

DISTRIBUCIÓN DEL INGRESO SEGÚN GÉNERO UN ENFOQUE SEMIPARAMÉTRICO¹

Dra. Juana Z. Brufman* Dr. Heriberto L. Urbisaia** Lic. Luis. A. Trajtenberg***
Instituto de Investigaciones en Estadística y Matemática Actuarial
Facultad de Ciencias Económicas - Universidad de Buenos Aires.
Av. Córdoba 2122 - Ciudad de Buenos Aires – Argentina - C112AAQ
*brufman@econ.uba.ar, **heribertourbisaia@speedy.com.ar, ***luis@econ.uba.ar

Recibido 15 de noviembre de 2007, aceptado 27 de diciembre de 2007

Resumen

En este trabajo se aplican *técnicas semiparamétricas* para el análisis de la distribución del Ingreso Laboral según género. En un trabajo anterior (Brufman *et al*, 2006), se relacionaban *en forma separada* el ingreso con los años de escolaridad y la experiencia laboral, usando una regresión no paramétrica. Una limitación seria en esta técnica es que no se extiende bien en el caso de más de dos variables: esta es la conocida “maldición de la dimensionalidad”. En este artículo, se incorporan *simultáneamente* ambas variables como regresores, por lo que se impone la regresión semiparamétrica.

El modelo lineal parcial $y = X\beta + m(z) + \varepsilon$ es un ejemplo de un análisis semiparamétrico. La estimación se realiza ajustando y y X no paramétricamente como función de z . Los residuos de la primera estimación se regresan en los de la segunda para obtener $\hat{\beta}^{OLS}$. Finalmente se obtiene $m(z)$ ajustando no paramétricamente $y - X\hat{\beta}^{OLS}$ con z .

Palabras Clave: superficie de regresión, maldición de la dimensionalidad, modelo lineal parcial, modelo índice.

¹ Presentado en las Sextas Jornadas de Tecnología Aplicada a la Educación Matemática Universitaria, septiembre de 2006. CMA, Facultad de Ciencias Económicas, UBA.

DISTRIBUTION OF LABOR INCOME BY GENDER A SEMIPARAMETRIC APPROACH

Dra. Juana Z. Brufman* Dr. Heriberto L. Urbisaia** Lic. Luis. A. Trajtenberg***
Instituto de Investigaciones en Estadística y Matemática Actuarial
Facultad de Ciencias Económicas - Universidad de Buenos Aires.
Av. Córdoba 2122 - Ciudad de Buenos Aires – Argentina - C112AAQ
*brufman@econ.uba.ar, **heribertourbisaia@speedy.com.ar, ***luis@econ.uba.ar

Received November 15th 2007, accepted December 27th 2007

Abstract

In this paper, we apply semi-parametric techniques in order to analyse the distribution of the income labor by gender. In a preceding analysis (Brufman *et al.*, 2006), in a bivariate dimension, we have related separately Labor income, years of schooling and labor experience, using a nonparametric regression. A serious limitation of this technique is that it does not extend well to more than two variables: this is the well-known *course of dimensionality*. Now, we include simultaneously both variables as regressors. So, we turn to a semiparametric regression.

A partial linear model $y = X\beta + m(z) + \varepsilon$ is an example of a semi-parametric analysis. Estimation proceeds by fitting y and X non-parametrically as a function of z . The resulting residualized y is regressed on residualized X to get $\hat{\beta}^{OLS}$; finally $m(z)$ is obtained by fitting $y - X\hat{\beta}^{OLS}$ non-parametrically with z .

Keywords: regression surface, course of dimensionality, partial linear model, index model.

1. INTRODUCCIÓN

De acuerdo con los principios de la inferencia paramétrica, los modelos utilizados en el análisis económico deben ser totalmente especificados, salvo el vector de parámetros, que se estima a partir de información muestral. Los métodos de estimación y las correspondientes propiedades de las estimaciones dependen de una serie de supuestos subyacentes, no siempre realistas, que forman parte de la propia especificación. La violación de alguno de estos supuestos genera estimaciones ineficientes o aún inconsistentes, que pueden conducir a falsas interpretaciones sobre el fenómeno que se desea explicar.

La regresión no paramétrica, utilizada con frecuencia en el análisis econométrico, evita la fijación de supuestos sobre la forma funcional de la ecuación de regresión. Es un método más flexible y permite estimar, en el caso de *una única variable explicativa*, una *curva* en el plano bidimensional, representativa del valor esperado de y , condicionado a los valores de x .

La generalización para dos variables explicativas, genera una *superficie* de regresión. Sin embargo, el agregado de variables explicativas complica dramáticamente el proceso de estimación; por un lado, cada variable que se agrega incrementa el requerimiento de cálculo en forma exponencial. En segundo lugar, aparece el problema denominado “maldición del dimensionamiento” (*course of dimensionality*), que consiste en la rápida disminución de la precisión de las estimaciones cuando se incluyen varios regresores.

Estas razones han llevado a los investigadores a la búsqueda de técnicas que permitan, en alguna forma, una *reducción de la dimensionalidad*. Tales métodos combinan rasgos de técnicas paramétricas y no paramétricas, por lo que se los denomina “técnicas semiparamétricas”.

La regresión semiparamétrica permite entonces mantener la interpretación conocida de las estimaciones paramétricas, a la vez que flexibiliza la forma funcional de la ecuación de regresión. De allí que cualquier área científica ó técnica que usa el análisis de la regresión pueda beneficiarse de la regresión semiparamétrica.

En general, como esta regresión incorpora parte de ambas metodologías, consideramos conveniente comparar algunos conceptos y su formalización, debido a que resultan fundamentales para su extensión y generalización.

2. CONSIDERACIONES SOBRE TÉCNICAS DE REGRESIÓN.

La teoría de la regresión intenta indagar sobre presuntas relaciones causales entre diversas variables que afectan a un determinado fenómeno.

Así, para el caso del modelo bivariado:

$$y = m(x) + \varepsilon \quad (1)$$

$m(\cdot)$ es una función de forma matemática desconocida y ε el término de error independiente, que satisface: $E(\varepsilon/x) = 0$; $m(\cdot)$ representa el valor medio de y , condicionado a los valores de x :

$$E(y/x) = m(x) \quad (2)$$

2.1. Regresión paramétrica

En la regresión paramétrica, la forma funcional de $m(x)$ se especifica como hipótesis sobre la relación entre las variables x e y . El significado de las componentes de la (1) es la siguiente:

1. El primer término: $m(x)$ recibe el nombre de parte funcional de la ecuación; por ejemplo, puede admitirse una relación lineal entre x e y : $m(x) = \alpha + \beta x$, lo que constituye un *supuesto* previo respecto a la forma en que se vinculan ambas variables.

2. La segunda componente denominada término de error o componente aleatoria viene a resumir una serie de factores que alteran el comportamiento medio de y , tales como: i) Error de especificación de la función elegida; ii) Omisión de variables relevantes; iii) Errores de medición. Los parámetros α y β son desconocidos y se estiman a partir de las observaciones muestrales.

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.

El inconveniente de este enfoque radica en la necesidad de suponer demasiado: con frecuencia el gráfico de dispersión entre las variables es confuso, por lo que no se puede entrever una relación y por tanto, no sugiere forma funcional alguna. Por ejemplo, se admite una relación entre el nivel de educación y el ingreso, pero de allí a presuponer una relación lineal es arriesgar una hipótesis demasiado fuerte.

Dado que en el análisis econométrico suele intervenir gran cantidad de variables explicativas, se recurre a la regresión multivariada; el *modelo lineal general* con K variables explicativas, por ejemplo, especifica el valor medio de la variable y , como función lineal de las variables x_1, x_2, \dots, x_K .

$$E(y/x_1, x_2, \dots, x_K) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_K x_K \quad (3)$$

y en el enfoque paramétrico se trata de estimar los parámetros $\alpha, \beta_1, \dots, \beta_K$.

De la regresión paramétrica resulta una función analítica que permite estimar valores promedios de y para los cuales no existen observaciones muestrales, como también predecir valores fuera del espacio muestral.

2.2. La regresión no paramétrica

La regresión no paramétrica constituye una alternativa diferente respecto al enfoque anterior, y a causa de que no presupone estructura alguna para la forma funcional que se estima ni para la distribución del término de error. Se trata de un método intermedio entre el análisis gráfico y la inferencia paramétrica.

Retomando las expresiones (1) y (2):

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

$$E(y/x) = m(x)$$

la estimación no paramétrica \hat{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 con las observaciones de dicho intervalo más las de intervalos adyacentes; estas últimas se ponderan 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.

De la regresión no paramétrica resulta un gráfico definido sobre la muestra, que no responde a función analítica alguna y que, a lo sumo podrá superponerse a otro con iguales variables.

Para el caso de más de una variable explicativa, el modelo de regresión múltiple no paramétrico estima el valor medio condicional de la variable respuesta, como *función suave* de las variables predictoras:

$$m(x_1 \dots x_K) = E(y / x_1, x_2, \dots, x_K) = f(x_1, x_2, \dots, x_K) \quad (4)$$

La regresión no paramétrica relaja el supuesto de linealidad, sustituyéndolo por una restricción más débil sobre la forma de la distribución conjunta de los datos; por ejemplo, el valor medio condicional de y es *función continua* de los predictores x_k , $k = 1, 2, \dots, K$. El objetivo es estimar dicha función de la misma forma que la regresión paramétrica estima los parámetros β_k .

La expresión (4) se denomina *Modelo de regresión múltiple general o irrestricto*, dado que no se imponen restricciones en la forma de la función K-dimensional, salvo la de su *continuidad*.

2.2.1. La maldición de la dimensionalidad

La extensión de la regresión no paramétrica a mayores dimensiones constituye un problema conceptualmente sencillo; sin embargo, su implementación genera inconvenientes que tornan impracticable su aplicación. Entre estos se señalan: i) la creciente complejidad de cálculo por el agregado de variables regresoras; ii) la dificultad en lograr una adecuada representación de los resultados; iii) las propiedades estadísticas de los estimadores se deterioran rápidamente debido a que el volumen de

datos requeridos para mantener un grado de precisión tolerable crece más rápido que la cantidad de variables que se incluyen en el modelo.

Una forma simple de ilustrar este problema es la siguiente. Considérese la estimación de un histograma a partir de una muestra aleatoria simple de tamaño n , generada por una distribución uniforme K dimensional en el intervalo $(0,1)$.

Si se particiona el hipercubo unitario en celdas de lado igual a α , cada celda contendrá en promedio sólo un porcentaje de datos igual a α^K % .

Supóngase además, que para estimar el histograma con precisión tolerable, se requiere por lo menos 30 observaciones por celda; entonces una muestra apropiada deberá tener, un tamaño promedio de al menos $n = 30(1/\alpha)^K$.

Se indican en la tabla siguiente algunos cálculos sobre el tamaño promedio de la muestra, para $K = 1,2,\dots,5$ y $\alpha = 0.10$

	Cantidad de variables de x				
$\alpha = 0.10$	1	2	3	4	5
N° de celdas	10	100	1.000	10.000	100.000
$E[n]$	300	3.000	30.000	300.000	3.000.000

Dejando de lado el problema no trivial de la representación del histograma, en el caso de 5 dimensiones, la estimación sería demasiado imprecisa como para que resulte de utilidad, a menos que se disponga de una muestra gigante de 3.000.000 de observaciones.

2.2.2. El Modelo Aditivo

Como se sabe, una propiedad importante del modelo de regresión lineal es:

$y = \alpha + \sum_{k=1}^K \beta_k x_k + \varepsilon$ es la *aditividad*. Esta propiedad permite *separar* el

efecto de los diferentes regresores, e interpretar el coeficiente β_k como la derivada parcial del valor medio condicional de y con respecto a x_k .

Del mismo modo, una forma simple de regresión múltiple no paramétrica es la que se conoce como *modelo aditivo*, con esperanza condicionada:

$$m(x_1 \dots x_K) = E(y / x_1, \dots, x_K) = \alpha + f_1(x_1) + f_2(x_2) + \dots + f_K(x_K) \quad (5)$$

donde las f_k son funciones de las que sólo se especifica su continuidad.

Como $f'_k(x_k) = \partial f(x) / \partial x_k$, esta especificación mantiene la interpretación del efecto individual de cada regresor. El modelo aditivo es más restrictivo que el modelo general no paramétrico, a razón de que excluye el efecto interacción entre los predictores; no obstante, es más flexible que el modelo lineal de regresión y presenta la ventaja de reducir el problema de su estimación a una serie de regresiones *parciales no paramétricas* en dos dimensiones. Esta ventaja es importante desde el punto de vista del cómputo, como también por la interpretación de los resultados. La idea central se basa en el concepto de *residuo parcial* de la regresión.

Los gráficos de residuos parciales se utilizan habitualmente para diagnosticar una relación *no lineal* en un problema de regresión clásico. En efecto; suponiendo una relación funcional lineal:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_K x_K + \nu \quad (6)$$

y su estimación preliminar por MCC:

$$y = a + b_1 x_1 + b_2 x_2 + \dots + b_K x_K + u$$

El residuo parcial para el primer predictor x_1 , $u_{[1]}$, se define como:

$$u_{[1]} = y - a - b_2 x_2 - \dots - b_K x_K = u + b_1 x_1 \quad (7)$$

Si existe una componente *no lineal* en la relación entre x_1 e y , dicha componente aparecerá en los residuos M.C.C., de modo que un gráfico del residuo parcial $u_{[1]}$ contra x_1 puede revelar el tipo de relación parcial entre y y x_1 .

El Modelo Aditivo extiende la noción de residuos parciales, restando los ajustes potencialmente no lineales para los demás predictores:

$$u_{[i]} = y - a - f_2(x_2) - \dots - f_K(x_K) \quad (8)$$

y suaviza en forma no paramétrica $u_{[i]}$ contra x_1 para estimar f_1 .

Existen paquetes de cómputo para la estimación de este tipo de modelos. Si por simplicidad se considera el caso de dos predictores:

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon \quad (9)$$

el algoritmo denominado *backfitting* se basa en los siguientes principios:

$$E[y - f_1(x_1)/x_2] = f_2(x_2) \quad \text{y} \quad E[y - f_2(x_2)/x_1] = f_1(x_1) \quad (10)$$

Además, si \hat{f}_1 es una buena estimación de f_1 , entonces f_2 puede ser estimada mediante regresión no paramétrica del residuo $y - \hat{f}_1(x_1)$ sobre x_2 . Similar razonamiento opera para estimar f_1 .

Valores iniciales para f_1 y f_2 suelen fijarse en cero, o bien se utilizan estimaciones paramétricas, tales como la regresión lineal.

El *Backfitting* procede según los siguientes pasos:

1. Inicialización: selecciona estimaciones iniciales: f_1^0, f_2^0
2. Iteración: obtiene \hat{f}_1^i mediante regresión no paramétrica de $y - \hat{f}_2^{i-1}(x_2)$ sobre x_1 . Obtiene \hat{f}_2^i mediante regresión no paramétrica de $y - \hat{f}_1^{i-1}(x_1)$ sobre x_2 .
3. Convergencia: continúa el proceso iterativo hasta obtener resultados muy próximos entre estimaciones sucesivas.

3. REGRESIÓN SEMIPARAMÉTRICA

La modelización semiparamétrica, como su nombre lo sugiere, combina la forma paramétrica para algún componente del proceso generador de los

datos con restricciones no paramétricas débiles impuestas sobre el resto del modelo. Se trata, por tanto, de un híbrido de los enfoques paramétricos y no paramétricos analizados previamente.

Un modelo semiparamétrico implica que la densidad conjunta de los datos observados, condicionada a alguna información auxiliar, se halla completamente especificada mediante un vector de parámetros β , de dimensión finita y una función desconocida $\lambda(\cdot)$.

Los métodos semiparamétricos constituyen un compromiso atrayente para la construcción de modelos estadísticos. Al admitir supuestos de rigor intermedio entre los enfoques paramétrico y no paramétrico, se reduce el riesgo de una errónea especificación relativa al primero de estos enfoques, a la vez que evita, al menos en parte, los inconvenientes del segundo.

En Microeconometría, es habitual la aplicación de este tipo de modelos. En primer lugar, la Teoría Económica puede sugerir cierta estructura; por ejemplo: simetría y restricciones de homogeneidad en una función de demanda. Tal información es incorporada a una regresión no paramétrica. En segundo lugar, y con mayor frecuencia, los modelos de la Microeconomía incluyen gran cantidad de regresores, por lo que, el problema de la dimensionalidad excesiva, torna impracticable la regresión no paramétrica.

Existe gran variedad de modelos semiparamétricos, como también paquetes de cómputo para su estimación consistente. Mencionamos, entre otros:

Nombre	Modelo	Para- métrico	No Paramé- trico
Parcialmente Lineal	$E[y/x, z] = x'\beta + \lambda(z)$	β	$\lambda(\cdot)$
Índice Lineal	$E[y/x, z] = g(x'\beta)$	β	$g(\cdot)$
Parcialmente Aditivo	$E[y/x, z] = x'\beta + c + \sum_{k=1}^K g_k(z_k)$	β	$g_k(\cdot)$
Lineal Heteroscedástico	$E[y/x, z] = x'\beta; V(y/x) = \sigma^2(x)$	β	$\sigma^2(\cdot)$

4. PROPIEDADES DE LOS ESTIMADORES SEMIPARAMÉTRICOS

Se sintetizan a continuación, los principales resultados teóricos desde el punto de vista de la consistencia de los estimadores y su distribución asintótica.

En la regresión no paramétrica, la condición de *función suave* (continua) que se impone a λ y, en particular, la existencia de derivadas acotadas de esta función, garantizan la consistencia de su estimador; son también importantes en la determinación de la tasa de convergencia y su distribución asintótica. Si la función de regresión λ es suave, sus derivadas pueden ser estimadas en forma consistente, a veces, diferenciando el estimador de la propia función.

4.1. Sobre la tasa de convergencia

La tasa de convergencia de un estimador mide, en términos de tamaño de muestra, la rapidez con que se llega a la verdadera función de regresión. En otros términos, es la tasa a la cual tiende a cero la varianza de un estimador. Por ejemplo, si consideramos la media muestral como estimador de la media poblacional, se sabe que \bar{X} es insesgado y su varianza igual a σ^2/n . Por lo tanto: $var[\sqrt{n}(\bar{X} - \mu)] = 0$; con la simbología matemática utilizada en secuencias de variables aleatorias, estos resultados se expresan:

$$\bar{X} - \mu = O_p(n^{-1/2}); \quad (\bar{X} - \mu)^2 = O_p(1/n) \quad (11)$$

y donde el subíndice P hace referencia a la convergencia *en probabilidad*. Se dice también que \bar{X} es \sqrt{n} consistente.

En una especificación paramétrica, la tasa de convergencia no depende de la cantidad de variables explicativas del modelo. Supóngase: $y = \beta_0 + \beta_1 x + \varepsilon$; como $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados y sus varianzas y covarianzas convergen a 0 a la tasa $1/n$, se verifica que:

$$\int (\beta_0 + \beta_1 x - \hat{\beta}_0 + \hat{\beta}_1 x)^2 dx = \quad (12)$$

$$= (\beta_0 - \hat{\beta}_0)^2 \int dx + (\beta_1 - \hat{\beta}_1)^2 \int x^2 dx + 2(\beta_0 - \hat{\beta}_0)(\beta_1 - \hat{\beta}_1) \int x dx = O_p(1/n)$$

Para los estimadores no paramétricos, la convergencia cae dramáticamente al aumentar el número de variables regresoras, aún cuando esta condición se aminora cuando la función es diferenciable. La tasa óptima a la cual un estimador no paramétrico converge a la verdadera función de regresión está dada por la expresión:

$$\int [\hat{f}(\mathbf{x}) - f(\mathbf{x})]^2 dx = O_p \left(\frac{1}{n^{2d/(2d+K)}} \right) \quad (13)$$

donde d es el grado de diferenciación de la función de regresión y K es la cantidad de variables explicativas, es decir, la dimensión de \mathbf{x} . Por ejemplo, para una función diferenciable dos veces y una variable explicativa resulta $O_p(n^{-4/5})$ y con una variable explicativa $O_p(n^{-2/3})$.

Si el modelo es aditivo separable o parcialmente lineal, la tasa de convergencia del estimador óptimo depende de la componente no paramétrica de mayor dimensionalidad. Por ejemplo, el modelo aditivo separable:

$$y = f_1(x_1) + f_2(x_2) + \varepsilon \quad (14)$$

donde x_1 y x_2 son escalares, tiene igual tasa de convergencia que una regresión paramétrica con un sólo regresor. Lo mismo ocurre si el modelo es $y = x'\beta + \lambda(z) + \varepsilon$ donde x y z son escalares.

En general, para la función de regresión:

$$f(z, x_1, x_2, x_3) = z'\beta + f_1(x_1) + f_2(x_2) + f_3(x_3), \quad (15)$$

donde x_1, x_2 y x_3 son de dimensión d_1, d_2 y d_3 respectivamente, la tasa de convergencia óptima para la función de regresión total corresponde a la de un modelo no paramétrico con un número de variables igual al $\max\{d_1, d_2, d_3\}$.

4.2. Distribución asintótica de los estimadores

Para una gran variedad de estimadores no paramétricos, el estimador de la función de regresión *en un punto* se distribuye en forma aproximadamente normal. En una *colección de puntos*, la distribución conjunta es conjuntamente normal. La suma de residuos cuadráticos se halla también normalmente distribuida².

Para el modelo parcialmente lineal, pueden construirse estimadores de β que son \sqrt{n} consistentes: la varianza del estimador se reduce a la tasa $1/n$, y son asintóticamente normales.

Considérese, en general, un modelo semiparamétrico con componente paramétrico β y componente no paramétrico indicada por λ . Siguiendo a Robinson (1988b), sintetizamos a continuación, las propiedades del estimador $\hat{\beta} = \beta(\hat{\lambda})$, donde $\hat{\lambda}$ es un estimador no paramétrico de λ .

Idealmente, el estimador $\hat{\beta}$ es *adaptativo*³, en el sentido de que no se produce pérdida de eficiencia al haber estimado λ en forma no paramétrica, es decir:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N[\mathbf{0}, V_\lambda] \quad (16)$$

siendo V_λ la matriz de covarianzas de la función m . En el marco de la estimación máximo-verosímil, V_λ es la cota de Cramer-Rao. En el contexto de las condiciones de momentos de segundo orden, V_λ está fijada por el Teorema de Gauss-Markov.

² Las demostraciones pueden consultarse en Yatchew (2003).

³ Un estimador semiparamétrico es *adaptativo* si las componentes paramétricas y no paramétricas del modelo son ortogonales.

5. EL MODELO PARCIALMENTE LINEAL

El modelo parcialmente lineal es de la forma:

$$y = \mathbf{x}'\beta + \lambda(z) + \varepsilon \quad (17)$$

Del cual analizaremos enfoques alternativos para su estimación.

5.1. Modelo lineal por partes (*Peacewise linear model*)

Considérese, por razones de simplicidad, el caso en que \mathbf{x} y z son regresores escalares, admitiendo, además que:

$$E[y/x, z] = 0 \text{ y } V(\varepsilon/x, z) = \sigma_\varepsilon^2 \quad (18)$$

Supongamos que z representa años de experiencia laboral. Una posibilidad para resolver el problema sería incluir en la regresión clásica: z, z^2, z^3 , etc. Este criterio resulta muy sensible a valores altos de z cuando dichos valores son escasos en la muestra. Por otra parte, habría que decidir el grado del polinomio necesario para lograr una buena aproximación de $f(z)$.

Una solución a este problema consiste en aproximar $f(z)$ por una *función lineal por partes*. Por ejemplo si la variable z tiene un rango de 0 - 100, podemos reemplazar $f(z)$ por la función:

$$\gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 z_4 \quad (19)$$

donde:

$$\begin{aligned} z_1 &= z \text{ si } z \leq 20 \text{ y cero en los restantes casos} \\ z_2 &= z \text{ si } 20 < z \leq 40 \text{ y cero en los restantes casos} \\ z_3 &= z \text{ si } 40 < z \leq 60 \text{ y cero en los restantes casos} \\ z_4 &= z \text{ si } z > 60 \text{ y cero en los restantes casos} \end{aligned}$$

Otras divisiones de las observaciones son posibles. La ventaja de este enfoque es que admite su tratamiento como modelo lineal.

5.2. Estimador en diferencias

El estimador $\hat{\beta}_{dif}$ diseñado por A. Yatchew (2003), se obtiene estimando el modelo en primeras diferencias. El fundamento de este criterio es como sigue.

Supóngase que: *i)* z es acotada y las observaciones ordenadas en forma creciente, de modo que $z_1 \leq z_2 \leq \dots \leq z_n$; *ii)* la media condicional de x es función continua de z : $E(x/z) = g(z)$, *iii)* g' es acotada; *iv)* $V(x/z) = \sigma_u^2$. Se puede escribir, entonces:

$$x = g(z) + u. \quad (20)$$

Consideramos ahora el modelo en primeras diferencias:

$$y_i - y_{i-1} = (x_i - x_{i-1})\beta + (\lambda(z_i) - \lambda(z_{i-1})) + \varepsilon_i - \varepsilon_{i-1} \quad (21)$$

y se reemplaza $(x_i - x_{i-1})$ por su aproximación según la (20):

$$\begin{aligned} y_i - y_{i-1} &= (g(z_i) - g(z_{i-1}))\beta + (u_i - u_{i-1})\beta + (\lambda(z_i) - \lambda(z_{i-1})) + \varepsilon_i - \varepsilon_{i-1} \\ &\cong (u_i - u_{i-1})\beta + \varepsilon_i - \varepsilon_{i-1} \end{aligned} \quad (22)$$

Se observa que los efectos directos de la componente no paramétrica: $\lambda(z)$ y los indirectos $g(z)$, estos últimos a través de x , pueden removerse mediante la diferenciación.

Por lo tanto, para estimar $\hat{\beta}$ puede aplicarse Mínimos Cuadrados Clásicos a las observaciones diferenciadas, obteniéndose:

$$\hat{\beta}_{dif} = \frac{\sum (y_i - y_{i-1})(x_i - x_{i-1})}{\sum (x_i - x_{i-1})^2} \quad (23)$$

El estimador $\hat{\beta}_{dif}$ es consistente y de distribución asintóticamente normal:

$$\sqrt{n}(\hat{\beta}_{dif} - \beta) \xrightarrow{D} N\left(0, \frac{1.5\sigma_\varepsilon^2}{\sigma_u^2}\right) \quad (24)$$

Estimaciones consistentes de las varianzas σ_ε^2 y σ_u^2 se obtienen a partir de los residuos:

$$s_\varepsilon^2 = \frac{1}{2n} \sum_{i=2}^n \left[(y_i - y_{i-1}) - (x_i - x_{i-1}) \hat{\beta}_{dif} \right]^2 \cong \frac{1}{2n} \sum_{i=2}^n (\varepsilon_i - \varepsilon_{i-1})^2 \xrightarrow{P} \sigma_\varepsilon^2 \quad (25)$$

$$s_u^2 = \frac{1}{2n} \sum_{i=2}^n (x_i - x_{i-1})^2 \cong \frac{1}{2n} \sum_{i=2}^n (u_i - u_{i-1})^2 \xrightarrow{P} \sigma_u^2 \quad (26)$$

Este procedimiento puede generalizarse para el caso de mayor cantidad de variables explicativas paramétricas.

5.3. Estimador parcialmente lineal de Robinson

Dado el modelo parcialmente lineal:

$$y = \mathbf{x}'\beta + \lambda(z) + \varepsilon \quad (27)$$

Consideraremos nuevamente, por razones de simplicidad, el caso en que \mathbf{x} y z son escalares. Si $E[\varepsilon/x, z] = 0$, se verifica que

$E[y/x, z] = \mathbf{x}'\beta + \lambda(z)$, y por tanto:

$$\varepsilon = y - E[y/x, z] \quad (28)$$

Además, como $E[\varepsilon/x, z] = 0$ implica que $E[\varepsilon/z] = 0$

resulta que:

$$E[y/z] = E[x/z]' \beta + \lambda(z) \quad (29)$$

Restando (29) de (27) se obtiene:

$$y - E[y/z] = (x - E[x/z])' \beta + \varepsilon \quad (30)$$

Los momentos condicionados de la ecuación (29) se reemplazan por estimaciones no paramétricas que se indican con \hat{m}_{yi} y \hat{m}_{xi} .

Robinson propuso utilizar estimaciones tipo *Kernel*⁴ para \hat{m}_{yi} y \hat{m}_{xi} , que convergen suficientemente rápido, de modo que su inclusión en la (29) no altera la distribución asintótica del estimador $\hat{\beta}$. Finalmente, este último se obtiene por Mínimos Cuadrados Clásicos a partir de la ecuación:

$$y_i - \hat{m}_{yi} = (x_i - \hat{m}_{xi})' \beta + v_i \quad (31)$$

En síntesis, la estimación se lleva a cabo en las siguientes etapas:

1°. Etapa: se estiman \hat{m}_{yi} y \hat{m}_{xi} mediante regresiones no paramétricas de y_i sobre z_i y x_i sobre z_i

2°. Etapa: se calculan los residuos parciales de las regresiones no paramétricas de la etapa anterior.

3°. Etapa: se estima β mediante la regresión paramétrica M.C.C. entre ambos residuos parciales.

Por esta razón, este método se denomina *de doble residuo*.

Dada la independencia entre observaciones de la muestra, y suponiendo $\varepsilon_i \sim iid[0, \sigma_\varepsilon^2]$, el estimador parcialmente lineal $\hat{\beta}_{PL}$ así obtenido es \sqrt{n} consistente y de distribución asintóticamente normal:

$$\sqrt{n}(\hat{\beta}_{PL} - \beta) \xrightarrow{D} N\left[0, \frac{\sigma_\varepsilon^2}{\sigma_u^2}\right] \quad (32)$$

donde $\sigma_u^2 = V(x/z)$.

Finalmente, la estimación no paramétrica de $\lambda(z)$ se obtiene como diferencia:

$$\hat{\lambda}(z) = \hat{m}_{yi} - \hat{\beta} \hat{m}_{xi} \quad (33)$$

⁴ La estimación *Kernel* ha sido tratada y aplicada en Brufman *et al.* (2006) *op. cit.*

Si el modelo incluye más variables explicativas en la componente paramétrica, es decir, es de la forma:

$$y = \mathbf{x}'\beta + \lambda(z) + \varepsilon \quad (34)$$

donde \mathbf{x} es vector K-dimensional, el método del doble residuo requiere llevar a cabo regresiones no paramétricas separadas para la variable dependiente y para cada variable paramétrica.

Siguiendo un razonamiento similar al caso anterior, e indicando con \hat{m}_{yi} y \hat{m}_{xi} las estimaciones no paramétricas de $E[y/z]$ y $E[\mathbf{x}/z]$, se estima el vector β a partir de la ecuación de regresión:

$$y_i - \hat{m}_{yi} = (\mathbf{x} - \hat{m}_{xi})'\beta + v_i \quad (35)$$

El estimador $\hat{\beta}_{PL}$ resultante tiene una distribución asintótica normal:

$$\sqrt{n}(\hat{\beta}_{PL} - \beta) \xrightarrow{D} N\left[0, \sigma_\varepsilon^2 \sum_{\mathbf{x}/z}^{-1}\right] \quad (36)$$

siendo $\sum_{\mathbf{x}/z}$ la matriz de varianzas-covarianzas de \mathbf{x} condicionada a z . A partir de los residuos se estima σ_ε^2 en forma consistente:

$$s_\varepsilon^2 = n^{-1} \left[y - \hat{m}_y - (\mathbf{x} - \hat{m}_x \hat{\beta}) \right]' \left[y - \hat{m}_y - (\mathbf{x} - \hat{m}_x \hat{\beta}) \right] \quad (37)$$

y finalmente, la matriz de varianzas-covarianzas de $\hat{\beta}$ resulta:

$$\hat{\Sigma}_{\hat{\beta}} = s_\varepsilon^2 \left[(\mathbf{x} - \hat{m}_x)' (\mathbf{x} - \hat{m}_x) \right]^{-1} \quad (38)$$

5.4. Comparación entre los estimadores

Si bien el método de la diferenciación es relativamente simple y evita la estimación no paramétrica de los valores valores medios $E[y/z]$ y $E[\mathbf{x}/z]$, genera un estimador de mayor varianza que el correspondiente al

de Robinson. Comparando las expresiones (24) y (32), la eficiencia relativa resulta:

$$\frac{\text{var } \hat{\beta}_{PL}}{\text{var } \hat{\beta}_{dif}} = 1/1.5 = 2/3 \quad (39)$$

6. ANÁLISIS EMPÍRICO.

Se aplican, a continuación, las técnicas semiparamétricas para cuantificar los Retornos de la Educación, considerando separadamente la población de Varones y Mujeres; la información utilizada proviene de la Encuesta Permanente de Hogares, elaboradas por INDEC. Se llevaron a cabo dos ensayos. En el primero de ellos, se intentó seguir paso a paso, la construcción del estimador $\hat{\beta}_{PL}$ de Robinson, utilizando el procesador E-Views, versión 5.1.

En el segundo se utilizó el procesador STATASE 8, que ofrece la estimación semiparamétrica en forma inmediata, unida a una mayor capacidad de procesamiento de la información.

1° Caso: Efecto de la Educación sobre el Ingreso de las personas, considerando la Edad y Antigüedad en el Empleo como variables paramétricas del modelo.

Fuente de Información

Se utilizaron datos de la Onda Octubre/2000. Respecto al Ingreso, Nivel Personas, se consideraron perceptores con edades entre 15 y 60 años, procesándose las variables según los criterios siguientes:

1. I_h : **Ingreso por hora trabajada:** $10 \leq I_h \leq 30$.
2. $Ed.$: **Edad** medida por Años cumplidos;
3. $Ant.$: **Antigüedad en el empleo:** ≤ 40 años
4. $Niv.E.$: **Nivel de Educación;** se asignó la siguiente escala en años, según nivel de instrucción alcanzado: Primario: 7; Secundario: 12; Terciario: 14; Universitario: 17.

Modelo a estimar: $I_{hi} = \beta_1 Ed_i + \beta_2 Ant_i + \lambda(Niv.E_i) + \varepsilon_i$

Método de Estimación: Se utilizó el estimador de Robinson para los parámetros β_1 y β_2 , habiéndose efectuado los siguientes pasos:

i) Regresión no paramétrica de cada una de las variables I_h , Ed . y Ant . sobre $Niv.E$, obteniéndose: $\hat{m}_{I_h,i}$, $\hat{m}_{Ed,i}$ y $\hat{m}_{Ant,i}$. En esta etapa se utilizó el Kernel Epanechnikov, $h=1.5000$.

ii) Cálculo de los residuos:

$$res_{I_h} = I_h - \hat{m}_{I_h}; res_{Ed} = Ed - \hat{m}_{Ed}; res_{Ant} = Ant - \hat{m}_{Ant}$$

iii) Con las variables así depuradas del efecto educación, se estimaron $\hat{\beta}_1$ y $\hat{\beta}_2$ separadamente, por MCC, mediante sendas regresiones paramétricas entre: res_{I_h} y res_{Ed} por un lado, y res_{I_h} y res_{Ant} . por el otro.

iv) Finalmente, el efecto Educación, (Retornos de la Educación) resulta:

$$\hat{\lambda}(Niv.E) = \hat{m}_{I_h} - \hat{\beta}_1 \hat{m}_{Ed} - \hat{\beta}_2 \hat{m}_{Ant}$$

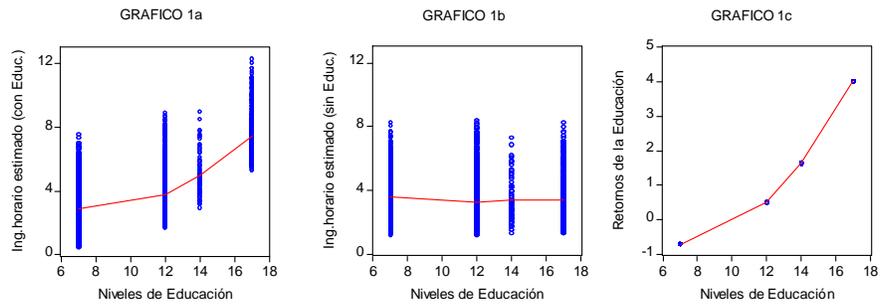
Análisis de Resultados

Para la población de varones se obtuvo:

$$\hat{I}_h^V = 0.073907 * Ed + 0.113583 * Ant + \hat{\lambda}^V(Niv.E)$$

$$\hat{\lambda}^V(Niv.E) = I_h^V - 0.073907 * Ed - 0.113583 * Ant$$

VARONES

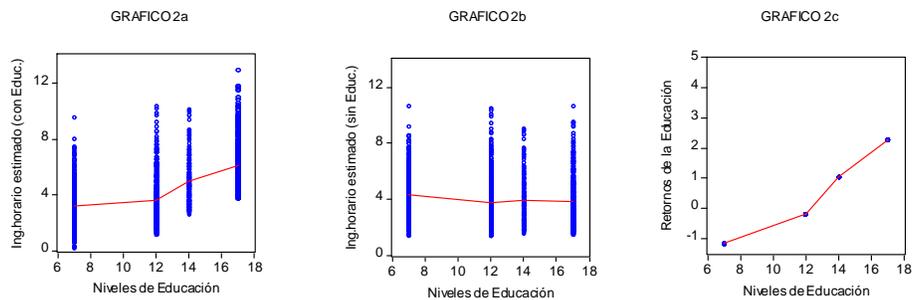


Para la población de mujeres se obtuvo:

$$\hat{I}_h^M = 0.086079 * Ed. + 0.147484 * Ant. + \hat{\lambda}^M (Niv.E)$$

$$\hat{\lambda}^M (Niv.E) = I_h^M - 0.086079 * Ed. - 0.147484 * Ant.$$

MUJERES



En síntesis, el ingreso por hora resulta explicado mediante un modelo aditivo, parcialmente lineal, en las variables Edad y Antigüedad en el Empleo, (ambas depuradas del efecto educación) más el retorno de la educación.

Los Cuadros del Apéndice sintetizan las características descriptivas de la variable analizada, Ingreso por hora, a partir de muestras de 2.045

observaciones en el caso de varones y 1.427 para mujeres. Del mismo modo, se indican similares estadísticos para los Retornos de la Educación estimados. Por último, se muestran valores porcentuales relacionando el Retorno con el Ingreso horario. Todos ellos tratando separadamente varones y mujeres, para cada uno de los Niveles de Educación considerados.

El análisis de ellos permite expresar los siguientes comentarios.

1. Contra lo que se hubiera esperado, el ingreso por hora promedio de mujeres supera al de los varones.

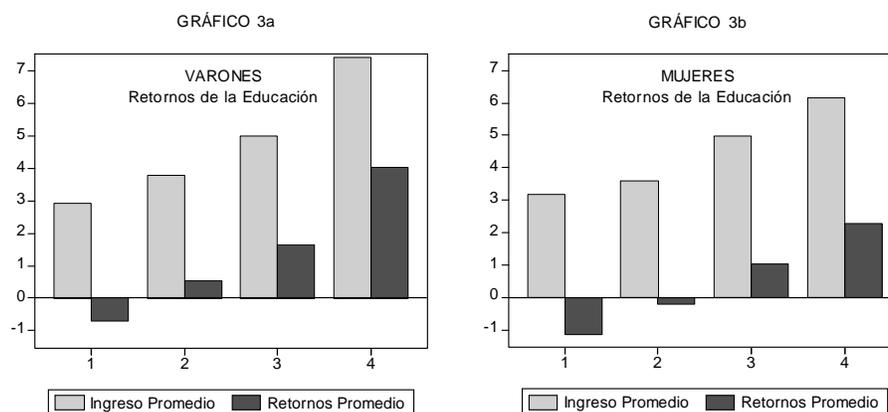
2. Si se analizan estas diferencias por niveles de educación, se observa que el ingreso promedio de mujeres es superior sólo para el caso del nivel primario; para los restantes niveles de educación, es superior el de los varones, incrementándose esta diferencia a medida que se pasa a niveles de instrucción superior. No obstante ello, lo que acontece en el nivel primario es de tal relevancia, que determina el comportamiento del promedio general del ingreso.

Analizando la componente Educación y su relación con el Ingreso se concluye:

3. Tanto para varones como mujeres, el Retorno aumenta con el nivel de educación. Sin embargo, mientras para los varones el efecto resulta significativo a partir del nivel secundario, para las mujeres este efecto es neutro en educación primaria y secundaria, mostrando cierta significación en los dos niveles superiores de instrucción.

4. Para el nivel universitario, el Retorno de la Educación en varones es superior al de mujeres: 54% contra 37%. En el nivel Terciario los guarismos son 32.7% y 20.9%, respectivamente.

Los gráficos 1a y 1b permiten visualizar lo expresado en estos comentarios.



2° Caso: Efecto de la Educación sobre el Ingreso de las personas, considerando la Edad y Edad² como variables paramétricas del modelo.

Se utilizaron datos de la Onda Octubre/2003, sobre la base de 3.796 observaciones.

El detalle de las variables del modelo es como sigue:

1. **Variable dependiente:** logaritmo del Ingreso horario ($\log I_h$). Con respecto a esta variable, se eliminan los registros que reportan ingresos horarios nulos.

2. **Regresores:** **Edad** medida por Años cumplidos; **Edad²**, para captar la incidencia del factor Experiencia en el nivel de Ingreso; **Género:** codificada mediante una variable binaria: 0 para varones; 1 para mujeres. Respecto a **Educación**, fue construida a partir de los niveles educativos que reportan las personas, combinado con el dato sobre la finalización, o no, de los respectivos ciclos. Se asignó la siguiente escala en años: Primario Incompleto: 5; Primario Completo: 7; Secundario Incompleto: 9; Secundario Completo: 12; Terciario Incompleto: 14; Terciario Completo: 15; Universitario Incompleto: 14; Universitario Completo: 17.

3. **La componente no paramétrica** se estimó por el método *Kernel*, con un ancho de banda de Silverman y ponderación "normal".

Análisis de resultados

La salida que brinda el procesador muestra la siguiente estimación:

$$\log \hat{I}_h = -1.00106 + 0.0550944 * Edad - 0.0005242 * -Edad^2$$

$$-0.059347 * (Gen^M = 1) + \hat{\lambda}(Educ);$$

$$R^2 = 0.2643$$

Se agrega el gráfico representativo del Retorno de la Educación.

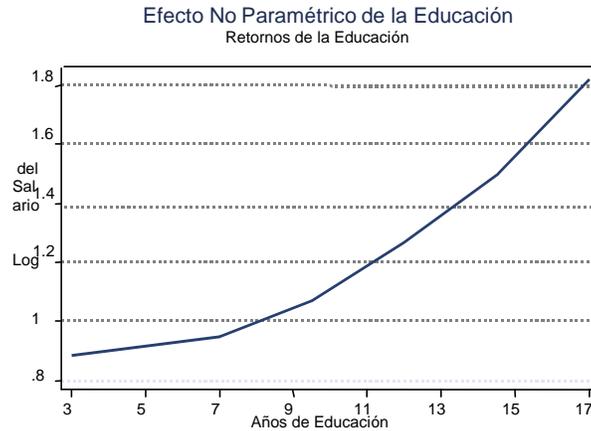


Gráfico 4. Efecto no paramétrico de la educación

Corresponde efectuar los siguientes comentarios:

- i. Los coeficientes estimados deben interpretarse como semielasticidades, dada la escala utilizada para la variable dependiente.
- ii. El coeficiente estimado para la variable Género resulta negativo, dada la codificación elegida (0 para varones; 1 para mujeres). Por lo tanto, para el año 2003 el salario promedio de las mujeres resulta inferior al de varones.
- iii. Como era de esperar, también es negativo el coeficiente de la variable $Edad^2$. Teniendo en cuenta el objetivo señalado para la inclusión de este regresor en el modelo, el signo negativo es coherente con el hecho comprobado de que, con el transcurso del tiempo, la Experiencia tiene

una incidencia decreciente sobre el nivel de ingreso: la representación gráfica de su trayectoria mostraría una curva cóncava hacia abajo.

iv. Las estimaciones resultaron altamente significativas.

v. La función representativa del Retorno muestra una pendiente más pronunciada a medida que se incrementan los años dedicados a la Educación.

	UNIVERSITARIO	TERCIARIO	SECUNDARIO	PRIMARIO	TOTAL
Mean	7.4400	5.0083	3.8086	2.9216	4.2020
Median	5.6800	4.1700	2.9800	2.5000	3.1300
Maximum	28.9100	19.4400	28.1300	21.8800	28.9100
<Minimum	1.0000	0.6300	0.1800	0.1400	0.1400
Std. Dev.	5.2725	3.3322	3.1699	2.0460	3.7183
Skewness	1.4906	1.6729	3.1828	3.2128	2.7212
Kurtosis	5.0981	6.6763	18.5571	21.7674	12.7855
Observations	376	77	871	721	2,045

Cuadro 1. Ingreso por hora Varones

	UNIVERSITARIO	TERCIARIO	SECUNDARIO	PRIMARIO	TOTAL
Mean	6.1389	4.9697	3.6091	3.1888	4.3236
Median	5.0000	4.5200	2.8500	2.5500	3.3950
Maximum	28.5700	17.7500	22.2200	16.6700	28.5700
Minimum	0.7400	1.0000	0.1200	0.1700	0.1200
Std. Dev.	4.6175	2.7359	2.7354	2.2130	3.4181
Skewness	2.0584	1.3966	2.7146	1.9697	2.5536
Kurtosis	8.0645	5.8057	14.4334	9.6358	12.7055
Observations	362	203	504	358	1427

Cuadro 2. Ingreso por hora Mujeres

	UNIVERSITARIO	TERCIARIO	SECUNDARIO	PRIMARIO	TOTAL
Mean VARONES	4.0207	1.6377	0.5157	-0.7026	0.7729
Mean MUJERES	2.2719	1.0369	-0.1953	-1.1528	0.3657

Cuadro 3. Retornos de la Educación

	VARONES			MUJERES		
	Inr. Horario	Retorno	retor/ingr.h.	Inr. Horario	Retorno	retor/ingr.h.
Universitario	7.44	4.02	0.5404	6.14	2.27	0.3701
Terciario	5.01	1.64	0.3270	4.97	1.04	0.2086
Secundario	3.81	0.52	0.1354	3.61	-0.20	-0.0541
Primario	2.92	-0.70	-0.2405	3.19	-1.15	-0.3615
Total	4.20	0.77	0.1839	4.32	0.37	0.0846

Cuadro 4. Relación Retorno / Ingreso Horario

BIBLIOGRAFÍA

- Barnett W. A.; Powell, J.; Tauchen, J. (Eds) (1991). *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Boston, Cambridge University Press.
- Brufman, J. Z.; Urbisaia, H. L.; Trajtenberg, L.A. (2006). "La distribución del ingreso según género. Un enfoque no paramétrico". *Cuadernos del CIMBAGE*. N° 8. pp.129-168.
- Cameron, A.C.; Trivedi, P.K. (2005). *Microeconometrics*. Boston, Cambridge University Press.
- DiNardo J.; Tobías, J. L. (2001): "Nonparametric Density and Regression Estimation". *Journal of Economic Perspectives*. Vol 15 N° 4 .pp. 11-28.
- DiNardo J.; Fortin, N. M.; Lemieux, T. (1996). "Labor Markets Institutions and the Distribution of Wages 1973-1993. A Semiparametric Approach". *Econometrica* Septiembre Vol. 64 N°5. pp. 1001-1044.
- Fortín N.M.; Lemieux, T. (2000). "Are Women Wages Gains Men's Loses?" *American Economic Review. Papers and Proceeding*. Mayo 2000. pp. 456-460
- Fox, J. (2000). "Non Parametric Simple Regresssion: Smoothing Scatterplots". *SAGE University Papers. Series Quantitative Aplicacions in the Social Sciences*. London, SAGE Publications Inc.
- Fox, J. (2000). "Multiple and Generalized Nonparametric Regression". *SAGE University Papers. Serie Quantitative Aplicacions in the Social Sciences*. London, SAGE Publications Inc.
- Härdle W. (1995). "Applied Nonparametric Regression". *Econometric Society Monographs*. Boston, Cambridge University Press
- Härdle W.; Linton, O. (1994). "Applied Nonparametric Methods". In Engle R.; McFadden D. (Eds.) *The Handbook of Econometrics*. Vol. IV Amsterdam. North Holland, pp. 2297-2334.
- Johnston J.; DiNardo, J. (1997). *Econometric Methods*. 4ª. Edition. Washington, McGraw Hill
- Pagan A.; Ullah, A. (1999). *Nonparametric Econometrics*. Boston, Cambridge University Press.

Powell, J. (1994). "Estimation of Semiparametric Models". In Engle R.; McFadden D. (Eds.) *The Handbook of Econometrics*. Vol. IV Amsterdam. North Holland, pp. 2444-2523.

Robinson, P. M. (1988a). "Root-N-Consistent Semiparametric Regression." *Econometrica* 56, pp.931-954.

Robinson, P. M. (1988b). "Semiparametric Econometrics: A Survey". *Journal of Applied Econometrics*, 3. pp. 35-51.

Ruppert D.; Wand M. P.; Carroll, R. J. (2003). *Semiparametric Regression*. Boston, Cambridge University Press.

Sattinger, M. (1993). "Assignment Models of the Distribution of Earnings". *Journal of Economic Literature*. Junio Vol. 31 N°2. pp.831-880.

Silverman B. W. (1996). *Density Estimation for Statistics and Data Analysis* Londres , Chapman & Hall.

Wooldridge J.M. (2002). "Econometric Analysis of Cross Section and Panel Data". Cambridge, MIT Press

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Boston, Cambridge University Press.