Servy. Elsa
Garcia, María del Carmen
Paccapelo, Valeria
Instituto de Investigaciones Teóricas y Aplicadas, de la Escuela de Estadística

REGRESIÓN NO PARAMÉTRICA: UNA APLICACIÓN

1.- INTRODUCCIÓN

La teoría clásica de la regresión se basa, en gran parte, en el supuesto que las observaciones son independientes y se encuentran idéntica y normalmente distribuidas. Si bien existen muchos fenómenos del mundo real que pueden modelarse de esta manera, para el tratamiento de ciertos problemas, la normalidad de los datos es insostenible. En el intento de eliminar esa restricción se diseñaron métodos que hacen un número mínimo de supuestos sobre los modelos que describen las observaciones.

La teoría de los métodos no paramétricos trata, esencialmente, el desarrollo de procedimientos de inferencia estadística, que no realizan una suposición explícita con respecto a la forma funcional de la distribución de probabilidad de las observaciones de la muestra. Si bien en la Estadística no paramétrica también aparecen modelos y parámetros, ellos están definidos de una manera más general que en su contrapartida paramétrica.

La regresión no paramétrica es una colección de técnicas para el ajuste de funciones de regresión cuando existe poco conocimiento a priori acerca de su forma. Proporciona funciones suavizadas de la relación y el procedimiento se denomina suavizado.

Los fundamentos de los métodos de suavizado son antiguos pero sólo lograron el estado actual de desarrollo gracias a los avances de la computación y los estudios por simulación han permitido evaluar sus comportamientos.

La técnica más simple de suavizado, los promedios móviles, fue la primera en usarse, sin embargo han surgido nuevas técnicas como la estimación mediante núcleos ("kernel") o la regresión local ponderada. Estos estimadores de regresión no paramétrica son herramientas poderosas para el análisis de datos, tanto como una técnica de estimación para resumir una relación compleja que no puede ser aprehendida por un modelo paramétrico, como para suplementar (o complementar) un análisis de regresión paramétrico.

En este este trabajo se presenta una aplicación de estos métodos para el ajuste de modelos de regresión para explicar el ingreso del jefe del hogar, a partir de información suministrada por la Encuesta Permanente de Hogares, relevada por el INDEC, para el aglomerado Rosario en la segunda onda de 2002. El mismo se realiza en el marco del proyecto "Métodos no paramétricos y semiparamétricos para el análisis de regresión con datos univariados y multivariados".

2. Regresión no paramétrica

En los análisis paramétricos se comienza haciendo supuestos rígidos sobre la estructura básica de los datos, luego se estiman de la forma más eficiente posible los parámetros que definen la estructura y por último se comprueba si los supuestos iniciales se cumplen.

La regresión no paramétrica, en cambio, desarrolla un "modelo libre" para predecir la respuesta sobre el rango de valores de los datos. Básicamente está constituida por métodos que proporcionan una estimación suavizada de la relación para un conjunto de valores (denominado ventana) de la variable explicativa. Estos valores son ponderados de modo que, por ejemplo, los vecinos más cercanos tengan mayor peso que los más alejados dentro de una ventana de datos. Se pueden utilizar diversas funciones de ponderación, que son los pesos en que se basan los estimadores. La combinación de la función de ponderación y el ancho de la ventana inciden sobre la bondad de la estimación resultante.

La mayor parte de las publicaciones sobre regresión no paramétrica consideran el caso de un solo regresor a pesar de que, a simple vista no pareciera de gran utilidad, ya que las aplicaciones más interesantes involucran varias variables explicativas. Sin embargo, la regresión no paramétrica simple es importante por dos motivos:

- En etapas preliminares del análisis de datos o en pruebas de diagnóstico se utilizan gráficos de dispersión en los cuales puede ser muy útil ajustar una "curva suavizada". Por ejemplo, para explorar la forma de la función respuesta, para confirmar una función respuesta en particular que haya sido ajustada a los datos, para obtener estimaciones de la respuesta media sin especificar la forma de la función respuesta, para estudiar el cumplimientos de supuestos, etc.
- Forma la base a partir de la cual se extienden los conceptos para regresión no paramétrica múltiple.

El análisis de regresión considera que una variable respuesta Y es función de un conjunto de variables explicativas X_1, X_2,…,X_p. En general, se asume que la relación entre las variables es de tipo lineal y $g(.)$ es una función que depende de parámetros (β_j)

$$y_i = g(\mathbf{x}_i^{'}) + \varepsilon_i ,$$

siendo,

y_i, i = 1, ..., n, i-ésima observación de la respuesta

$\mathbf{x'}_i$ el vector que incluye los valores observados (x_{ij}, j=1,…,p) de las variables explicativas.

ε_i es el error aleatorio con $E(\varepsilon_i)=0$, variancia constante σ^2 a través de i y no correlacionados para diferentes valores de i.

Estos supuestos conducen al modelo de regresión que se estima mediante el método de mínimos cuadrados. Se supone, además, que los errores siguen una distribución normal.

El principal objetivo es estimar la función de regresión $E(y_i) = g(x_{i1}, x_{i2},..., x_{ip})$ en base a una muestra.

Sin embargo, los supuestos realizados pueden ser muy fuertes y para muchos problemas del mundo real no son adecuados.

El análisis de regresión no paramétrico en lugar de estimar los parámetros tiene por objeto una estimación directa de $g(.)$ y sustituye el supuesto de linealidad por uno más débil que consiste en que el valor promedio $g(.)$ es una función de regresión suave.

No exigir linealidad implica mayor trabajo computacional y en algunas circunstancias, resultados más complejos pero a su vez se logra una estimación de $g(.)$ más adecuada. Por otra parte, en algunos casos suponer linealidad puede provocar resultados sin sentido.

Una de las desventajas de la regresión no paramétrica es que no intenta especificar una expresión analítica de la función de regresión $g(.)$. Identificar la forma de dicha función

no siempre resulta de interés y es suficiente determinar la existencia de una relación entre las variables explicativas y la respuesta.

2.1 Técnicas de suavizado

Muchos métodos de suavizado fueron desarrollados inicialmente para datos de series de tiempo en donde la variable explicativa indica períodos de tiempo equiespaciados y luego extendidos al caso de regresión simple y múltiple. La técnica de suavizado más popular es la de los promedios locales ponderados.

El *método de los promedios locales* considera que el tamaño de la muestra es grande y divide el rango de las variables explicativas en intervalos que pueden solaparse y se los denomina "ventanas". Los centros de esas ventanas se denominan focos.

Se puede estimar la función $g(\mathbf{x}_i)$ para un gran número de valores focales, que generalmente están igualmente distribuidos dentro del rango de valores observados de las variables explicativas o en los valores observados ordenados ($x_{(ij)}$).

Existen dos formas de elegir el ancho de la ventana:

1. Fijo: la amplitud (h) de los intervalos centrados en el valor focal $\mathbf{x}_0'$ es la misma para todos ellos

2. Ajustada: la amplitud de los intervalos se ajusta para que todos ellos contengan igual número de observaciones (m). Estas observaciones son las más cercanas al valor focal ($\mathbf{x}_0'$) de cada ventana.

La primera opción funciona bien si la variable explicativa está uniformemente distribuida. Si no fuera así y se considerasen intervalos de igual amplitud, los promedios en cada uno de ellos no serían estables y podrían quedar intervalos sin observaciones debido a la asimetría de los regresores.

A medida que el número de predictores aumenta, el número de puntos en el vecindario del punto focal tiende a declinar rápidamente. Cuando se usa la segunda opción, es decir, incluir un número fijo de observaciones en el ajuste local, los vecindarios pueden llegar a ser bastante extendidos. Un supuesto general de los ajustes locales es que las observaciones cercanas al punto focal $\mathbf{x}_0'$ son informativas acerca de $g(\mathbf{x}_0')$. Diversos autores han demostrado al incrementarse el tamaño de los vecindarios alrededor del punto focal disminuye la calidad de la estimación de $g(\mathbf{x}_0')$, aumentando el sesgo del estimador y que en los extremos del rango de valores observados de los regresores se pueden presentar problemas porque se produce un aplastamiento de la función de regresión. El método de promedios locales además del aplastamiento en los extremos, puede producir una estimación de $g(.)$ que no sea suave debido a pequeños saltos que se produce en la estimación al incorporar y quitar observaciones de una ventana.

Dos refinamientos de este método, se presentan a continuación por ser los más utilizados en regresión. Ellos son el método de estimación mediante núcleos y el método de regresiones locales.

2.1.1 Método de estimación mediante núcleos (" Kernel estimation")

El *método de estimación mediante núcleos* es una extensión del método de promedios locales que logra una estimación de $g(.)$ más suave que este. Considera intervalos que se

solapan a lo largo de todo el rango de valores observados de las variables explicativas.

Para la estimación de la función de regresión para un valor determinado de las variables explicativas $g(\mathbf{x_0'})$ con $\mathbf{x}_0' = (x_{01}, x_{02}, ..., x_{0p})$ se considera que las observaciones cercanas a $\mathbf{x}_0'$ son más importantes que las observaciones más distantes. Esta importancia se representa por los pesos que se asignan a las observaciones durante el proceso de estimación.

Sean h la amplitud de la ventana cuyo foco es $\mathbf{x_0'}$ y $\mathbf{z_i}$ el vector de distancias ajustadas por escala entre la *i*-ésima observación y el foco, siendo una componente $z_{ij} = \frac{x_{ij} - x_{oj}}{h}$,

j= 1,2,.., p. Sea $K(\mathbf{z})$ una función de Kernel que asigna pesos *w* a las observaciones de manera que:

- observaciones próximas a $\mathbf{x_0'}$ tienen mayores pesos
- observaciones con igual valor absoluto de *z* tienen igual peso
- observaciones lejanas a $\mathbf{x_0'}$ tienen menores pesos

Existen varias funciones de núcleo, algunas de ellas son:

Nombre	Fórmula	Características
Box	$K(z) = \begin{cases} 1 & \lvert z\rvert \le 0.5 \\ 0 & \lvert z\rvert > 0.5 \end{cases}$	
Normal	$K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$	El ancho de banda es el desvío de una normal centrada en x_0.
Parzen	$K(z) = \begin{cases} \frac{k_1 - z^2}{k^2} & \lvert z\rvert \le c_1 \\ \frac{z^2}{k_3} - k_4\lvert z\rvert + k_3 & c_1 < \lvert z\rvert \le c_2 \\ 0 & \lvert z\rvert > c_2 \end{cases}$	
Rectangular	$K(z) = \begin{cases} 1 & \lvert z\rvert < 1 \\ 0 & \lvert z\rvert \ge 1 \end{cases}$	Los pesos en una ventana son los mismos, equivale al método de *promedios locales sin ponderar*
Triángulo	$K(z) = \begin{cases} 1 - \frac{\lvert z\rvert}{c} & \lvert z\rvert \le \frac{1}{c} \\ 0 & \lvert z\rvert > \frac{1}{c} \end{cases}$	

Tricubo	$K(z)=\begin{cases}\left(1-\lvert z\rvert^3\right)^3 & \lvert z\rvert<1\\ 0 & \lvert z\rvert\geq 1\end{cases}$	

Una vez elegida la función $K(\mathbf{z})$ se calculan los pesos $w_i = K(z_i)$.

Existen dos formas de calcular las ponderaciones de cada respuesta:

1. Calcular los pesos marginales separadamente para cada variable explicativa y luego realizar el producto de los pesos marginales. Esto es, para el predictor j y la observación i-ésima calcular

$w_{ij} = K(z_{ij})$ j=1,…,p i=1,…n.

El peso final atribuído a la i-ésima observación es el producto de los pesos marginales

$w_i = w_{i1} w_{i2} ... w_{ip}$.

2. Medir la distancia $D(\mathbf{x'_i}\ \mathbf{x'_0})$ entre las variables explicativas $\mathbf{x'_i}$ para la observación i-ésima y el punto focal $\mathbf{x'_0}$. Los pesos se calculan directamente para estas distancias

$$\mathbf{w_i} = K\left(\frac{D(\mathbf{x'_i}\ \mathbf{x'_0})}{h}\right).$$

Como antes, el valor h puede ser ya sea fijo o ajustado para incluir un número constante de vecinos más cercanos al punto focal, es decir, que ésta varíe de modo que en cada ventana se consideren exactamente m observaciones. La fracción $\alpha = m/n$ se denomina "span" o parámetro de suavizado. A medida que α se aproxima a 1, más suave es la curva. Ocurre lo contrario a medida que α se acerca a 0.

. Existen varias formas de definir la distancia. La más utilizada es

$$D(\mathbf{x'_i}, \mathbf{x'_o}) = \sqrt{\sum_{i=1}^{p}(z_{ij} - z_{0j})^2}$$

El valor ajustado en $\mathbf{x'_0}$ se calcula como un promedio ponderado de las observaciones:

$$\hat{g}(x'_o) = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

Cuando la función de ponderación es del tipo de las presentadas en el cuadro anterior, sólo intervienen en el promedio aquellas observaciones tales que $K(w_i)>0$.

Como en el método de promedios locales, el estimador puede evaluarse para valores igualmente distribuidos a través del rango de las variables explicativas o para los valores ordenados.

En comparación con el método de promedios locales, el estimador de Kernel es más suave pero aún muestra aplastamiento en los extremos.

2.1.2 Regresión polinomial local ("LOESS").

Una de las aplicaciones más utilizadas de la regresión polinomial local es el método Loess[1] desarrollado por Cleveland (1979). A diferencia de la estimación de Kernel, no calcula promedios localmente ponderados sino que estima funciones de regresión ponderando las observaciones en las proximidades de distintos focos.

Para estudiar el proceso de ajuste, se comenzará considerando la evaluación de la función de regresión en una determinada ventana, cuyo foco llamaremos

$$\mathbf{x}_{\alpha}^{'} = (x_{\alpha 1}, x_{\alpha 2}, ..., x_{\alpha p})$$

El foco es un punto particular de las variables explicativas, y en su ventana supondremos que el modelo asumido es un polinomio de grado r

$$\begin{aligned} y_i = \beta_0 + \beta_{11}(x_{i1} - x_{\alpha 1}) + \beta_{12}(x_{i1} - x_{\alpha 1})^2 + ... + \beta_{1k}(x_{i1} - x_{\alpha 1})^r + ... + \\ \beta_{p1}(x_{ip} - x_{\alpha p}) + \beta_{p2}(x_{ip} - x_{\alpha p})^2 + ... + \beta_{pk}(x_{ip} - x_{\alpha p})^r + \varepsilon_i \end{aligned} \qquad (2.1.2.1)$$

En forma matricial, el modelo para un foco $\mathbf{x}_{\alpha}$ es,

$$\mathbf{Y} = \mathbf{X}_{\alpha}\boldsymbol{\beta}_{\alpha} + \boldsymbol{\varepsilon} \qquad (2.1.2.2)$$

Las estimaciones de los coeficientes se llevan a cabo mediante el método de los mínimos cuadrados ponderados. Según ese método, adjudicando ciertos pesos $\mathbf{w}_{i\alpha}$ a los valores de las respuestas, los estimadores de los coeficientes resultan de minimizar la expresión

$$\sum W_{i\alpha}\boldsymbol{\varepsilon}_i^2 , \qquad (2.1.2.3)$$

donde, $\boldsymbol{\varepsilon}_i$ simboliza el error que corresponde a la i-ésima observación de la regresión (2.1.2.2).

Los estimadores de los coeficientes de la regresión local son,

$$\mathbf{b}_{\alpha} = \hat{\boldsymbol{\beta}}_{\alpha} = (\mathbf{X}_{\alpha}^{'}\mathbf{W}_{\alpha}\mathbf{X}_{\alpha})^{-1}\mathbf{X}_{\alpha}^{'}\mathbf{W}_{\alpha}\mathbf{Y} \qquad (2.1.2.4)$$

$$\hat{\mathbf{Y}} = \mathbf{X}_{\alpha}(\mathbf{X}_{\alpha}^{'}\mathbf{W}_{\alpha}\mathbf{X}_{\alpha})^{-1}\mathbf{X}_{\alpha}^{'}\mathbf{W}_{\alpha}\mathbf{Y} = \mathbf{S}_{\alpha}\mathbf{Y} \qquad (2.1.2.5)$$

Los pesos están dados mediante una función de núcleo $w_{i\alpha} = K(z_{i\alpha}) = K(\frac{x_{ij} - x_{\alpha j}}{h})$,

del tipo descripto en la Sección anterior. Estos pesos son nulos para las observaciones que están fuera de la alfa-ésima ventana, de modo que (2.1.2.3) involucra sólo a las observaciones de esa ventana y (2.1.2.5) sólo tiene sentido para las coordenadas de $\hat{\mathbf{y}}$ que corresponden a la misma. En particular, (2.1.2.5) sirve para estimar la respuesta que corresponde al foco de la ventana,

$$\hat{y}_{\alpha} = s_{\alpha 1}y_1 + s_{\alpha 2}y_2 + ... + s_{\alpha n}y_{\alpha n}$$

Supongamos que el proceso se repite varias veces, tomando como focos a cada uno de

[1] Loess es un caso particular de la regresión polinomial local cuando la función de pesos K que se utiliza para el ajuste es la función tricubo.

los puntos observados de las variables explicativas. Se puede escribir,

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ s_{n1} & s_{n2} & \dots & s_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$$

La matriz nxn de la expresión anterior se denomina matriz de suavizado, se designa S, de modo que,

$$\hat{\mathbf{y}} = \mathbf{Sy}.$$

El grado del polinomio que se ajusta en cada una de las ventanas es r. Si r=0 la regresión polinómica local es equivalente al ajuste mediante núcleos. En cambio, cuando r=1 se realiza un ajuste lineal local. Se puede demostrar que a mayor grado del polinomio, mayor es la flexibilidad que tiene la curva y que, el aumento del valor de r refleja cierta reducción en el sesgo pero a la vez**,** un aumento en la variancia.

El ancho de banda h puede ser fijo o puede variar en función del foco. Cuando h define una "ventana de m observaciones vecinas" es conveniente especificar el grado de suavizado de la curva a través de la proporción de observaciones incluidas en las ventanas. Esta fracción es llamada "span" o constante de suavizado y se simboliza s. Así, el número de observaciones en cada ventana es $m = [n.s]$[2]. Es importante distinguir que mayor es el valor de s, más suave es la curva. Nuevamente surge el problema de aumento en el sesgo simultáneamente con disminución de variancia.

2.1.2.1 Selección de la constante de suavizado (span)

Para este apartado se considera que el ancho de ventana varía de acuerdo con los valores de las variables explicativas en que se realiza la estimación, de manera que cada ventana contenga m observaciones. En dicho caso, interesa determinar el valor de la constante de suavizado s o "span" es decir, de la proporción de observaciones a considerar por ventana.

A continuación se presentan diferentes propuestas para la elección de s. En todos los casos se supone, sin pérdida de generalidad, que se realiza un ajuste lineal local es decir, r=1.

- Prueba y error

El método de prueba y error consiste en un "ciclo" de búsqueda del menor valor de s que provee un ajuste suave. Se comienza tomando $s = 0.5$. Si la curva ajustada es demasiado suave se lleva a cabo un nuevo ajuste con un valor $s < 0.5$, en caso contrario se elige un valor $s > 0.5$. Se evalúa el grado de suavizado para la nueva curva ajustada, si es muy suave se ajusta otra curva con una constante de suavizado menor, si ocurre lo opuesto, el valor de s será mayor. Se continúan probando distintos valores de s hasta encontrar un valor para el cual la curva no sea ni demasiado sinuosa ni demasiado suave.

Este procedimiento de considerar a un ajuste como suave o sinuoso es muy subjetivo.

- Gráfico de dependencia de residuos

[2] *m* es la parte entera del producto entre el número de observaciones *n* y la fracción "*span"*

Se grafican los residuos $\hat{\mathbf{e}}_i = \mathbf{Y}_i - \hat{\mathbf{Y}}_i$ versus cada variable explicativa y se puede complementar este gráfico con una curva de suavizado. En este método de selección de la constante de suavizado, se busca el mayor valor de s que proporcione residuos no relacionados con la variable explicativa.

Si los datos se suavizan demasiado, el gráfico de dependencia mostrará que los residuos dependen en alguna manera de $\mathbf{X}$. Si no se lleva a cabo un suavizado excesivo, el promedio de los residuos será cero para cualquier valor de la variable explicativa, por lo tanto no dependen de ésta.

- Validación cruzada

El método para seleccionar s por validación cruzada se basa en la omisión de cada observación en la regresión local en el punto. La estimación de $E(y/\mathbf{X}=\mathbf{x}_i)$ que omite la i-ésima observación se simboliza $\hat{y}_{i/-i}$. La omisión de la i-ésima observación tiene una consecuencia muy importante: la estimación resultante no depende de la observación (y_i).

Llamando $\hat{y}_{-i}(s)$ a la estimación $\hat{y}_{i/-i}$ para una constante de suavizado s se calculan los residuos omitidos $r_{-i} = y_i - \hat{y}_{-i}(s)$.

La función de validación cruzada es:

$$CV(s) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{y}_{-i}(s)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[r_{-i}\right]^2 .$$

En la práctica, se calcula la función $CV(s)$ para un rango de valores de s y se elige aquél valor de s que brinda el menor valor de la función de validación cruzada.

A pesar de que el uso de la función CV es un método formal, se debe tener en cuenta que se trata de una estimación. Consecuentemente, se debe tener en cuenta la variación muestral; en particular cuando el tamaño de muestra es pequeño el método tiende a proveer valores de s muy pequeños.

- Prueba de hipótesis

Se puede comparar el ajuste de dos "modelos anidados" es decir, dos modelos tales que uno de ellos es un caso particular del otro. Por ejemplo, se pueden comparar los dos ajustes siguientes: una curva loess con $s=1$ (es decir, que usa todas las observaciones para ajustar un polinomio de grado p) y un ajuste que considere $0<s<1$ (modelo más general porque permite una forma más flexible que el polinomio de grado p). El procedimiento se verá más adelante.

2.1.2.2 Ajuste robusto

En situaciones en las que la distribución de la respuesta tiene colas pesadas o existen outliers no es conveniente utilizar para estimar los parámetros el método de mínimos cuadrados. Dado que el ajuste polinómico local se basa en éste, tampoco resulta un ajuste robusto.

Una solución es asignar menor peso a aquellas observaciones que sean posibles outliers, es decir aquellas que tengan residuos $\hat{e}_i = Y_i - \hat{Y}_i$ de mayor valor absoluto. Para llevar a cabo un ajuste robusto se procede de la siguiente manera

1. Se asignan pesos a los $\hat{e}_i$ a partir de una función W_i que dé menor peso a los residuos más distantes de cero y mayor peso a los más próximos. Las funciones de pesos más utilizadas en este caso se presentan en el siguiente cuadro.

Función	Expresión	Características
Bicuadrado	$W_i(e_i) = \begin{cases} \left[1-\left(\frac{e_i}{cS}\right)^2\right]^2 & \lvert e_i\rvert < cS \\ 0 & \lvert e_i\rvert \geq cS \end{cases}$	La constante c generalmente se considera igual a 6 pues, si la distribución fuera normal, este método resultaría tan bueno como el método de mínimos cuadrados. Menores valores de c producen mayor resistencia a los outliers.
Huber	$W_i(e_i) = \begin{cases} 1 & \lvert e_i\rvert \leq cS \\ \frac{cS}{\lvert e_i\rvert} & \lvert e_i\rvert > cS \end{cases}$	Esta función de pesos nunca toma el valor cero. Si la constante c se tomara igual a 2, el método resultaría tan bueno como el de mínimos cuadrados en caso de normalidad.
NOTA: S es una estimación robusta de la variancia de los $\hat{e}_i$ por ejemplo, $S = \text{mediana}\lvert\hat{e}_i\rvert$		

2. Se realiza un ajuste de regresión polinómica local a los residuos. En cada foco $\mathbf{x_0'}$ se minimiza la expresión $\sum_{i=1}^{n} w_i^2 W_i^2 \hat{e}_i^2$ donde w_i son los pesos asignados por cierta función de pesos K según la proximidad al foco.
3. Se calculan los residuos de este nuevo ajuste, se repiten los pasos 1 y 2.

Este proceso de iteración se continúa hasta que se satisface algún criterio de convergencia preestablecido para los valores estimados. Por lo general, se necesitan entre 2 y 4 iteraciones.

Cleveland (1979) denomina proceso lowess (locally weighted scatterplot smoothing) a la regresión local que utiliza la función tricubo para asignar pesos de acuerdo con la proximidad al foco y la función bicuadrado para los pesos robustos.

2.1.2.3 Inferencia estadística

La mayoría de las aplicaciones de regresión no paramétrica tienen como objetivo la visualización de una curva suave en un gráfico de dispersión. En estos casos, la inferencia estadística es de interés secundario. No obstante, en el manejo de regresión no paramétrica múltiple el interés en ésta es mayor. Según se vio en 2.1.2 los valores ajustados $\hat{y}_i = \hat{y}_{/x_i}$

son sumas ponderadas de las observaciones es decir:

$$\hat{y}_i = \sum_{j=1}^{n} s_{ij}\, y_j \qquad i = 1,2,...n$$

donde s_{ij} son los pesos que dependen de los valores de las variables explicativas y se pueden representar en la siguiente matriz denominada matriz de suavizado:

$$\mathbf{S} = \begin{bmatrix} s_{11} & \cdots & s_{1i} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ s_{i1} & \cdots & s_{ii} & \cdots & s_{i1} \\ \vdots & & \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{ni} & \cdots & s_{nn} \end{bmatrix}$$

Luego, $\hat{\mathbf{Y}} = \mathbf{SY}$ donde $\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_1 & \cdots & \hat{y}_i & \cdots & \hat{y}_n \end{bmatrix}'$ es el vector de valores ajustados y el vector de valores observados es $\mathbf{Y} = \begin{bmatrix} y_1 & \cdots & y_i & \cdots & y_n \end{bmatrix}'$. Nótese que la matriz $\mathbf{S}$ juega el mismo rol que la matriz "hat", simbolizada $\mathbf{H}$, en la estimación por mínimos cuadrados para la regresión lineal.

La matriz de covariancias de los valores estimados es $\text{Cov}\left(\hat{\mathbf{Y}}\right) = \mathbf{S}\text{Var}\left(\mathbf{Y}\right)\mathbf{S}' = \sigma^2\mathbf{SS}'$ y el vector de los residuos viene dado por la expresión $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \left(\mathbf{I} - \mathbf{S}\right)\mathbf{Y}$.

Para determinar la matriz $\mathbf{S}$ se debe recordar que $\hat{y}_i$ surge de una regresión dada en (2.1.2.1).

Las pruebas de hipótesis se pueden llevar a cabo ajustando modelos a los datos y comparando las sumas de los cuadrados de los residuos (SCE). Por ejemplo, para probar el efecto de un predictor particular x_j se puede omitir ese regresor del modelo. Sea SSE_1 la suma de cuadrados del modelo completo que tiene gl_1 grados de libertad y sea SSE_0 la suma de cuadrados del modelo con el regresor omitido, el cual tiene gl_0 grados de libertad.

La estadística de la prueba

$$F = \frac{(SCE_0 - SCE_1)/(gl_1 - gl_0)}{SCE_1/n - gl_1} \sim F_{gl_1 - gl_0, n - gl_1}$$

Cuando se estima por mínimos cuadrados una regresión lineal, las pruebas F comparan dos modelos alternativos que se encuentran anidados (un modelo es un caso particular del otro). La misma noción se extiende para este caso de regresión no paramétrica.

Fijado un nivel de significación del $100\alpha\%$, si la probabilidad asociada $p = P\left(F > F_{obs}\right) < \alpha$, se rechaza de la hipótesis nula.

Cuando se estima una función de regresión por mínimos cuadrados los grados de libertad pueden definirse de distintas maneras, todas equivalentes. Se tiene en cuenta la matriz "hat" que tiene la propiedad de ser simétrica e idempotente es decir, $\mathbf{H} = \mathbf{H}'$ y $\mathbf{H} = \mathbf{HH}$. En este caso, los grados de libertad para el modelo resultan:

- $\text{rgo}\left(\mathbf{H}\right) = \text{tr}\left(\mathbf{H}\right)$
- $\text{tr}\left(\mathbf{HH}'\right)$

- $\text{tr}\left(2\mathbf{H}-\mathbf{HH}'\right)$

Mientras que los grados de libertad del error para la misma situación vienen dados por:

$\text{df}_{\text{error}} = \text{rgo}\left(\mathbf{I}_n - \mathbf{H}\right) = \text{tr}\left(\mathbf{I}-\mathbf{H}\right) = n - \text{tr}\left(\mathbf{H}\right)$.

En forma análoga se obtienen los grados de libertad cuando se aplica regresión polinomial local, sustituyendo la matriz $\mathbf{H}$ por la matriz de suavizado $\mathbf{S}$. No obstante, los tres modos de cálculo de los grados de libertad del modelo, en general, no resultan equivalentes. A continuación se presenta una breve justificación de cada uno de ellos:

- $\text{df}_{\text{mod}} = \text{tr}\left(\mathbf{S}\right)$ es fácilmente calculable
- $\text{df}_{\text{mod}} = \text{tr}\left(\mathbf{SS}'\right)$ surge debido a que en modelos lineales ocurre que los grados de libertad del modelo con iguales a la suma de las variancias de los valores ajustados sobre la variancia del error. Por lo tanto:

$$\text{gl}_{\text{mod}} = \frac{1}{\sigma^2}\sum_{i=1}^{n} V\left(\hat{Y}_i\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\sigma^2\sum_{j=1}^{n} s_{ij}^2\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}^2 = \text{tr}\left(\mathbf{SS}'\right)$$

- $\text{df}_{\text{mod}} = \text{tr}\left(2\mathbf{S}-\mathbf{SS}'\right)$ se basa en el resultado que demostraron Hastie & Tibshirani (1990) que indica que $E\left(\text{SCE}\right) = \sigma^2\left[n - \text{tr}\left(2\mathbf{S}-\mathbf{SS}'\right)\right] + \text{sesgo}^2$. Si el sesgo fuera despreciable, un estimador de la variancia del error sería el cociente $\text{SCE}/\left[n - \text{tr}\left(2\mathbf{S}-\mathbf{SS}'\right)\right]$ sugiriendo que lo grados de libertad del error sean $\text{gl}_{\text{error}} = n - \text{tr}\left(2\mathbf{S}-\mathbf{SS}'\right)$ y por lo tanto, $\text{gl}_{\text{mod}} = \text{tr}\left(2\mathbf{S}-\mathbf{SS}'\right)$

2.1.2.4 Usos del método loess

El método loess tiene varias aplicaciones entre las que se destacan:

- Exploración de la forma de la función de regresión
- Confirmación de la forma de la ecuación de regresión
- Estimación de la ecuación de regresión sin la especificación de la forma funcional.

Los diagramas de dispersión permiten visulizar rápidamente la forma de la función de regresión. Sin embargo, en algunas ocasiones el gráfico resulta complejo y se vuelve dificultoso reconocer la forma de la relación. En estos casos es útil explorar la forma de la relación ajustando una curva loess.

Cuando se utilizan más de dos variables explicativas no es posible mostrar gráficamente la superficie de regresión ajustada y en consecuencia no se puede ver su apariencia. A diferencia de la regresión paramétrica no se suministra ninguna expresión de la superficie estimada.

Entonces, para confirmar la forma de la función de regresión elegida se estima una función de regresión no paramétrica y se comprueba si los valores estimados mediante loess caen dentro de la banda de confianza para la función de regresión paramétrica.

Los límites de confianza vienen dados por $\hat{y}_h \pm W\, s\{\hat{y}_h\}$,

donde

$\hat{y}_h$ es el valor estimado o ajustado de la respuesta en el punto $x_h^{'} = (1, x_{1h}, ..., x_{ph})$

$\hat{S}^2(\hat{y}_h) = CME\ (x_h^{'}(X'X)^{-1}x_h)$

$W^2 = p\ F_{(1-\alpha;\ p,\ n-p)}$.

Para el caso de regresión simple la validación se puede realizar gráficamente, dibujando la curva no paramétrica conjuntamente con el intervalo de confianza paramétrico.

3. Aplicación

Los métodos de estimación mediante núcleos y regresión polinómica local ("loess") se usan para obtener una imagen suavizada de la relación entre dos variables sin hacer supuestos rígidos sobre la distribución de las mismas. También se usan para complementar análisis realizados paramétricamente y detectar fallas en su aplicación

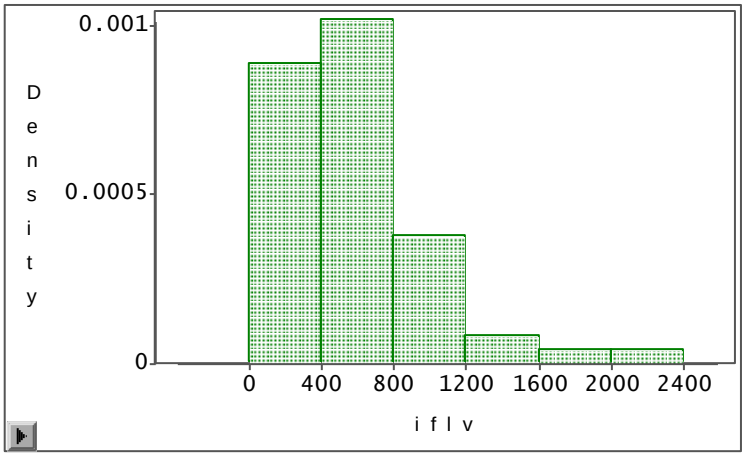
A continuación se va a estudiar la relación entre el ingreso del hogar y el nivel educativo del jefe del hogar y su cónyuge. La información es la relevada por la Encuesta Permanente de Hogares que lleva a cabo el Instituto Nacional de Estadística y Censos (INDEC).

La estimación de modelos de regresión que explican los ingresos individuales presenta algunas dificultades, entre las que se pueden destacar la no normalidad de la variable respuesta. Por lo tanto, no es aconsejable utilizar métodos de regresión paramétricos.
En este trabajo se realiza una estimación paramétrica y otra no paramétrica (mediante núcleos y "Loess") para relacionar la variable ingreso con la educación. Se postulan dos modelos con una y dos variables explicativas, el nivel educativo del jefe de hogar y de éste y su cónyuge, respectivamente.

1. Regresión simple

En el gráfico 3.1 se visualiza la distribución de la variable ingreso del hogar

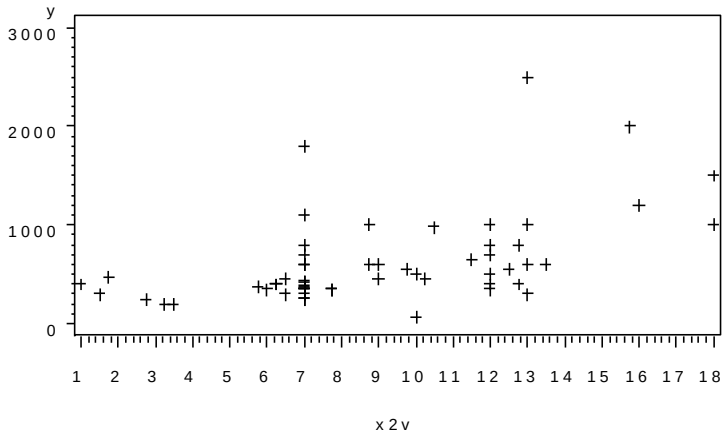
Gráfico 3.1 Distribución de variable ingreso



La distribución es asimétrica y no exiten valores negativos, por lo que puede suponerse que la variable tiene distribución lognormal (Gráfico 3.1).

El diagrama de dispersión para las variables ingreso y años de escolaridad del jefe de hogar se presenta a continuación.

Gráfico 3.2



Ingresos versus años escolaridad del jefe de hogar

En el gráfico se observa que:

- la forma de la relación entre las variables no es clara. Resulta difícil indicar si es lineal o curvilínea,
- presencia de valores atípicos (algunos puntos no parecen concordar con el resto de las observaciones),
- la variancia parece no ser constante.

Si bien se sospecha que no se cumplen los supuestos para realizar un análisis de regresión clásico, se estima el siguiente modelo.

$$\hat{y}_i = 75.398 + 60.446\, x_{2Vi}$$

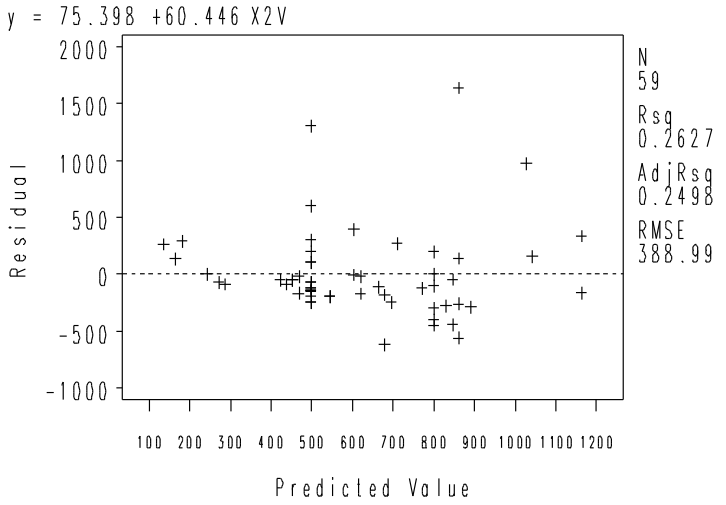
(13.413)

El valor entre paréntesis es el error estándar.

La variable años de escolaridad ayuda a explicar la respuesta ($p<.0001$), en el sentido que es mejor incluirla en el modelo que no incluirla.

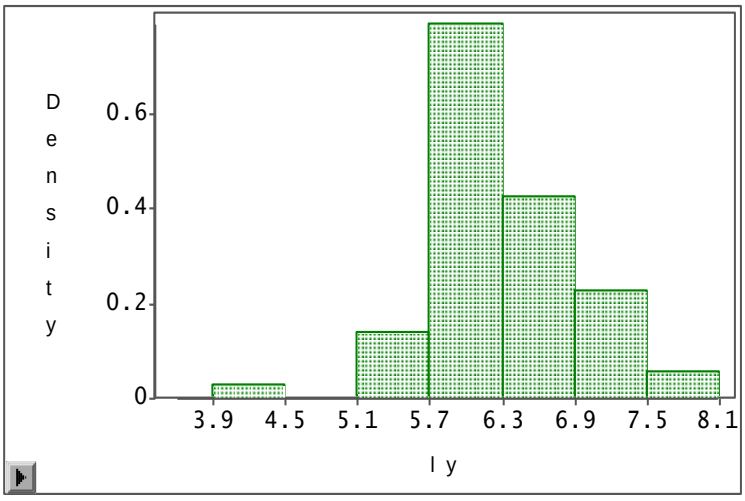
El gráfico de residuos de la regresión mínimo cuadrática refleja que no es adecuado el ajuste anterior.

Gráfico 3.3 Residuos versus valores ajustados



En el contexto de una regresión paramétrica se puede aplicar una transformación logarítmica a la variable ingreso para atenuar el problema de variancia no constante. Esta transformación, a su vez, hace que la distribución de la variable se vuelva más simétrica (gráfico 3.4).

Gráfico 3.4 Distribución de la variable logaritmo del ingreso



Al transformar la variable ingreso, mediante el logaritmo, se logra que la distribución se vuelva más simétrica que la variable sin transformar, pero aún es asimétrica.

Para evitar el uso de transformaciones y realizar un análisis exploratorio del diagrama de dispersión con el fin obtener una aproximación de la forma de la relación se utilizan métodos de suavizado.

Primero se computa una regresión mediante núcleos. El ancho de la ventana es elegido automáticamente por el software mediante validación cruzada (GCV), siendo su valor 2.167. El gráfico 3.5 presenta la curva paramétrica (línea entera) y la estimada mediante núcleos (línea de puntos). Se usan, además, otras amplitudes de ventana diferentes (band-

width=1.1 y 3.3) para mostrar el cambio en la forma de la curva (gráfico 3.6). Se observa que a medida que la amplitud disminuye la gráfica se vuelve más irregular.

Gráfico 3.5 Ajuste paramétrico y mediante núcleos óptimo (h=2.167)

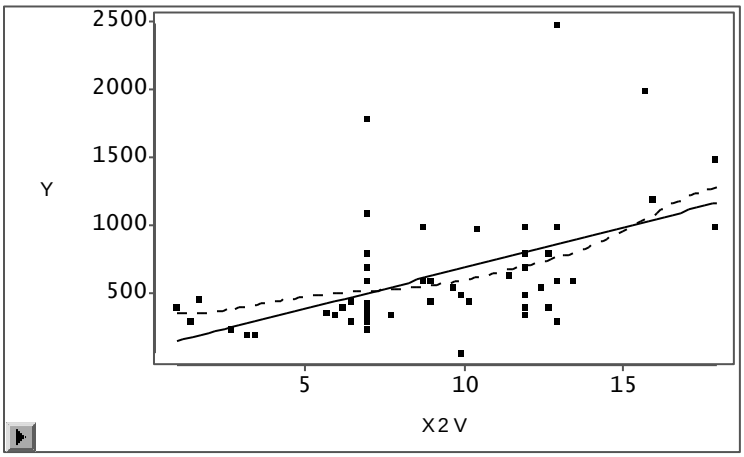
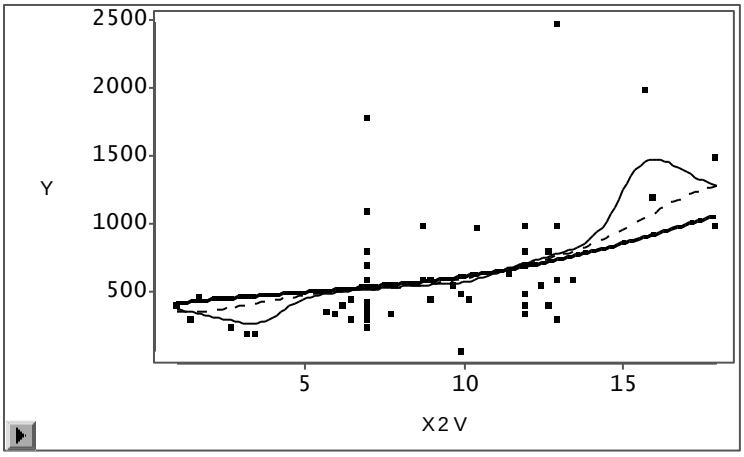


Gráfico 3.6 Ajuste mediante núcleos óptimo y con dos anchos de ventana diferentes



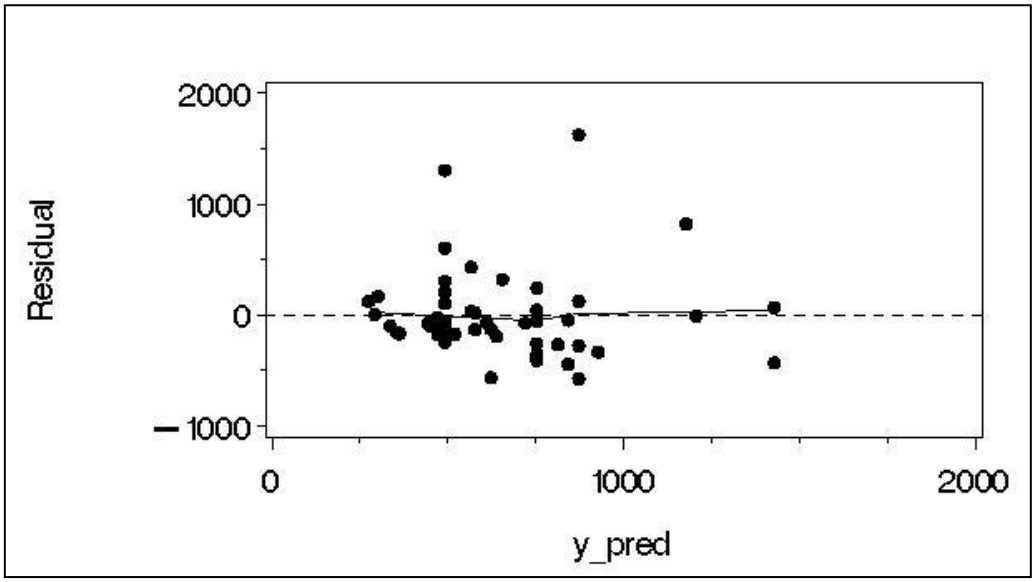
Kernel Fit

Curve	Weight	Method	C Value	Bandwidth	DF	R-Square	MSE	MSE(GCV)
- - - - - -	Normal	GCV	0.9796	2.1670	3.444	0.2865	150228.921	159542.673
———	Normal	C	0.5000	1.1060	6.296	0.3452	145332.401	162694.799
———	Normal	C	1.5000	3.3181	2.207	0.2370	157166.994	163274.882

Posteriormente se estima la curva mediante el método de regresión local. A diferencia de la estimación mediante núcleos, no calcula promedios localmente ponderados sino que estima funciones de regresión ponderando las observaciones en las proximidades de distintos focos. Este procedimiento se usa tanto para explorar la forma de la relación como para confirmarla.

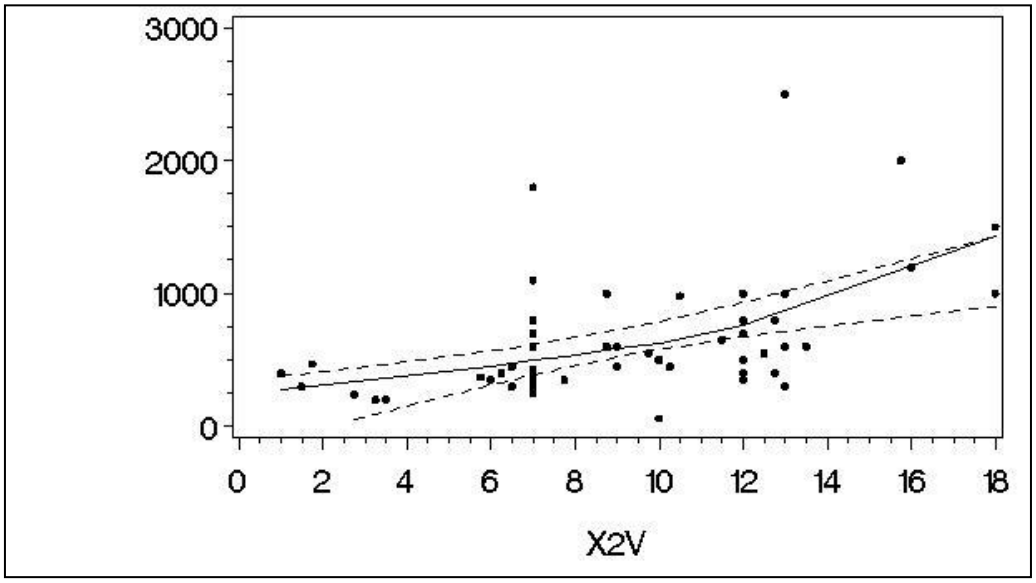
La constante de suavizado óptima es s=0.8. El gráfico de residuos revela ciertas asimetrías.

Gráfico 3.7 Residuos versus valores ajustados



Para juzgar el comportamiento de la regresión paramétrica, se calcularon los intervalos de confianza para la misma, y se los representó gráficamente, junto con la curva obtenida mediante el método de regresión local. Si esta curva cae dentro de la banda de confianza, debe interpretarse que la regresión paramétrica es aceptable.

Gráfico 3.8 Curva loess e intervalos de confianza paramétricos



Se observa que la curva “loess” cae dentro de los límites de confianza paramétricos, sugiriendo que la forma de la curva de regresión es aproximadamente lineal. Sin embargo los

residuos no permiten aceptar plenamente ese veredicto. Por lo tanto se recurre al procedimiento de incorporar otra variable explicativa al estudio del ingreso: los años de escolaridad de la mujer del jefe del hogar.

2. Regresión múltiple

Con las variables años de escolaridad del hombre y la mujer se realiza la estimación mínimo cuadrática de la función de regresión, la cual viene dada por

$$\hat{y}_i = -10.583 + 30.060\, x_{2Vi} + 39.358\, x_{2Mi}\,,$$
$$\qquad\qquad (18.249) \qquad (16.709)$$

siendo los valores entre paréntesis los errores estándares.

La falta de distribución normal de la variable respuesta indica que una superficie de regresión no paramétrica, tal como la que se obtiene mediante regresiones locales podría ser apropiada para este conjunto de datos. Además, el método"loess" se utilizará para avalar el ajuste paramétrico.

El paso preliminar consiste en ajustar la superficie para un rango de parámetros de suavizados tentativos: 0.3, 0.5, 0.7 y 1. Luego se selecciona el parámetro de acuerdo a algunos de los criterios previamente presentados.

El parámetro óptimo elegido automáticamente utilizando el criterio de validación cruzada resultó 0.449.

Con el objeto de evaluar el valor obtenido usando los criterios presentados en el punto 2.1.2.1 se realiza el examen de los gráficos de dispersión de los residuos. Para ello se grafican los residuos obtenidos para cada valor de suavizado versus la variable respuesta ajustada (Grafico 3.9) y versus las distintas variables explicativas (Graficos 3.10 y 3.11). La determinación del mejor parámetro de suavizado resulta del modelo con el parámetro que provea información más clara del ajuste en los residuos.

Gráfico 3.9 Residuos versus valores ajustados

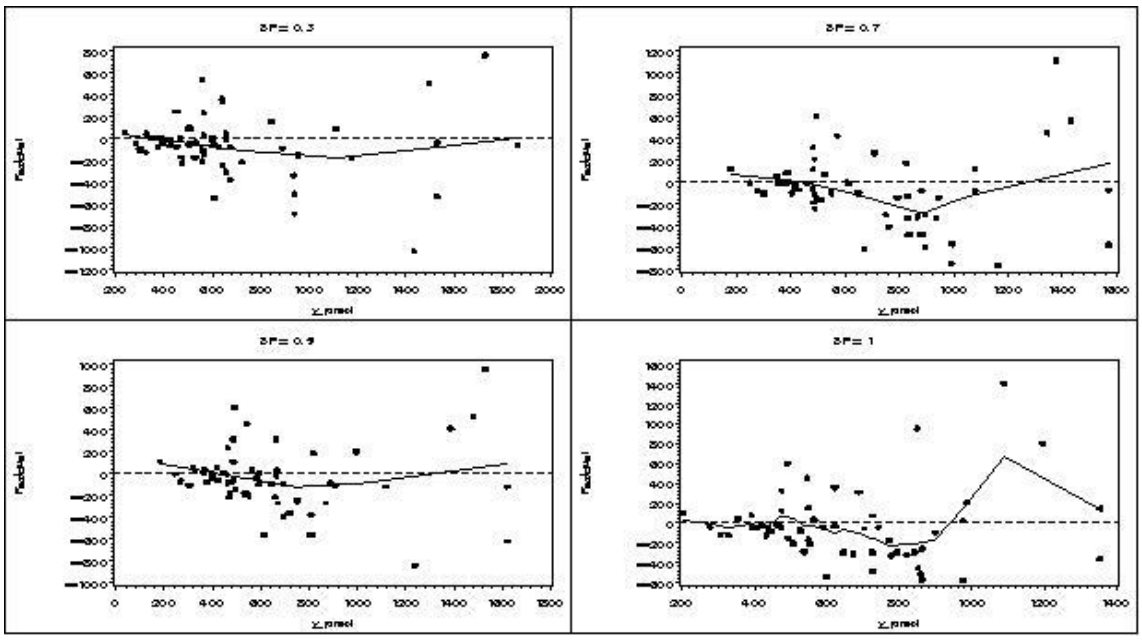


Gráfico 3.10 Residuos versus años escolaridad del varón

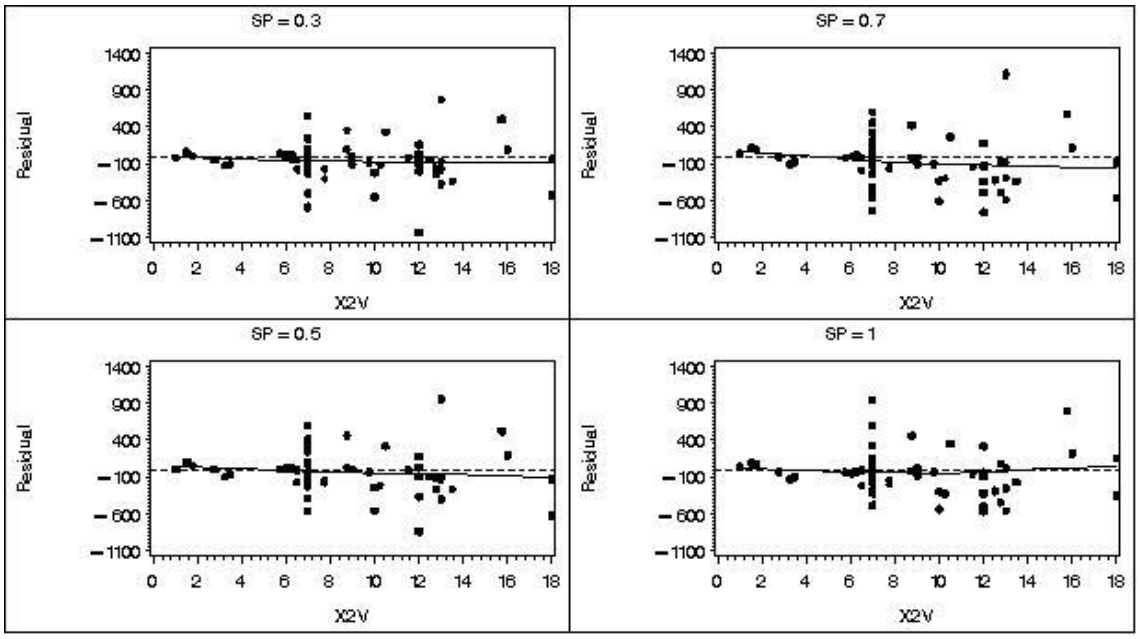
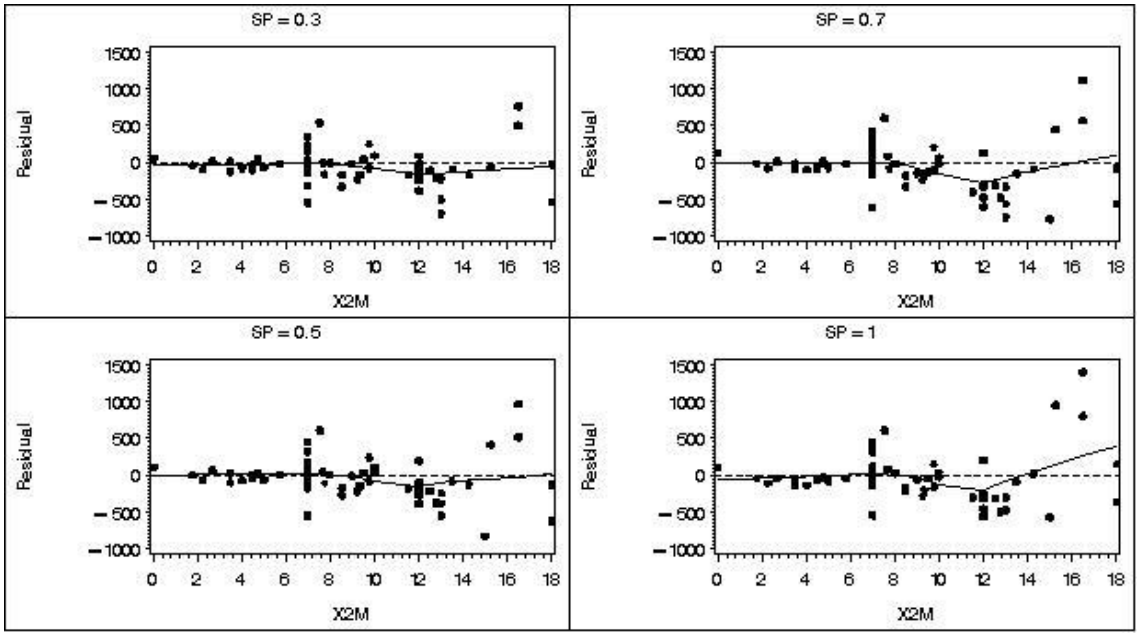


Gráfico 3.11 Residuos versus años escolaridad de la mujer



En los gráficos se visualiza que la constante de suavizado que proporciona residuos menos relacionados con las variables explicativas es s=0.5.

Como no se puede graficar la forma de la superficie de respuesta cuando se tienen dos predictores, no es posible visualizar su apariencia. A diferencia de la regresión paramétrica no se suministra ninguna expresión analítica de la superficie.

Por estos motivos, el ajuste de una superficie de regresión no paramétrica se puede usar para examinar si el modelo de regresión paramétrico ajustado es adecuado. Si la superficie no paramétrica ajustada cae dentro de los límites de confianza para la regresión paramétrica se concluye que el ajuste no paramétrico confirma que el paramétrico es correcto.

Para esta aplicación el 19% de los valores ajustados proporcionados por el método **"**loess"' no están contenidos dentro de los límites de confianza (Anexo 1). En consecuencia, se concluye que el ajuste paramétrico no es adecuado.

4. Discusión

En este trabajo se presenta el uso de la regresión no paramétrica en el contexto del análisis de regresión. Los métodos de estimación mediante núcleos y de regresión local se aplican en el ajuste de modelos con uno y dos regresores.

Las técnicas no paramétricas se utilizan para explorar y confirmar la forma de la función de regresión y para obtener estimadores de la respuesta media sin especificar la forma de la relación entre las variables.

En el caso de una variable explicativa (años de escolaridad del jefe del hogar) se visualizan las curvas no paramétricas y aunque la magnitud de los residuos es bastante grande, la regresión paramétrica lineal parece explicar, en parte, el ingreso. En cambio, para la regresión múltiple el ajuste paramétrico no resulta adecuado.

Para mejorar los ajustes se podría intentar usar polinomios de mayor grado en las regresiones locales, o incorporar otras variables explicativas a la regresión, o utilizar métodos robustos para estimar las regresiones locales.

5. REFERENCIAS

Altman, N.S. (1990). Kernel Smoothing of Data with Correlated Errors. *Journal of the American Statistical Association,* 85, 749-759.

Altman, N. S. (992) An introduction to Kernel and Nearest-neighbor Nonparametric Regression. The American Statistician 46, 175-185.

Browman, A.W., Azzalini, A. (1997). Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-plus illustrations. Oxford University Press.

Cleveland, W.S.(1993). *Visualizing data*. Hobart Press.

Fox, J. (2000a). *Non parametric Simple Regression: Smoothing Scatterplots*. Thousand Oaks C.A.: Sage.

Fox, J. (2000b). *Multiple and Generalized Non parametric Regression*. Thousand Oaks C.A.: Sage.

Montgomery, D., Peck, E. y Vining, G. (2004) *Introducción al análisis de regresión lineal*. CECSA.

Simonoff, J.S. (1996). Smoothing Methods in Statistics. Springer-Verlag.

ANEXO

Observación	Valor estimado loess	Intervalo confianza paramétrico para media		Observación	Valor estimado loess	Intervalo confianza paramétrico para media	
1	1478.41193	897.9519	1327	**31**	378.03353	194.7155	519.8267
2	1531.82663	826.6828	1233	**32**	695.26721	708.5575	996.4385
3	1118.79153	782.0660	1100	**33**	420.16416	256.9758	536.2831
4	1620.65555	976.6720	1501	**34**	459.11947	344.1636	576.4687
5	443.98128	253.8158	620.3249	**35**	542.07772	410.6216	645.2808
6	551.22203	488.2986	855.9772	**36**	400.01463	30.8201	559.1519
7	808.90530	499.8314	923.1610	**37**	566.77235	542.0554	749.9971
8	612.58824	424.9359	706.1166	**38**	817.31581	430.5789	820.7137
9	748.06455	656.6053	946.7473	**39**	469.55453	441.0737	686.7311
10	1388.69264	521.2042	1079	**40**	998.91515	718.4849	1167
11	1240.20531	767.2931	1114	**41**	487.66876	364.0666	586.6257
12	644.65922	702.8069	972.1291	**42**	597.03814	470.0710	679.5781
13	474.21101	395.5355	643.1718	**43**	659.74322	658.4389	920.5854
14	594.05123	695.0938	949.7822	**44**	487.66876	364.0666	586.6257
15	487.66876	364.0666	586.6257	**45**	318.13715	106.9617	502.6276
16	665.51296	552.3977	826.2682	**46**	265.34248	-9.0611	375.4290
17	667.49595	582.1568	865.9275	**47**	417.57163	92.7332	594.8092
18	891.52582	737.9835	1025	**48**	536.26403	481.7700	823.1474
19	491.39002	457.3733	729.4691	**49**	586.99729	422.9865	693.0359
20	808.90530	499.8314	923.1610	**50**	244.63695	-64.7337	346.6532
21	496.74720	480.3283	699.1237	**51**	530.49268	388.4673	607.3151
22	717.18607	718.6412	985.2723	**52**	462.92600	452.3251	714.8381
23	1620.65555	976.6720	1501	**53**	324.78497	141.5344	473.5294
24	669.96321	705.9049	984.0611	**54**	400.49523	226.3394	527.5612
25	492.75439	385.4643	604.5864	**55**	304.16156	44.2590	405.4747
26	695.26721	708.5575	996.4385	**56**	364.93339	207.1533	491.2764
27	462.92600	452.3251	714.8381	**57**	402.63590	276.8201	530.3863
28	663.77013	427.7990	733.3134	**58**	368.86876	222.7631	505.7266
29	908.50330	478.0975	877.3227	**59**	183.47646	-208.0955	277.1107
30	868.10336	523.7234	935.8243				