
Medidas de desempeño en regresiones no paramétricas de curvas de Engel

Performance Measurement of Nonparametric Engel Curves

Rodrigo García Arancibia^a
rgarcia@fce.unl.edu.ar

Resumen

El objetivo de este trabajo es comparar diferentes medidas de desempeño de regresiones no paramétricas comúnmente utilizadas en la literatura empírica. En particular se toma el caso de las denominadas curvas de Engel. Las mismas son estimadas por cuatro estimadores no paramétricos, tres basados en polinomios locales (local constante, lineal y cuadrático) y un suavizado *spline*. Para los mismos se computan una serie de medidas empíricas de desempeño basadas en los errores de estimación y predicción. Para evaluar los resultados de tales medidas, se realiza un ejercicio de simulación en el que se proponen tres especificaciones paramétricas tipo Working-Leser de curvas de Engel, computando los verdaderos errores de estimación y predicción, comparándolos con el *ranking* de estimadores más precisos propuestos por cada medida empírica. Adicionalmente, se presentan los resultados de una aplicación tomando una base de datos de gastos de hogares de Argentina.

Palabras clave: suavizados lineales, errores de estimación, errores de predicción, gastos de hogares.

Abstract

The aim of this paper is to compare different performance measurements commonly used in the empirical literature of nonparametric regressions, taking the case of so-called Engel curves. They are estimated by four nonparametric estimators, three based on local polynomials (local constant, linear and quadratic) and a smoothing spline. To them are computed a set of empirical measures of performance based on estimation and prediction errors. To evaluate the results of such measurements, is performed a simulation exercise using three parametric specifications type Working-Leser, computing the true errors. These are comparing with the more accurate ranking of estimators proposed by every empirical measure.

^aInstituto de Economía Aplicada Litoral. Facultad de Ciencias Económicas. Universidad Nacional del Litoral.

Additionally, it presents an application with a database of household expenditures of Argentina.

Key words: linear smoothers, estimation error, prediction error, households expenditures.

1. Introducción

Las curvas de Engel (CE) se definen desde el punto de vista teórico como las funciones que relacionan el gasto en bienes y servicios que desembolsa una determinada familia, con sus ingresos o recursos totales percibidos, así como otras variables que caracterizan la composición de la familia, dado los precios fijos. La teoría microeconómica no determina alguna forma funcional específica para las curvas de Engel pero establece criterios que esta debe cumplir en concordancia con la teoría del consumidor (Deaton 1980).

Desde el punto de vista estadístico, el término curva de Engel es usado para describir la dependencia empírica entre el gasto realizado en un determinado bien (o conjunto de bienes) y el ingreso o gasto total en una población de consumidores muestreada en un lugar y tiempo determinado (Lewbel 2006).

En general los trabajos empíricos han buscado cerrar esta brecha, de modo que las especificaciones econométricas cumplan con los requisitos básicos impuestos por el marco de la teoría neoclásica del consumidor. Los primeros trabajos que siguieron este enfoque adoptaron modelos paramétricos, para estimar una forma funcional específica de las CE. Entre estos se encuentran los influyentes trabajos realizados por Working (1943) y Leser (1964) quienes brindaron un conjunto de criterios metodológicos para especificar curvas de Engel mediante el estudio de varios modelos lineales, concluyendo con la especificación denominada Working-Leser (W-L), la cual presenta una relación lineal entre la proporción del gasto en un determinado bien sobre el gasto total y el logaritmo de este último. Posteriormente se ha trabajado con versiones extendidas de la especificación W-L, como ser cuadrática (Banks et al. 1997) o cúbica (Costa 2001, Blow 2003).

En las últimas décadas se ha extendido el uso de modelos no paramétricos y semiparamétricos para el estudio de las CE, puesto que los mismos permiten una forma funcional flexible, a la vez que posibilitan chequear la validez de diferentes formas funcionales específicas, contribuyendo de esta manera a la definición de nuevas curvas de Engel y sistemas de demanda paramétricos.

Dentro de las investigaciones que involucran estimaciones no paramétricas de CE, en general se utiliza un estimador, aunque no siempre se deja en claro el porqué de tal elección. De la literatura relevada, el estimador Nadaraya-Watson parece ser el más utilizado para la estimación no paramétrica y semiparamétrica de curvas de Engel (Banks et al. 1997, Delgado & Miles 1997, Blundell et al. 1998, Fousekis & Lazaridis 2001, Blundell et al. 2003, Barrientos 2009). Otros trabajos utilizan polinomios locales lineales o cuadráticos, como Deaton & Paxson (1998), Yatchew

(2003), Parpiev & Yusupov (2011) y Chakrabarty & Hildenbrand (2011). En los últimos tiempos también se ha extendido el uso de suavizados *spline*, aunque con menor extensión (Blow 2003, Yatchew 2003, Buseti & Miani 2006).

En base a todos los posibles estimadores que pueden aplicarse, es deseable encontrar aquel que proporcione un menor error de estimación y predicción. Para ello existen diversas medidas comúnmente utilizadas a fin de comparar diferentes estimadores no paramétricos o criterios de suavizados de los mismos. Muchas veces tales medidas no proporcionan un *ranking* consistente para la elección de un método de estimación en particular. Por ello, el objetivo de este trabajo es evaluar tales medidas de desempeño para el caso especial de las curvas de Engel. Para ello se realiza un ejercicio de simulación, generando los datos a partir de especificaciones tipo W-L, calibrados con base en parámetros obtenidos de estimaciones con datos de la Encuesta Nacional de Gasto de Hogares del período 1996-97 realizada por el Instituto Nacional de Estadísticas y Censos de Argentina (INDEC). Adicionalmente se analizan los resultados de aplicar tales medidas a la CE correspondiente al consumo de alimentos, donde se desconoce la forma funcional verdadera que adopta la CE.

Entre los estimadores no paramétricos a comparar se elijen: el estimador Nadaraya-Watson (W-W), dos estimadores basados en polinomios locales, el estimador local lineal y cuadrático local y por último el suavizado *spline*. Cabe mencionar que este trabajo forma parte de los resultados de una investigación más compresiva sobre modelos no paramétricos y semiparamétricos de CE (García 2012).

2. Metodología

2.1. Especificación y estimación de las curvas de Engel

La expresión estocástica general de una curva de Engel especificada a través de la participación en el gasto que tiene un determinado bien j en el presupuesto del hogar i , w_{ij} , como función del ingreso o gasto total G_i , (tomado en su logaritmo), viene dada por:

$$w_{ij} = m_j(\ln G_i) + \varepsilon_{ij} \quad (1)$$

donde $m_h \in \mathfrak{S}_h$, siendo \mathfrak{S}_h un espacio de funciones suaves. Para la estimación de $m_{h,j}(x) = E(w_{ij} | \ln G_i = x)$ existen métodos de regresión o suavizados no paramétricos con propiedades particulares que hacen que algunos estimadores se prefieran a otros, sobre la base de criterios que van desde su simplicidad computacional a su mejor poder predictivo. En las investigaciones econométricas aplicadas los estimadores más utilizados son aquéllos basados en núcleos o *kernel*, incluyendo a los polinomios locales; últimamente se ha extendido el uso de regresiones *spline*, aunque siguen predominando los primeros (Cameron & Trivedi 2005).

Por lo tanto, en este trabajo se considerarán cuatro estimadores: 1) Nadaraya-Watson o estimador local constante; 2) Polinomios locales lineales; 3) Polinomios

locales cuadráticos; y 4) Suavizados *spline*. Estos estimadores forman parte de una clase más general de suavizados lineales. Formalmente, se dice que un estimador \hat{m} de m es un suavizado lineal si existe una matriz \mathbf{L} , denominada matriz de suavizado, tal que:

$$\hat{m} = \mathbf{L}\mathbf{w} \quad (2)$$

Donde $\hat{m} = (\hat{m}(x_1), \dots, \hat{m}(x_n)) = (\hat{w}_1, \dots, \hat{w}_n) \equiv \hat{\mathbf{w}}$, y \mathbf{L} es una matriz de orden $n \times n$ con $[L]_{ij} = l_j(x_i)$, por lo que cada fila de \mathbf{L} es un vector de ponderaciones que se aplican a cada w_i para obtener el estimador $\hat{m}(x_1)$. La particularidad de cada estimador está en la forma que adopta la matriz de suavizados, la que depende de la ventana o *bandwidth* (h), para el caso de los estimadores basados en núcleos o del parámetro de penalización del problema de mínimos cuadrados penalizados correspondiente al suavizado *spline*. Sin embargo, para todos los suavizados lineales se tiene una medida en común que muestra la cantidad o monto de suavizado de un estimador, denominada grados de libertad del estimador (Loader, 2004, p.11). Los mismos pueden definirse por:

$$v = \text{tr}(\mathbf{L}) = \sum_{i=1}^n \text{lev}(x_i) \quad (3)$$

Donde *lev* representa los *leverage*, i.e. los elementos de la diagonal principal de la matriz de suavizado, y miden la sensibilidad de la curva ajustada $\hat{m}(x_i)$ en cada punto particular de los datos observados.

Para la elección de los suavizados se elige el criterio de convalidación cruzada generalizada (*generalized cross-validation*, GCV), el que da suavizados óptimos similares al criterio de *leave-one-out* de convalidación cruzada, con un menor costo computacional.

2.2. Medidas de desempeño

Dados los diferentes suavizados lineales propuestos, es necesario contar con alguna medida que indique el o los estimadores preferidos con base en algún criterio (estadístico), a fin de poder focalizar el estudio de las relaciones de Engel bajo una perspectiva más económica, una vez seguros de que el estimador elegido es el mejor entre una cierta clase de estimadores alternativos.

Sobre la base de esto, se proponen una serie de medidas de desempeño estadística que serán aplicados a los diferentes suavizados de curvas de Engel propuestos, toda vez que cada uno de ellos es estimado utilizando un suavizado óptimo. Con base en el error de predicción (EP), un estimador *naive* del mismo es el Error Cuadrático de Predicción Promedio en la muestra (ECPP), definido por:

$$ECPP = \frac{1}{n} \sum_{i=1}^n [w_i - \hat{m}_h(x_i)]^2 \quad (4)$$

Dicha medida es muy simple de computar y, con algunas variantes, es comúnmente utilizada en la literatura empírica para comparar el desempeño estadística de diferentes tipos de suavizados (Aydin 2007). Para el caso de curvas de Engel, Blundell et al. (1998) utilizan el ECPP a fin de comparar diferentes suavizados basados en núcleos. Esta medida si bien puede ser útil para confrontar diferentes estimadores no paramétricos, resulta muy deficiente como medida de desempeño para la búsqueda del suavizado óptimo, pues la minimización del (4) se logra cuando $w_i = \hat{m}_h(x_i)$ para cada i . De todas formas, en el presente contexto se computa una vez que ya fue elegido el suavizado óptimo mediante el criterio GCV. Una medida análoga a (4) que en ciertos casos permite ver mejor la diferencia de desempeño entre dos estimadores, es el Error Absoluto de Predicción Promedio (EAMP), el que puede escribirse de la forma:

$$EAPP = \frac{1}{n} \sum_{i=1}^n |w_i - \hat{m}_h(x_i)| \quad (5)$$

Adicionalmente a estas dos medidas basadas en el error de predicción, se calcula el riesgo de estimación C_p introducido por Mallows (1973) en el contexto de regresiones paramétricas. Considerando el error de estimación, dado por $\sum_{i=1}^n [\hat{m}_i(x) - m_h(x_i)]^2$, el criterio C_p propone un estimador insesgado del mismo definido por:

$$C_p = \frac{1}{\sigma^2} \sum [w_i - \hat{m}_h(x_i)]^2 - n + 2v_1 \quad (6)$$

Con $\sigma^2 = \text{Var}(\varepsilon_i)$. Una forma alternativa de expresarlo es (Wasserman 2006)

$$C_p = \frac{1}{n} \left\{ \sum [w_i - \hat{m}_h(x_i)]^2 + 2\sigma^2 v_1 \right\} \quad (7)$$

Engel y Kneip (1996) utilizan esta versión del C_p para comparar distintas estimaciones de CE no paramétricas, algunas de ellas basadas en suavizados óptimos locales. La expresión (7) permite ver que el estadístico C_p también puede ser entendido como una corrección del estimador *naive* ECPP, penalizando por el número de parámetros, lo que surge de tomar esperanza en (7). Específicamente, es fácil probar que $E(EAPP) = EP + \sigma^2 \text{tr}(\mathbf{L})n^{-1}$, lo que significa que el estadístico C_p resulta ser un estimador más sensible del error de predicción. Como se verá en la aplicación empírica y en el ejercicio de simulación, la corrección introducida por C_p puede ser crucial en la comparación de desempeño de estimadores no paramétricos¹.

De (6) y (7) se observa que C_p depende de la varianza. Un estimador consistente de la misma es:

$$\sigma^2 = \frac{1}{n - 2v_2 + u} \sum_{i_1}^n [w_i - \hat{m}(x_i)]^2 \quad (8)$$

¹El criterio C_p suele también utilizarse para la selección de la ventana óptima de los estimadores no paramétricos. Sin embargo, respecto al criterio GCV, puede que las diferencias sean despreciables, dado que el estadístico C_p puede expresarse como una aproximación del GCV (Wasserman 2006, p.70). Por lo tanto, en la presente investigación la medida C_p se usará para comparar los diferentes suavizados con los grados de libertad óptimos ya seleccionados.

Donde $v_2 = \text{tr}(\mathbf{L}^T \mathbf{L})$ es una definición alternativa de grados de libertad de un suavizado lineal (Loader, 2004).

De (8) se puede ver que la estimación de la varianza depende de los grados de libertad de la estimación no paramétrica, por lo que se tiene una estimación de σ^2 para cada ajuste no paramétrico. Sin embargo, para el análisis comparativo es necesario implementar (8), considerando el mismo estimador de σ^2 para los diferentes ajustes correspondientes a diferentes parámetros de suavizado. Cleveland & Devlin (1988) recomiendan estimar σ^2 usando el ajuste con el menor bandwidth (o mayor grado de libertad) entre los que se están considerando.

La última medida a considerar está basada en el criterio de información de Akaike, siguiendo la versión mejorada por Hurvich et al. (1998). La misma fue propuesta como criterio para la selección del suavizado. Sin embargo, dado un suavizado, puede utilizarse para comparar diferentes estimadores no paramétricos. La misma viene dada por:

$$AIC_c = \ln \left\{ \frac{1}{n} \sum_{i_1}^n [w_i - \hat{m}(x_i)]^2 \right\} + 1 + \frac{2[v+1]}{n-v-2} \quad (9)$$

2.3. Datos y simulación

La evaluación de las medidas de desempeño para comparar estimadores no paramétricos se realizará mediante un sencillo ejercicio de simulación, en donde las verdaderas curvas de Engel se corresponden con alguna versión de la forma funcional Working-Leser, sea lineal, cuadrática o cúbica. De esta manera, se generan datos a partir de la siguiente especificación polinómica general de una Working-Leser paramétrica:

$$w_i = b_{i0} + b_{i1}X_i + b_{i2}X_i^2 + b_{i3}X_i^3 + u_i \quad (10)$$

Con:

$$X_i \sim N(\mu_x, \sigma_x^2), \quad u_i \sim N(0, \sigma_u^2) \quad (11)$$

De (10)-(11) se pueden plantear tres modelos: lineal, cuadrático y cúbico. Los parámetros de los modelos son calibrados a partir de los datos de la Encuesta Nacional de Gasto de Hogares (ENGH) elaborada por el Instituto Nacional de Estadísticas y Censos de Argentina (INDEC) del período 1996-97, tomando el consumo de alimentos y bebidas (A) en la región Pampeana. De aquí se tienen los siguientes valores:

- Lineal: $b_{i0} = 1.13$, $b_{i1} = -0.11$, $b_{i2} = b_{i3} = 0$.
- Cuadrático: $b_{i0} = 0.95$, $b_{i1} = -0.06$, $b_{i2} = 0.005$, $b_{i3} = 0$.
- Cúbico: $b_{i0} = -2.36$, $b_{i1} = 1.58$, $b_{i2} = -0.27$, $b_{i3} = 0.014$.

A su vez, se fija $\mu_x = 6.3$ y $\sigma_x^2 = 0.75$. Dada las características de los datos microeconómicos, los términos de error se modelan heterocedásticos, en función de X . Para la especificación de la función de varianza se siguió a Cameron & Trivedi (2005), ajustando los valores de acuerdo a la calibración basada en la ENGH. De esta manera se toma:

$$u_i = \varepsilon_i \sqrt{0.015 - 0.2X_i}, \quad \varepsilon_i \sim N(0, 0.16) \quad (12)$$

Los datos se generan considerando tres tamaños de muestra (N): 100, 1000 y 5000. Sobre la base de los datos simulados se estima el modelo $E(w|X) = m(X)$, utilizando los cuatro estimadores propuestos: el N-W, polinomios locales lineales (PL1) y cuadráticos (PL2) y suavizados *splines* (SS), siguiendo el criterio de convalidación cruzada generalizada (GCV) para la elección del suavizado². A partir de tales estimaciones se computan las medidas empíricas de desempeño o evaluación estadística presentadas anteriormente: los errores cuadráticos ECPP y EAPP, el estadístico C_p y el AICc. A fin de evaluarlos, se los compara con el error cuadrático promedio y el error de predicción, i.e.:

$$ECP_{np} = n^{-1} \sum_{i=1}^n (\hat{m}_{np}(X_i) - m(X_i))^2 \quad (13)$$

Donde $m(X_i) = \sum_j b_{ij} X_i^j$ según los modelos lineal, cuadrático o cúbico se están postulando, y $\hat{m}(X_i)$ es el valor estimado de la función en X_i , para un determinado estimador no paramétrico np , con np =N-W, PL1, PL2, SS. Adicionalmente, se computa el error promedio de predicción:

$$EPP_{np} = n^{-1} \sum_{i=1}^n (\hat{m}_{np}(X_i^+) - m(X_i^+))^2 \quad (14)$$

Siendo $\{X_i^+\}_{i=1}^n$ una sucesión de nuevas observaciones. Puesto que en la simulación se realizan varias réplicas, si con $X_{i,k}$ se denota a la observación i muestreada en la etapa k , luego en la estimación del modelo con la réplica k se toma $\{X_i^+\}_{i=1}^n \equiv \{X_{i,k-1}^+\}_{i=1}^n$. Es decir, se toma como nuevo conjunto de variables para evaluar la predicción, al conjunto de co-variables de la réplica anterior. Específicamente, dados los estimadores no paramétricos de una réplica k , los mismos se utilizan para predecir con los valores muestreados en la etapa anterior, y a su vez, se calcula la predicción según el verdadero modelo.

De esta manera, se comparan los valores de las medidas empíricas de desempeño con los verdaderos errores de estimación y predicción dados por (13) y (14). Claro está que las conclusiones quedan acotadas a que la CE verdadera toma la forma simple dada en (10).

²Para la estimación se utiliza el *software* R. Para las funciones basadas en núcleos se usa el paquete *locfit*, con el núcleo tricúbico (por *default*) y el paquete *smooth.spline* para los suavizados *spline*.

A fin de analizar el comportamiento de las medidas de desempeño para datos reales, las mismas se aplican a los datos del consumo total de alimentos y bebidas de la región pampeana argentina según la ENGH. De aquí se analiza la convergencia/divergencia que existe en el *ranking* propuesto por cada medida de los estimadores con menores errores empíricos.

3. Resultados

En la Figura 1 se presentan las nubes de puntos generadas por los tres modelos simulados, para una determinada réplica de 1000 observaciones. Se observa que la forma funcional subyacente a cada una de ellas no es directamente visible. Se puede ver una mayor variabilidad en el extremo inferior de las X 's, lo que es compatible con los observado de la encuesta de gastos.

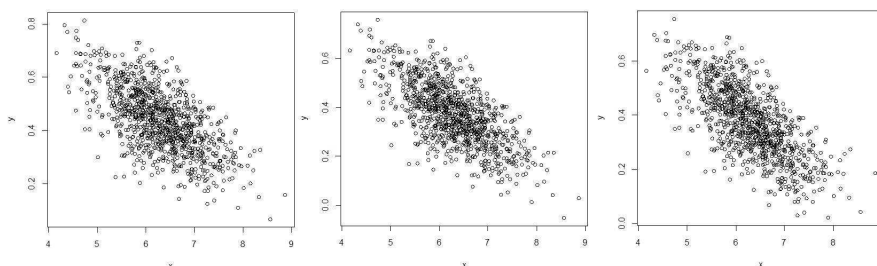


Figura 1: *Diagrama de puntos de los datos generados con $N = 1000$. Fuente: Elaboración propia.*

Para esta réplica las curvas estimadas por cada método, concatenadas con la verdadera curva, se presentan en las Figuras 2 a 4. Las mismas ya dan una idea sobre qué estimador es más exacto en cada caso; sin embargo los resultados pueden variar ante diferentes réplicas. Por ello, la simulación se lleva a cabo con 500 réplicas, para luego comparar los promedios de las medidas de evaluación de estimadores seleccionadas.

Los resultados de las 500 réplicas, para diferentes valores de N , son presentados en las Tablas 1-3, para las tres formas funcionales elegidas. Para el caso de que el verdadero modelo sea el lineal (Tabla 1) se observa que mientras el estimador N-W es el que muestra los menores valores del ECPP y del EAPP, para el AICc y el C_p este estimador arroja los mayores valores. Análogamente, aunque en forma inversa, con el suavizado *spline*. Solo coinciden para el caso del polinomio local lineal, colocándolo en el segundo lugar en términos del menor error. Al observar los valores del ECP y el EPP, el estimador con menor error es el suavizado *spline*, seguido del estimador local lineal. Por lo tanto, sólo las medidas empíricas AICc y C_p proponen un *ranking* de evaluación consistente con los errores de estimación y predicción reales (i.e. ECP y EPP).

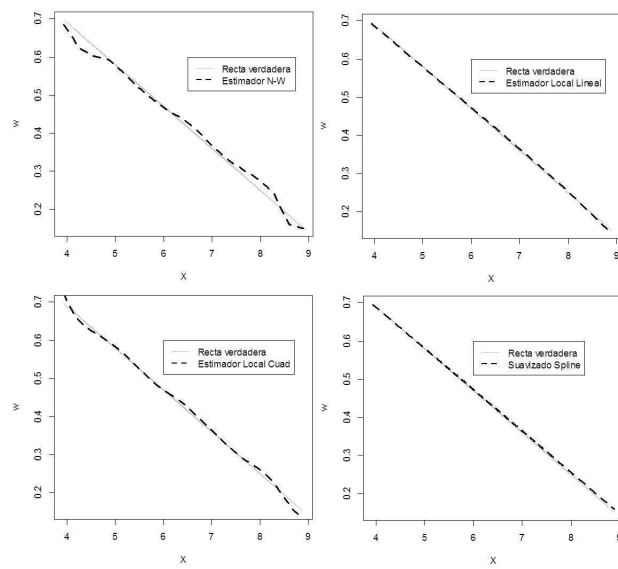


Figura 2: *Estimaciones no paramétricas CE lineal. Fuente: Elaboración propia.*

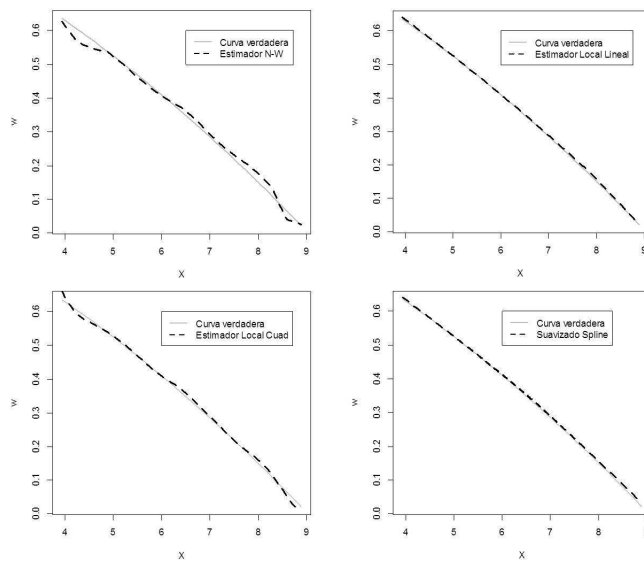


Figura 3: *Estimaciones no paramétricas CE cuadrática. Fuente: Elaboración propia.*

Si la curva verdadera es la cuadrática, los estimadores que arrojan el menor error (real) son los polinomios locales cuadráticos y, nuevamente, los suavizados *splines*.

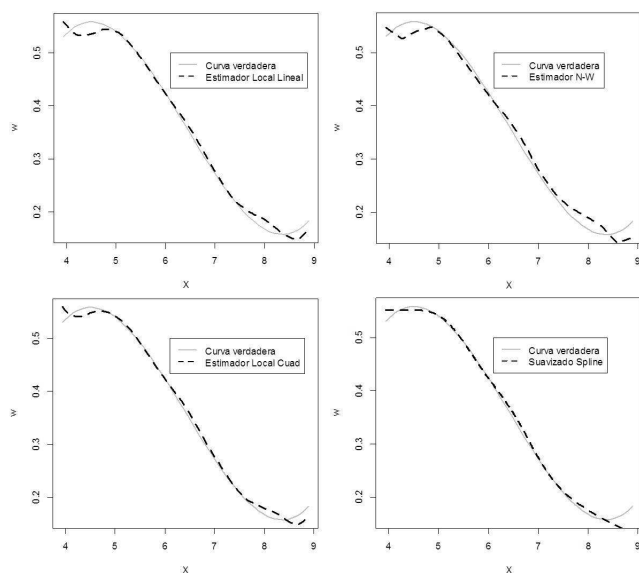


Figura 4: Estimaciones no paramétricas CE cúbica. Fuente: Elaboración propia.

Tabla 1: Medidas de desempeño de los estimadores CE lineal. Fuente: Elaboración propia.

		v	ECPP	EAPP	AICc	C_p	ECP	EPP
N=100	NW	5.47	0.00797	.0709	-3.7036	0.00888	0.00056	0.00056
	PLL	4.57	0.0079	0.07061	-3.7307	0.00867	0.00046	0.00041
	PLC	4.67	0.00795	0.07077	-3.7229	0.0087	0.00048	0.00049
	Spline	3.56	0.00804	0.07131	-3.7369	0.00863	0.00035	0.00028
N=1000	NW	9.95	0.00844	0.07313	-3.753	0.00861	0.00009	0.00009
	PLL	4.61	0.00847	0.07326	-3.76	0.00855	0.00005	0.00003
	PLC	5.22	0.00847	0.07327	-3.7587	0.00856	0.00005	0.00004
	Spline	3.95	0.00848	0.0733	-3.7605	0.00855	0.00004	0.00003
N=5000	NW	15.07	0.00850	0.07336	-3.761	0.00855	0.000026	0.000027
	PLL	4.88	0.00851	0.07343	-3.7634	0.00853	0.000009	0.000006
	PLC	5.61	0.00851	0.07343	-3.7631	0.00853	0.000011	0.000009
	Spline	3.94	0.00852	0.07344	-3.763	0.00853	0.000007	0.000005

Mientras que los estadísticos AICc y C_p muestran ser consistentes con ello, los errores medios empíricos, en términos cuadrático como absoluto, muestran resultados totalmente contrapuestos. Se observa de la Tabla 2 que los grados de libertad óptimos (v) del estimador N-W difieren suficientemente del resto, y puesto que el ECPP y el EAPP no corrige por los mismos (contrariamente del C_p y el AICc), la ineficacia de tales medidas se realza con ello.

Para el caso del modelo cúbico, las conclusiones son similares, aunque, mientras las medidas empíricas muestran valores que sugieren indiferencia entre la elección

Tabla 2: Medidas de desempeño de los estimadores CE cuadrática. Fuente: Elaboración propia.

		v	ECPP	EAPP	AICc	C_p	ECP	EPP
N=100	N-W	10.8	0.00764	0.06915	-3.616	0.00951	0.00107	0.00167
	PLL	5.7	0.00785	0.07019	-3.7149	0.00881	0.00064	0.00074
	PLC	4.7	0.00796	0.07077	-3.7229	0.00876	0.00048	0.00049
	<i>Spline</i>	4.8	0.00797	0.07101	-3.7199	0.00878	0.00047	0.00047
N=1000	N-W	18.8	0.0084	0.07292	-3.7402	0.00872	0.00018	0.00022
	PLL	9.7	0.00843	0.07302	-3.7556	0.00859	0.00009	0.0001
	PLC	5.2	0.00848	0.07327	-3.7588	0.00857	0.00005	0.00004
	<i>Spline</i>	6.5	0.00847	0.07324	-3.7574	0.00858	0.00006	0.00005
N=5000	N-W	23.7	0.00851	0.07339	-3.7569	0.00859	0.00006	0.00007
	PLL	13.6	0.0085	0.07335	-3.762	0.00855	0.00002	0.00003
	PLC	5.6	0.00852	0.07343	-3.7631	0.00854	0.00001	0.00001
	<i>Spline</i>	8	0.00851	0.07342	-3.7626	0.00854	0.00001	0.00001

de dos estimadores alternativos, las medidas reales de error muestran una clara superioridad del *spline* cuando $N=100$ y $N=1000$ (aunque una menor diferencia en este último caso), y para $N=5000$, donde los errores se reducen considerablemente, el polinomio local cuadrático supera al *spline*.

Tabla 3: Medidas de desempeño de los estimadores CE cúbica. Fuente: Elaboración propia.

		v	ECPP	EAPP	AICc	C_p	ECP	EPP
N=100	NW	5.98	0.00793	0.07068	-3.6971	0.00893	0.00059	0.000604
	PLL	4.9	0.0079	0.07049	-3.7255	0.00872	0.00054	0.000563
	PLC	4.91	0.00794	0.07067	-3.7198	0.00876	0.00053	0.000593
	<i>Spline</i>	4.1	0.00803	0.07127	-3.7271	0.00871	0.00045	0.00045
N=1000	NW	10.7	0.00844	0.0731	-3.7522	0.00862	0.0001	0.0001
	PLL	7.78	0.00845	0.07314	-3.757	0.00858	0.00008	0.00009
	PLC	6.58	0.00847	0.07322	-3.7574	0.00858	0.00007	0.00007
	<i>Spline</i>	7.02	0.00846	0.07323	-3.7567	0.00858	0.00006	0.00006
N=5000	NW	16.1	0.0085	0.07336	-3.7608	0.00856	0.00003	0.00003
	PLL	12	0.0085	0.07336	-3.7622	0.00854	0.000023	0.000024
	PLC	7.86	0.00851	0.07342	-3.7626	0.00854	0.000016	0.000016
	<i>Spline</i>	9.41	0.00851	0.07341	-3.7622	0.00854	0.000017	0.000017

Para las estimaciones de CE, del consumo total de alimentos y bebidas utilizando la base de datos de ENGH, en la Figura 5 se muestran las cuatro estimaciones en una misma gráfica. En general se observa que, a partir de un gasto total de 150 pesos (a precios de 1996-97), donde la participación del gasto en alimentos y bebidas alcanza un máximo, el comportamiento de las curvas es compatible con la denominada ley de Engel, caracterizándolos como bienes necesarios en la jerga de la teoría del consumidor. Sin embargo, las cuatro estimaciones muestran que para los hogares de menores recursos (menores valores del logaritmo del gasto total), la participación del gasto en alimentos aumenta con el incremento del poder adquisitivo, lo que a *prima facie* contradice la ley de Engel. Este hecho puede

asociarse a la deficiencia en canasta alimenticia de los hogares pobres, tanto en cantidad como en calidad (Barrientos 2009). Específicamente, la pendiente positiva de las líneas tangentes de la CE puede ser consistente con el comportamiento de un consumidor en la búsqueda de una mayor calidad para tales bienes, de forma tal que los convierte en bienes de lujo en el sentido de que incrementan su peso en el presupuesto del hogar (Deaton & Paxson 1998, Fleischer & Rivlin(Byk) 2009).

Al comparar los cuatro estimadores de la gráfica se revela que las diferencias son muy leves, con mayor divergencia en los niveles extremos del gasto total. Al comparar los estimadores locales lineales y cuadráticos, con el local constante (i.e. el N-W) se observa que los dos primeros reducen el sesgo en los niveles extremos. En general, el N-W siempre es más susceptible al sesgo y rara vez resulta ser el más adecuado en comparación con polinomios locales de mayor grado. Esto ocurre porque un incremento del grado de los polinomios locales disminuye el sesgo, aunque aumenta la varianza del estimador. En general los polinomios más utilizados en la literatura empírica son los lineales y cuadráticos, y en algunos casos los cúbicos, puesto que un grado mayor del polinomio no siempre resulta ser más beneficioso (Loader 1999, p.22). Adicionalmente se tiene que, en general, los polinomios con grado impar proporcionan un mejor ajuste respecto a los pares (Härdle et al. 2003).

En la Tabla 4, se presentan los resultados de las cuatro medidas de desempeño consideradas. Adicionalmente, se reportan los parámetros de suavizados y grados de libertad óptimos según el criterio GCV, así como también su *score* mínimo. Según el ECPP, el suavizado *spline* es el que da un menor error promedio, seguido del polinomio local lineal. A su vez, el estimador N-W es el que presenta el mayor valor del ECPP. Por su parte, el error absoluto de predicción coloca al polinomio local cuadrático en primer lugar (en términos del menor error), y en segundo lugar al *spline*. Los valores del C_p y el AICc, son coincidentes en cuanto al *ranking* que establecen de estimadores con menor error, colocando en primer lugar al suavizado *spline*, luego al polinomio local cuadrático, seguido del lineal, y por último al N-W. Este *ranking* también se da al observar los valores mínimos de la función GCV evaluada en los respectivos suavizados óptimos. Por lo tanto, existen bastantes coincidencias entre las diferentes medidas de desempeño.

Tabla 4: *Medidas de evaluación de estimadores no paramétricos. Total alimentos y bebidas. Fuente: Elaboración propia.*

Estimador	h/λ	v	GCV	ECPP	EAPP	C_p	AICc
N-W	0.41	9.14	0.02485	0.024791	0.125937	0.024851	-2.69459
Local lineal	0.66	7.56	0.02484	0.024787	0.125851	0.024837	-2.69518
Local cuadrático	1.33	6.48	0.02483	0.024789	0.125838	0.024832	-2.69537
<i>Spline</i>	0.069	6.73	0.02483	0.024785	0.125846	0.024829	-2.69548

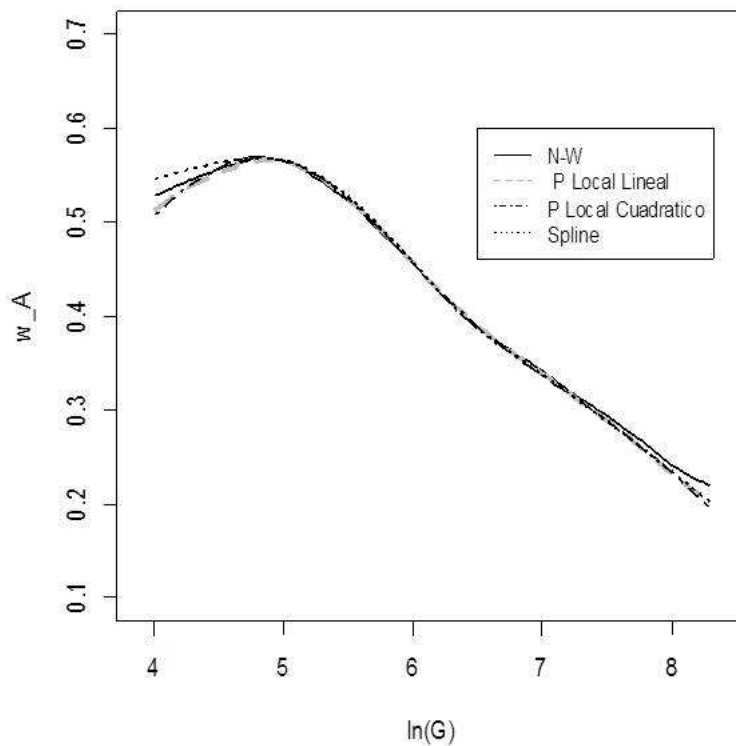


Figura 5: *Curvas de Engel de alimentos y bebidas (A) comparación de estimadores.*
Fuente: *Elaboración propia.*

4. Conclusiones

En este trabajo se evalúan medidas simples de desempeño estadística de un conjunto de estimadores no paramétricos, tomando el caso especial de las denominadas curvas de Engel. Tal evaluación se realiza por medio de un ejercicio de simulación, y luego se analizan los resultados de tales medidas cuando son aplicadas a datos reales.

Los resultados de la simulación muestran que las medidas más simples para evaluar el desempeño y comparar estimadores no paramétricos, el ECPP y el EAPP, no resultan muy recomendables si el modelo que se está estimando tiene una función muy suave y simple como ser lineal, cuadrática o cúbica. Sin embargo, las mismas son muy utilizadas en la literatura empírica, por lo que estos resultados buscan mostrar una alerta para las futuras investigaciones. Como contraparte, se observa que los estadísticos C_p y AICc muestran muy buen desempeño para tales fines.

Para una interpretación cautelosa de estos resultados deben tenerse en cuenta las siguientes consideraciones: 1) Para cada estimador se aplicó el criterio de convali-

dación cruzada generalizada (GCV) para la obtención del suavizado óptimo. Por lo tanto, debe tenerse presente que otro criterio podría arrojar diferentes resultados para las medidas de evaluación. 2) Las CE que se proponen como verdaderas adoptan una forma quizás muy simple y suave para evaluar el desempeño de las medidas empíricas. Sin embargo, tales formas funcionales son las más vistas como especificaciones paramétricas en la literatura micro-econométrica. 3) Quizás sea más informativo evaluar las medidas empíricas para estimadores con los mismos suavizados o grados de libertad. Pero si se quiere evaluar la decisión del investigador aplicado, en elegir el método y luego suavizarlo con algún criterio automático, es más informativo mirar el output que tendría de su elección.

Para los datos reales, se observan mayores coincidencias entre las medidas de desempeño, por lo que su uso complementario puede ser recomendable, y más aún cuando no puede probarse estadísticamente que las CE se corresponden con alguna forma simple conocida.

Recibido: 07 de septiembre de 2012

Aceptado: 22 de noviembre de 2012

Referencias

- Aydin, D. (2007), 'A comparison of the nonparametric regression models using smoothing spline and kernel regression.', *World Academy of Science, Engineering and Technology* **36**, 253–257.
- Banks, J., Blundell, R. & Lewbel, A. (1997), 'Quadratic engel curves and consumer demand', *Journal of Economics and Statistics* **79**(4), 527–539.
- Barrientos, M. J. (2009), 'On the consumer behavior in urban colombia: The case of bogotá', *Ensayos sobre Política Económica (Colombia)* **27**(59), 46–82.
- Blow, L. (2003), Demographics in demand analysis. Working Paper 03/18. The Institute of Fiscal Studies.
- Blundell, R., Browning, R. & Crawford, I. (2003), 'Nonparametric engel curves and revealed preference', *Econometrica* **71**(1), 205–240.
- Blundell, R., Duncan, A. & Pendakur, K. (1998), 'Semiparametric estimation and consumer demand', *Journal of Applied Econometrics* **13**, 435–461.
- Busetti, F. & Miani, C. (2006), Testing of polynomial specification in semiparametric regression. Unpublished manuscript, Research Department, Bank of Italy.
- Cameron, A. C. & Trivedi, P. K. (2005), *Microeconometrics: Methods and Applications*, New York: Cambridge University Press.

- Chakrabarty, M. & Hildenbrand, W. (2011), 'Engel's law reconsidered', *Journal of Mathematical Economics* **47**(3), 289–299.
- Cleveland, W. & Devlin, S. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting.', *Journal of the American Statistical Association*. **83**, 596–610.
- Costa, D. (2001), 'Aggregation, rank and some historical engel curves', *Economic Letters* **70**, 183–186.
- Deaton, A. (1980), 'Quality, quantity, and spatial variation of price', *The American Economic Review* **78**(3), 418–430.
- Deaton, A. & Paxson, C. (1998), 'Economies of scale, household size, and the demand for food', *The Journal of Political Economy* **106**(5), 897–930.
- Delgado, M. A. & Miles, D. (1997), 'Household characteristics and consumption behaviour: A nonparametric approach', *Empirical Economics* **22**(3), 409–429.
- Fleischer, A. & Rivlin(Byk), J. (2009), 'More or better? quantity and quality issues in tourists consumption.', *Journal of Travel Research*. **47**(3), 285–294.
- Fousekis, P. & Lazaridis, P. (2001), 'Food expenditure patterns of the urban and the rural households in greece. a kernel regression analysis', *Agricultural Economic Review* **2**(1), 5–19.
- García, R. (2012), Evaluación de curvas de engel no paramétricas y semiparamétricas. una aplicación al consumo de alimentos en la región pampeana, Tesis de maestría en estadística aplicada, Universidad Nacional de Córdoba.
- Härdle, W., Müller, M., Sperlich, S. & Werwatz, A. (2003), *Nonparametrics and Semiparametrics Models. An Introduction.*, Berlin: Springer. (e-book: <http://fedc.wiwi.huberlin.de/xplore/ebooks/html/spm/>).
- Hurvich, C., Simonoff, J. & Tai, C. L. (1998), 'Smoothing parameter selection in nonparametric regression using improved akaike information criterion.', *Journal of the Royal Society, Series B*. **60**, 271–293.
- Leser, C. B. V. (1964), 'Forms of engel functions', *Econometrica* **32**(4), 694–703.
- Lewbel, A. (2006), *The New Palgrave Dictionary of Economics*, 2nd edn, Palgrave Mcmillan, chapter Engel Curves.
- Loader, C. (1999), *Local Regression and Likelihood.*, Berlin: Springer.
- Mallows, C. L. (1973), 'Some comments on c_p .', *Technometrics*. **15**, 661–675.
- Parpiev, Z. & Yusupov, K. (2011), 'Testing household economies of scale in uzbekistan', *Eurasian Journal of Business and Economics* **4**(7), 25–51.
- Wasserman, L. (2006), *All of Nonparametric Statistics.*, New York: Springer.

- Working, H. (1943), 'Statistical laws of family expenditure', *Journal of American Statistical Association* **38**, 43 – 56.
- Yatchew, A. (2003), *Semiparametric Regression for the Applied Econometrician*, Cambridge University Press.