

DISTRIBUCIÓN DEL INGRESO SEGÚN GÉNERO **Un enfoque no paramétrico**

Juana Z. Brufman*, Heriberto L. Urbisaia**, Luis. A. Trajtenberg***
Instituto de Investigaciones en Estadística y Matemática Actuarial
"Prof. Dr. Fausto I. Toranzos"
Facultad de Ciencias Económicas
Universidad de Buenos Aires
Av. Córdoba 2122

C1120AAQ – Ciudad de Buenos Aires - Argentina

*brufman@econ.uba.ar, **heribertourbisaia@speedy.com.ar, luis@econ.uba.ar***

Recibido 12 diciembre 2005, aceptado 15 de febrero 2006

Resumen

Las técnicas no paramétricas constituyen herramientas esenciales para el Análisis de Datos, sin imponer supuestos previos. El propósito de este trabajo es aplicar dichas técnicas al estudio de la Distribución del Ingreso Laboral, según género.

En la primera parte se aborda el análisis univariado: se ajustan funciones de densidad *kernel* a las respectivas distribuciones empíricas. Esta técnica constituye un afinamiento de los clásicos histogramas, caracterizados por presentar saltos o discontinuidades y ser sensibles a la forma y amplitud con que se definen los intervalos. El estimador de densidad *kernel* reemplaza los "rectángulos" del histograma por "protuberancias" suavizadas, mediante el uso de funciones de ponderación, denominadas *kernels*. Permite captar diferencias en tramos específicos de la distribución, según el interés del investigador.

En la segunda parte se efectúa un análisis bivariado. Se utiliza la Regresión No Paramétrica como alternativa de los modelos tradicionales de la Econometría: no presupone estructura alguna para la distribución del término de error o para la forma funcional que se estima. Los métodos no paramétricos aplicados al estudio del Ingreso Laboral de hombres y mujeres separadamente intentan detectar la existencia de la tan mentada "desigualdad salarial" según género. La variable es analizada según quintiles y se la cruza con variables relevantes, como son Educación y Experiencia Laboral, a los efectos de estimar funciones de regresión e inferir relaciones de causalidad. Los datos provienen de la Encuesta Permanente de Hogares, Ondas 1999 a 2003.

Palabras claves: histograma, amplitud del intervalo, función de densidad kernel, ancho de banda, parámetro de suavizado.

DISTRIBUTION OF LABOR INCOME BY GENDER A nonparametric approach

Juana Z. Brufman*, Heriberto L. Urbisaia**, Luis. A. Trajtenberg***
Instituto de Investigaciones en Estadística y Matemática Actuarial
"Prof. Dr. Fausto I. Toranzos"
Facultad de Ciencias Económicas
Universidad de Buenos Aires
Av. Córdoba 2122

C1120AAQ – Ciudad de Buenos Aires - Argentina

*brufman@econ.uba.ar, **heribertourbisaia@speedy.com.ar, luis@econ.uba.ar***

Received 12 november 2005, accepted 15 february 2006

Abstract

Nonparametric techniques have become essential tools for Data Analysis without imposing prior assumptions. The goal of this paper is to apply these methods to the study of the distribution of Labor Income for male and female separately.

In the first part an univariate analysis is performed; we estimate *kernel* density functions to fit to empirical distributions. This technique is a refinement of classical histograms, characterized by the presence of jumps or discontinuities and sensitiveness to the form and amplitude of the intervals or "bins". The *kernel density estimator* replaces the rectangulars of the histogram by smoothed "bumps", using weighting functions named *kernels*. This method allows to capture differences in specific sections of the distribution, according to the interest of the researcher.

In the second part we perform the analysis in a bivariate dimension. We use nonparametric Regression as alternative to traditional econometric models: it does not assume a particular structure for the error term or for the functional form of the model.

All these nonparametric techniques are applied for the study of Labor Income of male and female workers. The variable is analyzed by quintiles and is crossed with relevant variables as Education and Labor Experience, with the aim of estimate regression functions and infer causality relations between them. The data come from the EPH (Encuesta Permanente of Hogares), Waves 1999-2003.

Keywords: Histogram, Bin, Kernel density function, Bandwith, Smothed parameter.

PRIMERA PARTE: ANÁLISIS UNIVARIADO

INTRODUCCIÓN

En el presente trabajo se analizan en forma exploratoria y gráfica las similitudes y diferencias del ingreso laboral según género (varón-mujer), con el objetivo de arrojar luz sobre las características diferenciadas de sus distribuciones. Se aplican técnicas de análisis e inferencia no paramétricas, que requieren el uso intensivo de la computación. Posteriormente se comparan las conclusiones con algunos estadísticos, en especial, de orden. En esta primera parte se efectúa el estudio en dimensión univariada.

El análisis propuesto se basa preponderadamente en estadísticos ordinales: mediana, cuantiles y algunas de sus derivaciones, como la distancia intercuartil. La estimación de funciones de densidad completa el estudio y permite detectar diferencias en las distribuciones analizadas. La información procesada proviene de la Encuesta Permanente de Hogares (EPH), elaborada por INDEC, ondas Octubre/1999, Octubre/2001 y Octubre/2003.

El análisis es meramente descriptivo, omitiendo cualquier supuesto que sustente una evaluación inferencial. De esta manera se pone énfasis en el análisis exploratorio de la información como paso previo a la especificación de modelos con fines de inferencia.

Se efectuarán primeramente algunas consideraciones sobre los aspectos siguientes:

- Antecedentes de la utilización de métodos no paramétricos
- Análisis Exploratorio de Datos

- Estadísticos a utilizar
- Construcción de gráficos: de tallo y hojas y de caja-bigotes
- Estimación de funciones de densidad.

1. ANTECEDENTES DE LA UTILIZACIÓN DE MÉTODOS NO PARAMÉTRICOS

Tradicionalmente los métodos no paramétricos se aplicaron a la estimación y contraste de hipótesis de parámetros, sin sujetarse a supuestos previos sobre la distribución poblacional; se valieron de estadísticos ordinales: mediana, cuantiles, distancia intercuartil, etc.

De esta manera se diseñaron una serie de tests para contrastar hipótesis sobre parámetros, sobre aleatoriedad, y construcción de intervalos de confianza. Se substituyó la inferencia paramétrica ligada a la distribución normal por la de “distribución libre”.

La aparición de métodos de cómputo electrónico permitió ampliar esta metodología aplicándola a la estimación de funciones de densidad univariadas y, posteriormente, bivariadas.

La generalización de las técnicas de suavizado permitieron visualizar funciones de densidad, en lugar de histogramas, como así también comportamientos conjuntos de variables, sin sujeción a hipótesis pre-especificadas.

El denominado Análisis Exploratorio de Datos se complementa con una nueva generación de gráficos, como son los de “caja y bigotes” y fijación de criterios para identificar valores atípicos o “*outliers*” en el conjunto de

observaciones. Estas nuevas técnicas enriquecen el análisis cuantitativo de la información estadística y ayudan en la etapa de especificación de los Modelos Econométricos.

2. SOBRE EL ANÁLISIS EXPLORATORIO DE DATOS (AED)

El estadístico y matemático norteamericano John W. Tukey (1915-2000), fue el creador y figura central en la aplicación de esta metodología. Tukey comparó el Análisis Exploratorio de Datos con la actividad de un detective, consistente en recolectar evidencia y cuestionar supuestos para generar un *caso*, que sería luego contrastado formalmente en la *Corte* de la Inferencia Estadística.

El desarrollo de esta técnica fue posible gracias a la disponibilidad creciente de computadoras electrónicas; en este caso, por la rapidez y eficacia de las representaciones gráficas, que apuntan a descubrir estructuras subyacentes en los datos, a la vez que detectar desviaciones importantes respecto de ellas.

El AED no requiere la especificación previa de supuestos ó modelo estadístico alguno; comienza con el examen del conjunto de datos a partir de diferentes tipos de representaciones: gráfico de tallo y hojas (*stem-and-leaf display*), diagrama de caja y bigotes (*box and whisker plots*), etc., basados en estadísticos ordinales.

3. ESTADÍSTICOS A UTILIZAR

Para el análisis exploratorio de datos y construcción de diagramas, se utilizan estadísticos ordinales, robustos ante la presencia de valores

extremos. En primer lugar: la mediana, (M_e), el primer cuartil, (Q_1) y el tercer cuartil, (Q_3), que representan características de posición en la distribución. En la terminología de Tukey, Q_1 y Q_3 se denominan “bisagras” (*hinges*).

En segundo lugar, como medida de variabilidad, se utiliza el rango intercuartil: diferencia entre el tercer y primer cuartil: $[Q_3 - Q_1]$.

Para el diagrama de caja, se definen además:

a.- “Peldaño” o “escalón” (*step*): valor resultante de tomar 1.5 veces el rango intercuartil: $1.5[Q_3 - Q_1]$.

b.- “Vallas Interiores” o “Umbrales superior e inferior” (*inner fences*):
 $U_{sup.} = Q_3 + 1 \text{ peldaño}$; Umbral inferior: $U_{inf.} = Q_1 - 1 \text{ peldaño}$.

c.- “Vallas Exteriores” (*outer fences*): se obtienen tomando 2 veces el peldaño.

d.- “Valores adyacentes”: son dos valores, uno inferior y otro superior, los más cercanos al límite de las vallas interiores, pero *dentro* de éstas.

e.- “Valores atípicos” (*outside values ó outliers*): todo dato que excede los límites que determinan las “vallas interiores”.

f.- Valores fuera de las “vallas exteriores” se denominan “marcadamente atípicos”. (*far out values*).

4. REPRESENTACIONES GRÁFICAS

4.1. Gráfico de tallo y hojas

Es un procedimiento semi-gráfico para presentar la información de variables cuantitativas, útil en especial, cuando la cantidad total de datos es pequeña (menor que 50).

Para construir el gráfico se procede de la siguiente manera:

a.- Se redondean los datos a dos ó tres cifras significativas, expresándolos en unidades convenientes.

b.- Se los presenta en una tabla de dos columnas, de manera que:

b.1.- Si los datos son de dos dígitos, se escribe en la primer columna los dígitos de las decenas —que forman el tallo — y en la segunda columna los dígitos de las unidades. Por ej. el dato 76 se escribiría: $7|6$.

b.2.- Si los datos son de tres dígitos, los correspondientes a las centenas y decenas se escriben en la columna izquierda —que constituye el tallo — y los dígitos de las unidades en la columna derecha.

c.- Cada *tallo* define una *clase* y se escribe una sola vez. El número de *hojas* es representativo de la *frecuencia* de cada *clase*.

El gráfico de tallo y hojas permite siempre reconstruir la información de origen; con el histograma, en cambio, se pierde información en la medida en que se incrementa la amplitud del intervalo de clase.

A título de ejemplo: supóngase la información correspondiente al Ingreso Anual de 16 familias, expresado en miles de pesos:

113.57 125.42 113.84 124.31 142.12 152.13 133.00 113.00
172.06 127.10 134.55 161.43 121.62 127.21 134.20 146.98

Redondeando la información para evitar decimales resultan los datos:

114 125 114 124 142 152 133 113
172 127 135 161 122 127 134 147

y el gráfico de tallo y hojas muestra la información de la manera siguiente:

Decenas	Unidades
11	443
12	54727
13	354
14	27
15	1
16	1
17	1

Tabla 1. Gráfico de tallo y hojas

4.2. Diagrama de caja y bigotes

Se trata de una representación semi-gráfica del conjunto de observaciones, construida sobre la base del resumen de *cinco valores* vinculados a las características de posición de la distribución que son mediana, primer y tercer cuartil, y sus extremos: los valores mínimo y máximo de los datos.

El diagrama de *caja y bigotes* es una representación simple de estos cinco números, que sintetizan suficiente información acerca de la distribución de la variable, de modo que permite detectar características de forma y observaciones atípicas. No requiere agrupar o promediar datos, como ocurre en el histograma y es robusta ante la existencia de valores atípicos.

Sin reemplazar al *histograma*, constituye un buen sustituto del mismo, ya que permite al analista visualizar la distribución “de un golpe”.

Un diagrama de *caja y bigotes* se construye de la siguiente forma:

1.- Se ordenan los datos según su magnitud y se determinan: el valor máximo, el mínimo, la mediana y cuartiles primero y tercero.

2.- Se dibuja un rectángulo de base igual a la *diferencia intercuartil* y altura convencional; se indica la posición de la M_e mediante una línea divisoria dentro del rectángulo.

3.- Se calculan los umbrales superior e inferior:

$$U_{sup.} = Q_3 + 1.5[Q_3 - Q_1] ; U_{inf.} = Q_1 - 1.5[Q_3 - Q_1]$$

4.- Se trazan líneas desde cada extremo del rectángulo central hasta los valores adyacentes inferior y superior (estas líneas son los *bigotes* de la caja)

5.- Se marcan los datos que están fuera de los umbrales inferior y superior, como valores atípicos.

Gráficos de este tipo se incluyen en el Anexo Gráfico del trabajo.

5. ESTIMACIÓN DE FUNCIONES DE DENSIDAD KERNEL.

El ajuste de funciones de densidad *kernel* a las respectivas distribuciones empíricas permite captar diferencias en tramos específicos de la distribución, según el interés del investigador. Esta técnica constituye un afinamiento de los clásicos histogramas.

Como se sabe, el concepto de función de densidad de una variable aleatoria X puede representarse mediante la expresión:

$$f(x) = \lim_{h \rightarrow 0} \frac{1}{2h} P(x-h < X < x+h) \quad (1)$$

siendo el *histograma* la forma más simple de estimación no paramétrica: si se dispone de n datos agrupados en intervalos de amplitud $2h$, es posible aproximarnos a $f(x)$ mediante:

$$\hat{f}(x) = \frac{1}{2h} \frac{(\text{frecuencia de datos en } x \pm h)}{n} \quad (2)$$

Si bien se trata de una estimación simple, tiene algunos inconvenientes: *i)* es sensible a la elección del origen a partir del cual se definen los intervalos; *ii)* es constante dentro del intervalo; *iii)* considera sólo los datos dentro de cada intervalo, ignorando los adyacentes, por próximos que éstos sean.

Este último inconveniente puede evitarse otorgando cierto peso a los datos de intervalos contiguos al que se estima, lo que conduce a una estimación *más suave*.

Para aproximarnos al concepto de estimación de densidad *kernel*, introducimos los siguientes cambios en materia de terminología y notación:

i) La amplitud del intervalo del histograma ($2h$) se denomina “*bin*” en el contexto de la estimación kernel.

ii) La semi-amplitud del intervalo (h) se denomina ancho de banda o “*bandwidth*” y constituye el parámetro de suavizado, asociado al *Kernel*

iii) Se indica con $z = \frac{x - x_i}{h}$, la diferencia entre el punto genérico x y el centro del intervalo x_i , estandarizada en unidades h , es decir en unidades de la semi-amplitud del intervalo.

iv) Se introduce una función de ponderación w , la cual, para el caso del histograma, se define de la siguiente manera:

$$w(z) = h \text{ si } |z| < 1; \quad w(z) = 0 \text{ en cualquier otro caso}$$

Con esta notación el histograma (2) puede expresarse:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} w\left(\frac{x - x_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} w(z) \quad (3)$$

Nótese que la función de ponderación $w(z)$ utilizada implica que solo las observaciones que caen en el intervalo de amplitud h se toman en cuenta para la estimación de la frecuencia relativa en ese intervalo.

La forma de este estimador “ingenuo” no depende de la elección del origen de los intervalos; mantiene, no obstante, la ponderación constante dentro de cada intervalo; por tanto, no es aún un estimador *suave*; su gráfica resulta de aspecto *desdentado*, con saltos en los bordes de los intervalos y *derivadas nulas* fuera de cada intervalo.

Al decir de Silverman, el estimador “ingenuo” o estimador “kernel rectangular” es el *primer paso para llegar a un histograma donde “cada punto es el centro de un intervalo muestral”*.

Una representación *suavizada* de la distribución empírica requiere ponderar las observaciones en forma decreciente, a medida que aumenta la diferencia $z = (x - x_i)/h$. Para alcanzar este objetivo, se utiliza como ponderador, una *función continua*, K , denominada *Kernel* o *función núcleo*, simétrica, centrada en cero, tal que:

$$\int_{-\infty}^{\infty} K(x) dx = 1 \quad (4)$$

Se trata de una función *suave*, con *derivadas no nulas*, que asigna a cada observación un peso positivo menor que la unidad, decreciente a medida que aumenta z , es decir, la diferencia: $(x - x_i)/h$. Por lo tanto la (3) resulta:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K(z)$$

Veamos cómo opera el nuevo estimador. Supóngase una variable x , de la que se disponen n observaciones: x_1, x_2, \dots, x_n , cada una de ellas con frecuencia relativa $1/n$. La función de densidad $f(x)$ en un punto x_0 , se estima como suma ponderada de las frecuencias $1/n$ de todas las observaciones; la ponderación para la observación x_i se determina por la ordenada de la función *kernel* elegida, evaluada en $(x_i - x_0)/h$.

Considerando, por ejemplo, la distribución normal como función *kernel*, es evidente que para un h fijo, la mayor ponderación corresponderá a la observación que coincida con x_0 ; el peso será cada vez menor, a medida que se consideren observaciones x_i más alejadas de x_0 .

Veamos con un ejemplo lo explicado hasta aquí.

El histograma del Gráfico 1 representa la distribución del Ingreso Per Cápita Familiar, de 200 familias, según datos de la EPH, onda Octubre 1999. Obsérvese el aspecto discontinuo y *desdentado*, generado por ausencia de frecuencias en intervalos correspondientes, en este caso, a valores altos de la variable.

El Gráfico 2 representa la estimación kernel, apreciándose el grado de suavizado logrado.

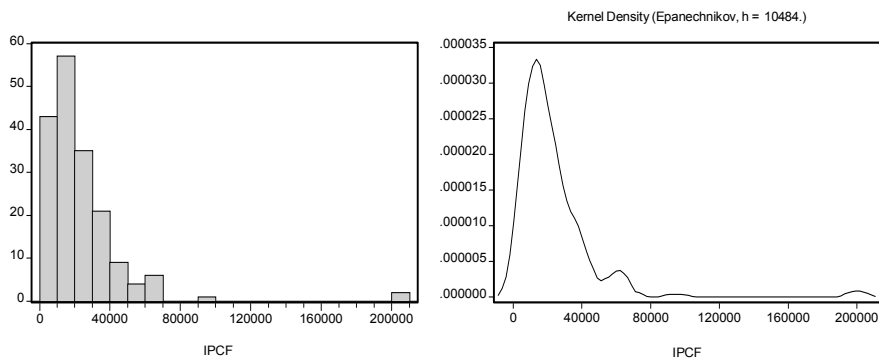


Gráfico 1: Histograma

Gráfico 2: Funciones de densidad estimada

Intuitivamente la idea central del estimador de densidad *kernel* es reemplazar los rectángulos del histograma por “*protuberancias*” suavizadas. En el *histograma*, la altura de cada rectángulo es proporcional

a la frecuencia relativa *dentro* del mismo y se asigna al *punto medio* del intervalo de clase. Las “*protuberancias*” suavizadas, resultado de la estimación por *kernel*, quedan determinadas sobre todo el recorrido de la variable en estudio.

Dos aspectos claves diferencian el estimador *kernel* del *histograma*:

En primer lugar, en el *histograma*, los intervalos *no se superponen*; el estimador *kernel*, en cambio, admite *bins solapados*. Este aspecto quiebra, por así decirlo, el eslabón existente entre *tamaño* y *centro* del intervalo, propio del *histograma*.

En segundo lugar, el *histograma* asigna igual ponderación a todas las observaciones pertenecientes al intervalo; el estimador *kernel* pondera los datos en forma decreciente, a medida que las observaciones se alejan del *centro* del *bin*.

Si se tiene presente la vinculación entre amplitud de un intervalo (en el *histograma*) y ancho de banda (*bandwidth*) del *kernel*, resulta que:

i) en el *histograma*, el mayor o menor detalle en la representación se determina controlando la amplitud del intervalo de clase, a mayor amplitud, menor detalle, *ii)* en el *kernel*, el grado de suavizado se controla seleccionando el parámetro *h*, mayor suavizado exige menor amplitud de banda.

En estudios empíricos, se utilizan preferentemente las siguientes funciones *kernel*, indicadas en la Tabla 2.

Epanechnikov	$\frac{3}{4}(1-z^2)I(z \leq 1)$
Normal (Gaussiana)	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right)$
Biweight	$\frac{15}{16}(1-z^2)^2 ; z < 1;$ 0 en cualquier otro lado

Tabla 2. Principales Funciones *Kernels*

Debe tenerse presente que el resultado de la estimación no depende tanto del *kernel* elegido, sino de la acertada elección del parámetro h . Este parámetro se denomina *amplitud de banda* y cumple la función de controlar el proceso de suavizado de los rectángulos característicos del *histograma*.¹ A mayor valor de h , menor será el grado de suavizado. La elección de este parámetro es de crucial importancia, habiéndose sugerido diferentes criterios para su determinación. La opción de Silverman es la siguiente:

$$h = 0.90 pn^{-0.20} \min(s, R/1.34)$$

donde n es la cantidad de observaciones, s es la desviación estándar y R es el rango intercuartil de la serie. El factor p ajusta el *ancho de banda*, de modo que el nivel de suavizado sea homogéneo entre diferentes funciones *kernel*.

¹ Decía Fisher al respecto: "Para alcanzar una verdadera curva, no sólo se debería ubicar una cantidad infinitamente grande de observaciones en cada clase, sino que el número de clases en que se divide la población debe ser infinito". Fisher, R.A. (1922), pp.312

En síntesis, el estimador de densidad *kernel* reemplaza los “rectángulos” del histograma por “protuberancias” suavizadas; el *histograma suavizado* constituye el puente entre el conjunto de observaciones (x_1, x_2, \dots, x_n) y el concepto teórico de una función de densidad $f(x)$. La función de densidad estimada es el resultado de un *trade-off* entre bondad de ajuste a los datos y grado de suavizado elegido.

SEGUNDA PARTE: ANÁLISIS MULTIVARIADO

En esta segunda parte continuamos el análisis no paramétrico sobre la distribución del ingreso laboral según género, abordando el problema en dimensión multivariada: se cruza el Ingreso con variables relevantes tales como Nivel de Educación y Experiencia Laboral, a los efectos de estimar funciones de regresión no paramétricas e inferir relaciones de causalidad.

6. CONSIDERACIONES GENERALES SOBRE LA REGRESIÓN

La teoría de la regresión intenta indagar sobre presuntas relaciones causales entre diversas variables que afectan a un determinado fenómeno. Para el caso de dos variables y en forma general:

$$y = m(x) + \varepsilon \quad (5)$$

$m(.)$ es una función de forma matemática desconocida que representa el valor medio de y , condicionado a los valores de x :

$$E(y/x) = m(x) \quad (6)$$

y en donde ε es un término de error.

En la regresión paramétrica, el significado de ambas componentes es el siguiente:

1.- La primera de ellas $m(x)$ recibe el nombre de parte funcional de la ecuación, con forma algebraica pre-especificada. Por ejemplo, suponer entre x e y una relación funcional lineal: $m(x) = \alpha + \beta x$ implica fijar un supuesto previo respecto a la forma en que se vinculan ambas variables. Otras funciones se obtienen mediante transformaciones de las variables (logaritmos, inversas, etc.)

2.- La segunda componente denominada término de error y más modernamente componente aleatoria, viene a resumir una serie de factores, como pueden ser:

- Error de especificación de la función elegida.
- Omisión de variables relevantes.
- Errores de medición.

Los parámetros α y β son desconocidos y se estiman a partir de las observaciones muestrales, mediante el uso de estimadores.

Para asegurar las propiedades deseables de los estimadores $\hat{\alpha}$ y $\hat{\beta}$, se imponen condiciones respecto al comportamiento del término aleatorio, que generalmente se engloban dentro de lo que se denomina Supuestos de Gauss-Markov.

Se señala como inconveniente del método anterior, la necesidad de fijar demasiados supuestos: a menudo el diagrama de dispersión entre las variables es confuso, máxime cuando se dispone de muchas

observaciones, que no permiten entrever una relación y por tanto, no sugiere forma funcional alguna. Así por ejemplo, se admite una relación entre el Nivel de Educación y el Ingreso, pero de allí presuponer una relación lineal es arriesgar una hipótesis demasiado fuerte.

7. SOBRE LA REGRESIÓN NO PARAMÉTRICA

La Regresión No Paramétrica constituye una alternativa diferente respecto al enfoque anterior; difiere de los modelos tradicionales de la Econometría pues no presupone estructura alguna para la distribución del término de error o para la forma funcional que se estima.

Retomando las expresiones (5) y (6):

$$y = m(x) + \varepsilon \quad (5)$$

$$E(y/x) = m(x) \quad (6)$$

la estimación no paramétrica m^* se obtiene mediante técnicas de suavizado aplicadas localmente a los pares de observaciones (x_i, y_i) , $i = 1, 2, \dots, n$; el procedimiento es similar al utilizado en la estimación de funciones de densidad univariada: el valor medio condicional para un intervalo pequeño de x se estima, no sólo con las observaciones de dicho intervalo, sino con las de intervalos adyacentes; esta información se pondera en forma decreciente a medida que es mayor la distancia de la observación respecto al centro del intervalo; como puede apreciarse, esto implica una gran carga de cálculo.

Otras características diferenciales son:

- De la regresión paramétrica resulta una función analítica, que además de su representación gráfica, permite estimar valores promedio de y para los cuales no existen observaciones muestrales, como también predecir para valores fuera del espacio muestral.
- De la regresión no paramétrica, resulta un gráfico definido sobre la muestra, que no responde a expresión analítica alguna y que, a lo sumo, podrá superponerse a otro con iguales variables.
- La regresión no paramétrica es un método intermedio entre el análisis gráfico y la inferencia paramétrica. Actualmente, junto con el análisis exploratorio de datos, constituye el punto de partida de cualquier estudio cuantitativo; recién cuando se ha logrado formar una idea acerca del comportamiento de los datos, se inicia la búsqueda de mayores conclusiones mediante la estimación paramétrica.

7.1. Regresión Lineal Local (Loess)²

La idea básica en el método de regresión no paramétrica es la de regresión local y en particular, la regresión local lineal.

Suponiendo la relación (5):

$$y = m(x) + \varepsilon$$

uno de los métodos flexibles para estimar $m(x)$, consiste en obtener estimaciones locales de la función, para una cantidad importante de valores de x , partiendo de una aproximación lineal de la función a estimar. Describimos el método comenzando por la estimación $\hat{m}(x_0)$ en un punto x_0 .

Si se admite que $m(x)$ es una función “suave” y diferenciable en x_0 , puede aproximarse en el entorno de dicho punto, por la función lineal:

$$m(x) \approx \alpha_0 + \beta_0(x - x_0) \quad (7)$$

siendo $\alpha_0 = m(x_0)$ y β_0 la derivada de $m(x)$ en el punto x_0 . El método utiliza las observaciones (x_i, y_i) , que corresponden a valores de x_i cercanos a x_0 , de modo de poder estimar los parámetros α_0 y β_0 del modelo:

$$y_i = \alpha_0 + \beta_0(x_i - x_0) + \varepsilon_i \quad (8)$$

² Del alemán *loess*, sigla de “*lokal regression*”; a menudo traducido como *lowess* (**L**ocal **W**eighted **S**catter plot **S**moother)

en el cual ε_i representa un término de error, dado que la expresión lineal es solo una aproximación de la función a estimar.

Si el punto x_0 está presente entre las observaciones, la observación i_0 será $x_{i_0} = x_0$; entonces:

$$E(y_{i_0} / x_{i_0}) = m(x_{i_0}) = \alpha_0 \quad (9)$$

Por lo tanto, la estimación de α_0 se interpreta como la estimación de la función $m(x_{i_0})$.

La aproximación lineal será más precisa para valores de x_i próximos a x_0 , lo que requiere ponderaciones diferenciadas de las observaciones: mayor ponderación para las más cercanas a x_0 y menor peso para las alejadas de x_0 . Por tanto, para estimar α_0 y β_0 de la ecuación (8) se utilizará mínimos cuadrados ponderados en lugar de mínimos cuadrados clásicos: La función a minimizar resulta:

$$\sum_i w_i [y_i - \alpha_0 - \beta_0(x_i - x_0)]^2 \quad (10)$$

Observaciones demasiado alejadas de x_0 , reciben una ponderación $w_i = 0$, ya que dichas observaciones no suministran información confiable sobre $m(x_0)$. Observaciones x_i para las cuales la diferencia $|x_i - x_0|$ es menor que cierto umbral prefijado, se incluyen en la regresión local, con ponderaciones a determinar por el investigador, dentro de una gama de opciones disponibles.

El método descrito se generaliza al considerar la posibilidad de aproximar funciones polinómicas de orden superior a la lineal: cuadráticas, cúbicas, etc., en el entorno de x_0 .

Las ponderaciones w_i pueden determinarse según los siguientes enfoques:

i) Especificando una cantidad fija de pares de observaciones (x_i, y_i) , que se correspondan con los valores x_i más cercanos a x_0 ; esta cantidad fija, controlada mediante el parámetro λ , se incluye en cada regresión local independientemente de cuán distante se halle x_i respecto a x_0 .

ii) Fijando un ancho de banda fija $|x_i - x_0| = h$, para todas las regresiones locales. El parámetro h caracteriza la función kernel seleccionada para el cálculo de las ponderaciones w_i . En este caso, cada regresión local incluye una cantidad variable de observaciones, dependiendo de la distribución de x_i .

7.2. Método de Regresión Local con Ajuste de las Observaciones más cercanas (*Local Regression with Nearest Neighbour Fit*)

Este método corresponde al primero de los enfoques señalados. El parámetro λ , comprendido entre 0 y 1, se denomina *span* y representa la fracción de las n observaciones (x_i, y_i) que se incluyen en cada regresión local; se eligen λn observaciones con valores x_i más cercanos a x_0 . Habitualmente se toma $\lambda = 0.3, 0.6$ ó 0.7 . El parámetro λ controla el grado

de suavizado: a mayor λ , se incluye mayor cantidad de observaciones en cada regresión y por tanto, mayor es el suavizado logrado.

Respecto a las ponderaciones se utiliza preferentemente la función tricúbica:

$$w_i = (1 - d_i^3)^3; \quad (11)$$

siendo: $d_i = \frac{|x_i - x_0|}{D}$; y D la distancia máxima $|x_i - x_0|$ que se presenta en las λn observaciones de cada regresión. Por lo tanto $0 \leq d_i \leq 1$. La ponderación máxima ocurre cuando $d_i = 0$, ($x_i = x_0$) y decrece gradualmente hacia 0, cuando $d_i \rightarrow 1$

Repitiendo el procedimiento descrito para una variedad de valores de la variable explicativa, es posible obtener una estimación completa de $m(x)$.

7.3. Método de Regresión Kernel

Este método corresponde al segundo de los enfoques mencionados; se caracteriza por utilizar una banda fija de amplitud h para todas las regresiones locales. Nótese, sin embargo que, al fijar $|x_i - x_0|$ para todas las regresiones locales, cada una de ellas procesa una cantidad variable de observaciones (x_i, y_i) . Las ponderaciones resultan de la aplicación de una función kernel cuyo parámetro de suavizado es precisamente $h = |x_i - x_0|$.

Formalmente, el estimador de regresión lineal local $m(x_0)^*$ se indica como α_0^* , el cual minimiza la función:

$$\sum_{i=1}^n \left[K \left(\frac{x_i - x_0}{h_n} \right) (y_i - \alpha_0 - \beta_0 (x_i - x_0))^2 \right] \quad (12)$$

Repetiendo el procedimiento descripto para una variedad de valores de la variable explicativa, es posible obtener una estimación completa de $m(x)$.

Al igual que lo señalado para la estimación de funciones de densidad kernel, la aplicación del método descripto requiere:

- a) fijar la “amplitud de banda” o *bandwidth*, que controla el grado de suavizado que se pretende al efectuar la estimación.
- b) elegir la función kernel que determinará las ponderaciones w_i . Cabe reiterar lo expresado en esa oportunidad, en el sentido de la baja incidencia que esta elección tiene en el resultado de la estimación, siendo en cambio importante una buena calibración de h .

Para ilustrar los conceptos expresados, presentamos los siguientes ejemplos.

Se generaron por simulación 100 observaciones de las series:

$$x \sim U(0,1); \varepsilon \sim N(0,1); y = \text{sen}(2\pi(1-x)^2) + x\varepsilon$$

Se trata de estimar la función $y_i = m(x_i) + \nu_i$ siendo ν_i la perturbación aleatoria del modelo.

En los gráficos 3, 4 y 5 – *scatterplots* – se muestra la nube de puntos de las observaciones y diferentes estimaciones logradas aplicando los métodos descriptos. Se incluye, con fines comparativos, una regresión paramétrica de ajuste lineal.

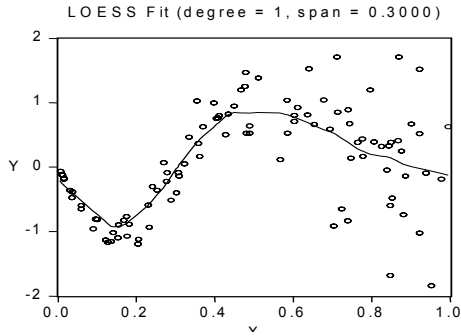


Gráfico 3

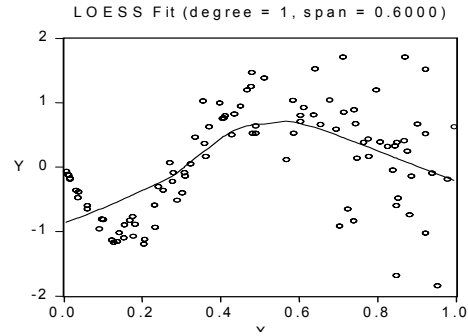


Gráfico 4

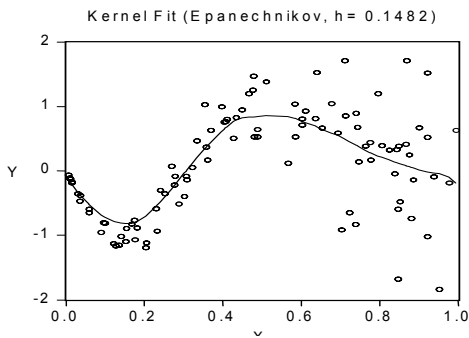


Gráfico 5

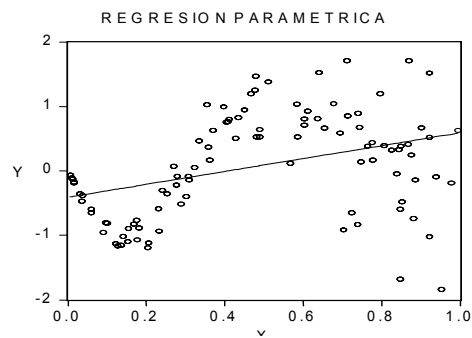


Gráfico 6

Obsérvese en los gráficos 3 y 4, el cambio operado en la calidad del ajuste, al pasar de un *span* 0.3 a 0.6: es evidente el sobre-suavizado, especialmente en los valores bajos de la variable x .

Por otra parte, si bien en todos los casos se aplicó una aproximación lineal, la estimación paramétrica resulta en este caso notablemente inferior, dada la curvatura de la función sinusoidal que generó las observaciones.

Veamos con un ejemplo lo explicado hasta aquí. Con datos de la EHP, Onda Octubre de 1999, correspondientes a 200 familias de Capital y Gran Buenos Aires, se estimaron regresiones no paramétricas, relacionando el Ingreso per-cápita familiar, en miles de pesos (IPCF en miles de pesos) con Edad (H12). Se muestran gráicamente los resultados obtenidos.

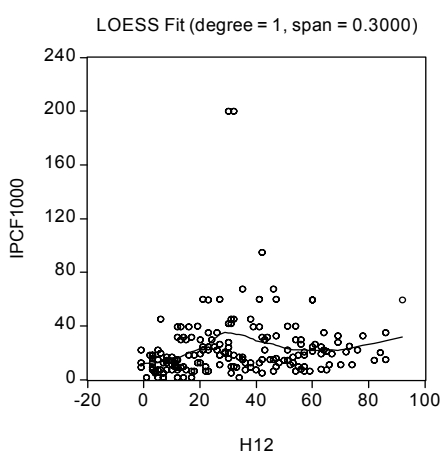


Gráfico 7

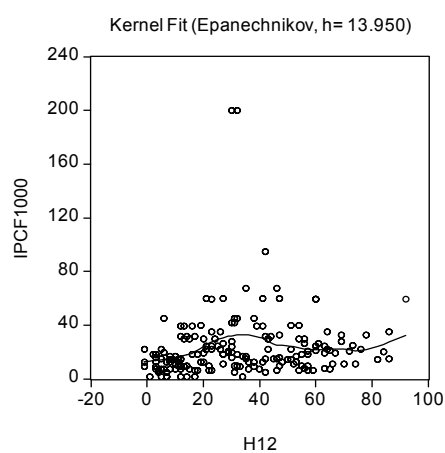


Gráfico 8

8. ANÁLISIS EMPIRICO

Aplicación de las técnicas no paramétricas en el estudio de la distribución del ingreso según género

A continuación se analizan y comparan los resultados del estudio propuesto, utilizando la información de la Encuesta Permanente de Hogares, Ondas 1999 a 2003.

El análisis de los resultados se efectuó en un doble plano numérico y gráfico. El numérico se basa en la comparación de los estadísticos ordinales; un primer objetivo fue detectar coincidencias y/o diferencias en los perfiles de la distribución de ingresos varón-mujer dentro de *cada* onda; en segundo lugar se pasó al análisis *entre* ondas para ver, de esta manera lo acontecido en la Onda de Mayo 2003, cuando ya se registran los efectos de la pesificación.

En el análisis de distribuciones asimétricas, la diferencia entre media y mediana es tradicionalmente explotada a los efectos de entrever la intensidad de la asimetría y ubicación de las colas de la distribución.

Si se adoptara como umbral de pobreza la mediana, puede afirmarse que: de dos distribuciones de igual mediana, una de ellas con asimetría positiva y la otra negativa, es preferible, desde el punto de vista del bienestar, la distribución asimétrica positiva, por cuanto en ella la cola extendida del 50% de observaciones corresponde a los valores mayores de la variable en estudio.

Al plano cuantitativo, le siguió el análisis gráfico mediante los diagramas anteriormente descriptos con el fin de visualizar y mejorar la apreciación de los resultados numéricos.

Se aclara que este tipo de análisis es meramente descriptivo. No se intenta sentar hipótesis alguna y menos aún efectuar inferencias a partir de los hechos observados. De cualquier manera el análisis permite diagnosticar y puntualizar aspectos de la distribución de los ingresos que pueden ser considerados al especificar modelos para el diseño de políticas distributivas.

8.1.- Resultados Numéricos

La tabla 2 sintetiza las características de las Distribuciones de Ingreso: Total, Varones y Mujeres, para cada una de las ondas consideradas.

	Oct.-99	Oct.-01	May.-03
Ingreso Medio			
Total	4,48	4,57	4,44
Varones	4,50	4,63	4,66
Mujeres	4,44	4,49	4,16
Mediana			
Total	3,13	3,13	3,00
Varones	3,13	3,00	2,92
Mujeres	3,47	3,33	3,075
Desvio Standard			
Total	4,43	4,86	5,30
Varones	4,61	5,19	6,32

Mujeres	4,15	4,35	4,62
Asimetría			
Total	4,35	4,42	6,32
Varones	4,11	4,26	6,15
Mujeres	4,76	4,63	6,80

Tabla 2. Distribuciones completas. Estadísticos resumen

El análisis de los guarismos hallados permite efectuar los siguientes comentarios:

- 1.- Para las ondas de 1999 y 2001, el ingreso medio de los varones resulta levemente superior al de las mujeres. En la onda de 2003 esa diferencia se profundiza llegando al 10%.
- 2.- Como toda distribución de ingresos, ambas son marcadamente asimétricas, con mayor preponderancia de esta característica en la de mujeres. En el año 2003 se agudiza la asimetría tanto para varones como mujeres lo que implica un aumento de la desigualdad.
- 3.-La mediana para ambos géneros y en todos los casos resulta inferior a la media poniendo de relieve la dirección *positiva* de la asimetría.
- 4.-El desvío estándar para ambos géneros aumentó significativamente en el 2003 detectándose el impacto de la crisis del año anterior sobre la distribución de todos los preceptores

Para profundizar el estudio en tramos diferenciales del ingreso, se analizaron las distribuciones por quintiles, poniéndose énfasis en la comparación media y mediana y omitiendo diferencias no significativas.

	Oct. 99		Oct. 01		May. 03	
Quintiles Totales	media	mediana	media	mediana	media	mediana
Primero	1,29	1,36	1,16	1,25	1,25	1,39
Segundo	2,28	2,31	2,15	2,08	2,09	2,07
Tercero	3,23	3,13	3,13	3,13	2,97	3,00
Cuarto	4,75	4,69	4,77	4,69	4,59	4,48
Quinto	10,96	8,33	11,57	9,38	11,35	8,68
Quintiles-Varones						
Primero	1,33	1,41	1,17	1,25	1,23	1,34
Segundo	2,26	2,29	2,09	2,08	2,09	2,08
Tercero	3,14	3,13	2,98	3,03	2,96	2,98
Cuarto	4,67	4,44	4,63	4,44	4,62	4,44
Quinto	11,72	9,30	12,39	10,00	13,11	9,47
Quintiles-Mujeres						
Primero	1,27	1,28	1,17	1,25	1,32	1,50
Segundo	2,36	2,43	2,30	2,29	2,12	2,08
Tercero	3,49	3,47	3,42	3,33	3,09	3,13
Cuarto	5,00	5,00	5,00	5,00	4,71	5,00
Quinto	10,30	8,33	10,60	9,00	9,58	8,13

Nota: En cada par media-mediana, se indica en negrita el valor mayor.

Tabla 3. Distribuciones por quintiles: Estadísticos resumen

Los resultados se consignan en el Tabla 3, de los que surgen las consideraciones siguientes:

1.- Para todas la ondas y para ambos géneros, el primer quintil registra un ingreso medio inferior a la mediana, lo que revela una asimetría *negativa*.

- 2.- En el segundo quintil el comportamiento es errático: en algunos casos ambos estadísticos son iguales y en otros cambian la dirección de la desigualdad; se identifican así los estratos más proclives a cambios en la distribución de ingresos.
- 3.- En el tercer quintil, la distribución de ingresos de varones es relativamente simétrica en las tres ondas; la distribución mujeres se torna simétrica recién en el cuarto quintil. En estos tramos de la variable, por tanto, se produce un comportamiento rezagado del ingreso de mujeres respecto al de varones.
- 4.- El cuarto y último quintil de la distribución varones muestra una profunda asimetría; en la de mujeres este rasgo aparece sólo en el último quintil.
- 5.- La salida de la convertibilidad distorsionó marcadamente la distribución de ingresos de las mujeres; en los quintiles tercero y cuarto se aprecia un cambio en el sentido de la asimetría: de positiva a negativa; sólo el último quintil mantuvo la asimetría positiva.
- 6.-En general podemos decir, que en los tramos centrales de ingresos, la media y mediana de la distribución mujeres superan a la de varones. Sin embargo la desigualdad del extremo superior en favor de los varones es tan marcada, que vuelca el promedio general a favor de estos últimos.

8.2. Análisis Gráfico

8.2.1. Gráficos de Cajas

El análisis de los gráficos de cajas correspondientes a las distribuciones analizadas es coherente con lo expresado más arriba: valores atípicos o

outliers se presentan en la escala *superior* de ingresos de los varones y en los tramos *inferiores* del ingreso de mujeres; para éstas últimas ello pone en evidencia el cambio en la asimetría apuntado.

8.2.2. Funciones de densidad Kernel

Los gráficos de las funciones de densidad *kernel* muestran en todas las ondas:

i) un desplazamiento de la distribución de ingresos de mujeres hacia la derecha de la de varones, en los *tramos centrales* de la distribución; de ahí que los ingresos de las mujeres pueden considerarse mejores, en el tramo medio de la variable en estudio.

ii) mayor proporción de varones en el extremo superior de los ingresos y mayor proporción de mujeres en el tramo inferior.

En la regresión no paramétrica se relacionó separadamente: Ingreso con Años de Escolaridad e Ingreso con Experiencia Laboral. En ambos casos se aplicó el kernel Gaussiano.

Respecto al Ingreso se consideraron perceptores con edades entre 18 y 65 años.

Para la variable Años de Escolaridad se consideró la siguiente escala, según nivel de instrucción alcanzado:

i) tres años: Primario Incompleto; *ii)* 7 años: Primario Completo; *iii)* 9 años: Secundario Incompleto; *iv)* 12 años: Secundario completo; *v)* 14.5 años: Universitario Incompleto; *vi)* 17 ó más: Universitario completo.

La experiencia laboral se mide también en años, de acuerdo con la combinación:

Años de Experiencia Laboral = Edad + Años de Escolaridad - 6.

Respecto a la primera de las regresiones, la función estimada por el método kernel capta la incidencia creciente de los Niveles de Instrucción sobre el Ingreso, a tasas también crecientes. Si se observa el diagrama de dispersión de las observaciones, está claro que un ajuste paramétrico lineal no hubiera puesto en evidencia este hecho.

Para la variable Experiencia Laboral, la estimación kernel muestra un tramo inicial de efecto ascendente sobre el Ingreso, el cual se agota en torno a los 35 años; a partir de allí se inicia un marcado descenso. Esta trayectoria se explica si se tiene en cuenta la forma en que se calcula la variable: tanto la Edad como Años de Escolaridad intervienen aditivamente en la determinación de la Experiencia Laboral; en los años iniciales de incorporación al mercado laboral, ambos términos de la ecuación crecen y generan mayor Experiencia; a partir de cierto momento del ciclo de vida, el factor Escolaridad se estanca: el aumento de la Experiencia sólo se debe al aumento de la Edad, precisamente en la etapa en que la mayor edad incide negativamente sobre los niveles de ingreso.

REFERENCIAS

- [1] Dinardo J.; J. L. Tobías (2001). "Nonparametric Density and Regression Estimation". *Journal of Economic Perspectives*. Vol 15 N° 4, pp.11-28.

- [2] Dinardo J.; N.M. Fortin; T. Lemieux (1996). "Labor Markets Institutions and the Distribution of Wages 1973-1993. A Semiparametric Approach". *Econometrica* N° 64 Vol.5. pp.1001-1044.
- [3] Fortin N.M.; T. Lemieux (2000). "Are Women Wages Gains Men's Loses?" *American Economic Review. Papers and Proceeding.* pp.456-460.
- [4] Fisher, R.A. (1922). "On the mathematical foundations of theoretical statistics". *Philosophical Transactions of the Royal Society A.* 222. pp.309-368.
- [5] Fox, J. (2000). *Nonparametric Simple Regression: Smoothing Scatterplots.* SAGE University Papers. Series Quantitative Applications in the Social Sciences. Londres.
- [6] Fox, J. (2000). *Multiple and Generalized Nonparametric Regression.* SAGE University Papers. Series Quantitative Applications in the Social Sciences.
- [7] Härdle W. (1995). *Applied Nonparametric Regression.* Econometric Society Monographs. Cambridge University Press.
- [8] Johnston J.; J. Dinardo (1997). *Econometrics Methods.* 4ª Edición. McGraw Hill
- [9] Sattinger, M. (1993). "Assignment Models of the Distribution of Earnings". *Journal of Economic Literature* N°31, Vol 2, pp.831-880.
- [10] Silverman B. W. (1996). *Density Estimation for Statistics and Data Analysis* Chapman & Hall. Londres.

ANEXO GRAFICO