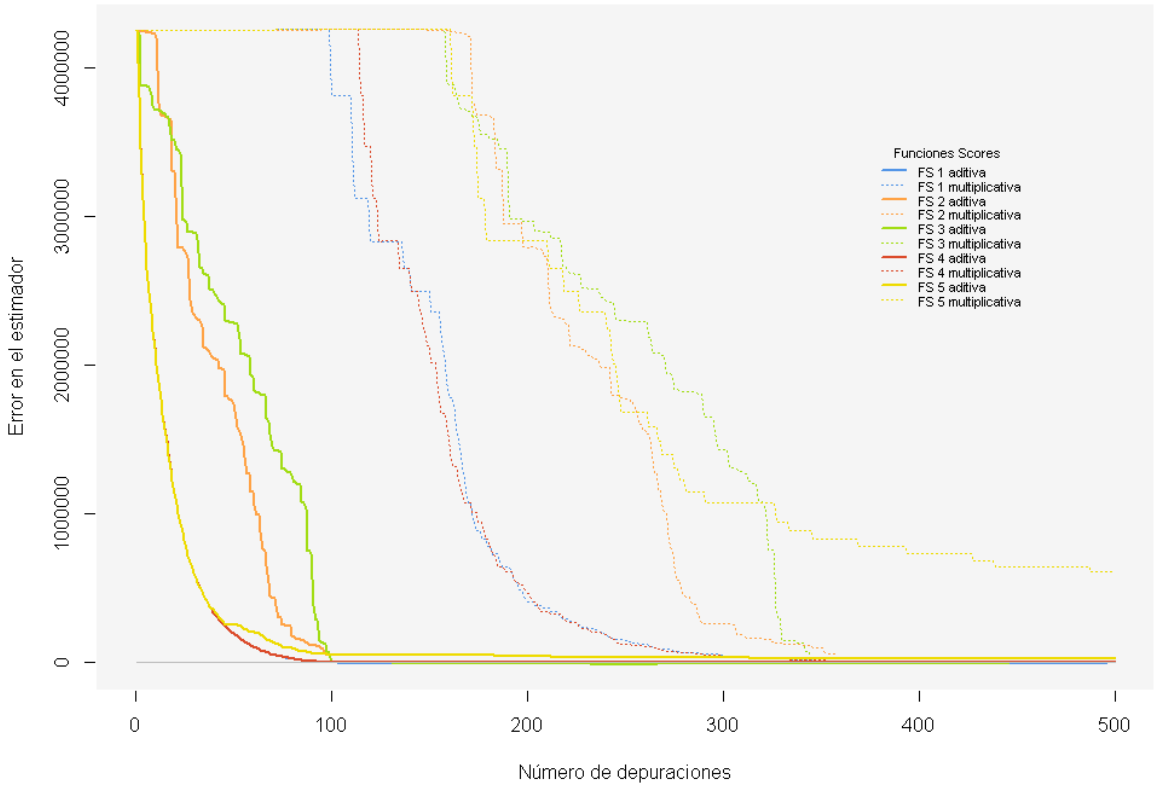
Resultados de la simulación

Una vez calculada para cada empresa su puntuación con la función "score" se han ordenado dichas puntuaciones. Las empresas con una mayor puntuación serán las primeras seleccionadas para depurar.

El siguiente gráfico muestra cómo evoluciona el error del estimador, en nuestro caso la media de la facturación, a medida que se depura empresas. Este error se calcula como la diferencia entre la media de la facturación sin errores y la media de la facturación con errores.

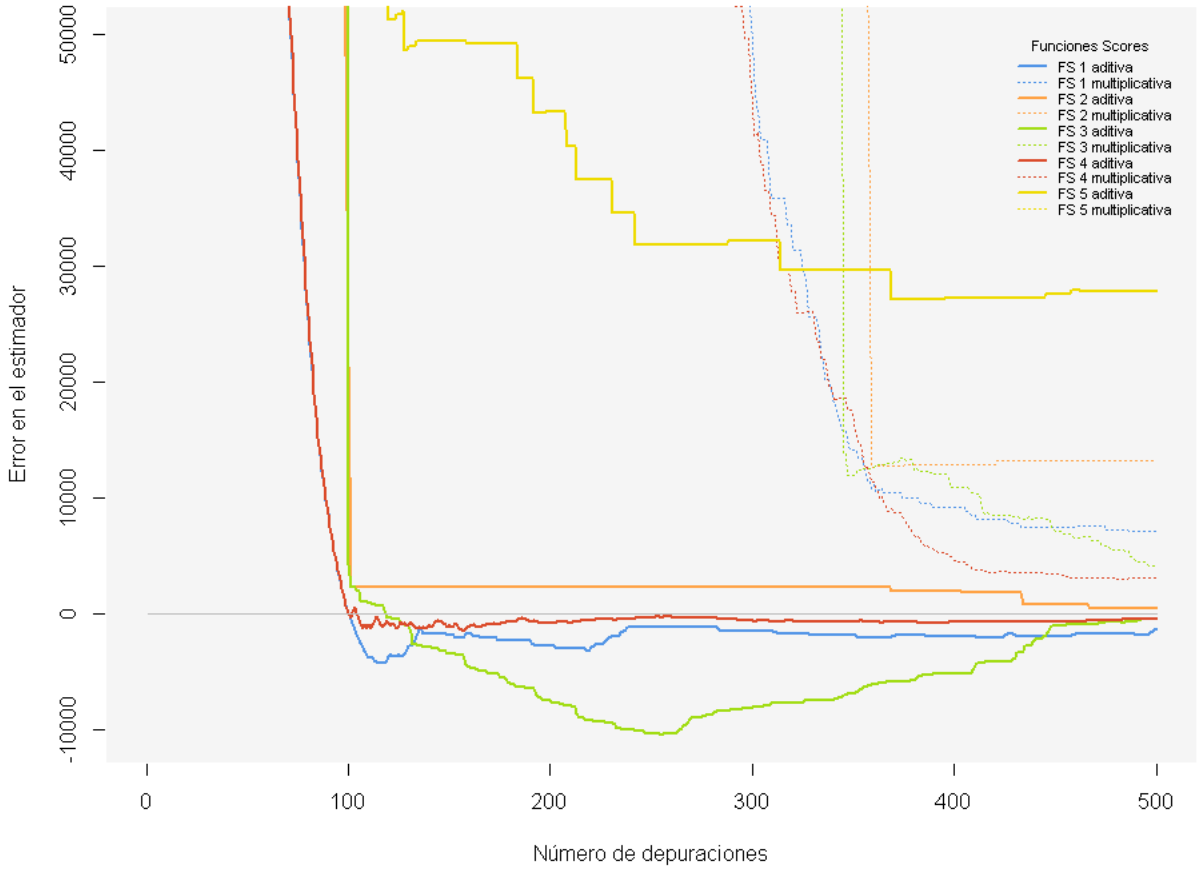
Se puede ver que las funciones "score" que usan la escala multiplicativa en el riesgo, son las menos eficientes ya que tardan más en disminuir el error en el estimador.

Entre las funciones "score" aditivas parece que son mejores la 1, 4. Estas funciones son aquellas funciones cuyo término de "influencia" tiene en cuenta la contribución esperada de la facturación de cada empresa sobre el estimador total.



En el siguiente gráfico se ha hecho un zoom sobre el eje del error en el estimador para poder comparar cómo se comportan las funciones “score” una vez eliminados los errores más importantes en la facturación.

Viendo el gráfico la función “score” 4 aditiva es la que antes se acerca al error nulo y la que más cercana se mantiene a dicho valor en casi todo momento. Por lo tanto, en nuestra simulación la función “score” que mejor se comporta es aquella cuyo término de “influencia” tiene en cuenta la contribución esperada de la facturación de cada empresa sobre el estimador total y cuyo término de “riesgo” se calcula teniendo en cuenta el ratio de la facturación con el número de empleados.



Capítulo

6

6. Macro SAS

A continuación, se presenta las macros de SAS que va a permitir realizar depuración selectiva en una base de datos.

Para más información mirar el "Manual de usuario de la aplicación SAS: Depuración Selectiva".

Macro *FUNCION_SCORE*

La macro de SAS *FUNCION_SCORE* permite calcular para una variable tres funciones "score" en cada registro.

- Función "score" tipo I: Toma como referencia el valor del período anterior y la influencia del registro en el estrato que definamos.

- Función "score" tipo II: Toma como referencia el valor de la mediana del estrato que se defina y la influencia del registro en ese estrato.

- Función "score" tipo III: Toma como referencia el valor estimado a través de regresión robusta en el estrato que se defina y la influencia del registro en ese estrato.

- Función "score" tipo IV: Toma como referencia el valor de la mediana y como divisor el rango intercuartílico del estrato que se defina y la influencia del registro en ese estrato.

Datos de entrada

Se precisa de un dataset SAS que contenga al menos:

- La variable que se quiere depurar en el momento t.
- La variable a depurar en el momento anterior t-1.
- Opcional: Una variable que se asocie con la de depuración para poder hacer regresión.

- Opcional: Variables que identifiquen los estratos donde se calculará bien su mediana o bien la influencia de cada registro para cada función "score".

Sintaxis de la macro

Esta es una breve descripción de los argumentos necesarios:

- dataset = dataset SAS donde realizar la depuración selectiva.

- var = La variable que se quiere depurar en el momento t.
- var_ant = La variable a depurar en el momento anterior t-1.
- var_reg = Una variable que se asocie con la de depuración para poder hacer regresión.
- tipoIestrato = Estrato donde se calculará la influencia del registro para la función "score" tipo I.
- tipoIIestrato = Estrato donde se calculará la mediana y la influencia del registro para la función "score" tipo II.
- tipoIIIestrato = Estrato donde se estimará la regresión robusta y se calculará la influencia del registro para la función "score" tipo III.
- tipoIVestrato = Estrato donde se calculará la mediana, el rango intercuartílico y la influencia del registro para la función "score" tipo IV.
- varpositiva = Se debe definir si la variable sólo toma valores positivos o no para su estimación en la regresión. Por defecto toma el valor "T", que significa que es positiva. Su opuesto sería "F".
- n_estrato = Número de registros mínimos en cada estrato para poder estimar la regresión robusta.
- fscore_1 = Se da la opción de calcular o no la función "score" tipo I. "T"= si y "F" =no.
- fscore_2 = Se da la opción de calcular o no la función "score" tipo II. "T"= si y "F" =no.
- fscore_3 = Se da la opción de calcular o no la función "score" tipo III. "T"= si y "F" =no.
- fscore_4 = Se da la opción de calcular o no la función "score" tipo IV. "T"= si y "F" =no.

Macro *FS_GLOBAL*

La macro de SAS *FS_GLOBAL* permite calcular la función "score" global, *fs_global,* que combina las funciones "score" locales previamente calculadas.

Datos de entrada

Se precisa de un dataset SAS en el que se haya calculado con la macro *FUNCION_SCORE* al menos las funciones "score" locales para una variable.

Sintaxis de la macro

Esta es una breve descripción de los argumentos necesarios:

- dataset = dataset SAS donde realizar la depuración selectiva.

- var = La variable que se quiere depurar en el momento t.
- vartext = Literal de la variable.
- max = Calcula el máximo de las tres funciones “score”. Por defecto toma el valor “T”. Para calcular la suma de las tres funciones “score” bastaría con poner max = “F”.
- w1 = Peso para función “score” tipo I. Por defecto 1.
- w2 = Peso para función “score” tipo II. Por defecto 1.
- w3 = Peso para función “score” tipo III. Por defecto 1.
- w4 = Peso para función “score” tipo IV. Por defecto 1.
- pesos = peso de cada observación en la población. Por defecto 1.

Capítulo

7

7. Implementación práctica en la operación Estadística de Servicios

La nueva operación Estadística de Servicios del Instituto Vasco de Estadística se elabora, por un lado, con datos procedentes de un cuestionario dirigido directamente a establecimientos y, por otro, con datos procedentes de tres registros administrativos: el Registro Mercantil, el Registro de Cooperativas y el Registro de Asociaciones y Fundaciones.

La utilización de información registral ha permitido obtener mejores estimaciones al poder disponer de una mayor cantidad de información. Pero a su vez, el tener que trabajar con un volumen tan grande de información ha supuesto una serie de dificultades de diferente índole. En especial, con el Registro Mercantil, que es la fuente que más datos proporciona; en concreto, en la Estadística de Servicios de 2012 se ha utilizado información proveniente de 16.251 sociedades mercantiles.

Es por ello que se hace necesario trabajar con procedimientos de depuración más eficientes y fiables, como es el caso de la depuración selectiva de bases de datos.

Implementación práctica de la depuración selectiva

Se ha aplicado la macro de depuración selectiva a la base de datos que combina la información procedente del cuestionario dirigido directamente a establecimientos y, por otro, la información procedente de los registros administrativos: Registro Mercantil, Registro de Cooperativas y Registro de Asociaciones y Fundaciones.

Las variables que se han utilizado han sido el Importe Neto de la Cifra de Negocios, el Valor Añadido a Coste de Factores y el Coste de Personal, que se han considerado las principales variables disponibles y que además son las variables que están más correlacionadas con la variable personal ocupado.

- **Depuración selectiva de la variable Importe Neto de la Cifra de Negocios**

A la hora de aplicar la macro se optó por las siguientes opciones. En primer lugar, se consideró que, para esta variable y en este caso, las funciones "score" más relevantes eran la tipo II y III, que buscan, respectivamente, registros que se separan de la mediana del estrato o del valor estimado de la regresión robusta, siendo el empleo la variable de ajuste.

Al no tener datos para cada establecimiento del año anterior, no se calculó la función "score" tipo I que es la que tiene en cuenta la desviación con respecto al periodo anterior. Se optó también por no calcular la función "score" tipo IV que es la que tiene más en cuenta la variabilidad en cada estrato.

El estrato donde se decidió calcular las funciones "score" fue la combinación del código CNAE a dos dígitos junto con su estrato de empleo. Para poder estimar la regresión robusta se decidió un mínimo de 20 establecimientos en cada registro y la combinación de funciones "score" en una global se hizo como el máximo entre ellas.

En la siguiente tabla se muestran, a modo de ejemplo, los primeros 10 establecimientos del output de la macro de depuración.

Tabla 7.1. Depuración selectiva de la variable Importe Neto de la Cifra de Negocios

Establecimiento	Razón	Empleo	CNAE09	Estrato Empleo	Importe Neto de la Cifra de Negocios	Mediana	Estimación	fs global
a1	2.- valor estrato	3	5229	1	3.470.393	25.344		1,530
a2	3.- valor regresión	1	7022	2	20.881.251	105.883	79.446	0,486
a3	2.- valor estrato	2	5221	1	1.244.845	25.344		0,479
a4	3.- valor regresión	1	7112	1	423.372	44.925	45.666	0,178
a5	3.- valor regresión	1	6910	1	308.531	41.729	40.776	0,128
a6	3.- valor regresión	1	7111	1	284.882	44.925	45.666	0,111
a7	3.- valor regresión	1	6910	1	288.304	41.729	40.776	0,110
a8	2.- valor estrato	1	6621	1	170.406	32.962		0,104
a9	3.- valor regresión	2	7022	2	11.482.794	105.883	119.914	0,097
a10	3.- valor regresión	2	7022	2	11.443.789	105.883	119.914	0,096

Como se puede observar en la tabla 7.1., estos establecimientos se separan mucho de la mediana de su estrato o de lo que se esperaría en su estrato con ese número de empleados. El Importe Neto de la Cifra de Negocios en estos establecimientos es muy superior a lo que ocurre en establecimientos del mismo sector y similar empleo, o incluso ajustando por el número de empleados supera ampliamente lo esperado.

En aquellos casos donde el número de establecimientos en el estrato no llegaba a 20 no se estimó ningún valor para el Importe Neto de la Cifra de Negocios.

- **Depuración selectiva de la variable ratio Importe Neto de la Cifra de Negocios por persona**

En este apartado se depurará el ratio del Importe Neto de la Cifra de Negocios entre el número de empleados del establecimiento teniendo en cuenta únicamente la mediana del estrato y dándole pesos iguales a todos los establecimientos.

Tabla 7.2. Depuración selectiva de la variable Importe Neto de la Cifra de Negocios por persona

Establecimiento	Razón	Empleo	CNAE09	Estrato Empleo	Ratio Importe Neto de la Cifra de Negocios por persona	Mediana	fs global
b1	2.- valor estrato	1	7022	2	**20.881.251**	65.193	0,213
b2	2.- valor estrato	3	5229	1	**1.156.798**	25.344	0,062
b3	2.- valor estrato	18	9001	4	**807.432**	45.749	0,031
b4	2.- valor estrato	2	5221	1	**622.423**	25.344	0,017
b5	2.- valor estrato	2	7022	2	**5.741.397**	65.193	0,016
b6	2.- valor estrato	2	7022	2	**5.721.895**	65.193	0,016
b7	2.- valor estrato	1	7112	2	**5.311.000**	70.454	0,014
b8	2.- valor estrato	108	9312	7	**438.467**	40.003	0,014
b9	2.- valor estrato	1	7219	2	**1.305.609**	65.612	0,010
b10	2.- valor estrato	1	6820	2	**4.266.072**	45.036	0,010

Se observa que el Ratio Cifra de Negocios por persona en estos establecimientos es muy diferente a lo que ocurre en su estrato, y por lo tanto se deberían revisar.

- **Depuración selectiva de la variable Valor Añadido a Coste de Factores**

Para depurar esta variable se optó por las mismas opciones que en el caso del Importe Neto de la Cifra de Negocios. En primer lugar, al no tener datos para cada establecimiento del año anterior, no se calculó la función "score" tipo I que es la que tiene en cuenta la desviación con respecto al periodo anterior. También se optó por no calcular la función "score" tipo IV que es la que tiene más en cuenta la variabilidad en cada estrato.

De igual manera, se consideró que para el Valor Añadido las funciones "score" más relevantes eran la tipo II y III, que buscan, respectivamente, registros que se separan de la mediana del estrato o del valor estimado de la regresión robusta, siendo el empleo la variable de ajuste.

El estrato donde se decidió calcular las funciones "score" fue la combinación del código CNAE a dos dígitos junto con su estrato de empleo. Para poder estimar la regresión robusta se decidió un mínimo de 20 establecimientos en cada registro y la combinación de funciones "score" en una global se hizo como el máximo entre ellas.

En la siguiente tabla se muestran los primeros 10 establecimientos del output de la macro de depuración.

Tabla 7.3. Depuración selectiva del Valor Añadido a Coste de Factores.

Establecimiento	Razón	Empleo	CNAE09	Estrato Empleo	Valor Añadido a Coste de Factores	Mediana	Estimación	fs global
c1	3.- valor regresión	1	6820	1	**537.319**	23.443	26.655	3,116
c2	3.- valor regresión	1	7112	1	**925.311**	27.002	25.698	2,466
c3	2.- valor estrato	2	5221	1	**718.272**	17.668		1,428
c4	3.- valor regresión	2	7022	2	**11.021.163**	63.658	76.697	0,271
c5	3.- valor regresión	1	4725	1	**142.580**	8.187	7.633	0,255
c6	3.- valor regresión	1	6910	1	**262.088**	30.000	29.363	0,173
c7	2.- valor estrato	1	9001	1	**285.340**	26.355		0,172
c8	3.- valor regresión	1	6910	1	**251.182**	30.000	29.363	0,158
c9	2.- valor estrato	1	7220	1	**60.703**	20.234		0,142
c10	2.- valor estrato	1	5320	1	**44.390**	30.717		0,117

- **Depuración selectiva del Valor Añadido a Coste de Factores por persona**

En este apartado se depurará el ratio del Valor Añadido entre el número de empleados del establecimiento teniendo en cuenta únicamente la mediana del estrato.

Tabla 7.4. Depuración selectiva del Valor Añadido a Coste de Factores por persona.

Establecimiento	Razón	Empleo	CNAE09	Estrato Empleo	Ratio Valor Añadido a Coste de Factores por persona	Mediana	fs global
d1	2.- valor estrato	2	7022	2	**5.510.582**	39.980	0,066
d2	2.- valor estrato	2	5221	1	**359.136**	16.517	0,043
d3	2.- valor estrato	1	7112	1	**925.311**	26.220	0,041
d4	2.- valor estrato	5	4742	3	**515.179**	9.097	0,034
d5	2.- valor estrato	1	6820	1	**537.319**	22.294	0,030
d6	2.- valor estrato	2	7022	2	**3.419.244**	39.980	0,025
d7	2.- valor estrato	108	9312	7	**407.720**	32.938	0,021
d8	2.- valor estrato	7	4764	3	**382.244**	9.097	0,019
d9	2.- valor estrato	50	5210	6	**842.255**	54.571	0,016
d10	2.- valor estrato	10	6820	4	**580.367**	69.618	0,016

Se observa que el Ratio Valor Añadido por persona en estos establecimientos es muy diferente a lo que ocurre en su estrato, y por lo tanto se revisarán para ver si su valor es correcto o si es necesario depurarlo.

- **Depuración selectiva del Coste Personal por persona**

En este caso se optó directamente por utilizar el Ratio entre el Coste de Personal y el empleo de cada establecimiento.

Tabla 7.5. Depuración selectiva del Coste Personal por persona.

Establecimiento	Razón	Empleo	CNAE09	Estrato Empleo	Ratio Coste Personal por persona	Mediana	fs global
e1	2.- valor estrato	108	9312	7	**340.707**	30.095	0,078
e2	2.- valor estrato	23	9312	5	**340.707**	27.364	0,032
e3	2.- valor estrato	65	9312	6	**144.770**	22.734	0,025
e4	2.- valor estrato	32	9312	5	**267.197**	27.364	0,019
e5	2.- valor estrato	246	8220	7	**80.607**	19.478	0,017
e6	2.- valor estrato	30	9312	5	**218.031**	27.364	0,013
e7	2.- valor estrato	106	9312	7	**144.770**	30.095	0,012
e8	2.- valor estrato	6	9102	3	**79.146**	31.681	0,012
e9	2.- valor estrato	7	4764	3	**340.707**	27.338	0,011
e10	2.- valor estrato	108	9312	7	**340.707**	30.095	0,078

En esta tabla se pueden ver establecimientos donde el Ratio Coste de Personal por empleo supera ampliamente la mediana de su estrato.

Resumen de la implementación de la depuración selectiva

La macro programada en SAS de Depuración Selectiva desarrollada en el Instituto ha servido para depurar la información económica obtenida a través de las distintas fuentes, después de validar dicha información, integrarla y contrastarla con el Directorio de Actividades Económicas de Eustat.

Las variables principales que se han utilizado han sido el Importe Neto de la Cifra de Negocios, el Valor Añadido a Coste de Factores y el Coste de Personal, que se han considerado las principales variables disponibles y que además son las variables que están más correlacionadas con la variable personal ocupado.

El análisis que proporciona la macro ha permitido detectar establecimientos con valores extremos e influyentes dentro de los estratos de elevación (actividad y estrato de empleo) teniendo en cuenta su influencia dentro de dichos estratos. Así, se ha podido valorar la información de una manera eficiente y fiable, especialmente teniendo en cuenta el volumen de los datos con el que se ha trabajado.

Capítulo

8

8. Conclusiones

Por último, en este capítulo se presentan un resumen y las principales conclusiones de este trabajo sobre la depuración selectiva de bases de datos que ha sido realizado durante el disfrute de la beca de formación e investigación en metodologías estadístico-matemáticas.

Resumen y conclusiones sobre la depuración selectiva

Disponer de métodos eficientes de depuración es fundamental para los organismos estadísticos ya que una de las partes que más tiempo lleva y que resulta más cara en el proceso de mejorar la calidad de los datos es la depuración manual o interactiva de los datos.

Se ha demostrado que el número de registros a depurar se puede reducir en gran medida, ya que para muchos registros, la depuración manual tiene una influencia insignificante en los estimadores de los principales parámetros de interés.

En este contexto, se precisa de una estrategia de selección que separe los registros en dos partes: una crítica con aquellos registros que supuestamente contienen errores influyentes y, otra con registros cuya depuración no se espere que cambie los resultados a publicar.

La depuración selectiva es aquella estrategia en la que sólo se depuran aquellos registros cuya corrección tiene una influencia significativa en los resultados a publicar, reduciendo por tanto costes y plazos de entrega.

La función "score" es el principal instrumento de la depuración selectiva. Esta función asigna una puntación a cada registro para cada variable analizada. Dicha puntuación da una indicación del efecto esperado sobre el parámetro a estimar en caso de ser depurado. Registros con una puntuación alta serán los que primero sean seleccionados para depurar.

Además se pueden calcular diferentes tipos de funciones "score" dependiendo de cuál sea el valor de referencia esperado o, también, si se está depurando varias variables conjuntamente. Es por ello que para poder seleccionar un registro entero y así depurarlo, se precisa de un valor que combine la información de las diferentes

funciones “score”. Este valor se conoce como puntuación global o “score” global. Esta puntuación tiene que reflejar la importancia de depurar el registro por completo.

En este trabajo se ha demostrado que la función “score” es un instrumento válido y eficiente para la selección de registros anómalos e influyentes. En primer lugar, y en un marco más teórico, se usó una base de datos simulada en la que se pudo observar el comportamiento de diferentes tipos de funciones “score” a la hora de seleccionar registros a depurar.

Posteriormente, se aplicaron dichas técnicas en un marco real, concretamente en la nueva operación Estadística de Servicios del Instituto Vasco de Estadística. La macro programada en SAS ha servido para depurar la información económica obtenida a través de las distintas fuentes, después de validar dicha información, integrarla y contrastarla con el Directorio de Actividades Económicas de Eustat.

Cabe resaltar, que tanto las técnicas de depuración aquí descritas como las macros de SAS programadas, pueden ser aplicadas y modificadas con relativa facilidad en cualquier tipo de base de datos que precise de ser depurada. Las macros SAS son parametrizables pudiéndose utilizar usando una única variable a depurar o varias variables conjuntamente. Se puede elegir entre cuatro funciones “score” y combinarlas dándoles diferentes pesos a cada una de ellas. El cálculo de cada función “score” se puede realizar también en diferentes estratos. Además, el peso de cada registro puede ser distinto o el mismo, según interese.

Las macros SAS se podrían ofrecer en la web del EUSTAT a aquellas instituciones, institutos de estadística o investigadores interesados en implementarlas.

Por último este cuaderno técnico se ha basado principalmente en la metodología aplicada por el Instituto Nacional de Estadística de Holanda y que ha sido publicada en los cuadernos técnicos (Hoogland, van der Loo, Pannekoek and Scholtus, 2011) y (de Wall, 2008) y en el modelo descrito en el proyecto europeo EDIMBUS (Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys 2007).

Capítulo

9

Bibliografía

BELLISAI, D.; DI ZIO, M.;GUARNERA, U. AND LUZZI, 0. (2009)

A selective editing approach based on contamination models: An application to an ISTAT business survey. Working Paper No. 27, UN/ECE Work Session on Statistical Data Editing, Neuchatel.

DE WAAL, T. (2008)

An overview of statistical data editing. Statistics Netherlands, The Hague/Heerlen.

DE WAAL, T., PANNEKOEK, J. AND SCHOLTUS, S. (2011)

Handbook of statistical data editing and imputation. Wiley.

DI ZIO, M.; GUARNERA, U. AND LUZZI, 0. (2008)

Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data. Working Paper No. 22, UN/ECE Work Session on Statistical Data Editing, Vienna

EUREDIT PROJECT (2004a)

Towards Effective Statistical Editing and Imputation Strategies. Findings of the Euredit Project, Volume 1. Disponible en: http://www.cs.york.ac.uk/euredit/results/results.html

EDIMBUS (2007)

Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. Manual prepared by ISTAT, Statistics Netherlands and SFSO.

GHOSH-DASTIDAR, B. AND SCHAFER, J. L. (2006)

Outlier Detection and Editing Procedures for Continuous Multivariate Data. Journal of Official Statistics 22, pp.487-506.

GRANQUIST, L (1995)

Improving the Traditional Editing Process. Business Survey Methods. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson,M.J. Colledge, and P.S. Kott, eds. John Wiley & Sons, New York, 385–401.

GRANQUIST, L AND KOVAR (1997)

Editing of Survey Data: How Much Is Enough? Survey Measurement and Process Quality. L.E. Lyberg, P. Biemer, M. Collins, E.D. De Leeuw, C. Dippo, , N. Schwartz, and D. Trewin, eds. John Wiley & Sons, New York, pp. 415–435.

HEDLIN, D (2003)

Score Funtions to Reduce Business Survey Editing at the U.K. Office for National Statistics. Journal of Official Statistcs 19, 177-199

HEDLIN, D (2008)

Local and Global Score Functions in Selective Editing. Working Paper No. 31, UN/ECE Work Session on Statistical Data Editing, Vienna.

HIDIROGLOU, M. A. AND BERTHELOT, J. M. (1997)

Statistica Editing and Imputation for Periodic Business Surveys. Survey Methodology 12, pp. 73-78.

HOOGLAND, J (2002)

Selective editing by means of Plausibility Indicators. UNECE Work Session on Statistical Data Editing, Helsinki, working paper no. 33.

HOOGLAND, J; VAN DER LOO, M; PANNEKOEK, J. AND SCHOLTUS, S. (2011)

Data editing. Detection and correction of errors. Statistics Netherlands, The Hague/Heerlen.

LATOUCHE, M. AND BERTHELOT, J. M. (1992)

Use of a Score Function to prioritise and Limit Recontacts in Editing Business Surveys Data editing. Journal of Official Statistics 8, pp. 389-400

LAWARENCE, D. AND MCKENZIE, R. (2000)

The General Application of Significance Editing. Journal of Official Statistics 16, pp. 243-253

VAN LANGEN, S (2002)

Selective Editing by Using Logistic Regression. Report, Statistics Netherlands, Voorburg.

Organismo Autónomo del

www.eustat.eus