Panel de datos del Impuesto sobre la Renta de las Personas Físicas del Instituto de Estudios Fiscales

César Pérez López
Instituto de Estudios Fiscales. Unidad de Estadística

Introducción y objetivo

El Instituto de Estudios Fiscales (IEF) ha elaborado un panel con datos del Impuesto sobre la Renta de las Personas Físicas que realiza el seguimiento las declaraciones de los individuos y hogares desde el año 1999 hasta el año 2007 siendo ampliable cada año con la correspondiente ola. Este panel constituye una evolución de otro panel anterior existente en el IEF para el periodo 1982/1998 aportando un nuevo diseño y nuevas capacidades de análisis y explotación.

El panel de Renta tiene como objetivo general facilitar información para realizar estudios de carácter longitudinal sobre la renta de las personas físicas y su distribución y concentración a partir del seguimiento de individuos a lo largo del tiempo. Con este panel se dispondrá de información de rentas fiscales de personas y hogares de una población representativa de sujetos obligados a tributar en el Impuesto sobre la Renta de las Personas Físicas (IRPF), en el denominado Territorio Fiscal Común (total nacional excluyendo la Comunidad Foral de Navarra y el País Vasco).

Los datos de panel

La característica esencial de un panel es su dimensión temporal, dado que se forma con observaciones sobre gran variedad de individuos tomadas en distintos momentos del tiempo, generalmente años. No obstante, el hecho diferenciador de un panel es que habitualmente las observaciones provienen de lo mismos individuos en los diferentes momentos del tiempo. Además, cuando la fuente de información que alimenta el panel son registros administrativos no existe prácticamente falta de respuesta, lo que permite formar paneles compactos ideales para el análisis exentos de valores perdidos y datos atípicos sin ser necesaria la aplicación de métodos de imputación de la información faltante. La etapa del análisis exploratorio de datos, previa a cualquier aplicación con datos de panel y decisiva en los resultados, se ve así fuertemente favorecida.

El desarrollo de una base de datos como es la que contiene la información del panel de IRPF, con datos cada vez más ricos en cuanto a números de unidades recogidas y características de estas, unido al aumento de la potencia de los recursos informáticos para su tratamiento, hace que últimamente haya proliferado el uso de los modelos de microsimulación con vistas a la estimación de los efectos de las políticas públicas.

Paneles puros y paneles expandidos

Un panel puro sigue en todo el período temporal a los mismos individuos. Ello provoca problemas de pérdida de representatividad de la muestra a medida que avanza el tiempo, dado que hay individuos que pueden abandonar la muestra por distintos motivos. Se produce de esta forma un desgaste del panel según avanza el tiempo (attrition) que puede llegar a invalidarlo.

Este problema suele paliarse tomando la decisión de utilizar paneles "expandidos" en contraposición con la idea más habitual de paneles puros. La razón de esta decisión es que estos últimos no constituyen una fiel representación transversal de la población de la que se extraen porque sobre la misma inciden, a lo largo del tiempo, un flujo de entradas y salidas que, en el caso del panel de IRPF o de cualquier otro impuesto, se materializaría, respectivamente, en los nuevos contribuyentes y en los que dejan de serlo en cada ejercicio. Estadísticamente ello significa que dichos paneles puros no constituirían, cada año, una muestra aleatoria de los contribuyentes del año. Ello tendrá efectos perniciosos en dos vertientes, cuando menos. Por una parte, los estudios tranversales realizados sobre el panel exigirían el desarrollo de estimadores estadísticos distintos cada año basados en técnicas de reescalamiento muestral, es decir, de asignación de pesos específicos anuales a cada individuo del panel. Evidentemente, como es natural, dichos estimadores serán tanto más complejos cuanto menor sea el grado de similitud entre el panel puro y el total de declarantes del ejercicio bajo consideración. En consecuencia, las diferencias entre los estimadores de uno u otro año dependerán del grado o nivel de similitud (panel/población de contribuyentes) previamente aludido.

Por otra parte, los paneles puros sobre impuestos tampoco permitirían analizar las peculiaridades específicas de los nuevos contribuyentes ni de los que dejan de serlo en cada ejercicio, ya que, por su propia naturaleza, no podrían contener

ninguno de ellos. En este sentido, esta limitación restaría operatividad al panel, máxime si tenemos en cuenta que en los últimos años los nuevos contribuyentes en cualquier impuesto suelen representar en cada ejercicio un porcentaje bastante significativo del total y que además, por lógica, deberán existir discrepancias significativas entre los nuevos contribuyentes y los antiguos.

El panel "expandido" surge como una vía que trata de eliminar los perniciosos efectos antes aludidos. Permite sin embargo, llevar a cabo todos los estudios asociados a los paneles puros, ya que contiene uno de ellos como subpanel. Consiste, en esencia, en muestras representativas de contribuyentes que están solapados y que pueden ser extraídos de forma recurrente, una vez seleccionada la primera, mediante la incorporación (o expansión) de una submuestra adicional de los nuevos contribuyentes de cada año, frente a los anteriores, con una afijación o tamaño al de la muestra ya existente frente al total de declarantes antiguos. Obviamente, los nuevos contribuyentes incorporados al panel continuarán siendo observados desde el momento de su incorporación en adelante. Las bajas surgirán, de hecho, como resultado de dicha observación, es decir, aparecerán sobre el panel de manera natural cuando no sean encontrados entre los declarantes del año. De esta forma, se dispondrá de muestras representativas de cada ejercicio en las que tendremos identificados los nuevos declarantes y los antiguos así como los que han causado baja en un determinado año, constituyendo la parte común de las mismas el ya mencionado panel puro subcontenido en ellas.

Muestras anuales y combinaciones de cortes transversales (pool de datos)

Por otro lado, siempre que no hay dificultades de selección, como es en el caso de los registros administrativos, suelen utilizarse también muestras anuales. Una muestra anual (o corte transversal) aislada ofrece información muy rica para un momento dado del tiempo. No obstante carece de dimensión temporal, lo que limita bastante su utilidad. Sin embargo, cuando se dispone de varias muestras obtenidas en diferentes años de modo independiente (denominadas combinaciones de cortes transversales independientes o pool de datos) la riqueza de los análisis aplicables aumenta exponencialmente.

Al igual que el panel, una combinación de cortes transversales independientes (pool de datos) tiene dimensión temporal, dado que se forma con observaciones sobre gran variedad de individuos tomadas en distintos momentos del tiempo, generalmente años. No obstante, el hecho diferenciador de un pool de datos es que las observaciones no provienen de los mismos individuos en los diferentes momentos del tiempo.

No obstante, muchos estudios sobre individuos, familias, empresas, etc. se repiten a intervalos regulares, a menudo anuales. Si se extrae una muestra aleatoria en cada período, el combinar los resultados de ellas nos da un pool de datos. Al combinar muestras aleatorias extraídas de la misma población, pero en distintos momentos del tiempo, obtenemos estimadores más precisos y estadísticos de prueba más potentes derivados sobre todo de los elevados tamaños muestrales que se manejan. No obstante, hay que tener presente que es necesario que se mantengan en el tiempo las relaciones entre la variable dependiente y alguna de las independientes en los modelos que se apliquen. Estadísticamente los problemas de un pool de datos pueden derivarse de que la población tal vez tenga distintas distribuciones en diferentes períodos, lo que suele solucionarse mediante la introducción de variables ficticias. A su vez también pueden aparecer problemas de variabilidad no constante en el tiempo (muchos individuos y varios períodos hacen casi imposible el mantenimiento de variabilidades constantes, habitualmente exigidas en la modelización económica), pero estos problemas también son resolubles estadísticamente.

Las combinaciones de cortes transversales también son de mucha utilidad para evaluar el impacto de ciertos sucesos o políticas. Generalmente se utilizan cuando se dispone de dos o varios conjuntos de datos de corte trasversal, recopilados antes y después de la ocurrencia de un evento, para determinar el efecto de los resultados económicos de dicho evento. Por ejemplo, para evaluar el impacto de un cambio en el impuesto sobre el tabaco sobre su consumo se pueden obtener dos muestras aleatorias en distintos momentos del tiempo, de modo que en uno de los momentos no se haya producido el cambio impositivo y en el otro sí. La mayoría de las estimaciones de cambio y las estimaciones en ocasiones sucesivas son factibles razonablemente a partir de un pool de datos.

Sin embargo, el formato panel es, en lo que respecta a la simulación de reformas fiscales, muy superior a un pool de muestras anuales. Dos son los argumentos más potentes que justifican esta afirmación:

1. El panel permite la evaluación de reformas fiscales desde una perspectiva dinámica. En concreto, permite analizar el impacto de cualquier cambio impositivo en un mismo individuo (o unidad familiar o fiscal) a lo largo del tiempo. Es decir, permite un análisis estructural de las reformas. Esta es una cuestión fundamental en lo que a evaluación de reformas fiscales se refiere ya que los individuos alteran su comportamiento en respuesta a los cambios impositivos. Y muchos de esos cambios no se producen en el instante posterior a la reforma sino uno o varios períodos después, por lo que la dimensión temporal del panel para introducir la dinámica es esencial.
2. Los efectos de una reforma se pueden medir de modo más fiable cuando comparamos unidades homogéneas (individuos o unidades familiares o fiscales). Por ejemplo, cuando analizamos el impacto de una reforma sobre la renta neta de los individuos más pobres (los de la primera decila de renta) comparamos al mismo grupo de individuos. De aquí lo esencial de trabajar con observaciones de los mismos individuos en distintos momentos del tiempo.

Pero tampoco despreciemos las muestras anuales. Una sola muestra anual aislada puede no tener mucha fuerza, pero cuan-

do se va construyendo un pool de muestras anuales sucesivas, la riqueza para el análisis aumenta exponencialmente e incluso permite constrastar los resultados obtenidos mediante análisis de panel. No olvidemos que las técnicas econométricas sobre panel no son fáciles de implementar, mientras que sobre pool de datos no hay tantas dificultades. Lo evidente es que, en el caso del IRPF, como se dispone de registros administrativos, el muestreo estratificado con criterios geográficos y algún otro tipo adicional de criterio (tramos de renta o fuentes de renta) es sencillo de realizar. Ello lleva a que sea muy adecuado disponer, tanto de sucesivas muestras anuales, como de un panel.

El panel de IRPF 1999/2007 del Instituto de Estudios Fiscales: población objetivo, marco, ámbito y unidad de muestreo

Se trata de disponer de un panel con información de rentas fiscales de personas y hogares de una población representativa de los sujetos pasivos de IRPF en el Territorio de Régimen Fiscal Común a lo largo del tiempo. Este panel debe responder al concepto de panel expandido; es decir que anualmente debe incluirse una representación de las altas que se produzcan controlando también las bajas.

La población objeto de estudio son los declarantes del impuesto sobre la renta de las personas físicas en los sucesivos años siendo el año inicial del panel 1999, el año base 2003 y el año final a determinar según la *attrition* (desgaste) del panel. Para obtener la muestra de cada año se utilizará muestreo estratificado aleatorio de los declarantes de IRPF de ese año (ámbito temporal) en el territorio de régimen fiscal común (ámbito geográfico). Se utilizan tres niveles de estratificación: comunidad autónoma, nivel de renta y fuente de renta. La muestra de cada año se va incorporando al panel de renta en los sucesivos años.

Dado el objetivo perseguido y la información disponible, se considera que la opción más adecuada es la utilización de los individuos como unidad muestral al ser esta, y no las declaraciones, una unidad homogénea a lo largo del tiempo. Para los individuos que resulten seleccionados según el diseño de muestreo se suministrará la información de las declaraciones presentadas por ellos constituyendo el marco de lista de individuos del que se seleccionará la muestra los modelos 100.

Las variables directas de estudio serán las contenidas en los modelos de declaración hasta nivel de 3 dígitos que no sean confidenciales y como variables derivadas se generarán todos los agregados habituales en el estudio del IRPF (mínimos, distintos tipos de rentas, índices de concentración y redistribución, etc.). Como variables de clasificación se utilizarán los tramos de rendimientos, las fuentes de renta, el sexo, etc. Es notorio que se incorporan todas las variables disponibles que permiten los análisis desde la perspectiva de género.

Así mismo, para poder llevar a cabo análisis referido a hogares, se suministrará el mismo conjunto de información

referida a los cónyuges de los individuos seleccionados para la muestra, siempre que se disponga de la información que permita su identificación como tales cónyuges.

Las variables que se utilizarán para estratificar son:

- Comunidad autónoma de residencia.
- Rentas brutas, aproximada por los Ingresos íntegros, sin deducir por tanto ni los gastos ni las reducciones [con la excepción de las Rentas de actividades económicas en la que se tomarán los Rendimientos Netos]. En términos de las casillas de la declaración la variable para el año base 2003 vendría definida por la suma de las casillas: 01, 02, 03, 07, 12, 13, 18, 20, 21, 22, 23, 24, 25, 26, 27 y 28.
- Proporción de ingresos del trabajo sobre el total de rentas.

Se trata de variables adecuadas para la estratificación porque están muy correlacionadas con las características en estudio y que permiten dividir convenientemente la población en estratos homogéneos. Las Comunidades Autónomas legislan parte del IRPF, lo que permite considerar cada comunidad como un grupo homogéneo. Asimismo los contribuyentes del mismo tramo de renta tienen propiedades parecidas, por lo tanto esta variable también produce grupos homogéneos. Por último, las fuentes de renta suelen agrupar también a individuos homogéneos.

Diseño y selección de la muestra

Diseño muestral

Se considera como año base del panel 2003 porque es el primer año en que la Agencia Tributaria graba los datos hasta tres dígitos incluidos en los modelos de declaración que se usan como marco muestral.

Del marco de lista que incluya las declaraciones de todos los individuos se extraerá la muestra en base a un muestreo estratificado aleatorio, siendo las variables de estratificación y subestratificación las siguientes:

- Comunidad autónoma de residencia (son las 15 CCAA del Territorio de Régimen Fiscal Común, además de Ceuta y Melilla que se considerarán como una única comunidad autónoma).
- Nivel de rentas brutas. Los subestratos propuestos son:
  - Negativas y 0.
  - Inferiores a 3.000 €.
  - Superiores a 3.000 € e inferior o igual a 6.000 €.
  - Superiores a 6.000 € e inferior o igual a 12.000 €.
  - Superiores a 12.000 € e inferior o igual a 18.000 €.
  - Superiores a 18.000 € e inferior o igual a 30.000 €.
  - Superiores a 30.000 € e inferior o igual a 60.000 €.
  - Superiores a 60.000 € e inferior o igual a 120.000 €.
  - Superiores a 120.000 € e inferior a 240.000 €.
  - Superiores a 240.000 €.
- Origen de las rentas, con dos valores posibles: proporción de ingresos del trabajo mayor del 50%; y proporción de ingresos del trabajo menor o igual que el 50%.

Tamaño y afijación

Se utilizará afijación de mínima varianza. La justificación de este hecho proviene de la gran desigualdad de variabilidades de la renta en los distintos estratos.

El tamaño de muestra viene definido por un error relativo de muestreo menor del 1,5%, con un nivel de confianza adicional del 3 por mil (entre 300.000 y 400.000 individuos por año).

Selección de la muestra

En el año base (2003)

En cada CCAA (16 estratos) los individuos se agruparán según el tramo de ingresos íntegros que le corresponda. El resultado serán 176 subestratos (si se mantienen finalmente los 11 tramos de tamaño propuestos para cada una de las 16 CCAA).

En cada uno de los subestratos definidos según el tramo de renta, se separaran en dos grupos los individuos según el origen de dichas rentas (más del 50% de los ingresos son del trabajo y el 50% o menos provienen del trabajo). El resultado serán 176 x 2 = 352 estratos de último nivel en cada uno de los cuales se realiza la extracción aleatoria simple.

Una vez seleccionados los individuos que formarán parte de la muestra y cuya información, por tanto, vendrá acompañada del correspondiente factor de elevación, a efecto de aná-

Estimadores y errores

El estimador de cualquier total poblacional X en muestreo estratificado aleatorio es la suma de los estimadores del total en cada uno de los L estratos. Se tiene:

$$\hat{X}_{st} = \sum_{h=1}^{L} \hat{X}_h = \sum_{h=1}^{L} N_h \bar{x}_h = \sum_{h=1}^{L} \frac{N_h}{n_h} x_h = \sum_{h=1}^{L} fe_h x_h$$

$\bar{x}_h$ = *media muestral en el estrato h*
x_h = *total muestral en el estrato h*
N_h = *tamaño poblacional del estrato h*
n_h = *tamaño muestral del estrato h*
fe_h = *factor de elevación del estrato h*

Por lo tanto, para estimar cualquier total poblacional se suman los productos de los factores de elevación fe_h por los totales muestrales en cada estrato x_h.

El estimador de cualquier media en muestreo estratificado aleatorio es la media ponderada de los estimadores de la media en cada estrato, siendo los coeficientes de ponderación $W_h = N_h/N$ de suma unitaria (N_h es el tamaño poblacional del estrato y N es el tamaño de la población).

$$\hat{\bar{X}}_{st} = \bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h = \sum_{h=1}^{L} \underbrace{\frac{N_h}{N}}_{W_h} \frac{1}{n_h} x_h =$$

$$= \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_h} x_h = \frac{1}{N} \sum_{h=1}^{L} fe_h x_h$$

Por lo tanto, para estimar cualquier media poblacional se suman los productos de los factores de elevación por los totales muestrales en cada estrato y se divide por el tamaño poblacional.

Las varianzas de los estimadores y sus estimaciones son ($f_h = n_h / N_h$):

$$V\left(\hat{X}_{st}\right) = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{S_h^2}{n_h},$$

$$V(\bar{x}_{st}) = V\left(\sum_{h=1}^{L} W_h \bar{x}_h\right) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

$$\hat{V}\left(\hat{X}_{st}\right) = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h},$$

$$\hat{V}\left(\bar{X}_{st}\right) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}$$

S_h^2 = cuasivarianza poblacional en el estrato h,

$\hat{S}_h^2$ = cuasivarianza muestral en el estrato h

Los errores relativos estimados se calculan mediante las expresiones:

$$\hat{C}_v\left(\hat{X}_{st}\right) = \frac{\sqrt{\hat{V}(\hat{X}_{st})}}{\hat{X}_{st}} \qquad \hat{C}_v(\bar{x}_{st}) = \frac{\sqrt{\hat{V}(\bar{x}_{st})}}{\bar{x}_{st}}$$

lisis se seleccionarán sus cónyuges, hayan realizado declaración (conjunta o individual) o no la hayan realizado pero se disponga de información.

En años posteriores al año base (2004 y siguientes)

Cada año posterior se debe extraer una muestra de las altas de individuos (sea por presentación de declaración sea por información suministrada por los retenedores). Para ello es preciso:

- Detectarlas por comparación con la población de individuos del año anterior.
- Dividir las rentas en el caso de las declaraciones conjuntas.
- Muestrear el conjunto
  Se propone una estratificación por comunidades autónomas y un muestreo aleatorio simple de los individuos con las rentas previamente ordenadas según su cuantía. Para ello se tomará la variable equivalente cada año a la definida en el año base.

De los individuos seleccionados se suministrarán sus factores de elevación y toda la información tal y como se ha establecido para el año base.

Se buscará la existencia de cónyuges en el conjunto total de individuos del año de referencia, extrayéndose también la correspondiente información.

Este colectivo se tratará como una subpoblación independiente, con sus correspondientes factores de elevación.

Muestra en años anteriores al año base (1998-2002)

El muestreo en los años anteriores al año base será simétrico al definido para años posteriores. En una primera aproximación puede decirse que se trata de disponer de la información de los años anteriores de los individuos seleccionados en el año base (t).

Para hacer el procedimiento correspondiente al año (t-1) se deben conocer:

- Las altas habidas en el año 2003 (individuos de los que tenemos información en 2003 y no en 2002).
- Las bajas del año 2002 (los que están en el 2002 y no están en el 2003) que deberán muestrearse con los mismos criterios que los señalados para los años posteriores y utilizando la variable equivalente para la estratificación. Igualmente, deberá tratarse como una subpoblación independiente.

Implicaciones de política fiscal

Los modelos de microsimulación sobre datos de panel provenientes de registros administrativos no sólo permiten evaluar los efectos de las políticas públicas actuales, sino también las de sus posibles reformas, a través de la proyección de los probables cambios normativos sobre una base de datos representativa en el tiempo de la población afectada. Para comparar una misma población en dos situaciones distintas en el tiempo (la inicial y la que resultaría de la hipotética aplicación de los cambios previstos) es ineludible disponer de la dimensión temporal que ofrecen los paneles de datos y de la medición de las observaciones sobre los mismos elementos de la población en los distintos momentos del tiempo.

A través de los datos de panel de IRPF es posible evaluar con fiabilidad los factores que explican los movimientos en la recaudación a lo largo del tiempo, o la medición del impacto sobre el tipo medio efectivo del impuesto derivado de cambios legales que afecten a la tarifa y a las deducciones del IRPF, o el efecto redistributivo del IRPF y el grado de progresividad del gravamen a través de los índices habituales. La medición de los efectos recaudatorios y redistributivos de una reforma y sus implicaciones sobre el bienestar social se ven muy favorecidos cuando se dispone de datos de panel y se aplican las técnicas adecuadas de microsimulación sobre ellos.

Si además el panel se diseña con estratificación geográfica, es posible realizar estudios de ámbito territorial. De esta forma será posible, por ejemplo, estudiar las distintas alternativas de descentralización del IRPF como instrumento de financiación autonómica y los efectos redistributivos de la cesión del impuesto a las Comunidades Autónomas. Se pueden así medir los cambios distributivos de la renta en las CCAA derivados de sucesivas reformas, así como simular escenarios de descentralización de la imposición sobre la renta personal.

Para saber más…

- Progresividad y redistribución en el IRPF español: panel 82-98. Onrubia, Rodado, Sarralde, Pérez (Papel de trabajo IEF 23/2006).
- Panel de Declarantes: Castañer *et al.* (1999, 2000 y 2001), Onrubia y Rodado (2000).
- Microsimulación mediante fusión de phogue y panel de declarentes para evaluar reformas fiscales: fidel picos. Revista de economía aplicada, Nº 41 volumen 14 (2006).
- Modelos de microsimulación: aplicaciones a partir del panel de declarantes por IRPF del Instituto de Estudios Fiscales. Ayala, Onrubia y Ruiz Huerta. ICE nº 68 (2004).
- Documentos de trabajo IEF 15/05 y 20/06 La muestra de declarantes de IRPF 2002/2003: descripción general y principales magnitudes. Fidel Picos Sánchez, María Antiqueira Pérez, César Pérez López, Alfredo Moreno Sáez, Carmen Marcos García y Santiago Díaz de Sarralde Miguez.
- Técnicas de muestreo estadístico. Teoría, práctica y aplicaciones informáticas. César Pérez López. Rama (1999).
- Muestreo estadístico. Conceptos y problemas resueltos. César Pérez López. Pearson Educación – Prentice Hall (2005)
- Panel de renta del Instituto de Estudios Fiscales 1999/2007: César Pérez, Fidel Picos y Jorge Onrubia (IEF-2010).
- Técnicas de muestreo estadístico. César Pérez López. Garceta (2010).