UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Contraste entre los modelos de regresión paramétricos, no paramétricos y semiparamétricos utilizando la Encuesta Nacional de Ingresos y Gastos de los Hogares

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Matemático

P R E S E N T A :

Miguel Angel Monroy Cruz

DIRECTOR DE TESIS:
Doctor Vicente Antonio García Moreno
CIUDAD DE MÉXICO, 2020

1. Datos del alumno
Monroy
Cruz
Miguel Angel
5540775927
Universidad Nacional Autónoma de México
Facultad de Ciencias
Matemáticas
306293119

2. Datos del tutor
Dr.
Vicente Antonio
García
Moreno

3. Datos del sinodal 1
Dra.
Lizbeth
Naranjo
Albarrán

4. Datos del sinodal 2
M. en C.
Antonio
Soriano
Flores

5. Datos del sinodal 3
M. en C.
José Salvador
Zamora
Muñoz

6. Datos del sinodal 4
Act.
Francisco
Sánchez
Villareal

7. Datos del trabajo escrito:
Contraste entre los modelos de regresión paramétricos, no paramétricos y semiparamétricos utilizando la Encuesta Nacional de Ingresos y Gastos de los Hogares
77 p
2020

1. Introducción

2. Metodología

2.1. Regresión paramétrica

2.1.1. Estimación por mínimos cuadrados

2.1.2. Varianza de los estimadores

2.1.3. Ejemplo

2.2. Regresión no paramétrica

2.2.1. Promedio local

2.2.2. Suavizamiento tipo núcleo o tipo kernel

2.2.3. Polinomios locales

2.2.4. Estimador de Nadayara-Watson

2.2.5. Estadística inferencial de la regresión por polinomios locales

2.2.6. Ejemplo

2.3. Modelos aditivos

2.3.1. Ejemplo

2.4. Regresión semiparamétrica

2.4.1. Selección de h por medio de validación cruzada

2.4.2. Selección del tipo de núcleo

2.4.3. Método de selección Backfitting

2.4.4. Regresión parcial lineal

2.4.5. Ejemplo

1. Introducción

El principal objetivo de este trabajo es describir las diferencias entre los modelos de regresión paramétrico, no paramétrico, semiparamétrico, así como una breve descripción de los modelos aditivos, mostrando las bondades y desventajas de cada uno. Para ejemplicar cada uno de los modelos y lograr una explicación más dinámica, se realizará un ejercicio para cada una de las regresiones con datos reales. Con datos de la *Encuesta Nacional de Ingreso y Gasto de los Hogares (ENIGH)*. Además, el enfoque que se le dará a los ejercicios se encamina a analizar el comportamiento del ingreso, gasto y ahorro de los hogares en México.

El primer modelo que se abordará es el de regresión paramétrica, es el más utilizado y analizado por las ciencias sociales y económicas. Por ejemplo un fenómeno que ha sido de interés es la relación entre el salario con el nivel de estudios *(Denison, 1964)*. En esta sección, además de definir la regresión paramétrica se buscará modelar el ingreso por trabajo en función de los años de escolaridad. Para esto, se seguirá el modelo propuesto por *Mincer(1974)* el cual permite medir el beneficio en el ingreso en función de la acumulación del capital humano.

El siguiente modelo a describir es el de regresión no paramétrica, para este modelo no se determina una función previa para la estimación de la regresión. Por ejemplo, en el modelo paramétrico, si los datos presentaron un comportamiento de U invertida, como es el caso del primer ejercicio, entonces se supone una parábola para estimar la curva, en el modelo no paramétrico este supuesto no es necesario, debido a la libertad de respetar el comportamiento natural de los datos. Sin embargo, hay algunos inconvenientes, el primero es que la función debe ser derivable, es decir, las variables deben ser continuas. Otro problema de esta regresión, es que cuando modelan regresiones con múltiples variables, se puede tener más dimensiones que puntos, a este problema se le llama "*maldición*

*de la dimensionalidad*Ruppert (2004), lo que conlleva a errores en la estimación. Para ejemplificar este modelo, se propone modelar el gasto corriente, en función de la edad del padre de familia y la escolaridad. La regresión no paramétrica se puede estimar con diversos procedimientos, para el caso práctico de esta tesis, únicamente se abordarán los métodos de media móvil, kernel y polinomios locales.

Como se mencionó, se hablará brevemente del modelo no paramétrico aditivo, este tipo de regresión evita la "*maldición de la dimensionalidad*za que agrega la condición de aditividad, es decir, se tienen tantas funciones como variables predictivas se tengan. Si bien, se ataca el problema de las dimensiones, se debe seguir trabajando con variables continuas, por lo que es poco útil para modelar fenómenos sociales, debido a que usualmente éstos cuentan con variables categóricas, como el sexo de los individuos. Por tal motivo, es necesario proponer un modelo más flexible, es decir, que permita la incorporación de variables discretas, como el semiparamétrico.

La última regresión a exponer es el modelo semiparamétrico, este se compone por dos partes, una paramétrica y otra no paramétrica, en donde la parte no paramétrica, pueden incluir variables a las que se les quiere otorgar la libertad de comportarse como los datos mismos, mientras que el la parte paramétrica se puede integrar por variables discretas o que tengan un comportamiento claramente definido, por tal motivo, será la regresión utilizada para estimar el ejercicio más amplio de esta tesis.

Con el último modelo descrito se analizará el ahorro de los hogares, con el fin de mostrar cómo se distribuye siguiendo la teoría del ciclo de vida propuesta por *Modigliani y Ando (1961)*, el cual sugiere que los individuos ahorran en el punto medio de su vida laboral para poder suavizar su consumo en la vejez. Debido a que en México no se cuenta con información pública que siga a los

individuos en el tiempo, se decidió construir una base tipo pseudo-panel [1] con ayuda de la *ENIGH*. De acuerdo con *Deaton (1985)*, ésta es una técnica que puede ser implementada si las encuestas siguen un diseño similar en el tiempo.

Además de estimar la curva de regresión del ahorro, se buscará modelar el gasto y el consumo, como se mencionó se utilizará un modelo semiparamétrico, particularmente el propuesto por *Speckman (1988)*, siguiendo los modelos de la teoría del ciclo de vida de *Modigliani y Ando (1961)*. En México hay dos trabajos similares al presentado en esta tesis *Mejia (2013) y Owen (2014)*, los cuales encuentran que los mexicanos no ahorran para suavizar su consumo, es decir, en el momento más productivo de los hogares, se tiene mayor gasto.

[1] Un pseudo-panel es una base de datos que observa grupos en lugar de individuos. Ésta se construye con datos tipo panel que se publican en determinados periodos de tiempo como la $ENIGH$ y debido a que en la encuesta no necesariamente se encuesta a los mismos individuos es necesario agruparlos por características similares, en este caso la edad.

2. Metodología

En este capítulo se describirá con mayor detalle los modelos de regresión que se busca analizar *paramétrico, no paramétrico y semiparamétrico*, con el fin de contrastar sus ventajas y desventajas.

2.1 Regresión paramétrica

Se considera un modelo de regresión paramétrica simple en donde Y la variable dependiente está relacionada con variables independientes $X_1, X_2, ..., X_k$, esta relación se presenta en la ecuación (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_k X_k + \epsilon = X\beta + \epsilon \tag{1}$$

En donde β_0,...,β_k son los parámetros asociados a cada $_k$, estos parámetros son desconocidos.

Esto es, ninguna variable ingresada al modelo debe de estar correlacionada con los errores de la regresión. De acuerdo con lo descrito en la ecuación (1), para calcular los errores se realiza la siguiente operación:

$$\epsilon = Y - E(Y|\mathbf{X} = x) \tag{2}$$

Lo cual se puede reescribir como:

$$E(Y|\mathbf{X} = x) = Y + \epsilon \tag{3}$$

En donde se supone:

- $Var(\epsilon_i) = \sigma_\epsilon^2 < \infty$ *para toda* $i = 1, 2, ...$

- $E(\epsilon_i) = 0$

- $Cov(\epsilon_i, \epsilon_j) = 0$ *para toda* $i \neq j$

Existen dos métodos para estimar los parámetros de una regresión: 1) máxima verosimilitud; y 2) mínimos cuadrados ordinarios. En este trabajo se describirá brevemente la segunda metodología.

2.1.1 Estimación por mínimos cuadrados

Uno de los métodos empleados para la estimación de los parámetros, es decir, los coeficientes de la regresión es el método llamado *Mínimos Cuadrados*. La idea es minimizar la suma de los cuadrados de los errores ϵ_i, y proponer estimadores para β, los cuales minimicen los errores. Para ejemplificar, únicamente se considera un modelo con 2 parámetros, sin embargo, esta metodología se puede generalizar.

Se tiene que minimizar la siguiente expresión:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i + ... + \beta_k x_i))^2 = \sum_{i=1}^{n} (y_i - E(Y|\mathbf{X} = x))^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{4}$$

De manera matricial se tiene lo siguiente:

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} y_1 & - & \hat{y_1} \\ y_2 & - & \hat{y_2} \\ & \vdots & \\ y_n & - & \hat{y_n} \end{bmatrix} = y - \hat{y}$$

y cuando se considera la definición de $\hat{y}$, se puede reescribir la ecuación como sigue:

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} y_1 & - & \beta_0 + \beta_1 x_{1,1} + ... + \beta_k x_{k,1} \\ y_2 & - & \beta_0 + \beta_1 x_{1,2} + ... + \beta_k x_{k,2} \\ & \vdots & \\ y_n & - & \beta_0 + \beta_1 x_{1,n} + ... + \beta_k x_{k,n} \end{bmatrix} = y - \hat{y}$$

por lo tanto:

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{1,1} & \cdot & x_{k,1} \\ 1 & x_{1,2} & \cdot & x_{k,2} \\ & & \vdots & \\ 1 & x_{1,n} & \cdot & x_{k,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = y - X\beta$$

Por lo tanto, la varianza residual se puede expresar de la siguiente forma:

$$n\sigma^2 = u^T u = (y - X\beta)^T (y - X\beta) \tag{5}$$

Por lo tanto, se tiene:

$$\Phi(b) = \sum_{i=1}^{n} (y_i - y\hat{} - i) = u^T u$$

Para obtener la estimación de los parámetros se tiene que cumplir la condición para que tenga un mínimo de la siguiente manera:

$$\frac{\partial \Phi(b)}{\partial b} = 0$$

Aplicando la derivada parcial a la función antes definida, se obtiene lo siguiente:

$$\frac{\partial \Phi(b)}{\partial b} = \frac{\partial (y - Xb)^T (y - Xb)}{\partial b} = -2X^T Y + 2X^T X\beta$$

Al igualar esta ecuación a cero y despejando, se obtiene:

$$X^T Y = X^T X\beta$$

Para despejar β, se multiplica ambos lados de la ecuación por $(X^T X)^{-1}$

$$(X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T X\beta$$

Entonces:

$$\beta = (X^T X)^{-1} X^T Y$$

Con esto se obtiene el estimador del parámetro β

La colinealidad de las variables se puede detectar analizando la matriz de correlación, una alta correlación entre las variables puede ocasionar problemas con el cálculo de la inversa de (XX^T).

Se ha probado que el estimador por mínimos cuadrados está dado por $\hat{\beta} = (X^T X^{-1}) X^T Y$, y como $\hat{Y} = X\hat{\beta}$ entonces: $\hat{Y} = X(X^T)^{-1} X^T Y$, si se define:

$$H = (X^T X)^{-1} X^T$$

Entonces se tiene que $\hat{Y} = HY$. La matriz $\hat{H}$, tiene la propiedad de mapear al vector de valores observados en el vector de valores ajustados, esta matriz

tiene las siguientes particularidades, entre las más importantes, se encuentran lo siguiente:

H es simétrica

H es idempotente ($HH = H^2 = H$)

El vector de residuales e pueden expresarse como $e = (I - H)Y$

2.2.2 Estimación de σ^2

Idealmente el estimador de σ^2 no debería depender del ajuste del modelo. Esto es posible solamente cuando hay varias observaciones de y para al menos un valor de x o cuando se tiene una opinión *a priori* sobre σ^2. Cuando no se tiene ninguna de las situaciones anteriores, el estimador de σ^2 se obtiene de la suma de cuadrados del error.

$$\mathbf{RSS} = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

para la varianza, se utiliza el siguiente estimador:

$$\hat{\sigma}^2 = \frac{\mathbf{RSS}}{n - k - 1} \tag{6}$$

Con estos estimadores, se pueden calcular los intervalos de confianza y pruebas de hipótesis, para mayor detalle véase *Wackerly, Mendenhall y Scheaffer (2008).*

2.1.3 Ejemplo

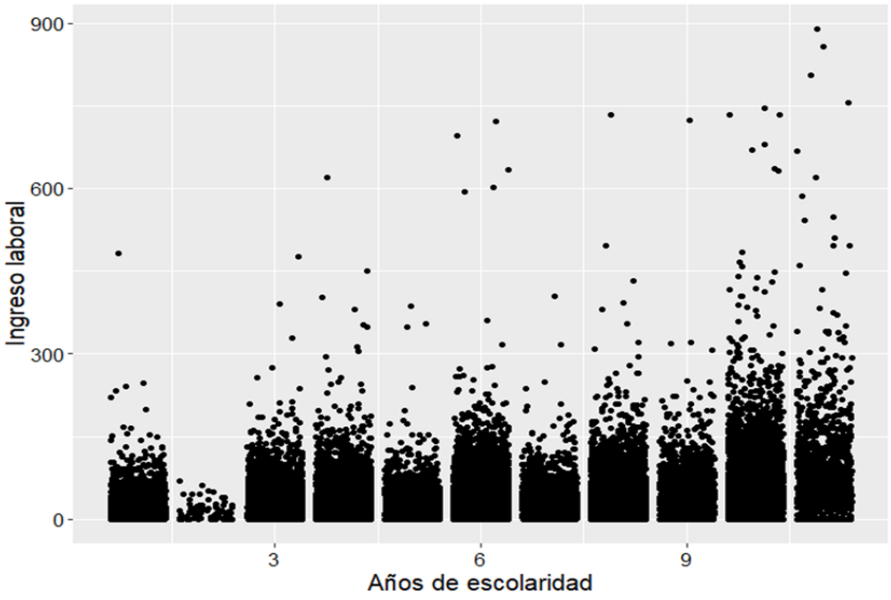
Para ejemplificar, se estimará un modelo de regresión lineal simple siguiendo la metodología de *Mincer (1974)*, en el cual se busca relacionar el ingreso por trabajo en función de los años de escolaridad. Para este ejercicio se utilizará la *ENIGH (2018)*, considerando únicamente a los hogares que reportaron un

ingreso por trabajo mayor a 0.

En la siguiente gráfica se muestra la distribución del ingreso por trabajo (IT) de acuerdo al nivel de escolaridad del padre de familia. Se puede observar que en los dos últimos grados de escolaridad (10 y 11) la distribución del ingreso es más alargada, es decir, en promedio el ingreso es superior en estos grupos que en los hogares en donde el jefe o jefa de familia cuentan con 2 años de escolaridad, por ejemplo.

Gráfica 1. Distribución del ingreso laboral por escolaridad de la madre o padre de familia

miles de pesos



Fuente: ENIGH

La regresión propuesta para describir un posible impacto de la escolaridad en el ingreso por trabajo es la siguiente:

$$E(IT|\textbf{Escolaridad}) = \beta_0 + \beta_1 Escolaridad \tag{7}$$

En la tabla 1 se pueden observar los coeficientes en donde se puede apreciar

Tabla 1. Estimación de los parámetros de la regresión paramétrica

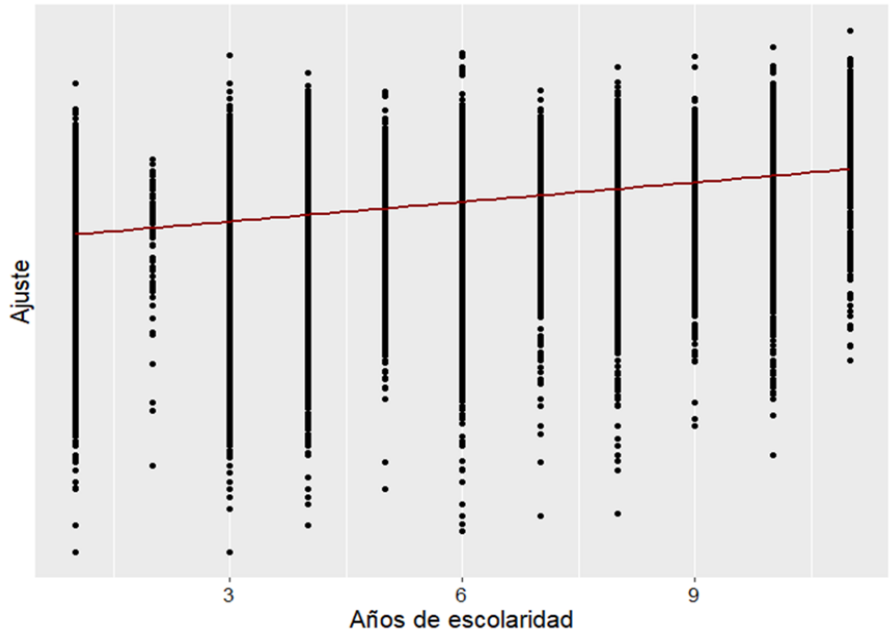
Concepto	Beta	Error estándar	t-value	Pr(>|t|)
Intercepto	2,763.7	365.1	7.6	3.84e-14
Escolaridad	4,697.0	60.1	78.0	<2e-16

Fuente: Cálculos propios con datos de la ENIGH
Significancia: 0 TT

que la variable utilizada es representativa, es decir, la escolaridad es un factor importante en el ingreso por trabajo. Según el modelo, en promedio, un hogar en donde el jefe de familia cuenta con un cero años de escolaridad, ingresa al hogar por trabajo un monto de $2,764$ pesos y por un año más de escolaridad de este el hogar ingresa $4,697$ pesos.

En los datos resultantes de la regresión se puede apreciar que la variable utilizada es representativa, es decir, la variable tiene un *p-value* menor a 0,05, por lo tanto se puede decir que la escolaridad es un factor importante en el ingreso laboral de las personas.

Gráfica 2. Ajuste de la regresión simple



Fuente: ENIGH

Como se pudo observar en la ecuación propuesta, se supuso que una línea recta ajustaba de una buena forma al modelo, esto es una desventaja ya que es poco flexible, es decir, esto lo hace relativamente rígido, debido a que pocos fenómenos tienen comportamientos claramente definidos por una función conocida, como lo es la recta o una parábola. Sin embargo, su interpretación es sencilla, ya que con un sólo parámetro se capta el efecto que se busca explicar. Por tal motivo, se buscan modelos que puedan lograr un mejor ajuste de los datos y que sean sencillos de interpretar.

2.2 Regresión no paramétrica

Como se mencionó, en algunas ocasiones los modelos de regresión paramétrica se ven limitados para modelar fenómenos que presenten irregularidades en su comportamiento. Sin embargo, una desventaja respecto a los modelos paramétricos es la interpretabilidad ya que no se cuenta con un parámetro que capture el efecto de cada variable en el modelo. En esta sección se hablará de los modelos no paramétricos, los cuales no suponen una distribución en el comportamiento de los datos. Para estimar estos modelo existen diversas metodologías, algunas de éstas se mencionarán en este trabajo. La estimación de regresiones no paramétrica es relativamente nueva, ya que sus primeros estudios comenzaron en 1970, en donde autores como: *de Boor, C. (1978), Silverman, B. (1984), Kimeldorf, G. y Wahba, G. (1970)* los introdujeron.

2.2.1 Promedio local

El modelo más simple de regresión no paramétrica es el de medias móviles, el cual calcula promedios con una banda definida en el periodo t hasta llegar a $t+n$, la unión de los promedios será la estimación de la regresión. Este modelo se define por la siguiente ecuación:

$$\hat{f}(x_i) = \frac{1}{k} \sum_{j=\underline{i}}^{\overline{i}} y_i \tag{8}$$

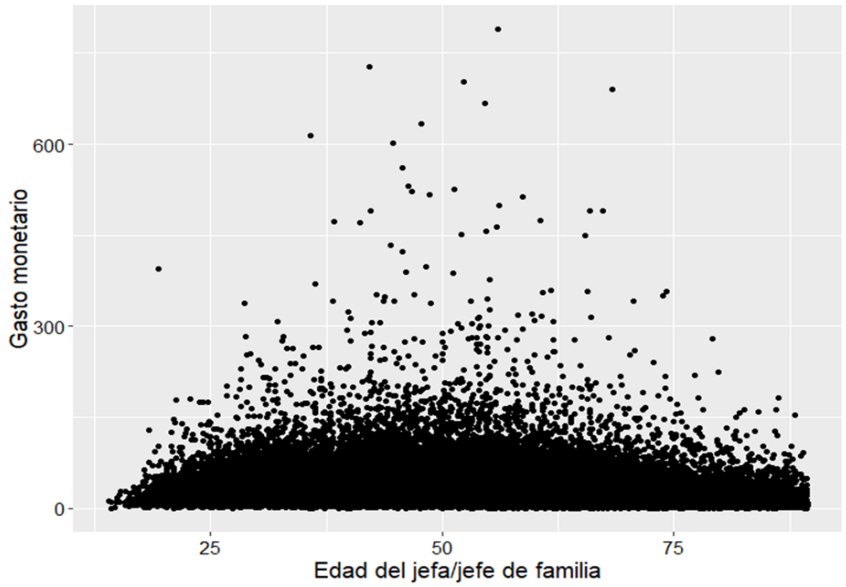
En dónde $\underline{i} = i - (\frac{k-1}{2})$, $\overline{i} = i + (\frac{k-1}{2})$ y k representa el número de observaciones o de intervalos. Si se amplía o se reduce k, también incrementamos o reducimos el número de datos incluidos en el intervalo.

El método de promedio local se puede apreciar de manera más intuitiva por medio de un ejemplo. Para este ejercicio se analizará el gasto monetario mensual en función de la edad del padre de familia.

La distribución que se busca modelar es la siguiente:

Gráfica 3. Distribución del gasto monetario y la edad de la madre o padre de familia

milles de pesos



Fuente: ENIGH

Los promedios locales se utilizarán de la siguiente manera. Se divide a la población en cohortes de 2, 5 y 10 años de acuerdo a la edad del padre de familia y se promedia el gasto por cada grupo formado, con la unión de los puntos resultantes se tendrá una gráfica que dará la relación entre la edad y el gasto corriente.

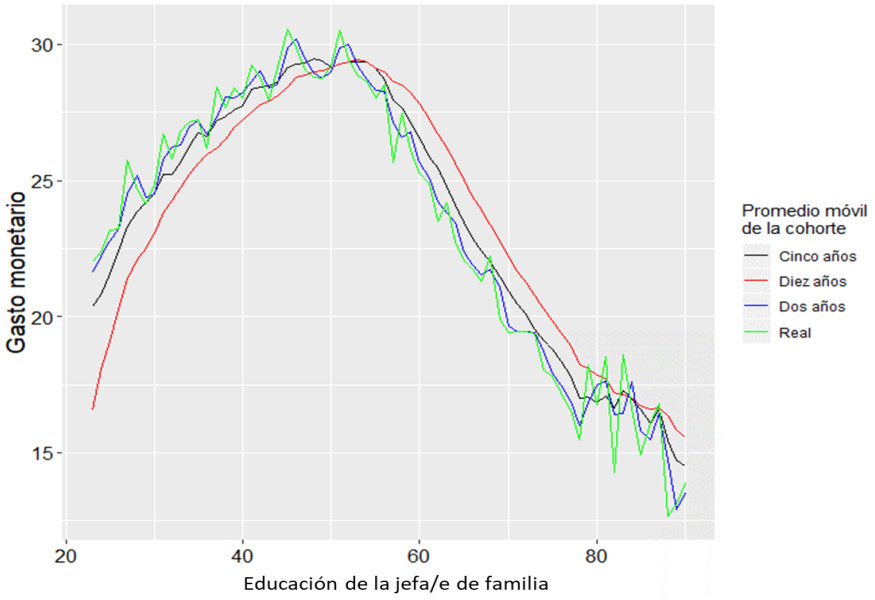
El estimador está dado por:

$$\hat{f}(edad_i) = \frac{1}{k} \sum_{j=\underline{i}}^{\overline{i}} gastoi \tag{9}$$

Considerar el promedio local de cada valor de x, es el modelo más simple de estimación no paramétrica. Este modelo es llamado *suavizamiento de medias*

móviles. La simplicidad de este modelo es su mayor ventaja, sin embargo, tiene grandes deficiencias incluso cuando se tiene un gran número de observaciones para cada valor de x, pues podría tener gran varianza y ser impreciso.

Gráfica 4. Promedio móvil del gasto monetario por edad
milles de pesos, mensual



Fuente: ENIGH

Cuando se reduce el ancho del intervalo, por ejemplo a 2 años, se tiene un menor número de puntos, lo que ajusta a los datos de una manera más precisa, aunque con mayor variabilidad. Si el intervalo es más grande, por ejemplo de 10 años, el número de datos considerados es mayor y se tiene una curva más suave, es decir, con menos variabilidad pero con mayor sesgo, como se muestra en la gráfica 2. Por tal motivo es indispensable contar con una técnica para seleccionar el ancho de banda óptimo y así armonizar el sesgo y la varianza.

De la gráfica 3 se puede observar que hasta los 30 años de edad del jefe o jefa de familia, la pendiente de la curva en el gasto monetario del hogar tiene

una pendiente positiva. También se puede apreciar que cuando el jefe o jefa de familia ronda los 50 años, el gasto del hogar alcanza su máximo, mientras que a partir de los 60 años, este comienza a disminuir hasta llegar a gastos menores que en inicio del ciclo de vida laboral.

2.2.2 Suavizamiento tipo núcleo o tipo kernel

El suavizamiento tipo núcleo es un método de regresión no paramétrica, que redefine el suavizamiento por medio de pesos promedio. Este método estima la relación entre $x_1, x_2, ..., x_k$ e y de manera local, a diferencia del promedio móvil, en donde se le daba el mismo peso a x y a y, no importando la distancia a la que se encontraran del punto que se quería estimar. Para corregir este inconveniente, *Nadayara (1964) y Watson (1964)* propusieron una función que le asignara un mayor peso a las observaciones de acuerdo a su cercanía con $x_1, x_1, ..., x_1$, con esto se tiene una mejor estimación y. Para ejemplificar este método de regresión, se estimará el ingreso corriente (IC) del hogar en función de la edad del padre de familia. En esta regresión es necesario definir una función que le de mayor peso a las observaciones cercanas a x_0, ésta se define como sigue:

$$z_i = \frac{(x_i - x_0)}{h}$$

En dónde z_i es la distancia entre x_i y x_0. El parámetro h es llamado el ancho de banda. En el ejemplo anterior, se mostró que al modificar el ancho del intervalo, la estimación cambia su suavizamiento y variabilidad, esta condición se explicará más adelante.

Para este modelo de regresión, se construye una función que le asigne pesos a los datos, a la que se le llama función kernel $K(z)$, ésta define la forma en la que se ponderan las observaciones dentro de cada banda definida. Las estimaciones tipo kernel son utilizadas para estimar variables de densidad aleatorias $f(x)$, o en regresiones tipo kernel para estimar la esperanza condicional de la variable aleatoria, *Silverman (1986), Wand y Jones (1995)*. En general cualquier función

que cumpla con las siguientes condiciones se puede utilizar como un kernel.

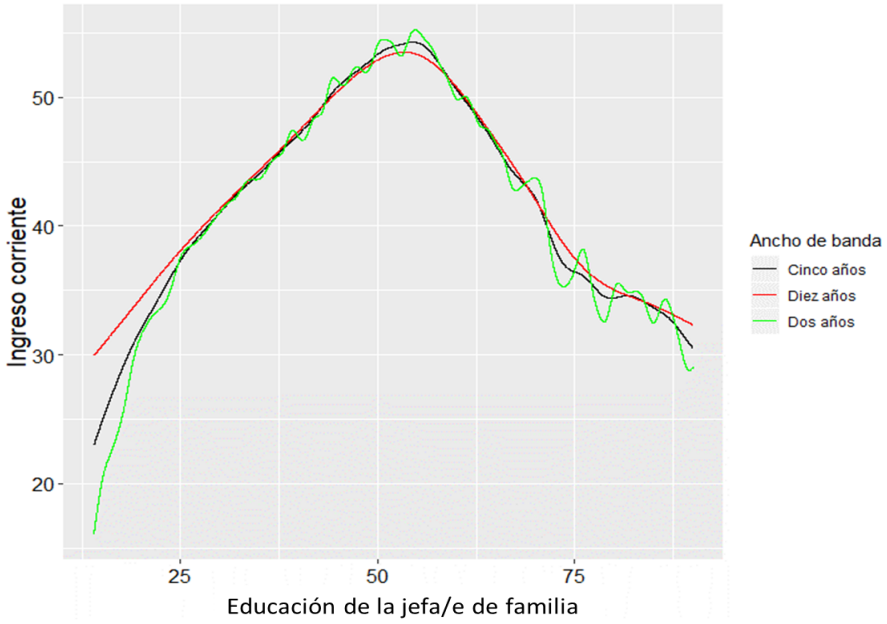
- $k(u) \geq 0$ y $\int K(z) = 1$
- ser simétrico alrededor del origen $\int K(z) = K(-z)$
- $lim_{x-\to\infty} k(u) = lim_{x+\to\infty} k(u) = 0$

Para dicha función es habitual el uso de algunas de las siguientes densidades:

- **Triangular**: $k(u) = (1 - |u|)1_{|u|\leq 1}$
- **Epanechnikov**: $k(u) = \frac{3}{4}(1 - u^2)1_{|u|\leq 1}$
- **Biponderado**: $k(u) = \frac{15}{16}(1 - u^2)^2 1_{|u|\leq 1}$
- **Gaussiano**: $k(u) = (2\pi)^{-1} exp(\frac{-x^2}{2})1_{|u|\leq 1}$

Para ejemplificar cómo cambia la estimación con los diferentes núcleos, se realizará la misma regresión utilizando un núcleo Epanechnivov. Se considerarán la edad del jefe o jefa de familia y el ingreso corriente del hogar, con un ancho de banda de 5 años, para ambos casos.

Gráfica 5. Distribución de ingreso corriente por edad
milles de pesos, mensual



Fuente: ENIGH

En la gráfica de arriba se aprecia que ambas regresiones siguen la misma tendencia, sin embargo, al utilizar el núcleo normal, se obtiene una curva de mayor suavidad.

Al aplicar función kernel a cada z_i, es decir, a la función de distancia antes definida, se obtiene un mayor peso para los datos cercanos a x_0. Esta función se define como sigue:

$$w_i = K(z_i) = K\left[\frac{(x_i - x_0)}{h}\right] \tag{10}$$

El peso w_i es usado para calcular el promedio ponderado local, como sigue:

$$\hat{f}(w_0) = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{11}$$

En la siguiente ecuación se estima el ingreso corriente, de acuerdo con la

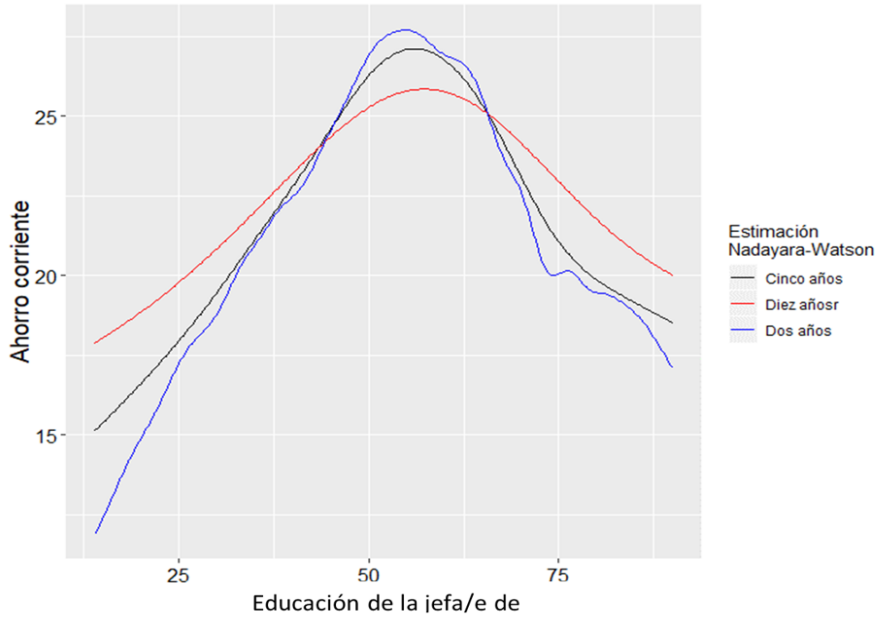
edad de los padres de familia, en donde f es una función que ajusta la curva, sin embargo, es desconocida que estima la distribución de IC de manera no paramétrica.

$$IC = f(edad) + \epsilon \tag{12}$$

De manera similar al de promedios móviles, para la estimación tipo kernel y en general para las estimaciones no paramétricas y semiparamétricas que se verán en esta tesis, es necesario seleccionar un ancho de banda. Para ejemplificar cómo la selección del ancho de banda de 2, 5 y 10 años para observar cómo estos influyen en la estimación de la curva de regresión.

Gráfica 6. Distribución del ahorro corriente por edad

milles de pesos, mensual



Fuente: ENIGH

En la gráfica 6, se observa como al reducir el ancho de banda la varianza de

la curva aumenta, sin embargo, el sesgo disminuye. Cuando el ancho de banda es 2, se observa un incremento en el ingreso corriente en el hogar, después de que el jefe o jefa de familia cumplen 60 años, esto podría ser debido a la devolución de su subcuenta de vivienda, aunque se tendrían que revisar los datos con mayor detalle. El pico que se observa después de los 60 años se suaviza al incrementar el ancho de banda.

Es importante mencionar que, estos modelos se utilizan primordialmente para hacer un análisis visual, el cual ayuda a generar un primer acercamiento a los datos, no obstante, su metodología es poco robusta. Por esto, en las próximas subsecciones se analizarán modelos con mayor complejidad estadística. Además, se explicará la metodología para seleccionar un ancho de banda óptimo utilizando validación cruzada.

2.2.3 Polinomios locales

Los polinómios locales son otra metodología con la que se puede estimar regresiones no paramétricas, a continuación se explica brevemente en qué consiste la estimación. Sea $(x_1, Y_1), ..., (x_n, Y_n)$ una muestra aleatoria bivariada de la que se tiene que estimar una función de regresión no conocida $f(x) = E(Y|X = x)$. Usando series de Taylor, se puede aproximar $f(x)$, en donde x es un punto cercano x_0, como se describe a continuación:

$$f(x) \approx f(x_0) + f^1(x_0)(x - x_0) + \frac{f^2(x_0)}{2!}(x - x_0)^2 + ... + \frac{f^p(x_0)}{f!}(x - x_0)^f$$

$$=\mathrm{f}(\mathrm{x}_0) + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + ... + \beta_p(x - x_0)^p \quad (13)$$

En donde $f^{p+1}(.)$ es una función desconocida continua. Lo antes descrito

garantiza la existencia de p derivadas por ser un polinomio de grado p. Con los datos de x e Y se puede minimizar la función descrita arriba. Este es un problema de regresión de polinomio local en el cual se utilizan los datos para estimar que el polinomio de grado p con la mejor aproximación de $f(x)$ en una pequeña vecindad al rededor del punto x_0, es decir se minimiza $\beta_0, \beta_1, ..., \beta_p$ con respecto a la función.

$$\sum_{i=1}^{n}(Y_1 - \beta_0 - \beta_1(x_i - x_0) -, ..., -\beta_p(x_i - x_0)^p)^2 K(\frac{x_i - x_0}{h}) \tag{14}$$

Este es un problema de mínimos cuadrados ordinarios en donde los pesos están dados por $K(\frac{x_i - x_0}{h})$. Por lo que es conveniente definir los siguientes vectores y matrices, en donde la estimación se centra en x_0

$$X_{x_0} = \begin{pmatrix} 1 & x_1 - x_0, & ... & , (x_1 - x_0)^p \\ 1 & x_2 - x_0, & ... & , (x_2 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_0, & ... & , (x_n - x_0)^p \end{pmatrix}$$

$Y = (Y_1, Y_2, .., Y_n)^T$

$\beta = (\beta_0, \beta_1, ..., \beta_p)^T$

$$W_{x_0} = \begin{pmatrix} K((x_1 - x_0)/h) & 0 & , ..., & 0 \\ 0 & K((x_2 - x_0)/h) & , ..., & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & , K((x_n - x_0)/h) \end{pmatrix}$$

Como el núcleo en simétrico [2], se puede escribir el argumento de K como $(x_1 - x_0)/h$. Sin embargo, la notación usada es para enfatizar el hecho de que la regresión del polinomio local es una regresión ponderada usando los datos centrados en x_{x0}. Para encontrar β es necesario minimizar la suma ponderada de la función de mínimos cuadrados.

$$(Y - X_{x0}\beta)^T W_{x0}(Y - X_{x0}\beta) \tag{15}$$

Con respecto al parámetro β. La solución es:

$$\hat{\beta} = (X_{x0}^T W_{x0})^{-1}(X_{x0}^T W_{x0} Y) = [\sum_{i=1}^{n} X_i X_i^{-1} K_h(x_i - x_0) Y_i]^{-1} \sum_{i=1}^{n} X_i Y_i^{-1} K_h(x_i - x_0) \tag{16}$$

[2]En la sección de suavizamiento tipo núcleo se pidió que fuera simétrico, en polinómios locales se seguirá la misma línea.

Tabla 2. Distribución del gasto según edad y escolaridad

GC	Edad	Escolaridad
8.39	32	6
7.87	34	2
7.47	50	5
9.10	53	12
9.10	54	15
8.18	56	17

Fuente: Cálculos propios con datos de la ENIGH

Proporcionada por $(X^T WY)$ la cual no es una matriz singular. La medida de $f(x_0)$ es estimada por el ajuste de intercepto del parámetro (es decir, por $\hat{\beta}_0$) esto define la posición estimada de la curva del polinomio local en el punto x_0. Variando el valor de x_0, se puede construir una estimación de $f(x)$ sobre el rango de los datos. Deduciendo lo siguiente:

$$\hat{f}(x_0, 0, h) = e_1^T (X_{x0}^T W_{x0})^{-1} (X_{x0}^T W_{x0} Y) = \sum_{i=1}^{n} X_i K_h(x_i - x_0) Y_i / \sum_{i=1}^{n} X_i K_h(x_i - x_0) \tag{17}$$

en donde el vector e_1 es de magnitud $p+1$ y tiene 1 en la primera posición y $0Ts$ en las otras.

El estimador local para $(p = 1)$

$$\hat{f}(x_0, 1, h) = n^{-1} \sum_{i=1}^{n} \frac{(s_2(x,h) - s_1(x,h)(x_i - x_0)) K_h(x_i - x_0) Y_i}{(s_2(x,h) s_0(x,h) - s_1(x,h))^2} \tag{18}$$

en donde

$$s_r(x,h) = n^{-1} \sum_{i=1}^{n} (x_i - x_0)^r K_h(x_i - x_0) \tag{19}$$

Ruppert y Wand (1994) estudian el sesgo y la varianza condicional del estimador mostrado.

Para ejemplificar se consideraron 6 observaciones seleccionadas al azar, extraídas de la $ENIGH(2018)$, *gasto corriente, edad y escolaridad*. En este modelo no se da el mismo peso a todas las observaciones, es decir, para estimar el *gasto corriente* de una persona con 54 años de edad y 15 años de escolaridad se tomarían personas con características similares para estimar la curva, por ejemplo, un individuo con 53 años y 12 años de escolaridad, ya que una persona con 2 años de escolaridad y 34 años de edad, estaría muy alejada de las características del primer individuo.

Como se ha mencionado, la definición del estimador esta determinado por tres parámetros: ancho de banda h, la función tipo núcleo K y el grado del polinomio p, en los siguientes párrafos se definirán dichos parámetros.

El **ancho de banda** se define como un parámetro positivo que determinará la amplitud del intervalo. Sin embargo, la selección del ancho de este es uno de los aspectos cruciales del procedimiento de estimación. Las propiedades del estimador dependerán en gran medida de la elección que se haga de dicho parámetro. El ancho de banda compensa el sesgo y la varianza, es decir, si se elige un ancho de banda pequeño, solamente las observaciones muy cercanas al punto serán tomadas en cuenta para el cálculo, describiendo con precisión el comportamiento local, pero obteniendo una curva muy variable, si se elije un ancho de banda grande, las estimaciones en cada punto se verán afectadas por observaciones muy alejadas del punto, esto dificulta la captación de comportamientos locales, esto conlleva a sesgos grandes, aunque con poca variabilidad.

La **función tipo núcleo**, define las ponderaciones que se asignan a cada observación en el entorno considerado.

Para la elección del **grado del polinomio** se tendrá que buscar una compensación entre sesgo y varianza, es decir, si se elije un ajuste que tenga un grado de polinomio *cero* o *uno*, se tiene una curva suave, pero con mucho sesgo, si se ajusta con grados mayores a *tres* esto permite mayor adaptabilidad (menor sesgo), pero con mayor varianza. *Ruppert (2004)* argumenta que al elegir un ajuste de grado *dos*, se tiene mayor bondad que al elegir un grado de polinomio *tres* o mayor. Sin embargo, para una mejor comprensión se expondrá el modelo más sencillo en donde el polinomio local es de grado 0, el cual ha sido estudiado ampliamente por *Nadayara, Watson (1964).*

2.2.4 Estimador de Nadayara-Watson

Se supone que se busca estimar el modelo (X) de la sección anterior, en donde $m(\bullet)$ se expresa en términos de una función de densidad $f(x, y)$ como sigue:

$$m(x) = E(Y|X = x) = \int yf(y|x)dy = \frac{\int yf(x, y)dy}{\int f(x, y)dy} \tag{20}$$

Se quiere estimar el numerador y el denominador de manera separada usando estimadores tipo núcleo. En primer lugar, para la estimación de la densidad conjunta se utiliza el producto del núcleo, es decir:

$$\hat{f}(x, y) = \frac{1}{nh_xh_y}\sum_{i=1}^{n} K(\frac{x - x_i}{h_x})K(\frac{y - y_i}{h_y}) = \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x-x_i)K_{h_y}(y-y_i) \tag{21}$$

En donde se tiene:

$$\int y\hat{f}(x,y)dy = \frac{1}{n}\int y\sum_{i=1}^{n} K_{h_x}(x-x_i)K_{h_y}(y-y_i) \tag{22}$$

Se tiene que $\int yf(y|x)dy = y_i$. Se puede escribir como:

$$\int y\hat{f}(x,y)dy = \frac{1}{n}\int \sum_{i=1}^{n} K_{h_x}(x-x_i)y_i \tag{23}$$

Ahora se estimará el denominador de la siguiente forma:

$$\int \hat{f}(x,y)dy = \frac{1}{n}\int \sum_{i=1}^{n} K_{h_x}(x-x_i)\int K_{h_y}(y-y_i)dy = \frac{1}{n}\int \sum_{i=1}^{n} K_{h_x}(x-x_i) = \hat{f}(x) \tag{24}$$

Entonces, el estimador Nadaraya-Watson de una función de regresión desconocida esta dado por:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_{hx}(x-x_i)y_i}{\sum_{i=1}^{n} K_{hx}(x-x_i)} = \sum_{i=1}^{n} W_{hx}(x,x_i)y_i \tag{25}$$

En donde la función de peso es $W_{hx}(x,x_i)y_i = \frac{K_{hx}(x-x_i)}{\sum_{i=1}^{n} K_{hx}(x-x_i)}$. Notemos que $\sum_{i=1}^{n} W_{hx}(x,x_i)y_i = 1$. Este estimador fue propuesto por Nadayara (1964) y Watson (1964).

2.2.4 Estadística inferencial de la regresión por polinomios locales

La estadística inferencial se centra en estimar errores estándares para los parámetros del modelo. En el análisis, éstos se utilizan en conjunto con la distribución para construir intervalos de confianza y realizar pruebas de hipótesis. Para la regresión no paramétrica también se puede estimar bandas de confianza al ajuste de la curva.

Grados de libertad

El concepto de grados de libertad es más complejo en los modelos no paramétricos, ya que en estos no se estiman parámetros. No obstante, los grados de libertar de estos modelos se pueden ver como una generalización de los modelos paramétricos. En os modelos lineales, los grados de libertad son igual al número de parámetros estimados, lo cual coincide con:

$$Rango(H)$$

$$Traza(H) = traza(HH^T) = traza(2HHH^T)$$

Análogamente, los grados de libertad de un modelo no paramétrico se obtienen de la matriz H, sustituyendo esta por la matrix de suavizamiento S que juega el mismo papel, las aproximaciones a los grados de libertad se pieden definir como sigue:

$$Traza(S) = v$$

$$Traza(SS^T) = \widehat{v}$$

$$Traza(2S - SS^T)$$

ver *Hastie y Tibshirani (1990)* para mayor detalle.

Bandas de confianza

El proceso para estimar bandas de confianza es similar a estimar intervalos de confianza para predictores de la regresión lineal *Fox (2002).*

$$\hat{y}_i = \sum_{j=1}^{n} s_{ij}(x_i) y_i \tag{26}$$

El valor ajustado de $\hat{y}_i$ resulta de la estimación del peso local por mínimos cuadrados de y_i. Se tiene que $V(y_i) = \sigma^2$, por lo que la estimación resultaría de la siguiente manera:

$$V(\hat{y}_i) = \sigma^2 \sum_{j=1}^{n} s_{ij}^2(x_i) \tag{27}$$

Se reescriben las ecuaciones de manera matricial, para facilitar la notación.

$$\hat{y} = \mathbf{S}\mathbf{y} \tag{28}$$

en donde $\mathbf{S}$ es una matriz de suavizamiento de nxn para (i, j) elementos de $s_{ij}^2(x_i)$. Ahora se describe la varianza de los valores ajustados, de la siguiente manera:

$$V(\hat{y}) = SV(y)S^T = \sigma^2 SS^T \tag{29}$$

Ahora se requiere el estimador de σ^2, para el modelo de regresión paramétrica, se estiman por mínimos cuadrados, como sigue:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} \tag{30}$$

en donde $n-2$ son los grados de libertad y $e_i = \hat{y}_i - y_i$ es el residuo de la observación i. La estimación de σ^2 de un modelo no paramétrico se realiza de una manera similar a la de un modelo paramétrico. Es decir, una analogía con el modelo paramétrico para calcular dicho estimador, como ya se mencionó.

En el contexto de la regresión no paramérica, $\mathbf{S}$ es equivalente a $\mathbf{H}$, por lo que se usa $\mathbf{S}$ para calcular los grados de libertad. Para un modelo de regresión no paramétrico, $tr(\mathbf{S})$ es el número de parámetro utilizado en el suavizamiento, por lo que los grados de libertad se calculan como se describió anteriormente, para este ejercicio se va a utilizar $tr(\mathbf{S})$ como estimador de los grados de libertad.

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - tr(\mathbf{S})} \tag{31}$$

Asumiendo normalidad en los errores, y una banda de confianza del 95 %, se tiene lo siguiente:

$$\hat{y}_i \pm 2\sqrt{\hat{V}(\hat{y}_i} \tag{32}$$

como $\hat{y} = Sy$, entonces:

$$\hat{V}(\hat{y}) = \sigma^2 SS^T \rightarrow \hat{V}(\hat{y}_i) = \hat{\sigma^2} SS^T_{ii}$$

2.2.6 Ejemplo

Para ejemplificar, se utilizará la edad y el ahorro corriente, estas variables se reportan en la $ENIGH(2018)$. Para el cálculo del ahorro corriente (AC) de los hogares se resta el gasto corriente (GC) al ingreso total (IT) y se descartan los hogares que no tengan ahorro, esta definición es la utilizada por *Székely (1998)*, y se abordará de forma más detallada en el próximo capítulo. La ecuación se define como sigue:

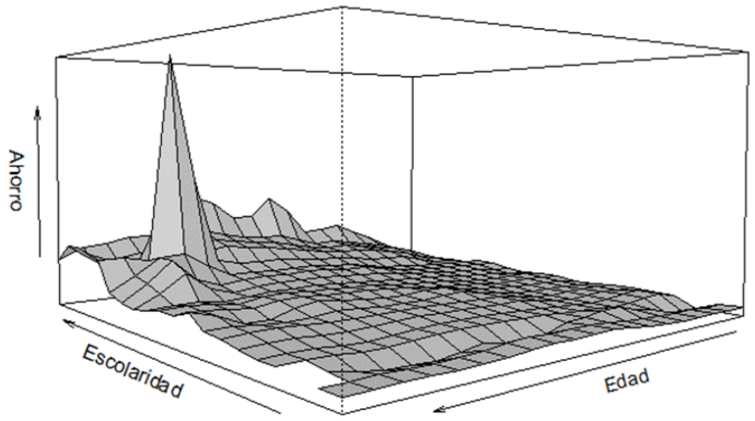
$$AC = IIC - GC \tag{33}$$

y la ecuación que se estimó se representa como sigue:

$$AC = f(edad, escolaridad) + \epsilon \tag{34}$$

Para este caso, se considera la edad y la escolaridad del padre o madre de familia dentro de la misma función desconocida. Además, se supone que ambas variables son continuas ya que como se mencionó, una de las limitantes de esta regresión es que sólo puede trabajar con este tipo de variables. Por otro lado, la curva de la regresión se estima de forma conjunta, es decir, resulta una sábana en tres dimensiones, como se muestra en la siguiente gráfica.

Gráfica 7. Distribución del ahorro corriente por edad y escolaridad
milles de pesos, mensual



Fuente: ENIGH

En la gráfica se puede observar que el ahorro máximo de los hogares se presenta cuando el padre de familia tiene alrededor de 60 años y cuando se tiene una escolaridad de 12 años o más, en el resto de la gráfica se aprecia un ahorro prácticamente nulo. Esto es, en la mayor parte de los puntos, se observa que los hogares mexicanos no ahorran, incluso el ingreso no les es suficiente para solventar sus gastos, ya que como se mencionó únicamente se consideró a los hogares con un ingreso mayor a cero.

Como se ha discutido, los modelos de regresión no paramétrica tiene dos limitaciones serias, la primera es la interpretación cuando la dimensión es mayor a dos. Y la segunda es que incluso si se pudiera visualizar el resultado, la estimación local de $k-dimensiones$ es un problema. Por ejemplo, para un valor grande de k el número de vecinos a un punto podría ser menor que el número

de dimensiones.

Realizar una extensión de un modelo no paramétrico de más de dos variables, requiere el supuesto de aditividad, con este supuesto el modelo se vuelve más restrictivo que un modelo de regresión no paramétrico multivariante, este supuesto es uno de lo más comunes en la regresión paramétrica, por tal motivo antes de hablar de modelos semiparamétricos, se definen los *modelos aditivos*.

2.3 Modelos aditivos

En la sección anterior se discutieron los modelos de regresión no paramétrica, en los que se supone la relación entre dos variables o más variables de manera conjunta. Los resultados de las regresiones no paramétricas son complejas de interpretar, ya que si se consideran más de dos variables gráficamente no sería posible visualizarlas. Por tal motivo, se buscó un modelo que pudiera aislar de forma individual el efecto de cada componente. Los modelos aditivos heredan una propiedad importante de los modelos de regresión paramétrica, *la aditividad*, lo que permite estimar de manera no paramétrica cada una de las variables. Sin embargo, se sigue contando con la desventaja del uso exclusivo de variables continuas *Ruppert (2003)*.

De acuerdo a *Brufman et al (2007)* una de las propiedades importantes de un modelo de regresión paramétrica es: la aditividad $y = \alpha + \sum_{k=1}^{k} \beta_k x_k + \varepsilon$. 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 de x_i.

Esta propiedad es heredada por los modelos de regresión múltiple no paramétrica y se le conoce como modelo aditivo, esta regresión se expresa de la siguiente forma:

$$m(x_1, ..., xk) = E(y/x1, ..., xk) = \alpha + f_1(x_1) + f_2(x_2) + ... + f_k(x_k) \qquad (35)$$

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

Como las $f_k(x_k) = \delta m(x)/\delta x_k$, esta especificación mantiene la interpretación del efecto individual de cada regresor. El modelo aditivo es más restrictivo que el modelo general no paramétrico, debido a la exclusión del efecto interacción entre los predictores *Ruppert (2003)*; 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 es importante desde el punto de vista del cómputo, como también de la interpretación de los resultados.

2.3.1 Ejemplo

Para ejemplificar esta regresión de una manera más intuitiva, se consideran únicamente dos predictores, por lo que la ecuación estaría representada como sigue:

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon \tag{36}$$

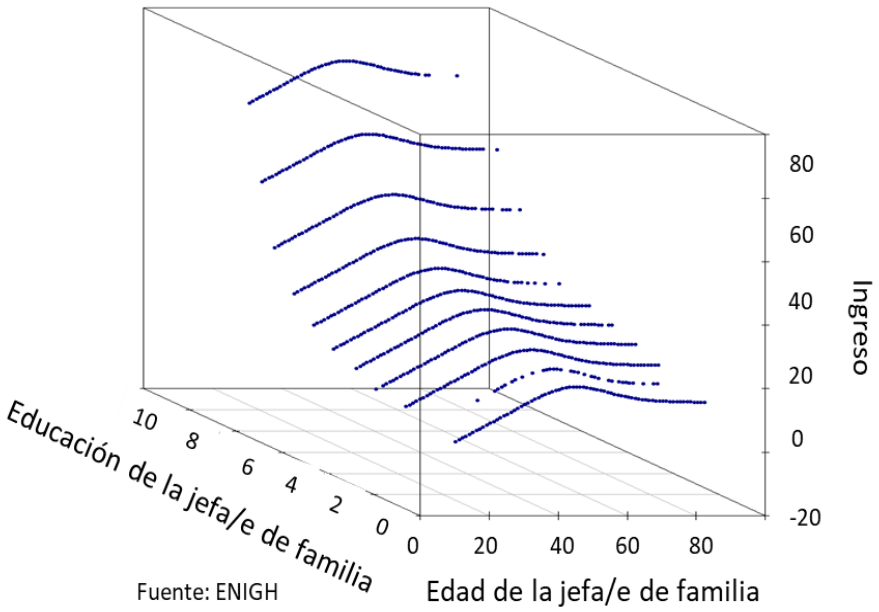
En el ejercicio con datos de la $ENIGH(2018)$, se considera el ingreso laboral (IL) en función de la edad y escolaridad de la madre o padre de familia. Como se ha mencionado, se supone que los dos regresores son continuos.

$$IL = \alpha + Edad(x_1) + Escolaridad(x_2) + \varepsilon \tag{37}$$

En la siguiente gráfica se observa como el ingreso por trabajo incrementa cuándo la educación incrementa, además se aprecia que al inicio de la vida laboral el crecimiento en el ingreso es prácticamente lineal, mientras que al pasar el nivel máximo tiene una pendiente pronunciada hasta llegar suavemente a tener un ingreso menor que el inicial.

Gráfica 8. Ingreso laboral

miles de pesos

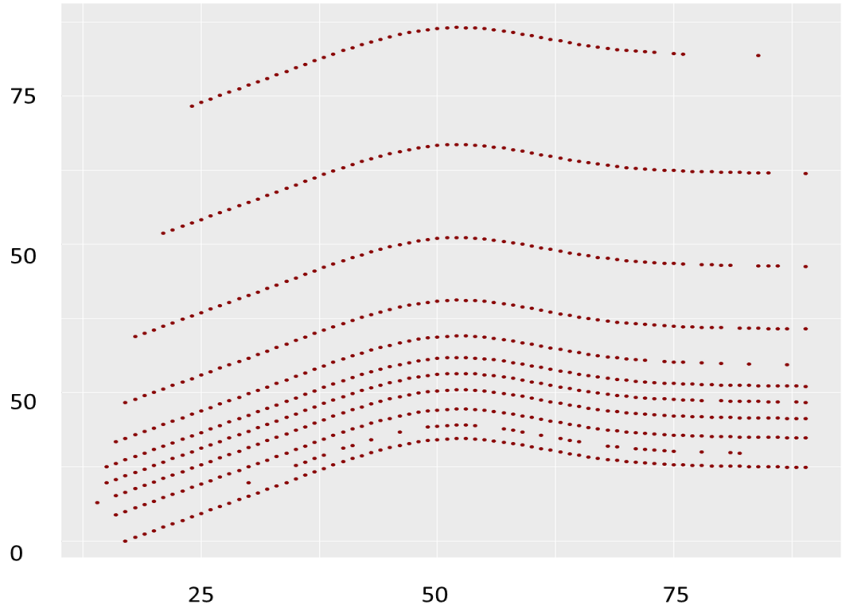


Fuente: ENIGH

Además, sí se realiza un corte del ingreso por trabajo y edad, se puede ver claramente como la escolaridad de la madre o padre de familia impacta positivamente en el ingreso laboral. Cuando la cabeza del hogar cuenta con una escolaridad hasta un nivel medio superior, existe diferencia,pero no es tan grande. No obstante, cuando la madre o padre de familia alcanzan un nivel superior, la brecha en el ingreso por trabajo cada vez es más grande, y cuando se llega a un nivel de posgrado, ésta se acentúa notablemente.

Gráfica 9. Distribución del ingreso laboral por edad y nivel de escolaridad

milles de pesos, mensual

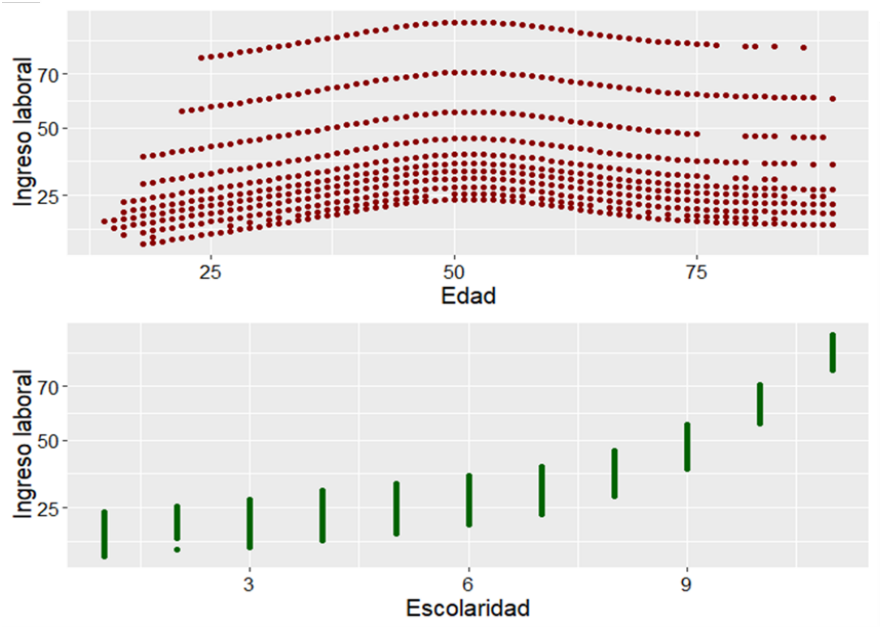


Fuente: ENIGH

Al analizar por separado las gráficas, se puede apreciar que la del ingreso laboral y la edad, muestran una especie de campana, con un crecimiento máximo alrededor de los 50 años, para cada uno de los años de escolaridad. En la gráfica del ingreso laboral y la edad, se observa un cambio significativo cuando la madre o padre de familia tiene más de 9 años de escolaridad.

Gráfica 10. Inreso laboral y edad

milles de pesos, mensual



Fuente: ENIGH

Como se mostró, estos modelos son útiles cuando se tienen variables continuas, no obstante, en la mayoría de las áreas de estudio se cuenta con distintos tipos de variables, por lo que es necesario proponer modelos con mayor flexibilidad que puedan incluir tanto variables continuas como discretas.

2.4 Regresión semiparamétrica

El modelo semiparamétrico sirve para estimar la relación tanto paramétrica y no paramétrica entre variables dependientes e independientes , si estas fueran continuas el modelo aditivo podría dar un buen ajuste, no obstante en la mayoría de las áreas de estudio social, existen variables discretas como el sexo, por tal motivo el modelo aditivo no se podría ajustar a este tipo de datos, además, se tendría otro inconveniente, si la relación entre X e Y sigue un comportamiento marcado se podría captar la relación con un parámetro, por lo que no habría razón para estimar parámetros adicionales. Por lo tanto, este modelo es de utilidad ya que incorpora términos de estimación paramétricos y no paramétricos.

$$Y = \alpha + f_1(X_1) + ... + f_k(X_k) + \beta_{k+1}X_{k+1} + ... + \beta_n X_n + \varepsilon \qquad (38)$$

Este tipo de regresión se asume que las primeras k covariables tienen un efecto

no paramétrico sobre Y y las covariables $k+1$ a n tienen un efecto paramétrico sobre Y. En la parte paramétrica del modelo, se permiten variables discretas, categóricas, ordinales y continuas, mientras que en la no paramétrica, únicamente variables continuas.

Para estimar la parte no paramétrica del modelo es necesario encontrar el ancho de banda que minimice el sesgo y la varianza en la estimación, así como la selección de la función tipo núcleo y en el caso de polinomios locales el grados de éste. La selección de la función tipo núcleo y el grado del polinomio ya se discutió en secciones anteriores y se eligió un núcleo Epanechnicov por considerar que es la densidad que cumple de una forma más adecuada todas las características necesarias. Por otra parte, para seleccionar el grado del polinomio se utilizará la función Nadayara-Watson, la cual supone un polinomio de grado 0. Sin embargo, la metodología para la selección del parámetro correspondiente al ancho de banda se realiza por un tipo de remuestreo (validación cruzada), al cual vale la pena introducir brevemente, ya que es utilizado para modelos más sofisticados de aprendizaje de máquinas, como árboles de decisión y random forest Breiman (2001).

2.4.1 Selección de h por medio de validación cruzada

Para suavizar un modelo por medio de polinomios locales, se requieren estimar un ancho de banda que minimice el sesgo y la varianza, en éste trabajo se describirá el método de validación cruzadas, el cual es una técnica de remuestreo. Esta metodología es utilizada en diversos métodos estadísticos, como en la regresión tipo LASSO *Zou, H. y Hastie (2005)*, el cual recae en la suma de los cuadrados de los residuos o por sus siglas en inglés (RSS) como una medida del modelo ajustado.

Como RSS es una medida con una habilidad de predicción no puede ser utilizado para la selección del modelo, pues esto implicaría que se utilizaran los mismos valores de Y para predecir Y. Con la validación cruzada, se evita que sean predicciones los valores de Y con ellos mismos *Hastie y Tibshirani (2013)*.

La selección del ancho de banda por validación cruzada, omite a i-ésima observación de los datos y se ajusta un modelo de regresión con ancho de banda h. La estimación se denota $\hat{f}_h(x_{-i})$, este proceso es repetido para las n observaciones, obteniendo una estimación de s cada vez que se repite el modelo.

La validación cruzada se calcula promediando las n diferencias al cuadrado.

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_h(x_{-i})]^2 = \frac{1}{n}\sum_{1}^{n}[\frac{y_i - \hat{f}_h(x_{-i})}{1 - S_{ii}}]^2 \tag{39}$$

El valor de h que minimiza la función es considerado el valor óptimo para aplicarlo al ajuste de la regresión. Este método es de una intensidad computacional grande, por ejemplo, en un conjunto de datos de 100 observaciones y 10 valores de h, implicaría una estimación por polinomios locales de unas 1000 veces. Esto resulta ineficiente para un gran número de observaciones.

La generalización de la validación cruzada o por sus siglas en inglés *(GCV). Craven y Wahba (1976)* desarrollaron este método, provee de una buena aproximación a la metodología expuesta anteriormente, con esto no es necesario estimar tantos modelos como datos se tengan, basta con ajustar únicamente la que hace intervenir todos los datos y calcular la traza de S_{ii}.

$$GCV(h) = \frac{\sum_{i=1}^{n}[y_i - \hat{f}_h(x_{-i})]^2}{(n - tr(S))^2} \tag{40}$$

En donde $tr(s) = df$ son los grados de libertad del modelo, la sustitución por $(n - df)^2$ en el cociente, se obtiene una puntuación al igual que en CV, se elige la h que minimice. En este modelo se tiene es supuesto implícito de que la función $f(x)$ es suave, este algoritmo provee de un parámetro que generaliza el modelo de la forma más óptima, es decir, si el ancho de banda fuera "pequeño", se sobreajustaría el modelo ya que pasaría por la mayoría de los puntos, esto podría generar errores de estimación cuando se quisiera aplicar a otros datos. Por otra parte, si el ancho de banda fuera "amplio" se tendría un mal ajuste, ya que prácticamente se estimaría una línea recta, *Lee (2003).*

Como se mencionó la CGV tiene la ventaja computacional de calcular la traza de S, lo cual resulta más sencillo que los elementos individuales de S_{ii}. además, en problemas de suavizamiento como el que se está presentanto, la validación cruzada generalizda puede evitar la tendencia que pasa en la metodología de CV, ya que en esta se tiene a sobreajustar el modelo, es decir, a seleccionar un ancho de banda pequeño *Hastie et al (2001)*

En este trabajo se utilizará el método de validación cruzada generalizada, para la selección del ancho de banda que minimice los errores. Una vez seleccionado el ancho de banda es necesario calcular los parámetros de la parte paramétrica de la regresión, para esta tesis se expondrá el método de *backfitting*

2.4.2 Selección del tipo de núcleo

Para determinar el kernel óptimo, es de utilidad estimar el *error cuadrático integrado medio* o por sus siglas en inglés *MISE* que se define como sigue:

$$MISE = E\int (f(x) - f(x))^2 dx \tag{41}$$

Además, se puede demostrar que la aproximación mínima del *MISE* se puede realizar escogiendo el ancho de banda como sigue:

$$h = n^{-1/5} k_2^{-2/5} [\int K^2(z)dz]^2 [\int (f^{''})^2(s)dx]^{-1/5} \tag{42}$$

Esta ecuación depende de una función de densidad desconocida f. Es importante mencionar que el ancho de banda óptimo para el *MISE* aproximadamente decrece con n en $n^{-1/5}$. Para un ancho de banda óptimo (aproximado), se tiene:

$$MISE \approx n^{-4/5} 1{,}25 C(K) [\int (f^{''})^2(s)dx]^{1/5} \tag{43}$$

en donde

$C(K) = k_2^{2/5} [\int K^2(z)dz]^4/5$

depende sólo del núcleo. Entonces el núcleo que es aproximadamente el óptimo para el *MISE* tiene el posible menor valor de $C(K)$ sujto a las restricciones en los momentos de K. Entonces, si se restringe a los núcleos, que son funciones de densidad de probabilidad, se obtiene que el núcleo óptimo es el núcleo Epanecchnicov.

$$K(z) = \frac{3}{45}(1 - z^2/5) \tag{44}$$

Este resultado y el de la ecuación 43, se puede realizar una comparación de la eficiencia de los distintos núcleos contra el Epanecchnicov, en donde se obtiene que, el núcleo triangular, rectangular, entre otros, tienen menor eficiencia como lo muestra *Zacchuini (2003)*

2.4.3 Método de selección Backfitting

Estimar modelos con múltiples términos no paramétricos es simple si ninguna de las variables están correlacionadas. Si todos los XTs son ortogonales, se podría estimar por Mínimos Cuadrados Ordinarios (MCO) para los componentes paramétricos y polinomios locales. En general es muy raro tener datos de este estilo, por lo que necesitamos un método que estime términos de un modelo aditivo o semiparamétrio. El algoritmo backfitting esta diseñado para tomar las correlaciones al estimar los términos paramétricos y no paramétricos *Keel,l.(2008)*.

El backfitting sugiere la idea de la función de regresión parcial lineal, en donde los regresores no son completamente independientes, para ejemplificar se consideran dos predictores.

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon \tag{45}$$

Para realizar este método, se asume que la forma de $f_2(\bullet)$ es conocida, pero no de $f_1(\bullet)$. Si esto fuera cierto, se puede ordenar la ecuación de tal forma que se puede resolver la ecuación, para que $f_1(\bullet)$ como sigue:

$$y - \alpha - f_2(x_2) = f_1(x_1) + \varepsilon \tag{46}$$

Suavizar $y - \alpha - f_2(x_2)$ contra x_1 produce una estimación de $f_1(x_1)$. Al conocer una parte de la regresión se puede estimar la otra parte. En realidad no se conoce ninguna de las funciones, pero se asumen valores iniciales para $f_1(\bullet)$, la función de regresión parcial sugiere que realizando una iteración se puede estimar un modelo aditivo o semiparamétrico.

2.4.4 Regresión parcial lineal

Para esta tesis, se utilizará una metodología particular de todo el conjunto de regresiones semiparamétricas, estos modelos se llaman parciales lineales y se caracterizan por tener dos componentes aditivos, uno paramétrico con múltiples variables y otro no paramétrico y se definen de la siguiente forma:

$$E(Y|\mathbf{X}_1, \mathbf{X}_2) = \mathbf{X}^T\beta + m(\mathbf{T}) + \epsilon \tag{47}$$

En donde X^T representa la parte paramétrica de la regresión, mientras que $\beta + m(\mathbf{X})$ se estimará de manera no paramétrica, además, ϵ denota al error con media cero y varianza finita y constante.

En donde $\beta = (\beta_1, ..., \beta_p)^T$ es un parámetro de dimensión finita y $\mathbf{m}(\bullet)$ es una función de suavizamiento. En este modelo se asume la descomposición de los vectores independientes como $\mathbf{X}$ y $\mathbf{X}$. El vector $\mathbf{X}_1$ denota a un vector usualmente categórico, el vector $\mathbf{T}$ es continuo y se modela por medio de técnicas no paramétricas.

Los modelos de regresión parcialmente lineales fueron estudiados ampliamente en diversas situaciones con datos reales. Éstos se propusieron por primera vez por *Engle eta al* (1986), para tratar de estudiar el efecto de las condiciones meteorológicas en la demanda de la electricidad.

Speckman (1998) realiza un experimento para determinar si el enjuague bucal de una marca de analgésico es efectiva para tratar una enfermedad de encías. Además, este mismo autor demuestra algunas propiedades de los estimadores. Uno de los métodos más empleados se basa en la combinación de la estimación por mínimos cuadrados ordinarios y la estimación tipo núcleo utilizado por el autor mencionado al principio de este párrafo.

En trabajos paralelos *Robinson (1988)* y *Speckman (1988)* proponen estimar β de la siguiente forma:

$$\hat{\beta}_b = (\widetilde{X}_b^T \widetilde{X}_b)^{-1} \widetilde{X}_b^T \widetilde{Y}_b \tag{48}$$

en donde $\widetilde{X}_b = (I - W_b)X$ e $\widetilde{Y}_b = (I - W_b)Y$, $W_b = (w_{n,b}(T_i, T_j))_{i,j}$ esta es la matriz de suavizamiento (nxn) y $b > 0$ el ancho de banda que controla el grado de suavizamiento, tal que $nb \longrightarrow \infty$ y $b \longrightarrow 0$ cuando $n \longrightarrow \infty$

Bajo ciertas condiciones, *Robinson (1988)* y *Speckman (1988)* demuestran que el estimador $\hat{\beta}_b$ es $\sqrt{n} - consistente$ para β y asintóticamente normal.

Asumiendo que $(w_{n,h}(\bullet, \bullet)$ es una función de pesos correspondientes a la estimación polinómica local de grado 0 ó 1, se tiene:

$$\widehat{m}(\bullet)_1 = \sum_{j=1}^{n} w_{j,h}(t, t_j)(T_j - T_j^T \beta) \tag{49}$$

En donde $w_{n,h}(\cdot, \cdot)$ es una función de pesos, caracterizada por una función kernel $K(\bullet)$ como se describió anteriormente.

A continuación se estima β por mínimos cuadrados ordinarios, a partir del modelo $Y_i = X_i^T \beta + w_{\beta,h}(i) + \epsilon_i$, de la siguiente manera:

$$\widetilde{X}_i = X_i - \sum_{j=1}^{n} w_{j,h}(t, t_j) X_j$$

y

$$\widetilde{Y}_i = Y_i - \sum_{j=1}^{n} w_{j,h}(t, t_j) Y_j$$

Con las estimaciones de X e Y, se puede calcular β

$$\widehat{\beta_h} = (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \widetilde{Y} \tag{50}$$

donde $\widetilde{X} = (\widetilde{X_1}, ..., \widetilde{X_n})$ e $\widetilde{Y} = \widetilde{Y_1}, ..., \widetilde{Y_n}$.

Finalmente, se obtiene el estimador $m(\bullet)$

$$\widehat{m}_h = \sum_j^n w_{n,h}(t, t_j)(Y_j - X_j^T \widehat{\beta}_h) \tag{51}$$

Robinson (1988) y *Speckman (1988)* demuestran que el estimador $\hat{\beta}_b$ es $\sqrt{n}-$ *consistente* para β y asintóticamente normal. Además, *Speckman (1988)* realiza una comparación con los estimadores del método basado en spline y mínimos cuadrados penalizados (denominados estimadores Green-Jennison-Seheult (GJS)), llegando a la conclusión de que el sesgo de β era menor cuando se utilizaba su método.

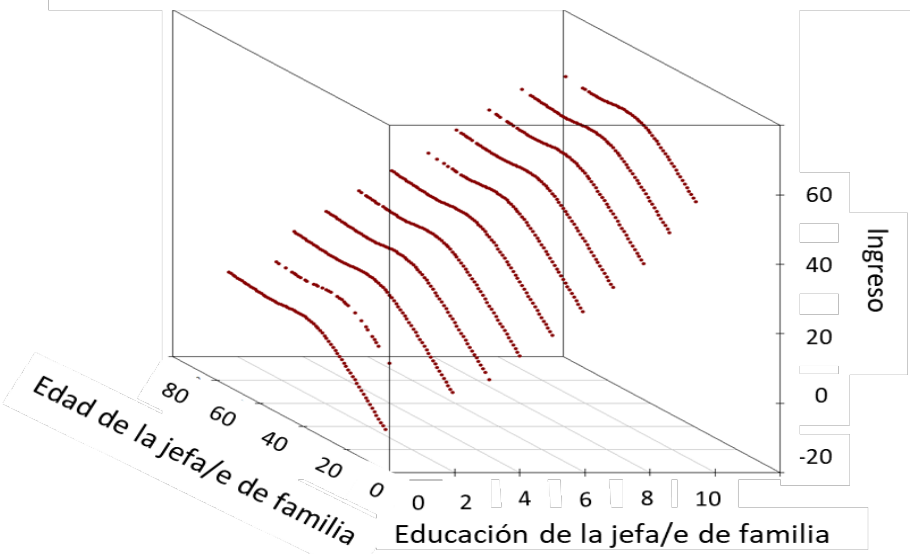
2.4.5 Ejemplo

Para ejemplificar se estimarán un modelo de regresión en donde se considera dejar que la edad se comporte con la naturaleza de los datos mismos y la escolaridad se restringirá para que tome una función determinada. Para contrastar el modelo, se estimará un modelo completamente paramétrico en donde se elevará la edad al cuadrado para estimar la regresión.

En el primer modelo se le dio la libertad de que la edad se comportara como los datos mismos, y considerando la escolaridad como una variable lineal. En esta se puede ver que el ingreso incrementa mientras mayor educación se tenga y se observa una campana en la edad.

Gráfica 11. Regresión paramétrica del ingreso laboral en función de la edad y escolaridad

miles de pesos

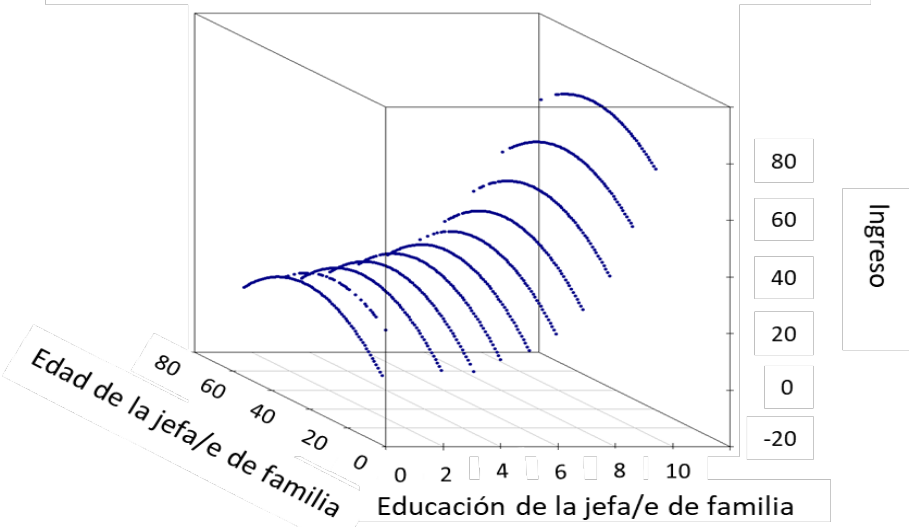


Fuente: ENIGH

En el segundo modelo se consideró la edad más la edad elevada al cuadrado, dando la libertar a la escolaridad de comportarse de manera no paramétrica. Se observa que a partir de los 6 años de escolaridad el ingreso tiene una mayor pendiente.

Gráfica 12. Regresión paramétrica del ingreso laboral en función de la edad y escolaridad

miles de pesos



Fuente: ENIGH

Al comparar los resultados de ambos modelos, se aprecia que en general, no existe una diferencia tan significativa en la media. Sin embargo, el comportamiento de los extremos de las gráficas se aprecia que en la regresión semiparamétrica la caída en el ingreso es más suave, mientras que en el modelo paramétrico la caída es muy marcada. Por los que se puede apreciar en las gráficas, el modelo semiparamétrico podría ser un mejor estimador de los valores extremos, no obstante, como se ha mencionado, su interpretación es más compleja y se debería de evaluar si los resultados son significativos.

3 Aplicación del modelo semiparamétrico

En este capítulo se revisará la literatura que hace referencia al ahorro, tanto trabajos realizados en México, como en el ámbito internacional, particularmente en Estados Unidos. Además, se revisarán las definiciones que se utilizarán para el ejercicio, así como la descripción de las variables utilizadas, la aplicación del modelo semiparamétrico y el análisis de éste.

3.1 Revisión de literatura

De acuerdo a la literatura económica, el ahorro es de suma importancia para prevenir incertidumbre en los hogares *Castellano y Garrido (2010)*, afirmando que el ahorro y crédito tienden a reducir la incertidumbre y suavizar el consumo, estos autores parten de la hipótesis del ingreso permanente de *Friedman (1957)*, la cual establece que el consumo corriente es proporcional al ingreso permanente y que ante cambios no anticipados del ingreso corriente, la tendencia del ahorro o crédito permitirá hacer frente a dichos eventos y mantener el nivel de consumo.

El modelo del ciclo de vida[3] de *Modigliani (1986)* retoma argumentos de la hipótesis del ingreso permanente y plantea que el ingreso laboral del hogar es positivo hasta antes del retiro, mientras la trayectoria del consumo es una línea recta, por lo que los agentes son ahorradores durante el periodo productivo y desahorradores durante su retiro. El modelo del ciclo de vida permite realizar un análisis del comportamiento del consumo y su financiamiento a lo largo del ciclo de vida de las familias; según el modelo, el ahorro evoluciona de una forma U-invertida a lo largo del ciclo.

En el caso de los Estados Unidos, *Férnandez-Villaverde y Krueger (2007)* estiman perfiles de consumo durante el ciclo de vida por medio de modelos semiparamétricos entre 1980 y 2001, controlados por efectos de edad, cohorte y tiempo. Los resultados indican que las pautas del consumo son similares a una U-invertida, tanto en bienes no duraderos, como en bienes duraderos, a diferencia de los modelos teóricos tradicionales los cuales afirman que el consumo se suaviza a lo largo del ciclo de vida. El consumo máximo en Estados Unidos se encuentran entre los 45 y 50 años de edad, después disminuyen hasta ser menores a los del inicio de ciclo.

Entre los trabajos que refieren los patrones de financiamiento del consumo de los hogares en México, *Attanasio y Székely (1999)* exploran cómo se comporta el ahorro de las familias entre 1984 y 1996; estos autores establecen relaciones del ahorro con los niveles de educación del jefe o jefa del hogar concluyendo que hogares con niveles altos de educación se correlacionan con mayores ingresos,

[3]Esta teoría explica cómo funcionan las finanzas de las personas en cada etapa de su vida y que factores le afectan a cada situación.

además, son hogares que poseen mayores niveles de ahorro. *Campos y Meléndez (2013)* realizan un análisis por hogar entre 1984 y 2010 acerca de las conductas del consumo y del ingreso en el ciclo de vida para México, con una metodología similar a la utilizada en este trabajo. *Fuentes y Villagómez (2001)*, analizan las tasas de ahorro para los hogares de menores ingresos con un tratamiento cercano al propuesto en este análisis; por medio de un pseudo-panel y una metodología paramétrica, en donde concluyen que los hogares de menos recursos no dejan de ahorrar hacia el final de la vida productiva, incluso presentan mayores tasas de ahorro que al inicio de éste. *Ceballos (2014)* realiza un estudio de los patrones de flujos de ahorro y pago de deuda a lo largo del ciclo de vida de los hogares mexicanos. En el que se implementó un modelo lineal parcial. Los resultados más sobresalientes indican que en lo hogares mexicanos, usan el ahorro y el crédito en distintos momentos del ciclo de vida. Al inicio del ciclo de su vida laboral utilizan el crédito para financiar su consumo, mientras que cuando los hogares alcanzan un ingreso máximo, esto es cerca de los 50 años de edad del jefe o jefa de familia, usan en mayor proporción el ahorro.

En este capítulo se buscará caracterizar el comportamiento del ahorro, consumo e ingreso de los hogares a lo largo del ciclo de vida, con un modelo semiparamétrico parcial lineal. Las tres estimaciones mencionadas serán en función de la edad del jefe o jefa del hogar como una aproximación del ciclo de vida familiar.

3.1.1 Ingreso de los hogares

El Ingreso familiar se compone de todo el ingreso ya sea monetario o en especie (bienes y servicios), que son recibidos por el hogar o por los miembros individuales de la familia en el año o en intervalos más frecuentes, pero son excluidas las ganancias imprevistas, los ingresos que son recibidos de manera irregular. El ingreso del hogar está disponible para el consumo corriente y no reducen el patrimonio neto de la familia a través de la reducción del dinero en efectivo. El ingreso del hogar puede estar compuesto por:

i) *ingreso de empleo* (auto empleo o remunerado);

ii) *renta de propiedades*;

iii) *ingreso por la producción de servicios del hogar para propio consumo*; y

iv) *transferencias recibidas.*

3.1.2 Gasto de los hogares

El gasto de consumo final en los hogares es el valor de mercado de todos lo bienes y servicios, incluidos los productos durables (tales como autos, computadoras personales y vivienda) comprados por los hogares. El gasto total de consumo en los hogares cubre todas las compras realizadas por los miembros del hogar (en casa o en el extranjero), este se caracterizará en dos rubros.

3.1.2.1 Gasto en bienes duraderos

Son aquellos en que la depreciación se observa en el mediano y largo plazo, mientras que los bienes no duraderos se deprecia en el corto plazo. Existen discrepancia al clasificar estos bienes entre algunos expertos, por ejemplo el concepto anterior es el utilizado por *Attanasio y Weber (1995)* y *Attanasio, Battistin e Ichimura (2007)*. Sin embargo, al agregar los bienes, por ejemplo, los autores antes mencionados agregan bienes semiduraderos dentro de su definición de bienes duraderos, como los gastos ejercidos en vestido y calzado; mientras que *Fernández-Villaverde y Krueger (2007)* no consideran al vestido y al calzado como bienes duraderos, sino que los excluyen de la definición de duraderos o no duraderos. Para este trabajo se sigue la metodología de *Campos y Meléndez (2013)*, quienes no consideran el calzado y vestido como bien duradero. Sin embargo, la educación si es considerada como un bien durable.

3.1.2.2 Gasto en bienes no duraderos

Son todas las mercancías cuya vida útil es menor a un año, demandados por los hogares, tales como alimentos, bebidas, combustibles *Banco de México*.

3.1.3 Ahorro de los hogares

Como se ha mencionado, el ahorro de los hogares es de interés para este trabajo, ya que se busca caracterizar el comportamiento a lo largo del ciclo de vida de las familias mexicanas. Sin embargo, existen diversas definiciones, por tal motivo se revisarán algunas. *Bultemann y Gallego (2001)*, definen al ahorro del hogar como la diferencia entre el ingreso y el gasto total. El ingreso corresponde a la percepción monetaria total del hogar, excluyendo el ahorro forzoso para el retiro y los impuestos, e incluyendo las transferencias públicas, privadas y los pagos de pensiones. Además, se incluyen tanto en el ingreso como en el gasto la renta imputadas de la vivienda ocupada. Cabe mencionar que no se incluyen como gasto ni como ingreso las transferencias no monetarias recibidas por los hogares, como los regalos.

Por otra parte *Székely (1998)* menciona dos formas de medir el ahorro de los hogares. Una forma general de medir el ahorro S de un hogar es restar el gasto corriente de los integrantes al ingreso corriente declarado. Esto se puede expresar como:

$$S_1 = Y - C \tag{52}$$

Donde S_1 representa al ahorro, Y al ingreso y C al gasto tanto en bienes duraderos y no duraderos.

Hurioka (1995) y Dagenais (1992) argumentan que C solamente debe incluir

el gasto de bienes no duraderos pues no sirven para transferir consumo hacia el futuro. Por tal motivo, se define a C_d como el gasto en bienes duraderos y a C_{nd} como el gasto en bienes no duraderos, de manera que $C = C_d + C_{nd}$. Con la definición antes mencionada se redefine el valor de S como:

$$S_2 = Y - C_{nd} \tag{53}$$

Para este trabajo se utilizarán ambas definiciones para estimar el ahorro por medio de una regresión semiparamétrica.

3.1.4 Teoría del ciclo de vida (TCV)

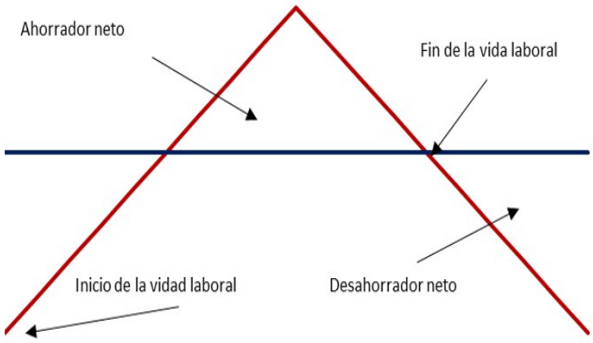
Esta teoría se encuentra asociada a los trabajos de *Modigliani y Brumberg (1954)*, *Ando y Modigliani (1963)*. Su planteamiento central es que el individuo realiza un plan de consumo para toda la vida y que el ahorro se debe fundamentalmente al deseo de las personas de garantizar su consumo.

Un aspecto importante a considerar en la teoría del ciclo de vida es que el ingreso usualmente varía a lo largo de la vida de una persona y que el comportamiento personal respecto al ahorro está determinado por la edad de la persona. Las disposición al ahorro, en cierta medida, están relacionadas con la posición de la persona a lo largo del ciclo de vida.

Ando y Modigliani (1963) argumentan que los ingresos laborales de un individuo tienen un perfil en el tiempo predecible en donde el máximo se encuentra en la edad adulta o mitad de la vida cuando la productividad es alta, y son menores en las etapas extremas de la vida, cuando la productividad es baja. Las etapas de poca productividad, corresponden a dos periodos en la vida de un individuo en que no puede ahorrar : los primeros años (prestario) y los últimos años (desahorrador). De acuerdo con esta teoría, en la edad adulta, los individuos pagan deudas anteriores y financian su jubilación.

La siguiente gráfica muestra el perfil del consumo, del ahorro y el desahorro a lo largo de la vida de un individuo, como lo sugiere la teoría del ciclo de vida: el individuo tiene un consumo constante a lo largo de su vida. La cantidad de consumo total es dada por el consumo de vida total. Durante la vida laboral activa el individuo financia su gasto de consumo con el ingreso corriente y ahorra acumulando activos.

Gráfica 12. Teoría del ciclo de vida



En la gráfica de arriba se representa el nivel máximo de los individuo en su vida activa. Al final de ella el gasto de consumo se financia con los ahorros que fueron acumulados durante la vida, y con las transferencias que reciben del gobierno (ya sea pensión del adulto mayor o por pensión por un trabajo formal) y en algunos casos de los hijos. Las áreas de ahorro y desahorro deben ser equivalentes dado que la TCV considera que el ahorro se debe fundamentalmente al deseo de las personas de prepararse para consumir en su vejez.

Si el ingreso permanente de un individuo varía, el individuo que posee activos además de las rentas permanente, planificará la forma de utilizar éstos para cambiar su consumo a lo largo de la vida.

3.2 Datos a utilizar

En el presente trabajo se utiliza la *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH)* publicada por el *Instituto Nacional de Estadística y Geografía (INEGI)*, la cual presenta los principales indicadores de ingresos y gastos en los hogares en México. Se utilizarán 14 encuestas para el análisis, las cuales resumen el comportamiento del ingreso y gasto de los últimos 5 años. Para ser específicos las bases de datos son de los siguientes años: 2010, 2012, 2014, 2016 y 2018. Las encuestas mencionadas cuentan con un diseño que asegura representatividad, además, son estrictamente comparables en términos de elaboración muestral entre el año 2010 y 2018, de acuerdo con *CONEVAL*.

La *ENIGH* se caracteriza por tener una amplia variedad de preguntas, recopiladas por al menos dos semanas consecutivas en los hogares seleccionados. Los temas recabados incluyen: características de las viviendas y de sus miembros; equipamiento del hogar; condición de ocupación y escolaridad de los individuos;

ingresos de las personas según fuente; gastos realizados por rubros; percepciones financieras y de capital, entre otras. Este trabajo utiliza variables similares a las utilizadas por *Campos y Meléndez (2013)*, en donde se clasifica como sigue:

i) *Educación.* El gasto en inscripciones, colegiaturas, enciclopedias, libros y revistas se incluye como gasto en bienes duraderos.

ii) *Vivienda.* El consumo en vivienda se dividió en bienes duraderos y no duraderos. Los gastos de renta o pagos de la vivienda, alquiler de terrenos e impuesto predial son considerados bienes duraderos.

iii) *Cristalería y blancos.* Este tipo de gastos son captados de manera trimestral y principalmente se refieren a vajillas, vasos, cubiertos, accesorios de plástico, ollas, herramientas para el hogar, colchones, cobertores, sabanas, toallas y cortinas. El gasto se convierte a gasto mensual.

iv) *muebles y enseres domésticos.* La encuesta cuantifica los gastos realizados durante los pasados seis meses en diversos muebles y enseres, entre los que se encuentran: refrigerador, lavadora, máquina de coser, recámara, comedor, sala y muebles de baño.

v) *Equipamiento del hogar.* Aparatos o equipos electrónicos también se consideran dentro de los bienes duraderos, por ejemplo, computadora, estéreos y televisores.

vi) *Vehículos.* En el último rubro de los bienes duraderos se tiene, la adquisición de autos, motocicletas, lanchas, remolques son considerados en la categoría de duraderos.

La clasificación de los bienes no durables se describe de la siguiente manera:

i *Alimentos.* Se incluyen 240 rubros de alimentos, entre los que se encuentran: cereales y derivados del maíz y trigo; consumo de carne e res, cerdo pollo y pescado; leche, huevo, quesos y otros derivados; aceites y grasas; tubérculos, verduras y legumbres; leguminosas y semillas; frutas frescas y procesadas; azúcares y mieles; especies y aderezos; bebidas alcohólicas y no alcohólicas; tabaco y otros alimentos.

ii) *Transporte.* En este concepto se incluyen los recursos destinados a: metro, autobús (local o foráneo), colectivo, taxi. También se incluyó el gasto en trasporte aéreo, ferroviario, cuotas de autopista.

iii) *Artículos de limpieza y servicios para el hogar.* Se incluyen aproximadamente 25 artículos, los cuales se obtienen de las preguntas realizadas a los miembros del hogar acerca de gastos realizados en: detergentes, blanqueadores, escobas, trapeadores, focos, recipientes, lavandería, tintorería, jardinería, entre otros.

iv) *Cuidado personal.* En la encuesta se pregunta por artículos que se compraron un mes antes del levantamiento respecto al cuidado personal, entre los que se encuentran la pasta dental, crema, papel sanitario, etc.

v *Entretenimiento.* La encuesta cuantifica los gastos durante el mes anterior en cines, teatros, conciertos, espectáculos deportivos, juegos de azar, entre otros artículos.

vi) *Comunicaciones y servicios para vehículos.* Gastos realizados en comunicaciones, como es el caso de teléfono fijo, celular e internet, así como reparaciones del versículo, combustibles y otros servicios.

vii) *Vivienda.* El consumo en vivienda se dividió en bienes duraderos y no duraderos. Los gastos contabilizados como no duraderos son: agua, energía eléctrica, gas, recolección de basura, vigilancia, administración, petróleo, diesel, carbón, leña, velas, veladoras, entre otros combustibles.

viii) *Vestido y calzado.* El gasto en este rubro incluye prendas de vestir para todos los integrantes del hogar. Esta definición es similar a la de *Alessie y De Ree (2009).*

ix *Gastos diversos.* Los gastos diversos se preguntan semestralmente, en los cuales se incluyeron: servicios profesionales de abogados, funerales, paquetes de fiestas, gastos turísticos, etc.

Adicionalmente, se usan otras características sociodemográficas como el tamaño de hogar, en nivel de educación, edad del jefe o jefa de familia, entre otras variables. Cabe mencionar que la unidad de análisis son los hogares y en su caso los jefes de familia, de los cuales se excluye a personas mayores a 90 años y menores a 17 años. Además, se realiza una desagregación del gasto para bienes duraderos y no duraderos. También se agruparon en cohortes de nacimiento por cada año de levantamiento de la encuesta. Todas las variables fueron construidas a precios constantes 2018, utilizando el *Índice Nacional de Precios al Consumidor (INPC).*

3.3 Motivación para el uso de modelos pseudo-panel

En la mayoría de los países no existen datos tipo panel para el análisis del ahorro, ingreso y consumo de los hogares, sin embargo, se pueden encontrar datos en el tiempo de una misma encuesta en la mayoría de éstos. Por ejemplo, en México no existe una encuesta que siga en el tiempo a los hogares, por tal motivo, no sería posible modelar el ciclo de vida de los hogares mexicanos. Sin embargo, el modelo propuesto en este trabajo da la posibilidad de seguir cohortes. Una cohorte es definida como un grupo de individuos que comparten alguna característica, el ejemplo más sencillo es el de la edad, por ejemplo todos los hombres mexicanos nacidos entre 1980 y 1985. En este trabajo se construyen cohortes clasificando a los jefes o jefas de hogar cada 5 años, según el año de nacimiento. Dando la posibilidad de seguir a distintos grupos a través del ciclo de

vida *Deaton (1985)*. Como ya se mencionó, en México se cuenta con la *ENIGH*, la cual es comparable en algunos periodos de tiempo, por ejemplo, la *ENIG (2018)* únicamente tiene comparabilidad con las encuestas realizadas a partir del 2010, de acuerdo con *CONEVAL*. Llevar acabo un método que convierta un conjunto de datos en datos tipo panel, es de gran ayuda para seguir el consumo, ingreso y ahorro de los hogares en México.

3.4 Aplicación del modelo semiparamétrico

De acuerdo con *Fernández-Villaverde y Krueger (2007)*, estimar el promedio de ahorro o pago de deuda de una cohorte en un momento de tiempo, requiere tener en cuenta que pueden haber tres efectos simultaneos: tiempo, cohorte y edad. El primer efecto representa el impacto de los ciclos económicos; las tasas de ahorro o crédito pueden variar de acuerdo al año debido al crecimiento de la economía. El segundo efecto refiere a las diferencias generacionales; es probable que las personas nacidas en 1960 se comporten diferente a las nacidas en 1995. El tercer aspecto recoge los efectos del ciclo de vida por medio de la edad. Como se mencionó anteriormente la flexibilidad de este modelo permite tener variables categóricas y continuas, esto resulta de gran utilidad, ya que la edad puede ser modelada de forma no paramétrica como se mostrará a continuación.

Para identificar los patrones de ahorro, ingreso y consumo a lo largo del ciclo de vida de las familias en México, se estima un modelo semiparamétrico como el propuesto por *Speckman (1988)*, ajustado a la metodología utilizada por *Fernández-Villaverde y Krueger (2004)*, el cual se describe a continuación.

3.5 Especificaciones de la elección del modelo

Los modelos no paramétricos que estiman curvas de regresión pretenden pedirle al modelo un mínimo de condiciones. Se han buscado diversos métodos para describir la relación entre las variables sin forzar las datos a una estructura de parametrización fija. *Fernández-Villaverde y Krueger (2004)* proponen documentar de manera empírica el perfil de ciclo de vida de los hogares que pueden ser usados para evaluar modelos teóricos.

Un modelo sin restricciones sería la opción óptima para estimar un modelo completamente no paramétrico, el cual sería de la forma:

$$C_{it} = M(cohorte_i, anio_t, edad_{it}, \varepsilon_{it}) \tag{54}$$

La estimación de la función M con los datos de la *ENIGH* no se podría llevar a cabo debido a que las cohortes y el año son variables categóricas, y por lo ya antes visto la estimación de modelos no paramétricos requiere de variables continuas. Además, se considera que la cohorte captura el efecto de los años, por tal motivo, no se considerarán para el modelo presentado.

3.6 Estimación del modelo parcial lineal

En este apartado se explicará cómo funciona el estimador de *Speckman (1988)*, dando mayor detalle de la aplicación del modelo a estimar.

$$c_{it} = \beta^T X + m(edad_{it}) + \varepsilon_{it} \tag{55}$$

Para una notación más sencilla, las variables categóricas se agruparan en X.
1. Primera estimación:

$$c_{it} = m_1(edad_{it}) + \varepsilon_{it} \tag{56}$$

En donde $\widehat{m_1}$ se calcula utilizando el estimador Nadaraya-Watson de la forma

$$\widehat{m_1}(edad) = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T} K_h(edad - edad_{it})c_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T} K_h(edad - edad_{it})} \tag{57}$$

Se tiene que $K_h(u) = 0{,}75h(1-(uh)^2)I(|uh| \leq 1)$ es un núcleo Epanechnikov y h es el ancho de la banda del parámetro. *Wand y Jones (1995)* utiliza el núcleo Epanechnikov ya que considera es el más eficiente, como se mencionó anteriormente.

2. Se define un a matriz de suavizamiento $\mathbf{S}$ como $\tilde{c_{it}} = \mathbf{S}y = m_1(edad_{it})$. Como el kernel es el promedio local, solo se necesita respaldar los pesos del promedio para encontrar la matriz $\mathbf{S}$. Esta matriz transforma el vector de observaciones y en valores ajustados $\tilde{y}$.

3. Se crea un vector parcial de residuos definido como $\tilde{c} = (I - \mathbf{S})cy\tilde{X} = (I - \mathbf{S})X$.

4. Se estima el parámetro β como:

$$\widehat{\beta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{c} \tag{58}$$

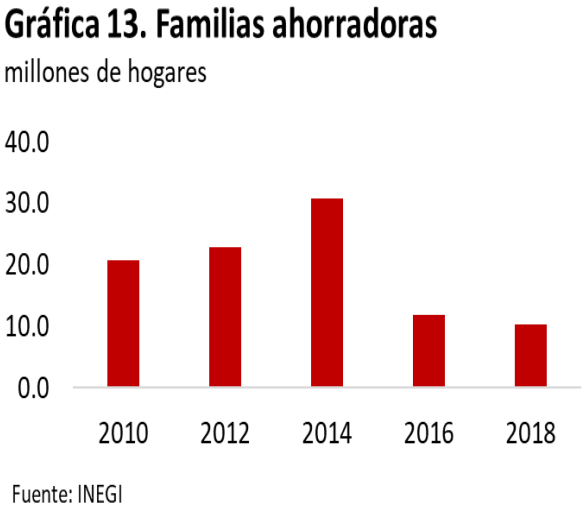
5. Finalmente, se estima la función $\widehat{m}(edad_{it})$ suavizando el núcleo usando como variable dependiente $\tilde{y} - \tilde{X}\tilde{\beta}$.

Speckman (1988) argumenta la motivación de usar este estimador, sus propiedades asintóticas y por que este método puede ser superior al de otros autores.Es importante decir que siguiendo a *Deaton (1997)*, se asume que los efectos del tiempo son ortogonales a la tendencia del tiempo y que su suma es normalizada a cero.

3.6 Resultados de la regresión parcial lineal

Como se ha mencionado, se consideró la $ENIGH$ entre los años 2010 y 2018, de los hogares con un gasto en bienes duraderos mayores a 0, la distribución de

las familias con ahorro duradero positivo por año se presenta en la siguiente gráfica.



En la gráfica 10 se puede observar que el número de familias con gasto de bienes duraderos alcanza el mayor número en 2014 (30.8 millones), mientras que en el 2018, se reduce a una tercera parte (10.3 millones).

Una vez, descrito el número de familias por cada año, se analizaran los resultados que se obtuvieron de estimar el consumo, ingreso y ahorro por medio de una regresión semiparamétrica, para la misma población.

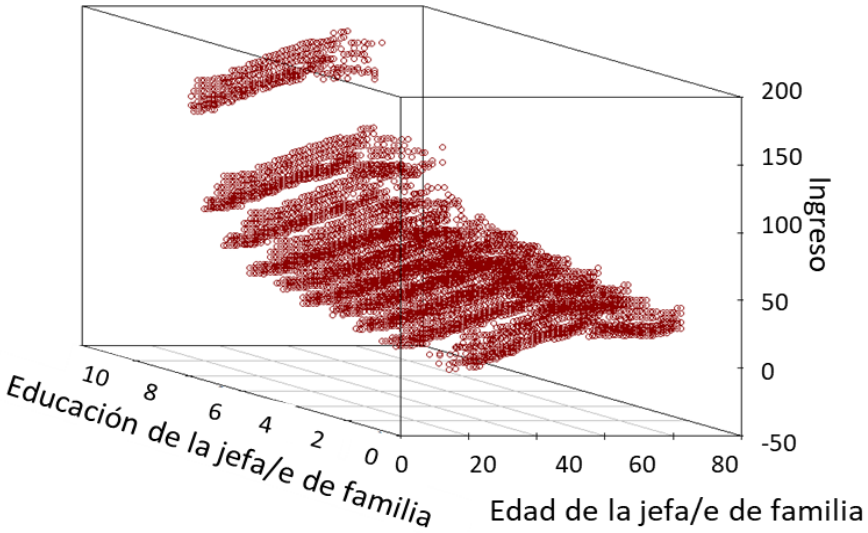
Para estimar el ingreso de los hogares en México se utilizaron las siguientes variables: 1) Cohorte (C); 2) Año en el que se realizó la encuesta (A) ; 3) Escolaridad de la madre o padre de familia (E); 4) sexo de la madre o padre de familia (S); y 5) Edad de la madre o padre de familia (Ed). Es impotante mencionar que la edad de la madre o padre de familia será la variable que se estimará de manera no paramétrica. La ecuación es la siguiente:

$$ingreso = f(Ed) + A + C + S + E + \epsilon \tag{59}$$

Dadas las estimaciones realizadas del ingreso de los hogares, se aprecia que éste alcanza su máximo entre los 45 años y 55 años, para después tener una caída que se pronuncia en la etapa del retiro.//

Gráfica 14. Ingreso de las familias

miles de pesos



Fuente: ENIGH

De acuerdo con lo argumentado por Fernández-Villaverde y Krueger (2004), para el caso del ingreso de los hogares en México, parece no haber un efecto en las cohortes construidas, es decir, que el jefe de hogar pertenezca a alguna cohorte no implica un mayor ingreso en el hogar. Sin embargo, el año en el que se realizó la encuesta tiene un resultado significativo, principalmente en el 2016 y 2018.

En la gráfica 11 se puede observar que el patrón de ingreso se mantiene independientemente del nivel educativo del padre de familia, y claramente se muestra que si el padre de familia tiene un nivel de posgrado, su ingreso es alto.

En la tabla 3 se presentan los resultados obtenidos de la regresión mencionada.

Tabla 3. Ingreso de los hogares en México

Variables	Coefientes	Error esttándar	t-valor	P-valor	
Intercepto	13,272	14,956	0.89	0.37	
2012	2,804	1,436	1.95	0.05	.
2014	1,353	1,735	0.78	0.44	
2016	12,228	2,435	5.02	0.00	***
2018	20,244	2,996	6.76	0.00	***
Mujer	- 4,247	1,026	- 4.14	0.00	***
Cohorte 2	- 3,320	7,140	- 0.47	0.64	
Cohorte 3	- 2,239	8,142	- 0.28	0.78	
Cohorte 4	- 1,213	9,472	- 0.13	0.90	
Cohorte 5	29	10,890	0.00	1.00	
Cohorte 6	8,696	12,224	0.71	0.48	
Cohorte 7	3,657	13,470	0.27	0.79	
Cohorte 8	5,140	14,641	0.35	0.73	
Cohorte 9	3,556	15,777	0.23	0.82	
Cohorte 10	2,179	16,908	0.13	0.90	
Cohorte 11	- 318	18,025	- 0.02	0.99	
Cohorte 12	- 3,166	19,203	- 0.17	0.87	
Cohorte 13	- 7,994	20,434	- 0.39	0.70	
Cohorte 14	- 9,153	21,742	- 0.42	0.67	
Preescolar	- 12	13,912	- 0.00	1.00	
Primaria incompleta	5,808	2,087	2.78	0.01	**
Primaria completa	13,174	2,099	6.28	0.00	***
Secundaria incompleta	18,125	2,916	6.22	0.00	***
Secundaria completa	21,124	2,072	10.20	< 2e-16	***
Preparatoria incompleta	26,105	2,976	8.77	< 2e-16	***
Preparatoria completa	31,987	2,301	13.90	< 2e-16	***
Profesional incompleta	47,083	3,065	15.36	< 2e-16	***
Profesional completa	68,422	2,251	30.40	< 2e-16	***
Posgrado	133,849	3,294	40.64	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(edad_jefe)	2.977	3.993	1.301	0.28

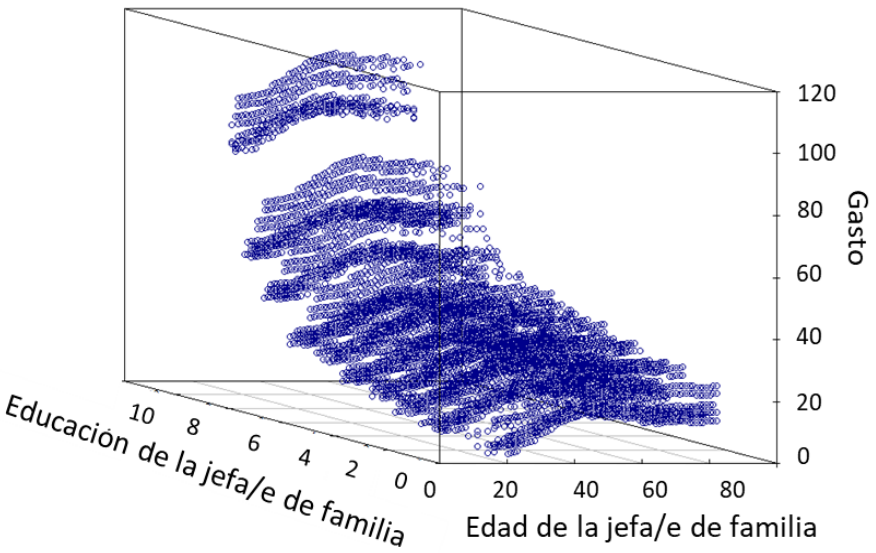
R-sq.(adj) = 0.0427 Deviance explained = 4.3%
-REML = 1.1973e+06 Scale est. = 1.8296e+13 n = 88742

De acuerdo con los resultados presentados, para el modelo estimado, la edad de la madre o padre de familia no es significativa en este caso, es decir, se pudo haber estimado de forma paramétrica. Por otra parte, se observa que si la cabeza del hogar es mujer, se tiene un ingreso promedio menor de $4,247$. Además, como ya se mencionó, el nivel educativo es relevante en el ingreso.

Para estimar el gasto corriente de los hogares en México se utilizaron las mismas variables que en el ejercicio anterior. En la gráfica 12 se puede apreciar que el gasto de los hogares, presenta el mismo comportamiento que la del ingreso. Este ejercicio, se contrapone a la hipótesis de *Ando y Modigliani (1963)*, quienes sugerían que los individuos mantienen un gasto constante durante toda su vida, para que en el retiro puedan seguir manteniendo el mismo consumo.

Gráfica 15. Gasto corriente de las familias

miles de pesos



Fuente: ENIGH

En la tabla 4 se muestran los coeficientes de la regresión, así como su nivel de significancia, para gasto corriente. Cuando se analizan los resultados, se aprecia que la edad muestra una significacia del 95 porciento. Por otro lado, las cohortes 8, 9 10 presentan coeficientes positivos y significativos, esto puede ser debido a que en estas cohortes se encuentran los adultos de la mediana edad con una expectativa de ingreso alta en el futuro. Por su parte, la escolaridad sigue teniendo gran relevancia en el gasto de los hogares.

Tabla 4. Gasto de los hogares en México

Variables	Coefientes	Error esttándar	t-valor	P-value	
Intercepto	7,094	3,685	1.93	0.05	.
2012	1,371	330	4.16	0.00	***
2014	- 1,165	398	- 2.93	0.00	**
2016	7,554	559	13.51	< 2e-16	***
2018	13,275	689	19.28	< 2e-16	***
Mujer	- 2,508	236	-10.64	< 2e-16	***
Cohorte 2	- 670	1,795	- 0.37	0.71	
Cohorte 3	7	2,260	0.00	1.00	
Cohorte 4	332	2,634	0.13	0.90	
Cohorte 5	2,608	2,942	0.89	0.38	
Cohorte 6	4,401	3,195	1.38	0.17	
Cohorte 7	4,789	3,427	1.40	0.16	
Cohorte 8	6,602	3,650	1.81	0.07	.
Cohorte 9	7,946	3,874	2.05	0.04	*
Cohorte 10	8,164	4,107	1.99	0.05	*
Cohorte 11	7,323	4,341	1.69	0.09	.
Cohorte 12	6,148	4,592	1.34	0.18	
Cohorte 13	4,212	4,856	0.87	0.39	
Cohorte 14	4,553	5,142	0.89	0.38	
Preescolar	1,400	3,194	0.44	0.66	
Primaria incompleta	3,233	479	6.75	0.00	***
Primaria completa	7,132	482	14.80	< 2e-16	***
Secundaria incompleta	9,707	670	14.50	< 2e-16	***
Secundaria completa	12,674	476	26.64	< 2e-16	***
Preparatoria incompleta	16,993	683	24.86	< 2e-16	***
Preparatoria completa	20,145	528	38.13	< 2e-16	***
Profesional incompleta	32,838	704	46.65	< 2e-16	***
Profesional completa	44,167	517	85.45	< 2e-16	***
Posgrado	75,518	756	99.84	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value	
s(edad_jefe)	5.428	6.776	5.573	2.71E-06	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

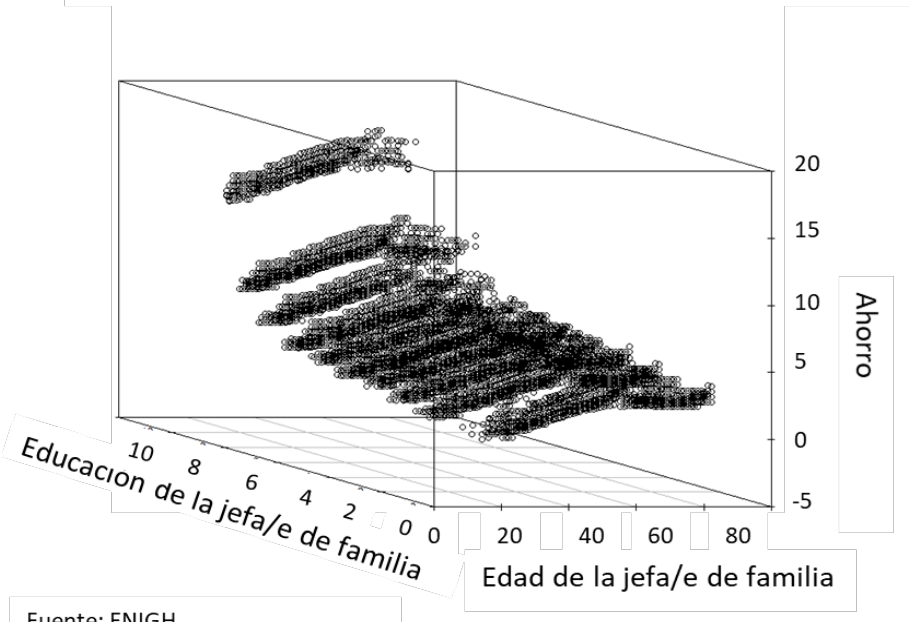
Como ya se había señalado para esta tesis se utilizará la definición de ahorro tanto de *Székely (1998)* como de *Hurioka (1995) y Dagenais (1992)*. Primero se estima la diferencia entre el ingreso total menos el gasto corriente, esta definición es la que utiliza el primer autor. Las variables usadas son las mismas que en los ejercicios anteriores.

En la gráfica se puede observar una pauta de ahorro similar a la del consumo e ingreso, es decir, a la edad de mayor bienestar también se cuenta con mayor ahorro. De acuerdo con la regresión, se tiene que el mayor ahorro en los hogares mexicanos se encuentra cuando la cabeza del hogar tiene cerca de 58 años.

Un resultado interesante es la edad en la que los hogares alcanzan un ahorro máximo con la primera definición, entre los 55 y 60 años de la madre o padre de familia. Esto se podría deber que al rededor de los 60 se obtiene recursos por parte del Instituto del Fondo Nacional de la Vivienda (INFONAVIT) o de algún

Gráfica 16. Ahorro de las familias

miles de pesos



Fuente: ENIGH

tipo de jubilación por parte de su empleo. No obstante, es importante analizarlo con mayor detalle.

Tabla 4. Ahorro de los hogares en México

Variables	Coefientes	Error esttándar	t-valor	P-valor	
Intercepto	17,960	14,962	1.20	0.23	
2012	2,453	1,425	1.72	0.09	.
2014	436	1,721	0.25	0.80	
2016	8,540	2,416	3.53	0.00	***
2018	15,839	2,974	5.33	0.00	***
Mujer	- 3,929	1,019	- 3.86	0.00	***
Cohorte 2	- 4,135	7,131	- 0.58	0.56	
Cohorte 3	- 4,264	8,219	- 0.52	0.60	
Cohorte 4	- 3,461	9,587	- 0.36	0.72	
Cohorte 5	- 2,867	11,008	- 0.26	0.79	
Cohorte 6	4,558	12,322	0.37	0.71	
Cohorte 7	- 1,290	13,538	- 0.10	0.92	
Cohorte 8	- 112	14,678	- 0.01	0.99	
Cohorte 9	- 1,922	15,785	- 0.12	0.90	
Cohorte 10	- 2,930	16,890	- 0.17	0.86	
Cohorte 11	- 4,819	17,984	- 0.27	0.79	
Cohorte 12	- 7,158	19,140	- 0.37	0.71	
Cohorte 13	- 12,032	20,350	- 0.59	0.55	
Cohorte 14	- 13,523	21,639	- 0.63	0.53	
Preescolar	- 143	13,804	- 0.01	0.99	
Primaria incompleta	5,417	2,070	2.62	0.01	**
Primaria completa	12,431	2,083	5.97	0.00	***
Secundaria incompleta	17,059	2,893	5.90	0.00	***
Secundaria completa	19,570	2,056	9.52	< 2e-16	***
Preparatoria incompleta	23,948	2,953	8.11	0.00	***
Preparatoria completa	29,302	2,283	12.84	< 2e-16	***
Profesional incompleta	42,370	3,041	13.93	< 2e-16	***
Profesional completa	61,847	2,233	27.70	< 2e-16	***
Posgrado	121,188	3,268	37.08	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Approximate significance of smooth terms:

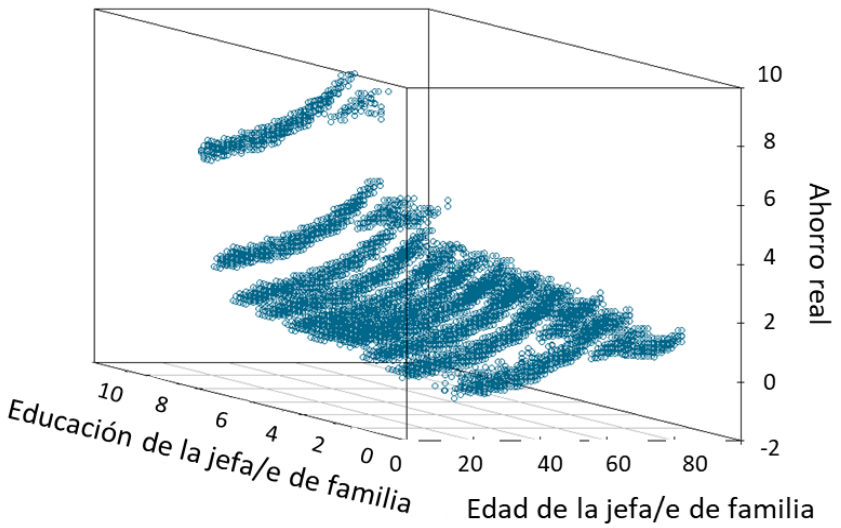
	edf	Ref.df	F	p-value
s(edad_jefe)	3.183	4.254	1.265	0.31

R-sq.(adj) = 0.035 Deviance explained = 3.53%
-REML = 1.1967e+06 Scale est. = 1.8013e+13 n = 88742

Cuando se analizan los coeficientes del ahorro de la primera definición, se observa que los años, muestran coeficientes positivos y significativos, esto implica que el ahorro ha aumentado en los últimos dos años principalmente. Por otra parte, el sexo presenta un coeficiente negativo, la interpretación que se le puede dar es que los hogares con jefa de hogar ahorran $3,929$ pesos menos trimestralmente. De la misma forma que en las regresiones anteriores, se aprecia que la educación de la cabeza del hogar, es un determinante de las pautas del ahorro.

Gráfica 17. Ahorro real de las familias

miles de pesos



Fuente: ENIGH

Para la gráfica 14 se utilizó la segunda definición de ahorro propuesta por *Hurioka (1995) y Dagenais (1992)*, en donde al ingreso únicamente se le resta el gasto corriente de bienes no duraderos, ya que se considera que adquirir un bien duradero es considerado como ahorro, ya que en el futuro se podría vender para obtener recursos. Como se puede observar que los hogares en donde la jefa o jefe del hogar tiene educación alta, en particular un posgrado, tiene una pauta de ahorro distinta a los demás hogares, mientras que en las gráficas anteriores, se notaba que a partir de una educación superior, las pautas de ingreso y consumo eran elevadas.

Tabla 5. Ahorro real de los hogares en México

Variables	Coefientes	Error esttándar	t-valor	P-valor	
Intercepto	7,486	14,172	0.53	0.60	
2012	1,422	1,361	1.05	0.30	
2014	2,516	1,644	1.53	0.13	
2016	4,655	2,308	2.02	0.04	*
2018	6,932	2,839	2.44	0.01	*
Mujer -	1,745	973	- 1.79	0.07	.
Cohorte 2 -	3,262	6,766	- 0.48	0.63	
Cohorte 3 -	3,256	7,716	- 0.42	0.67	
Cohorte 4 -	2,550	8,977	- 0.28	0.78	
Cohorte 5 -	3,537	10,320	- 0.34	0.73	
Cohorte 6	3,096	11,584	0.27	0.79	
Cohorte 7 -	2,776	12,764	- 0.22	0.83	
Cohorte 8 -	3,347	13,875	- 0.24	0.81	
Cohorte 9 -	6,115	14,951	- 0.41	0.68	
Cohorte 10 -	7,318	16,022	- 0.46	0.65	
Cohorte 11 -	8,646	17,080	- 0.51	0.61	
Cohorte 12 -	10,214	18,196	- 0.56	0.57	
Cohorte 13 -	13,275	19,362	- 0.69	0.49	
Cohorte 14 -	15,102	20,602	- 0.73	0.46	
Preescolar -	1,440	13,182	- 0.11	0.91	
Primaria incompleta	2,562	1,977	1.30	0.20	
Primaria completa	6,028	1,989	3.03	0.00	**
Secundaria incompleta	8,394	2,763	3.04	0.00	**
Secundaria completa	8,434	1,963	4.30	0.00	***
Preparatoria incompleta	9,090	2,820	3.22	0.00	**
Preparatoria completa	11,823	2,180	5.42	0.00	***
Profesional incompleta	14,197	2,904	4.89	0.00	***
Profesional completa	24,241	2,133	11.37	< 2e-16	***
Posgrado	58,319	3,121	18.69	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(edad_jefe)	2.978	3.995	0.805	0.529

R-sq.(adj) = 0.00785 Deviance explained = 0.819%

-REML = 1.1926e+06 Scale est. = 1.6427e+13 n = 88742

Al analizar los coeficientes resultantes de la regresión, se observan resultados similares a los reportados en las regresiones anteriores. Es decir, la variable que muestra un mayor impacto, es la educación de la cabeza del hogar. En este caso, el sexo de la jefa o jefe del hogar presenta una significancia del 90 porciento. La interpretación que se le podría dar, es que los hogares con jefas de familia ahorran las mismas cantidades incluso cuando, su ingreso es inferior.

4. Conclusiones

En este trabajo se realizó una revisión de 3 diferentes métodos para estimar regresiones (paramétrica, no paramétrica y semiparamétrica), para cada uno de los modelos se realizaron ejemplos, con datos reales obtenidos de la *ENIGH*. El fin de revisar los modelos antes mencionados fue introducir la regresión semiparamétrica, con la cual se llevó a cabo la caracterización del ingreso, ahorro y gasto de los hogares mexicanos. Siguiendo la metodología propuesta por *Fernández-Villaverde y Krueger (2004)*, *Campos y Meléndez (2013) y Ceballos (2014)*.

Al analizar la significacia de las variables propuestas por *Fernández-Villaverde y Krueger (2004)*, se observa que las cohortes no tienen un impacto significativo en el ingreso, gasto y ahorro de los hogares en México, es probable que esto se deba a que se utilizó un periodo de análisis corto y no se haya podido capturar el efecto. Por otra parte, en la mayoría de los ejercicio se observa que uno de los factores más importantes en las tres variables analizadas es el nivel educativo de la madre o padre de familia, principalmente si cuenta con estudios universitarios o de posgrado.

Se aprecia que el ingreso máximo de los hogares se tiene al rededor de los 55 años de edad del jefe o jefa de familia, aunque, a esta edad no presenta la tasa de ahorro máximo, por lo que se podría incentivar a las familias para que en la etapa de mayor ingreso, los hogares incrementen su ahorro y para así suavizar el consumo en el retiro.

Es importante mencionar que, para probar el desempeño de la regresión semiparamétrica, se particionó la muestra en dos partes; la primera para realizar el entrenamiento, en donde se consideró el 75 porciento de las observaciones, mientras que con los datos restantes, se probó la calidad del ejercicio. Para tener un punto de partida, se realizó la comparación con un modelo de regresión paramétrica en donde la edad de la madre o padre de familia se elevó al cuadrado, esto siguiendo la teoría del ciclo de vida. Los resultados obtenidos entre ambas metodologías no distan lo suficiente para poder determinar claramente que la regresión semiparamétrica es una mejor opción . No obstante, en en general estas regresiones registraron un mejor desempeño. Los resultados obtenidos se pueden revisar en el anexo.

Referencias

[1] Alessie, R., y De Ree (2009) *"Explaning the Hump in Life Cycle Consumption Profiles"*.
De Economist, 157(1), pp. 107-120

[2] Aneiros G, López A (2014), *PLRModels: Un paquete de R para realizar inferencia estadística en modelos de regresión parcialmente lineal.*
Universidad de Corña; Universidad de Vigo.

[3] Attanasio, P. Székely M (1999). *Ahorro de los hogares y distribución del ingreso.*
Economía Mexicana Nueva Época, vol VIII, núm 2, segundo semestre de 1999.

[4] Attanasio, P. (1993). *A cohort analysis of saving behavior by U.S. households.*
National Bureau of Economic Research.

[5] Bailey, R., y Addisom, T.,(2010) *"A smoothed-distribution form of Nadayara-Watson estimation"*.
Deparment of Economics Discusion Paper , pp. 10-30.

[6] Breidt, F.J. and Opsomer, J.D. (2000). *Local polynomial regression estimators in survey sampling.*
The Annals of Statistics, 28, 1026–1053.

[7] Brufman, J. (2008), *Distribución del ingreso según género; un enfoque semiparamétrico.*
Facultad de Ciencias Económicos. Buenos Aires, Argentina.

[8] Campos, R y Melendez, A. (2013). *Una estimación semiparamétrica de las pautas de consumo e ingreso a lo largo del ciclo de vida para Mexico.*
El trimestre economico, vol. LXXX(4), num. 320, octubre-diciembre de 2013, pp. 805-840

[9] Cleveland,William S. (1979) *Robust Locally Weigthed Regression and Smoothing Scatterplots.*
Journal of the American Statistical Association 74 (368): 829–836.

Consejo Nacional de Evaluación y la Política de Desarrollo (CONEVAL)
Consultado en https://www.coneval.org.mx/Paginas/principal.aspx.

[10] Deaton, A. (1985). *Panel Data from Time Series of Cross-Sections.*
Journal of Econometrics, 30,pp. 109-126.

[11] Fan, J., Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.*
New York: Chapman and Hall.

[12] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.*
New York: Springer

[13] Fernandez-Villaverde, L., D. Krueger (2004), *Technical Appendix of Consumption over the life cycle: Facts from Costumer Expediture Survey Data.*
Working paper

[14] Friedman, J., Hastie, T. and Tibshirani, R. (2000). *Additive logistic regression: a statistical view of boosting (with discussion).*
The Annals of Statistics, 28, 337–407

[15] Gimenez, O., Crainiceanu, C., Barbraud, C., Jenouvrier, S. and Morgan, B.J.T. (2006). *Semiparametric regression in capture-recapture modeling.*
Biometrics, 62, 691–698

[16] Green,P.J, y .W.Silverman. (1994). *Nonparametric Regression and Generalized Linear Models//*. Boca Raton, FL: Chapman Hall

[17] W (1990), *Applied Nonparametric Regression.*
Cambridge University Press

[18] Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). *Nonparametric and Semiparametric Models.*
New York: Springer Series in Statistics, Springer.

[19] Hastie,T.J y Tibshirani, R. J. (1990). *Generalized Additive Models.*
Vol. (43)of Monographs on Statistics and Applied Probability, Chapman and Hall, London

[20] Hastie,T., Tibshirani,R. , y Friedman,J. (2003). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.*
3rd edn. New York, NY: SpringerVerlag.

[21] Heim, S., Fahrmeir, L., Eilers, P.H.C. and Marx, B.D. (2007). *3D space-varying coecient models with application to diusion tensor imaging.*
Computational Statistics and Data Analysis, 51, 6212–6228

[22] Hothorn,T. y Everitt, B.S. (2006) *A Handbook of Statistical Analyses Using R*
Chapman Hall/CRC, Boca Raton, Florida, 2006

Instituto Nacional de Estadística y Geografía (INEGI)
Consultado en https://www.inegi.org.mx/programas/enigh/nc/2018/.

[23] Keele,L. (2008) *Semiparametric Regression for the Social Sciences*
Chapman Hall/CRC, Boca Raton, Florida, 2006

[24] Durbám,M. (2009) *An introduction to smoothing with penalties: P-spline*
Boletín de Estadística e Investigación Operativa

[25] Muller, M. (2001) *Estimation and testing in generalized partial linear models a comparative study.*
Statistics and Computing 11: 299-309

[26] Müller, M. (2009). *R material for "Nonparametric and Semiparametric Models..*
Online: www.marlenemueller.de/nspm.html.

[27] Müller, M. (2010). *Exploring data with non- and semiparametric models.*
Online: http://www.marlenemueller.de/talks/NSPMX.pdf.

[28] Muller, M. (2014). *An Introduction to the Estimation of GPLMs and Data Examples for the R.*
Working paper

[29] Ceballos, O. (2014). *Flujos de ahorro y pago de deuda en el ciclo de vida de los hogares mexicanos.*
Colegio de Mexico, Working paper

[30] Pedersen EJ, Miller DL, Simpson GL, Ross N (2019). *Hierarchical generalized additive models in ecology: an introduction with mgcv.*
PeerJ 7:e6876 DOI 10.7717/peerj.6876

[31] Robinson, P.M. (1988) *Root n-consistent semiparametric regression.*
56:931-954.

[32] Ruppert,D., Wand,M. P. y Carroll,R.J. (2003) *Semiparametric regression.*
New York: Cambridge University Press.

[33] Ruppert,D., Wand,M. P. y Carroll,R.J. (2009) *Semiparametric regression during 2003–2007.*
New York: Cambridge University Press.

[34] Sahagun, R. *Construccion y uso de mdelo de pseudo-panel aplicado al analisis de la propiedad y numero de autos por pare de los hogares.*
Gaceta Economica, año 16, Numero Especial, Tomo 1

[35] Schafer,C. y Wasserman,L. (2006) *Tutorial on Nonparametric Inference With R*
Mellon University

[36] Silverman, B.W. (1985). *Some Aspects of the Spline Smoothing Approach to NonparametricRegressionCurveFitting.*// Journal of the Royal Statistical Society, Series B 47: 1–53.

[37] Speckman, P. (1988). *Kernel Smoothing in Partial Lineal Models.*
Journal of the Royal Statistical Society, B(50),pp. 413-436.

[38] Wand, M.P y Jones M. C.,(1995) *Kernel Smoothing.*
Vol. 60 of Monographs on Statistics and Applied Probability, Chapman and Hall, London

[39] Wand, M.P. (2014). *Semipar: Semiparametric Regression*
R. package version 1.0-4.1.

[40] Wahba, G. (1990). *Spline Models for Observational Data*.
Philadelphia: SIAM.

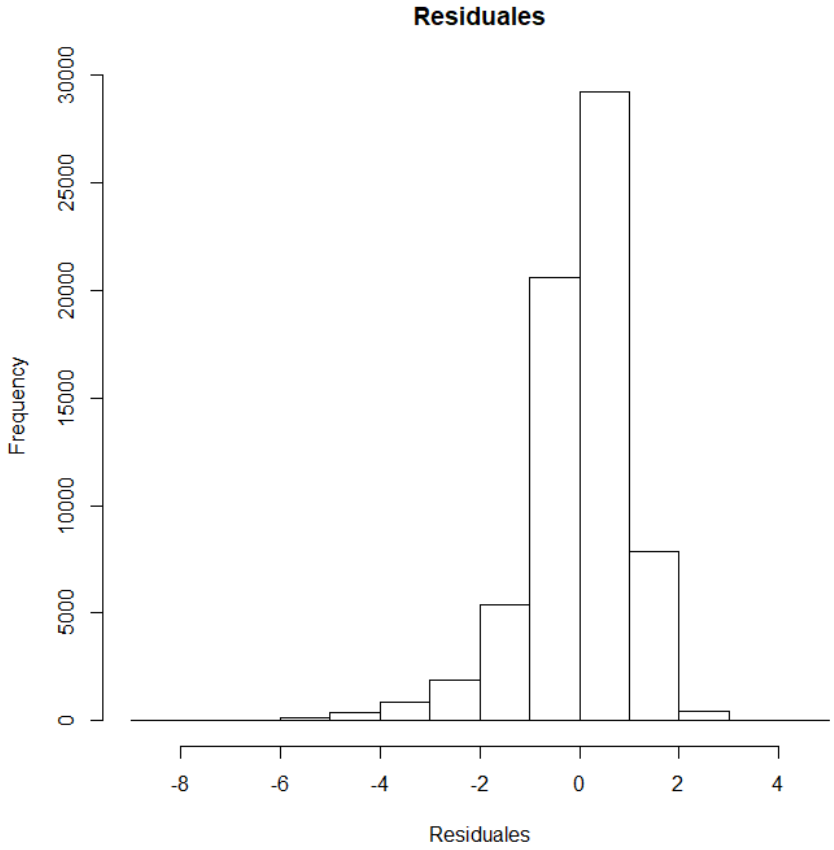
[41] Wood,S.N. (2006) *Generalized Additive Models: An Introduction with R*.
Chapman Hall/CRC, Boca Raton, Florida, 2006. ISBN 1-58488-474-6

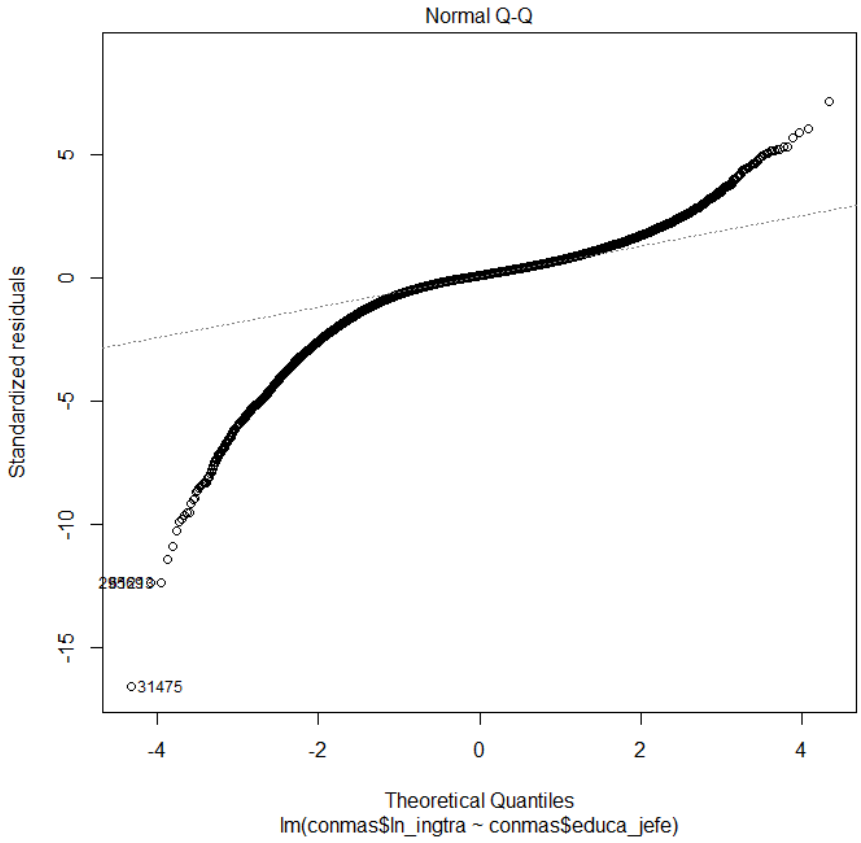
[42] Wood,S.N. (2016) *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*
R. package version 1.8-16.

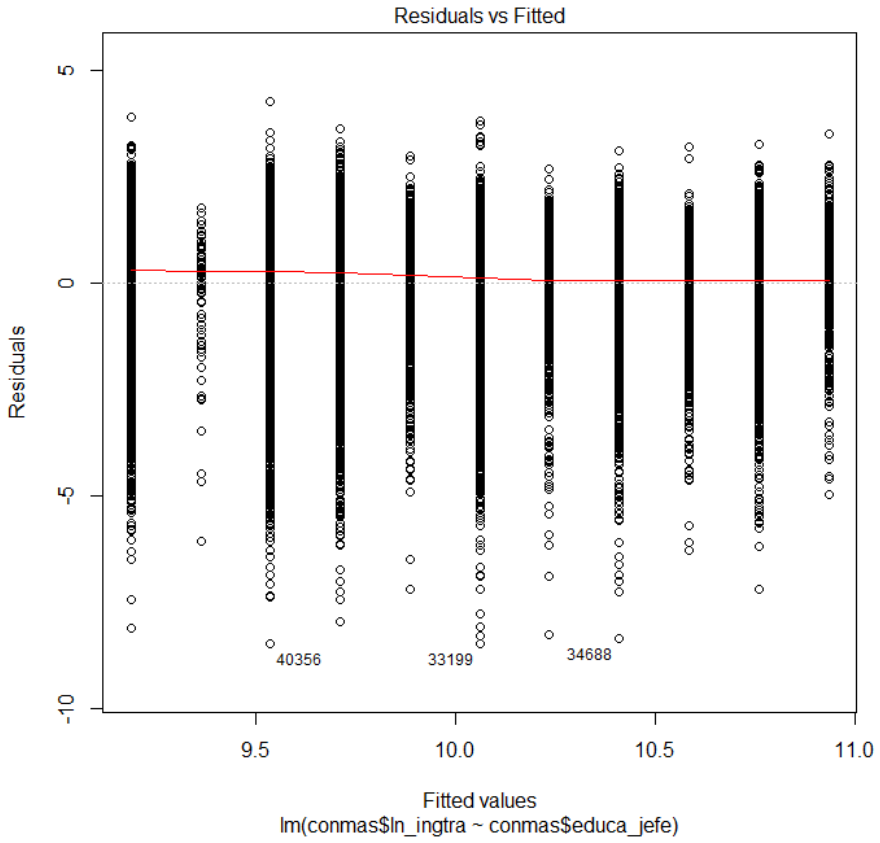
[43] Yatchew, A. (2003). *Semiparametric Regression for Applied Econometrician*.
Cambridge: Cambridge University Press.

5. Anexos

Pruebas para los errores del ejemplo de la regresión paramétrica




Normal Q-Q
Standardized residuals
5
0
-5
-10
-15
31475
-4
-2
0
2
4
Theoretical Quantiles
lm(conmas$ln_ingtra ~ conmas$educa_jefe)



Error cuadrático medio

Tabla 6. Error Cuadrático Medio (RMSE)

	Modelo semiparamétrico	Modelo paramétrico
Ahorro duradero	16,689.6	16,691.4
Gasto monetario	27,631.1	27,629.2
Ingreso corriente	52,914.4	52,915.8
Ahorro total	52,694.2	52,696.4
Ahorro real	52,770.0	52,772.9

6. Código

```
if(!require(c("multcomp","mvtnorm","survival","splines","car","doBy","reshape","
    install.packages(c("multcomp","mvtnorm","survival","splines","car","doBy","res

library("foreign")
library("car")
library("doBy")
library("reshape")
library("data.table")
library("stats")
library("dplyr")
library("fTrading")
library("KernSmooth")
library("mgcv")
library("scatterplot3d")
library("gplm")
library("survey")
library("gam")
library("forecast")
library("ggplot2")
rm(list = ls())
setwd("INF020_ENIGH2018")
memory.size()
con <- read.csv("Bases/concentradohogar.csv",as.is = TRUE)
head(con)
con$educa_jefe<-as.numeric(as.character(con$educa_jefe))
#Gr fica 1 Ingreso e ingreso laboral ENIGH 2018
g1<-con[which(con$ingtrab<900000),]

ggplot(g1, aes(x=g1$educa_jefe, y=g1$ingtrab, weight=g1$factor))+geom_point(posit
       labs(x="A os de escolaridad",y="Ingreso laboral")+
       theme(axis.text =element_text(size=12,color="black"),
             axis.title =element_text(size=15,color="black"))+
       scale_y_continuous(
         breaks = c(0,300000,600000,900000),
         label = c("0", "300", "600","900")#Breaks de los metros cuadrados
       )


conmas <- con[which(con$ingtrab>0 & con$edad_jefe<90),]
conmas$factor<-as.integer(conmas$factor)
conmas$ln_ingtra=log(conmas$ingtrab)
gg1 <- lm(conmas$ln_ingtra ~ conmas$educa_jefe, weights = conmas$factor)
```

```
educa.predict <- cbind(conmas, predict(gg1, interval = 'confidence'))



p <- ggplot(educa.predict, aes(x=educa.predict$educa_jefe, y=educa.predict$ln_ing
geom_smooth(method = "lm",color="darkred",formula=y~x)+
  labs(x="A os de escolaridad",y="Ajuste")+
  theme(axis.text =element_text(size=12,color="black"),
        axis.title =element_text(size=15,color="black"))+
  scale_y_continuous(
    breaks = c(0,500000,1000000,1500000),
    label = c("0", "500", "1,000","1,500")#Breaks de los metros cuadrados
  )
p

summary(gg1)
#Promedios m viles
g2<-con[which(con$gasto_mon<=900000 & con$edad_jefe<90),]
ggplot(g2,aes(x=g2$edad_jefe, y=g2$gasto_mon, weigh=g2$factor))+
          geom_point(position = position_jitter())+
          labs(x="Edad del jefa/jefe de familia",y="Gasto monetario")+
          theme(axis.text =element_text(size=12,color="black"),
              axis.title =element_text(size=15,color="black"))+
          scale_y_continuous(
            breaks = c(0,300000,600000,900000),
            label = c("0", "300", "600","900")
            )
#Promedio m vil

pm<-read.csv("Anios_gas_ENIGH_2016.csv")

pm1<-ggplot(pm, aes(x=edad, y=gastom))+geom_line(aes(color=PM))+labs(x="Edad del
pm1<-pm1+scale_color_manual(name="Promedio m vil\nde la cohorte",
                                          values = c("black",
                                           "red",
                                           "blue",
                                           "green"))
pm1+theme(axis.text =element_text(size=12,color="black"),
     axis.title =element_text(size=15,color="black"))+
    scale_y_continuous(
      breaks = c(0,15000,20000,25000,30000),
       label = c("0", "15", "20","25","30"))

##############################
#suavizamiento tipo Kernel
```

```
gkernel <- con[which(con$ing_cor >0 & con$edad_jefe<=90) ,]

nucleo1<-ksmooth(gkernel$edad_jefe,gkernel$ing_cor, kernel =
"normal", bandwidth = 5)
nucleo2<-ksmooth(gkernel$edad_jefe,gkernel$ing_cor, kernel =
"normal", bandwidth = 2)
nucleo3<-ksmooth(gkernel$edad_jefe,gkernel$ing_cor, kernel =
"normal", bandwidth = 10)
a<-data.frame(nucleo1$x,nucleo1$y)
b<-data.frame(nucleo2$x,nucleo2$y)
c<-data.frame(nucleo3$x,nucleo3$y)
a$type <- "Cinco a os"
b$type<-"Dos a os"
c$type<-"Diez a os"
names(a) <- names(b)
df <- rbind(a, b)
names(df)<- names(c)
df1<- rbind(df,c)

gker1<-ggplot(df1, aes(x=nucleo3.x, y=nucleo3.y),size=1)+geom_line(aes(color=type
gker1<-gker1+scale_color_manual(name="Ancho de banda",
                                values = c("black",
                                           "red","green"))+
  theme(axis.text =element_text(size=12,color="black"),
        axis.title =element_text(size=15,color="black"))
gker1+
  scale_y_continuous(
    breaks = c(20000,30000,40000,50000),
    label = c("20", "30", "40","50")
  )

################################################
#Regresi n Nadayara Watson
################################################
#Se har  la regresi n de del ahorro en funci n de la edad del padre de familia

con$gc<-con$ing_cor-con$gasto_mon
congc <- con[which(con$gc>0 & con$edad_jefe<=90),]
#gkernel <- con[which(con$ING_COR >0 & con$edad_jefe<=90) ,]

mh1 <- locpoly(congc$edad_jefe,congc$gc,degree=0,kernel="normal",bandwidth=2)
mh2 <- locpoly(congc$edad_jefe,congc$gc,degree=0,kernel="normal",bandwidth=5)
mh3 <- locpoly(congc$edad_jefe,congc$gc,degree=0,kernel="normal",bandwidth=10)

b<-data.frame(mh1$x,mh1$y)
a<-data.frame(mh2$x,mh2$y)
```

```
c<-data.frame(mh3$x,mh3$y)
a$type <- "Cinco a os"
b$type<-"Dos a os"
c$type<-"Diez a osr"
names(b) <- names(a)
df <- rbind(b, a)
names(df)<- names(c)
df1<- rbind(c,df)
#jpeg("Tesis_Final/TesisFinal/Graficas/g6.jpg", width = 350, height = "350")
gker2<-ggplot(df1, aes(x=mh3.x, y=mh3.y),size=1)+geom_line(aes(color=type))+labs(
gker2<-gker2+scale_color_manual(name="Estimaci n \nNadayara-Watson",
                                values = c("black",
                                           "red","blue"))
gker2+
  theme(plot.title=element_text(hjust=0.5))+
  theme(plot.subtitle = element_text(face = "italic")) +
  theme(plot.subtitle = element_text(hjust=0.5)) +
  scale_y_continuous(
    breaks = c(15000,20000,25000,30000),
    label = c("15", "20","25", "30")
  )+
  theme(axis.text =element_text(size=12,color="black"),
        axis.title =element_text(size=15,color="black"))
##dev.off()
##########################################
#Regresi n polinomial con 2 variables
#Se regresiona el ahorro total en funci n de la edad y escolaridad del
#padre de familia

bandwidth <- bandwidth.scott(cbind(congc$educa_jefe,congc$edad_jefe))
mh.biv <- kreg(cbind(congc$educa_jefe,congc$edad_jefe),congc$gc, bandwidth=bandw
Wind.grid <- unique(mh.biv$x[,1])
Temp.grid <- unique(mh.biv$x[,2])
o <- order(mh.biv$x[,2],mh.biv$x[,1]) ## order by 2nd column
mh2 <- matrix(mh.biv$y[o],length(Wind.grid),length(Temp.grid))
persp(Wind.grid,Temp.grid,mh2,xlab="Escolaridad",ylab="Edad", zlab="Ahorro",theta

#######################################
#Modelos Aditivos

mod.gam <- gam(conmas$ingtrab ~ s(conmas$edad_jefe) + s(conmas$educa_jefe), weigl
data=conmas)
summary(mod.gam)
fit.ing <- predict(mod.gam)
a<-data.frame(conmas$edad_jefe,fit.ing)
b<-data.frame(conmas$educa_jefe,fit.ing)
```

```
a<-a[order(conmas$edad_jefe),]
b<-b[order(conmas$educa_jefe),]
#jpeg("Tesis_Final/TesisFinal/Graficas/g8.jpg", width = 350, height = "350")
scatterplot3d(conmas$edad_jefe,conmas$educa_jefe, fit.ing,
              xlab="Edad del jefe de familia",ylab="Escolaridad",zlab="",
              #main="Gr fica 8. Distribuci n del ingreso laboral",
              angle=150, pch=20,color="Darkblue")

fit.1<-ggplot(a, aes(x=conmas.edad_jefe, y=fit.ing))+geom_point(color="Darkred")+
  labs(x="Edad",y="Ingreso laboral")

fit.1<-fit.1+
  theme(plot.title=element_text(hjust=0.5))+
  theme(plot.subtitle = element_text(face = "italic")) +
  theme(plot.subtitle = element_text(hjust=0.5)) +
  theme(axis.text =element_text(size=12,color="black"),
        axis.title =element_text(size=15,color="black"))+
  scale_y_continuous(
    breaks = c(25000,50000,70000),
    label = c("25", "50","70")
  )

#jpeg("Tesis_Final/TesisFinal/Graficas/g10.jpg", width = 350, height = "350")
fit.2<-ggplot(b, aes(x=conmas.educa_jefe, y=fit.ing))+geom_point(color="Darkgreen
  labs(x="Escolaridad",y="Ingreso laboral")

fit.2<-fit.2+
  theme(plot.title=element_text(hjust=0.5))+
  theme(plot.subtitle = element_text(face = "italic")) +
  theme(plot.subtitle = element_text(hjust=0.5))+
  theme(axis.text =element_text(size=12,color="black"),
        axis.title =element_text(size=15,color="black"))+
  scale_y_continuous(
    breaks = c(25000,50000,70000),
    label = c("25", "50","70")
  )
multiplot(fit.1,fit.2)

#dev.off()
##########################
#Modelo Semiparam trico
Modelo.sem <- gam(conmas$INGTRAB ~ conmas$edad_jefe+I(conmas$edad_jefe^2)+ lo(con
fit.ingS <- predict(Modelo.sem)
a<-data.frame(conmas$edad_jefe,fit.ingS)
b<-data.frame(conmas$educa_jefe,fit.ingS)
```

```
a<-a[order(conmas$edad_jefe),]
b<-b[order(conmas$educa_jefe),]
#jpeg("Tesis_Final/TesisFinal/Graficas/g11.jpg", width = 350, height = "350")
scatterplot3d(conmas$educa_jefe,conmas$edad_jefe, fit.ingS, grid = T,
              xlab="Escolaridad",ylab="Edad",zlab="",
              main="Modelo param trico",
              angle=150, pch=20,color="Darkblue")
#dev.off()
#Diferencias entre un modelo semiparam trico, con la edad considerada como
#Lineal.
#Modelo3
Modelo.sem3 <- gam(conmas$INGTRAB ~ lo(conmas$edad_jefe)+ I(conmas$educa_jefe),w
fit.ingS2 <- predict(Modelo.sem3)
a<-data.frame(conmas$edad_jefe,fit.ingS2)
b<-data.frame(conmas$educa_jefe,fit.ingS2)
a<-a[order(conmas$edad_jefe),]
b<-b[order(conmas$educa_jefe),]
scatterplot3d(conmas$educa_jefe,conmas$edad_jefe, fit.ingS2,grid = T,
              xlab="Escolaridad",ylab="Edad",zlab="",
              main="Modelo semiparam trico",
              angle=150, pch=20,color="Darkred")

#dev.off()
############
#Hacer las gr ficas con gam, si es posible, hacer las gr ficas resultantes del
#de la paqueter a kgplm
#En esta base se realiz  un pegado de las ENIG 2008-2018, para analizar el ingre
#largo del tiempo, considerando los precios del 2016.
tesis<-read.csv("BaseGastoDuradero_V002.csv")
head(tesis)
ta<- tesis[which( tesis$edad_jefe<=90),]
head(ta)
ta$anio<-as.factor(ta$anio)
ta$cohorte<-as.factor(ta$cohorte)
ta$educa_jefe<-as.factor(ta$educa_jefe)
ta$sexo_jefe<-as.factor(ta$sexo_jefe)
ta$t_hog<-as.factor(ta$t_hog)
ha<-bandwidth.scott(ta$edad_jefe)
#Modelo
mod.gam1 <- mgcv::gam(dura_real ~s(edad_jefe)+anio+cohorte,weights = factor ,data
summary(mod.gam1)
ta$fit.ing1 <- predict(mod.gam1)
#ejer<-data.frame(ta$a,ta$cohorte,ta$edad,fit.ing1)
plot(ta$anio,ta$fit.ing1)
plot(ta$edad_jefe,ta$fit.ing1)
#Con un  ngulo  de 60 la gr fica se ve bien
```

```
scatterplot3d(ta$edad_jefe,ta$anio, ta$fit.ing1,color="Darkblue",angle = 60)
scatterplot3d(ta$edad_jefe,ta$cohorte, ta$fit.ing1,color="Darkred",angle = 220)
#####################################

mod.gam2 <- mgcv::gam(dura_real ~s(edad_jefe)+anio+cohorte+educa_jefe,weights = f
summary(mod.gam2)
ta$fit.ing2 <- predict(mod.gam2)
#ejer<-data.frame(ta$a,ta$cohorte,ta$edad,fit.ing1)
plot(ta$anio,ta$fit.ing2)
plot(ta$edad_jefe,ta$fit.ing2)
#Con un  ngulo  de 60 la gr fica se ve bien
scatterplot3d(ta$edad_jefe,ta$educa_jefe, ta$fit.ing2,color="Darkgreen",angle = 1
#############
#Ahorro
mod.gam3 <- mgcv::gam(dura_real ~s(edad_jefe)+anio+sexo_jefe+cohorte+educa_jefe,v
summary(mod.gam3)
ta$fit.ing3 <- predict(mod.gam3)
#ejer<-data.frame(ta$a,ta$cohorte,ta$edad,fit.ing1)
plot(ta$anio,ta$fit.ing3)
plot(ta$edad_jefe,ta$fit.ing3)
plot(mod.gam3$residuals)
#Con un  ngulo  de 60 la gr fica se ve bien
o<-order(ta$edad_jefe)
plot(ta$educa_jefe,ta$fit.ing3)
scatterplot3d(ta$sexo_jefe,ta$edad_jefe[o], ta$fit.ing3,color="Darkblue",angle =
######################
#GASTO
mod.gam4 <- mgcv::gam(gasto_mon ~s(edad_jefe)+anio+sexo_jefe+cohorte+educa_jefe+t
summary(mod.gam4)
ta$fit.ing4 <- predict(mod.gam4)
#ejer<-data.frame(ta$a,ta$cohorte,ta$edad,fit.ing1)
scatterplot3d(ta$edad_jefe,ta$educa_jefe, ta$fit.ing4,color="Darkblue",angle = 16
#########
#GASTO
mod.gam4_1 <- mgcv::gam(gasto_mon ~s(edad_jefe),weights = factor ,data=ta,method
ta$fit.ing_4 <- predict(mod.gam4_1)
#######################33
#INGRESO
mod.gam5 <- mgcv::gam(ing_cor ~s(edad_jefe)+anio+sexo_jefe+cohorte+educa_jefe,we
summary(mod.gam5)
ta$fit.ing5 <- predict(mod.gam5)
plot(ta$edad_jefe[ta$ing_cor<300000],ta$ing_cor[ta$ing_cor<300000])
mod.gam5_1 <- mgcv::gam(ing_cor ~s(edad_jefe),weights = factor ,data=ta,method =
summary(mod.gam5_1)
ta$fit.ing5_1 <- predict(mod.gam5_1)
plot(ta$edad_jefe,ta$fit.ing5_1)
```

```
scatterplot3d(ta$edad_jefe,ta$educa_jefe, ta$fit.ing5,color="Darkred",angle = 160
################
#Ahorro total
ta$ahorro_t<-ta$ing_cor-ta$dura_real
mod.gam6 <- mgcv::gam(ahorro_t ~s(edad_jefe)+anio+sexo_jefe+cohorte+educa_jefe+ta
summary(mod.gam6)
ta$fit.ing6 <- predict(mod.gam6)
plot()
#ejer<-data.frame(ta$a,ta$cohorte,ta$edad,fit.ing1)
scatterplot3d(ta$edad_jefe,ta$educa_jefe, ta$fit.ing6,color="black",angle = 160)
########################################
#Primera gr fica estimacui n#
########################################
mod.gam6_1 <- mgcv::gam(ahorro_t ~s(edad_jefe)+ta$anio,weights = factor ,data=ta,
summary(mod.gam6_1)
ta$fit.ing_6 <- predict(mod.gam6_1)
par(mfrow=c(3,1))
#Ingreso
plot(ta$edad_jefe,ta$fit.ing5_1, xlab="", ylab="Ingreso total")
#Gasto
plot(ta$edad_jefe,ta$fit.ing_4,xlab="",ylab="Gasto corriente", col="darkcyan")
#Ahorro
plot(ta$edad_jefe,ta$fit.ing_6,xlab="Edad de la madre o padre de familia",ylab="A
########################################
#Ahorro
########################################
ta$ahorro_2<-ta$ing_cor-ta$gasto_mon
ta$t_hog<-as.factor(ta$t_hog)
mod.gam7 <- mgcv::gam(ahorro_2 ~s(edad_jefe)+anio+sexo_jefe+cohorte+educa_jefe+ta
summary(mod.gam7)
ta$fit.ing7 <- predict(mod.gam7)
#ejer<-data.frame(ta$a,ta$cohorte,ta$edad,fit.ing1)
scatterplot3d(ta$edad_jefe,ta$educa_jefe, ta$fit.ing7,color="deepskyblue4",angle
########################################
mod.gam7_1 <- mgcv::gam(ahorro_2 ~s(edad_jefe),weights = factor ,data=ta,method =
summary(mod.gam7_1)
ta$fit.ing_7 <- predict(mod.gam7_1)
plot(ta$edad_jefe,ta$fit.ing_7)

########################################
#La cohorte no fue significativa
#es probable que por el corto periodo de tiempo que se consider ,
ta$ahorro_t<-ta$ing_cor-ta$dura_real
ta$t_hog<-as.factor(ta$t_hog)
mod.gam8 <- mgcv::gam(ahorro_t ~s(edad_jefe)+anio+sexo_jefe+educa_jefe+t_hog,weig
summary(mod.gam8)
```

```
ta$fit.ing8 <- predict(mod.gam8)
#ejer<-data.frame(ta$a,ta$cohorte,ta$edad,fit.ing1)
scatterplot3d(ta$edad_jefe,ta$educa_jefe, ta$fit.ing8,color="deepskyblue4",angle
```