Estimación Robusta de la Ocultación de los Ingresos Personales en España

Juan Fco. Ortega Dato[*]
Fco. Javier Callealta Barroso[**]

Resumen

De la información que se desprende de la *Encuesta Básica de Presupuestos Familiares* (EBPF) y de la *Cuenta de Renta del Sector Hogares* (CRSH) de la *Contabilidad Nacional* (CN) se deduce la existencia de una gran diferencia referente a los respectivos datos agregados de *Ingresos*, la cual es denominada genéricamente *Ocultación*. A falta de más información que la de las EBPF y CRSH de la CN, en el trabajo "*Distribución Personal de la Renta en España*" (Pena y otros 1996) se propone utilizar la parte estrictamente creciente de un modelo parabólico para, una vez estimada globalmente esta *Ocultación*, distribuirla a nivel personal, en función de la *Renta Declarada*.

Mediante teorías robustas, es posible ajustar un modelo parabólico en su expresión más general en forma de "U", más acorde con la realidad, para la explicación de la *Ocultación* respecto de la *Renta Declarada*, utilizando los comportamientos atípicos de las observaciones extremas del modelo allí empleado. Así, en esta comunicación, como consecuencia del trabajo realizado en Ortega (2000) y utilizando un procedimiento original basado en técnicas robustas, se obtiene una función para cada región, que nos permite determinar una estimación de la *Ocultación* realizada por un individuo, respetando la información suministrada por la CN y el supuesto de dependencia parabólica.

Palabras Clave: Ocultación, Ingresos Personales, Robustez, Observaciones Atípicas.

[*] Área de Matemáticas de la Facultad de CC. Económicas y Empresariales de Albacete. Universidad de Castilla-La Mancha. Correo electrónico: jfortega@ecem-ab.uclm.es

[**] Departamento de Estadística Económica, Estructura Económica y O.E.I de la Universidad de Alcalá. Correo electrónico: franciscoj.callealta@uah.es

1.- Introducción

En el trabajo *"Distribución Personal de la Renta en España"* (Pena y otros 1996), donde se estudian temas relacionados con la renta y su distribución en nuestro país, se observa que existe una gran discrepancia entre la información obtenida mediante la *Encuesta Básica de Presupuestos Familiares* (EBPF) ya corregida, y la que se desprende de la *Cuenta de Renta del Sector Hogares* (CRSH) de la *Contabilidad Nacional* (CN) referente a los ingresos (rentas brutas disponibles), llamándose a esta diferencia *Ocultación*.

En el Capítulo V del citado trabajo, F. Javier Callealta plantea diferentes mecanismos mediante los cuales abordar la distribución personal de la citada *Ocultación*. De entre los métodos propuestos, y utilizando exclusivamente las informaciones manejadas en las EBPF y CRSH, se llega a la conclusión de que el más acertado correspondería a la utilización de una corrección con *Tasa de Ocultación Progresiva Lineal*. Para ello, llamando x a la *Renta Bruta Disponible Declarada* para un individuo determinado, proporcionada por la EBPF, y O_x a la *Ocultación* en que éste incurre, se define la función *Tasa de Ocultación*, denotándola por *TO*, por el cociente $TO = O_x / x$ de manera que, en el supuesto de *Tasa de Ocultación Progresiva Lineal*, el autor supone que:

$$TO(x) = a + m\,x$$

para constantes $a,m \in \Re$, las cuales son determinadas de manera que las ocultaciones agregadas por algún corte transversal (en el trabajo de Pena se utilizan los cortes transversales por región, categoría socio-profesional y hábitat) respeten la información agregada obtenida de la CRSH y, por otra parte, la pendiente m sea positiva. En concreto, en el estudio referido, m se elige como la pendiente promedio de las que proporcionan las dos situaciones extremas factibles, esto es, la de máxima y mínima progresividad.

De esta manera, la *Ocultación* (O_x) para un individuo que declara una determinada renta (x), se construye mediante una función parabólica llamada *Función de Ocultación* (*FO*), donde $O_x = FO(x)$, siendo en el trabajo citado de la forma:

$$FO(x) = (a + m\,x)\;x$$

y donde en este cálculo se ha respetado la información proporcionada por la CRSH al tiempo que ésta está en concordancia con las premisas básicas con las que el estudio prevé el fraude; es decir, los individuos con rentas de mayor valor declarado tenderán a ocultar más sus ingresos y además en mayor proporción.

Como advertían los autores del trabajo antes citado, a causa de la poca información de partida de la que se disponía, las ocultaciones obtenidas mediante unas *Tasas Progresivas Lineales* para rentas declaradas pequeñas serán cercanas a cero (siendo cero para rentas nulas), no siendo del todo coherentes con las previsiones racionales, ya que parece lógico pensar que ante rentas pequeñas los individuos tenderán a ocultarlas infradeclarando su valor. Por otra parte, una *Tasa de Ocultación* debería ser no acotada para estos valores, permitiendo corregir rentas declaradas casi nulas mediante rentas de supervivencia, como mínimo. Así, una mejor aproximación de las ocultaciones con respecto a las rentas declaradas, debería seguir un modelo parabólico general en forma de "U", de manera que la *Función de Ocultación* debería ser un polinomio de segundo grado en su expresión más general. Es decir;

$$FO(x) = A + B\ x + C\ x^2$$

para valores $A,B,C \in \Re$, de manera que la función *Tasa de Ocultación* sería de la forma;

$$TO(x) = (A / x) + B + C\ x$$

donde el nuevo modelo produciría una *Tasa de Ocultación* de forma hiperbólica cerca de 0 y convergente a una recta, de pendiente C, para valores grandes de x. Sin embargo, la determinación de la *Ocultación* mediante la inclusión de un término más en su definición, requiere imponer alguna otra condición o disponer de información externa adicional, por lo que, la falta de información hizo que esta propuesta no fuera considerada por Callealta en el trabajo de referencia.

Si nuestro propósito es el de distribuir la *Ocultación* dependiente de la *Renta Declarada* mediante una tal función parabólica, realizando una serie de transformaciones en las variables conocidas de la EBPF y la CN, es posible construir un modelo lineal donde la *Ocultación* será su error, y los comportamientos atípicos en los extremos de dicho modelo, detectados mediante las teorías robustas, pueden ser

utilizados como información extra para estimar la *Función de Ocultación* con forma parabólica.

En definitiva, en este artículo, vamos a presentar una alternativa para determinar unas nuevas ocultaciones mediante un procedimiento original con contenidos de estadística robustas, el cual utiliza la presencia de observaciones atípicas en un modelo lineal definido mediante unas variables generadas a partir de la información contenida en la EBPF, la CN y el trabajo de Pena y otros (1996), para la obtención de información adicional, fundamentalmente sobre las colas de la distribución, con la que construir una *Función de Ocultación* con forma parabólica con respecto a la *Renta Declarada* para cada región de España, respetando los supuestos propuestos en el citado trabajo de Pena y las que se desprenden de la CN.

El estudio realizado comienza con la presentación de la notación utilizada en el mismo, junto con las fuentes donde se pueden encontrar los datos de las diferentes variables utilizadas. En la siguiente sección, sección tres, se presenta el procedimiento, con sus diferentes pasos a seguir, que será utilizado sobre los datos de la EBPF y la CN en la sección cuarta. El trabajo acaba con unos comentarios a título de conclusiones sobre los estudios realizados.

2.- Notación y Fuentes

Es de vital importancia tener claros todos y cada uno de los conceptos que se van a utilizar y también su procedencia. Por ello debemos definir con rigor y a ser posible sin solapamientos los elementos necesarios, proporcionando información detallada de su ubicación.

2.1.- Notación

El concepto base en este estudio es el de *Renta Bruta Disponible*, que hará siempre referencia a la cantidad de dinero del cual dispone un individuo para el consumo o el ahorro, es decir, la suma de todos los ingresos que tiene un individuo, los cuales se componen de los *Ingresos Monetarios* (IM) y los *Ingresos no Monetarios* (InM),

restando las *Cotizaciones a la Seguridad Social* (Co), los *Impuestos* (Im) y los *Desembolsos* (De), donde estos últimos se calcula como la suma de los *Intereses Pagados* (*Ip*), las *Primas de Seguros* (Ps), y las *Transferencias* ya sean *Regulares a Hogares* (*Trh*), *Ocasionales a Hogares* (*Toh*) o *a Instituciones* (*Ti*). Además, a su vez los Ingresos Monetarios se dividen en seis tipos diferentes que son: *Ingresos por Cuenta Ajena* (*Ica*), *Ingresos por Cuenta Propia* (*Icp*), *Ingresos por Rentas del Capital* (*Irc*), *Ingresos por Transferencias Regulares* (*Itr*), *Ingresos por Transferencias Ocasionales* (*Ito*) y *Otros Ingresos* (*Io*).

En el referido trabajo *"La Distribución Personal de la Renta en España"*, se parte de la *Renta Declarada* (*rd*) como *Renta Bruta Disponible* de cada individuo obtenida mediante la EBPF, y, mediante el supuesto de *Tasa de Ocultación Progresivas Lineales*, se obtiene la variable *Ocultación* (*ocul*) que sumada a la variable *rd* genera un nuevo valor de *Renta Corregida* (*rc*), la cual está en concordancia con la que se desprende agregadamente de la CRSH.

En nuestro trabajo, ***rd*, *rc*** y *ocul* serán consideradas como variables estadísticas. Al ser *rd* la *Renta Bruta Disponible* declarada derivada de la EBPF, ésta se construye como suma de todos los ingresos que recibe el individuo, restándole los *Impuestos*, las *Cotizaciones* y los *Desembolsos* que realiza, corregida finalmente por el *Coeficiente de Cobertura* (Cob) que subsane las diferencias entre los ingresos declarados y los mínimos obtenidos al observar su consumo y ahorro, y que se calcula mediante la información de la EBPF, justificándose su construcción y uso en el trabajo de Pena. Por otra parte, ya que rc se obtiene como suma de rd y $ocul$, rd y rc serán variables construidas con los mismos elementos y por lo tanto, salvo por el valor de las ocultaciones, deberían ser valores idénticos para cada observación de la muestra.

Otros elementos utilizados para el cálculo de las ocultaciones fueron: las *Ocultaciones por Corte Transversal,* de las cuales nosotros utilizaremos sólo las *Ocultaciones por Región* (*OR*), y que se obtienen por comparación con la CRSH; y los *Factores de Elevación por Individuo* (*FE*), formados por el producto del *Factor de Elevación* y el *Número de Miembros para cada Familia* de la EBPF, al haber sido estimada la renta personal en el estudio de referencia como la renta per capita para cada miembro de las familias.

2.2.- Fuentes

Las fuentes de las que disponemos son: la *Encuesta Básica de Presupuestos Familiares* (EBPF), realizada por el INE; la *Cuenta de Renta del Sector Hogar* (CRSH), que se desprende de la *Contabilidad Nacional* (CN); y la base de datos utilizada y obtenida en el trabajo de Pena y otros (1996). El periodo de referencia utilizado será el último de los usados en el trabajo citado, y que coincide con el de la última EBPF entre 1990 y 1991 (EBPF-90/91).

En EBPF-90/91 contamos con una detallada información, en soporte informático, sobre 21.155 familias integradas por unos 80.000 individuos de nuestro país, de la cual utilizaremos sólo una parte contenida en los ficheros denotados en la misma encuesta por *Tipo-1* y *Tipo-3*. De la información del fichero *Tipo-1*, con título *Gastos Generales del Hogar*, hemos seleccionado las posiciones en la que encontramos las *Inversiones en Vivienda*, los *Prestamos Recibidos* y las *Amortizaciones de Prestamos* (posiciones 191-300) y las posiciones donde encontramos los *Gastos* e *Ingresos* (posiciones 765-780) que nos servirán para construir los *Coeficientes de Cobertura*. Del fichero *Tipo-3*, *Datos de los Miembros del Hogar*, hemos seleccionado los datos referentes a *Ingresos* en sus diferentes modalidades.

De la CN, utilizaremos los resultados de la *Cuenta de Rentas del Sector Hogares por Regiones*, los cuales han sido obtenidos y usados en el trabajo de Pena, y que nos permite, por comparación, determinar la *Ocultación por Región*.

Por último, los datos referidos a *Índices de Familias*, *Número de Miembros del Hogar*, *Factor de Elevación* y *Región* a la que pertenecen junto con los datos de *rd* y *rc*, son los también utilizados y proporcionados respectivamente por el trabajo de Pena.

3.- Modelización

Nuestro propósito es el de, mediante un procedimiento original y una serie de cálculos y estimaciones, llegar a determinar una nueva *Ocultación* que respete los supuestos sobre la forma parabólica de la *Función de Ocultación* y la información proporcionada por la CRSH.

El procedimiento original al que nos referimos consta de una serie de pasos en los que partiendo de la estimación mediante Mínimos Cuadros de un modelo lineal, su error residual (variable ocultación), se descompone como suma de otros dos errores: uno que denominaremos *error u ocultación lineal asumido por el modelo*, generado por la diferencia entre la variable estimada linealmente y la suma de las variables explicativas; y otro que denominaremos *error de regresión*, construido como diferencia entre el valor observado de la variable explicada y el estimado en el modelo lineal. Como nuestro propósito es el modelizar el error del modelo lineal (*Ocultación*) mediante una función parabólica, las observaciones atípicas en dicho modelo, determinadas y caracterizadas mediante procedimientos robustos, pueden ser utilizadas como información extra para conseguirlo, ya que, construyendo un modelo donde se estime el *error de regresión* mediante una función parabólica, y ponderando con mayor intensidad las observaciones anómalas en el modelo lineal inicial, su resultado producirá el "levantamiento" de las colas del *error de regresión*, de manera que al construir el error del modelo lineal como suma de dos errores; uno con estructura lineal, el *error asumido por el modelo*; y otro, el *error de regresión*, con forma parabólica, entonces el error del modelo lineal tendrá forma parabólica como pretendíamos.

Así, partiendo de una serie de transformaciones de las variables proporcionadas por EBPF y la *rd* dada por el trabajo de Pena, y mediante el citado procedimiento, determinaremos unas *Funciones de Ocultación* con forma parabólica, que nos proporcionen, para cualquier *Renta Declarada* de un individuo en una determinada región, una estimación de su *Ocultación* que respete los supuestos del trabajo de Pena y la información que se desprende de la CN.

3.1.- Consideraciones previas

Tomando como punto de partida el estudio de la renta en España realizado en el trabajo ya citado de Pena y otros (1996) para los años 90/91 y, en los casos que sea necesario, utilizando solamente el corte transversal para regiones, veamos algunas consideraciones previas.

a) Como se comenta en el trabajo de Pena, el montante de los Ingresos Monetarios en la construcción de la *Renta Declarada*, supone más del 85% del total de ésta, por lo que será éste el componente principal de la variable renta de nuestro estudio.

b) La *Renta Corregida* se obtiene como resultado del trabajo de Pena partiendo de la *Renta Declarada* y mediante el supuesto de *Tasas de Ocultación Progresivas Lineales*, donde esta *Renta Declarada* está corregida previamente por un *Coeficiente de Cobertura* (*Cob*) que intenta subsanar las diferencias en cada elemento de la encuesta entre los ingresos declarados y los obtenidos sobre consumo y ahorro. Para el caso de los años 90/91, este coeficiente se ha calculado de la siguiente forma:

$$Cob = \frac{\text{Gasto} + \text{Inversión en Vivienda} + \text{Amortización de Préstamos}}{\text{Ingresos} + \text{Préstamos Recibidos}}$$

3.2.- Procedimiento

Antes de comentar el procedimiento, realizaremos una serie de transformaciones de variables que simplifique la construcción de los modelos en él.

Así, para cada región podemos suponer que existe una relación entre la *Renta Declarada* (*rd)* y la *Renta Corregida* (*rc)* de manera que los errores observados, como consecuencia de este ajuste, corresponderían a las buscadas *Ocultaciones* (*ocul*). Planteemos dicha relación mediante la definición de unas nuevas variables.

De la EBPF se han podido aislar los elementos necesarios para construir el *Coeficiente de Cobertura* (*Cob*) según el supuesto b) citado en 3.1., para cada elemento de la encuesta, de manera que, conocida la estructura de *rd*, entonces;

$$rd = (IM + InM - Co - Im - De)\, Cob$$

Denotando por *IMc* a los *Ingresos Monetarios Corregidos por Cob* (es decir; *IMc=IM·Cob)*, ésta puede ser desglosada como suma de los seis ingresos monetarios de la forma:

$$IMc = (Ica + Icp + Irc + Itr + Ito + Io)\ Cob$$

donde, denotando por *Idc* a la diferencia entre *rd* e *IMc* entonces:

$$Idc = (InM - Co - Im - De) \cdot Cob = rd - IMc$$

Si ahora definimos las variables: I_1=Ica·Cob; I_2=Icp·Cob; I_3=Irc·Cob; I_4=Itr·Cob; I_5=Ito·Cob e I_6=Io·Cob, entonces se cumple que;

$$\sum_{k=1}^{6} I_k = IMc = rd - Idc$$

donde, siendo *ocul=rc-rd* la ocultación asignada a cada individuo en Pena y otros (1996), podemos definir una nueva variable *Y* de la forma:

$$Y = rc - Idc = rc - (rd - IMc) = (rc - rd) + IMc = IMc + ocul$$

y plantear un modelo a partir de seis variables independientes, los seis clases de Ingresos Monetarios (I_k para k=1,2,...,6) corregidos por el Coeficiente de Cobertura, y la variable *Y* dependiente linealmente de éstas, de la forma:

$$Y = \sum_{k=1}^{6} \beta_k I_k$$

donde se conoce para la variable multidimensional *(Y, I_1, I_2, I_3, I_4, I_5, I_6)* una muestra de tamaño N_j proporcionada por la EBPF para la región j-ésima, y donde la variable *ocul* estará directamente ligada al error de dichos modelos.

Presentemos ahora el procedimiento en sí.

Paso 1: Detección de Observaciones Atípicas.

Para cada región (subíndice j=1, 2,...,18), se construye un modelo de la siguiente forma:

***Modelo-1j*:**

$$Y = \sum_{k=1}^{6} \beta_k I_k$$

Sobre cada uno de estos modelos lineales se realiza un estudio, mediante teorías robustas, con el propósito de determinar unas ponderaciones para cada observación de la muestra y en cada región, que nos informen de lo apropiado de su inclusión en el modelo propuesto. Estas ponderaciones serán calculadas mediante las teorías robustas propuestas en Ortega (2000). Como resultado de dichas teorías, para cada observación en cada *Modelo-1j* se determina una distancia llamada *Distancia por Truncamiento* (DT_i^2) que nos proporciona información sobre lo adecuado que es suponer que la observación i-ésima es una observación atípica en el *Modelo-1j*.

El objetivo de este paso es exclusivamente determinar el grado de rareza de las observaciones frente a la linealidad del modelo. Sin embargo, la inclusión en los *Modelo-1j* de las ponderaciones de cada uno de las observaciones de la EBPF (que hemos denotado por *FE*) sería una gran carga para los cálculos (ya de por sí muy largos por la cantidad de observaciones y variables utilizadas). Es por este motivo pragmático por el que no se han incluido aquí las citadas ponderaciones, asumiendo que la diferencia de resultados sobre la determinación de las *Distancias por Truncamientos* no será determinante.

Paso 2: Descomposición de la Ocultación.

Con el propósito de estimar un error en los *Modelo-1j* no influido por las posibles observaciones atípicas, y posteriormente proceder a su descomposición en una parte lineal y otra no lineal, comenzaremos por construir unos nuevos modelos cuyas ponderaciones potencien la presencia de aquellas observaciones que son consideradas como genuinas en los anteriores modelos, y tengan menos en cuenta las posibles observaciones atípicas en éstos.

Así, se define un nuevo modelo lineal para cada región (subíndice j) donde cada observación en ella (subíndice i con $i \in N_j$) recibe una ponderaciones w_{2i} construidas mediante el producto de su *Factor de Elevación* correspondiente (FE_i), y, ya que cuanto mayor es el valor de la *Distancia por Truncamiento* en mayor medida se supone que la observación i-ésima es una observación atípica, la raíz cuadrada del cociente ($1/ DT_i^2$), de la forma:

Modelo-2j:

$$Y = \sum_{k=1}^{6} \beta_k I_k$$

Ponderaciones: $w_{2i} = FE_i \ (1/ DT_i^2)^{1/2} \qquad \forall i \in N_j$

Para cada *Modelo-2j* se obtiene, mediante Mínimos Cuadrados Ponderados, una estimación de *Y*, que denotaremos por _, y unos coeficientes $\hat{\beta}_k$ asociados a cada tipo de ingreso considerado. Con estas estimaciones, podemos construir la variable *ocul* como suma de dos errores de estimación: el primero, que denotaremos por ε', que hace referencia al *error u Ocultación lineal asumido por el modelo* y definido mediante los parámetros de regresión estimados; y el segundo, denotado por ε, que representa el error no lineal del mismo y definido como el *error de regresión* propiamente dicho.

Así, estos dos errores quedan definidos por:

$$\varepsilon' = \sum_{k=1}^{6} (\hat{\beta}_k - 1) I_k \qquad \varepsilon = Y - \_$$

donde la variable *ocul*, definida como la diferencia entre *rc* y *rd*, equivale a la suma de las dos variables anteriores, es decir;

$$ocul = rc - rd = (Y + Idc) - (\sum_{k=1}^{6} I_k + Idc) = Y - \sum_{k=1}^{6} I_k = \varepsilon' + \varepsilon$$

por lo que la *Ocultación* se redistribuye en una parte comúnmente explicada para todos los miembros de la muestra (ε') y una específica dependiente del individuo (ε). Es decir, hemos descompuesto la *Ocultación* propuesta por Callealta como suma de dos errores, uno de ellos representa la parte lineal del mismo mientras que el otro representa la parte no lineal.

Paso 3: Ajuste de la parte no lineal de la Ocultación.

En este paso, plantearemos un modelo para cada región donde la parte no lineal de la ocultación (ε) construida en el paso anterior, se ajusten mediante un modelo parabólico respecto de la *Renta Declarada*, de manera que sea posible obtener una redistribución de las ocultaciones iniciales dadas por Callealta. Para ello, supondremos que ε tiene forma parabólica con respecto a *rd* para cada región considerada, de manera que respeten la información proporcionada por CN, impuesta por una restricción, y donde los pesos utilizados en este caso (w_3) den mayor importancia a las observaciones atípicas en los *Modelo-1j*, es decir, para cada región las observaciones atípicas "ayuden" a conseguir la forma parabólica de la *Ocultación* respecto de la *Renta Declarada*.

Así, definimos los modelos para cada región (j=1,2,...,18), denotándolos por *Modelo-3j*, de la forma:

Modelo-3j:

$$\varepsilon = A_j + B_j\, rd + C_j\, rd^2$$

Restricción: $\sum_{i \in N_j} \hat{\varepsilon}_i\, FE_i = OR_j - \sum_{i \in N_j} \varepsilon'_i \cdot FE_i$

Ponderaciones: $w_{3i} = FE_i\ (DT_i^2)^{1/2} \qquad \forall i \in N_j$

donde $\hat{\varepsilon}$ representa la estimación buscada de ε, y OR_j representa la *Ocultación para la Región* j-ésima proporcionada por la Contabilidad Nacional.

Como resultado de cada *Modelo-3j*, mediante Mínimos Cuadrados Ponderados y considerando la restricción, se obtienen los coeficientes $\hat{A}_j$, $\hat{B}_j$ y $\hat{C}_j$ mediante los

cuales es posible construir una estimación de ε ($\hat{\varepsilon}$), que nos proporcione una redistribución de la *Ocultación* propuesta por Callealta (*reocul*) de la forma:

$$reocul = \varepsilon' + \hat{\varepsilon}$$

Paso 4: Ajuste de la Función de Ocultación global.

En el cuarto y último paso, utilizaremos la redistribución de las ocultaciones calculadas en el paso anterior (*reocul*) para estimar la forma de la *Función de Ocultación* para cada región, con el fin de poder predecir la *Ocultación* para un individuo de una determinada región con una determinada *Renta Declarada*.

Así, para la j-ésima región definimos:

***Modelo-4j**:*

$$\mathrm{r} = reocul = A'_j + B'_j\, rd + C'_j\, rd^2$$

Restricción: $\sum_{i \in N_j} \hat{r}_i\, FE_i = OR_j$

Ponderaciones: $w_{1i} = FE_i \qquad \forall i \in N_j$

donde $\hat{r}$ representa la estimación de la variable *reocul* de manera que se respete la información proporcionada por la CN (mediante la restricción), y donde se conserve la estructura de la EBPF (mediante las ponderaciones).

La estimación de los *Modelo-4j*, mediante Mínimos Cuadrados Ponderados y considerando la restricción, nos proporciona los coeficientes $\hat{A}'_j$, $\hat{B}'_j$ y $\hat{C}'_j$ con los que podremos definir la *Función de Ocultación* para la región j-ésima de la forma:

$$FO(rd) = \hat{A}'_j + \hat{B}'_j\, rd + \hat{C}'_j\, rd^2$$

4.- Resultados numéricos

En este apartado vamos a presentar los principales resultados obtenidos en la aplicación del procedimiento del apartado anterior sobre los datos conocidos.

Partiendo de los *Modelos-1j* (Paso 1) se obtienen las *Distancias por Truncamientos* utilizando las teorías propuestas en Ortega (2000), mediante las cuales es posible plantear los *Modelo-2j* (Paso 2) de los se obtienen, los coeficientes $\hat{\beta}_k$ (para k=1,2,...6) asociados a cada tipo de ingreso considerado, que recogemos en la *Tabla-1*.

J	Región	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
1	Andalucía	1.423	1.696	2.377	1.359	1.459	2.614
2	Aragón	1.264	1.455	1.539	1.285	0.725	3.150
3	Asturias	1.288	1.576	1.779	1.324	0.553	1.459
4	Baleares	1.317	1.463	1.630	1.323	1.096	1.260
5	Canarias	1.431	1.578	1.701	1.365	1.424	1.803
6	Cantabria	1.303	1.415	1.792	1.265	1.031	2.024
7	Castilla y León	1.425	1.500	2.181	1.373	0.952	1.921
8	Cast.-La Mancha	1.257	1.376	1.415	1.237	0.864	1.450
9	Cataluña	1.427	1.583	1.851	1.485	3.152	1.540
10	Com. Valenciana	1.372	1.674	2.774	1.402	1.026	1.011
11	Extremadura	1.422	1.581	2.552	1.292	2.483	1.388
12	Galicia	1.375	1.521	1.691	1.344	1.510	1.911
13	Madrid	1.366	1.788	2.158	1.523	1.189	1.262
14	Murcia	1.417	1.563	2.884	1.381	1.788	3.236
15	Navarra	1.427	1.603	0.895	1.634	0.748	0.983
16	País Vasco	1.338	1.428	1.799	1.363	0.933	1.890
17	Rioja, La	1.156	1.322	1.186	1.132	0.941	1.145
18	Ceuta y Melilla	1.429	1.587	3.251	1.372	0.545	0.000

Tabla-1

Con esta información se construye las nuevas variables ε e ε' para cada región, donde la *Ocultación* dada en el trabajo de Callealta (*ocul*) como diferencia entre *rc* y *rd*, equivale a la suma de las dos variables anteriores correspondientes al *error lineal asumido por el modelo* (ε') y al *error de regresión* (ε).

Imponiendo ahora que ε tenga forma parabólica en cada región con respecto a *rd*, construimos el *Modelo-3j* (Paso 3) para la región j-ésima, donde los pesos utilizados en este caso den mayor importancia a las observaciones atípicas, obteniendo como resultado los coeficiente $\hat{A}_j$, $\hat{B}_j$ y $\hat{C}_j$ recogidos en la *Tabla-2*.

j	Región	$\hat{A}_j$	$\hat{B}_j$	$\hat{C}_j$
1	Andalucía	106323.25	-0.4273252	2.286172e-7
2	Aragón	73715.05	-0.2641470	1.567991e-7
3	Asturias	294082.63	-0.7777980	3.925904e-7
4	Baleares	14436.37	-0.1668234	1.380428e-7
5	Canarias	100027.00	-0.4460632	3.276978e-7
6	Cantabria	139499.06	-0.4006829	1.993312e-7
7	Castilla y León	151896.15	-0.4689834	2.093744e-7
8	Cast.-La Mancha	57260.42	-0.2269121	1.194826e-7
9	Cataluña	186163.90	-0.3834806	6.812214e-8
10	Com. Valenciana	120469.50	-0.3620899	7.994175e-8
11	Extremadura	114108.60	-0.2878595	7.387357e-8
12	Galicia	123974.20	-0.3334034	1.001252e-7
13	Madrid	163425.20	-0.1368799	1.151335e-7
14	Murcia	206366.99	-0.5098239	1.096437e-7
15	Navarra	168614.20	-0.4654256	2.056789e-7
16	País Vasco	64688.80	-0.3057069	2.097891e-7
17	Rioja, La	75083.13	-0.1904307	7.305300e-8
18	Ceuta y Melilla	132800.00	-0.6478056	4.669945e-7

Tabla-2

Con estos coeficientes se construye $\hat{\varepsilon}$, de manera que podremos calcular una redistribución de la *Ocultación* propuesta por Callealta, que hemos denotado por *reocul*, de la forma:

$$reocul = \varepsilon' + \hat{\varepsilon}$$

Una vez calculada *reocul*, el siguiente paso consiste en determinar la llamada *Función de Ocultación* para cada región. Para ello, mediante los *Modelo-4j* (Paso 4) se obtienen los coeficientes $\hat{A}'_j$, $\hat{B}'_j$ y $\hat{C}'_j$, recogidos en la *Tabla-3*.

j	Región	$\hat{A}'_j$	$\hat{B}'_j$	$\hat{C}'_j$
1	Andalucía	163676.09	-0.082609778	2.335869e-7
2	Aragón	112204.14	-0.044363023	1.670398e-7
3	Asturias	309422.71	-0.395865720	3.205668e-7
4	Baleares	136118.60	-0.018509590	1.660659e-7
5	Canarias	133676.10	-0.051006450	2.967113e-7
6	Cantabria	187883.82	-0.167962882	1.823662e-7
7	Castilla y León	201044.94	-0.096210675	1.783013e-7
8	Cast.-La Mancha	111317.58	-0.038270498	1.139888e-7
9	Cataluña	314463.00	-0.077069260	6.918040e-8
10	Com. Valenciana	229750.90	-0.099610960	1.012344e-7
11	Extremadura	189001.80	-0.078395800	1.432575e-7
12	Galicia	184188.80	-0.014946650	7.920915e-8
13	Madrid	256248.00	-0.055859410	1.045752e-7
14	Murcia	253346.35	-0.114862153	9.960419e-8
15	Navarra	330849.10	-0.160398300	2.028252e-7
16	País Vasco	124103.88	-0.007942863	1.720471e-7
17	Rioja, La	92883.86	-0.036593460	5.473271e-8
18	Ceuta y Melilla	197664.30	-0.417897500	4.845621e-7

Tabla-3

En definitiva, con estos coeficientes se construye, para la región j-ésima, la *Función de Ocultación* con forma parabólica dependiente de la *Renta Declarada*, de la forma:

$$FO(rd) = \hat{A}'_j + \hat{B}'_j\, rd + \hat{C}'_j\, rd^2$$

De los resultados obtenidos se observa que en todos los casos, el coeficiente $\hat{A}'_j$ es mayor que cero, lo que nos asegura que para valores nulos de *rd*, la ocultación es mayor de cero. También en todos los casos $\hat{B}'_j$ es menor que cero, por lo que, como $\hat{C}'_j$ es siempre positivo, tendremos el mínimo de la *Función de Ocultación* para un valor de *rd* positiva. Por otra parte, como $\hat{A}'_j$ y $\hat{C}'_j$ son mayores que cero, la función *Tasa de Ocultación* cumple con lo esperado, tendiendo además a infinito para *rd* acercándose a cero y convergente a una recta de pendiente $\hat{C}'_j$ para *Rentas Declaradas* grandes.

En definitiva, para cada región, la *Función de Ocultación* tiene forma de "U" con mínimo para *rd* positivas, y por lo tanto la función *Tasa de Ocultación* tiene la forma propuesta.

5.- Conclusiones

La *Renta* es un indicador de la marcha de la economía, por lo que su estudio y modelización es de vital importancia para el perfecto conocimiento de nivel de bienestar de los individuos de un país. Su construcción está respaldada por las informaciones proporcionadas por dos fuentes, a saber: la que se desprende de la *Contabilidad Nacional* y la proporcionada por la *Encuesta Básica de Presupuestos Familiares* del INE.

De las discrepancias entre los datos proporcionados por la EBPF y los que se desprenden de la CRSH (CN) respecto a los *Ingresos*, parece necesaria la corrección de los datos de *Renta Declarada* en la EBPF mediante una variable que hemos llamado *Ocultación*. Así, en este artículo se realiza la corrección de dicha variable bajo el supuesto, que creemos más acorde con la realidad, de que la *Ocultación* depende de la *Renta Declara* mediante una forma parabólica.

Con estas premisas hemos construido un procedimiento, tomando como punto de partida y primera aproximación el trabajo realizado en Pena y otros (1996), y donde las teorías propuestas en Ortega (2000) para el estudio de la rareza de las observaciones atípicas, nos ha permitido ajustar una función de tipo cuadrática sobre la *Ocultación*. Realizando los cálculos sobre los datos correspondientes a los años 90/91 y los cortes a nivel regional, hemos obtenido los coeficientes que determinan dichas *Funciones de Ocultación* para el supuesto propuesto, respetando la información proporcionada por la EBPF y la CRSH de la CN.

Los resultado anteriores nos permiten determinar una estimación de la *Ocultación* previsiblemente realizada por un individuo con una determinada *Renta Declarada* en una determinada región de nuestro país.

Bibliografía

BARNETT, V. Y LEWIS, T. *Outliers in Statistical Data.* Ed. J. Wiley & Sons. 1994.

CHATTERJEE, S. Y HADI, A.S. "Influential Observations, High Leverage Points, and Outliers in Linear Regression". *Statistical Science*, Vol.1, No.3. 1986.

DAVIES, L. Y GATHER, U. "The Identification of Multiple Outliers". *J. of the American Statistical Association*, Vol.88, No.423. 1993.

ORTEGA, J.FCO. *Nuevas Familias de Estimadores Robustos y Detección de Observaciones Atípicas en Modelos Lineales*. Tesis Doctoral. Facultad de CC. Económicas y Empresariales de Albacete. Univ. de Castilla-La Mancha. 2000.

PENA, B.; CALLEALTA, F.J.; CASA, J.M.; MEREDIZ, A. Y MUÑOZ, J. *Distribución Personal de la Renta en España*. Ed. Pirámide. 1996.

PENNY, K.I. "Appropriate Critical Values when Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance". *J. Royal Statistical Society* (serie C), 45, No.1. 1996.

ROUSSEEUW, P.J. Y CROUX, C. "Alternatives to the Median Absolute Deviation". J. of American Statistical Association, Vol. 88, No.424. 1993.

ROUSSEEUW, P.J. Y LEROY, A.M. *Robust Regression and Outliers Detection*. Ed. J. Wiley & Sons. 1987.