Regresión Local

Pablo Ramsés Alonso Martín

Facultad de Ciencias de la Universidad de Oviedo
Grado en Matemáticas

17 de febrero de 2021

Índice

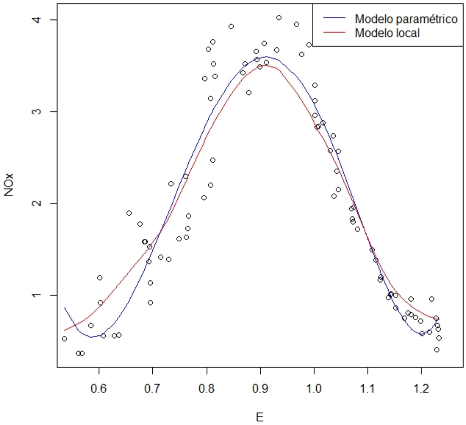
1. Motivación y objetivos
2. Preliminares
3. Estimadores tipo núcleo
4. Elección óptima de los parámetros
   - Núcleo óptimo
   - Ancho de banda óptimo
5. Aplicaciones de la Regresión Local tipo Núcleo
   - Detección de Outliers
   - Estimación del ruido de una Serie Temporal
   - Análisis de series temporales en tiempo real
6. Conclusiones

Motivación

- La relación entre diferentes sucesos siempre ha sido uno de los principales objetos de estudio de la estadística.
- En muchas ocasiones es difícil encontrar un modelo paramétrico adecuado.

Motivación

- La relación entre diferentes sucesos siempre ha sido uno de los principales objetos de estudio de la estadística.
- En muchas ocasiones es difícil encontrar un modelo paramétrico adecuado.



Objetivos

- Análisis teórico y empírico de las propiedades del estimador de Nadaraya-Watson.

Objetivos

- Análisis teórico y empírico de las propiedades del estimador de Nadaraya-Watson.
- Elecciones óptimas de los parámetros.

Objetivos

- Análisis teórico y empírico de las propiedades del estimador de Nadaraya-Watson.
- Elecciones óptimas de los parámetros.
- Aplicaciones de la regresión local tipo núcleo.

Objetivos

- Análisis teórico y empírico de las propiedades del estimador de Nadaraya-Watson.
- Elecciones óptimas de los parámetros.
- Aplicaciones de la regresión local tipo núcleo.
    1. Detección de Outliers.

Objetivos

- Análisis teórico y empírico de las propiedades del estimador de Nadaraya-Watson.
- Elecciones óptimas de los parámetros.
- Aplicaciones de la regresión local tipo núcleo.
  1. Detección de Outliers.
  2. Eliminación del ruido en series temporales.

Objetivos

- Análisis teórico y empírico de las propiedades del estimador de Nadaraya-Watson.
- Elecciones óptimas de los parámetros.
- Aplicaciones de la regresión local tipo núcleo.
  1. Detección de Outliers.
  2. Eliminación del ruido en series temporales.
  3. Análisis de series temporales a tiempo real.

Índice

1 Motivación y objetivos

2 Preliminares

3 Estimadores tipo núcleo

4 Elección óptima de los parámetros
- Núcleo óptimo
- Ancho de banda óptimo

5 Aplicaciones de la Regresión Local tipo Núcleo
- Detección de Outliers
- Estimación del ruido de una Serie Temporal
- Análisis de series temporales en tiempo real

6 Conclusiones

Planteamiento del problema

Problema clásico de regresión: describir el comportamiento de una variable Y a través de la observación de una variable predictora X .
Información: una m.a.s. $\{(x_i, y_i) : i = 1, ..., n\}$ de la variable (X, Y).
Modelo: $Y_i = \mu(X_i) + \epsilon_i$ donde ϵ_i representa el error no observable asociado a cada observación y μ es la transformación de X que mejor aproxima a Y.

Proposición

Sea (X, Y) una variable aleatoria, entonces la mejor aproximación de Y a través de una transformación de X, en términos del ECM, es $\mu(x) = E(Y|X = x)$.

Objetivo: estimar $\mu(x) = E(Y|X = x)$ a través de la muestra disponible.

Regresión Local

- La estimación por Regresión Local de $\mu(x)$ se caracteriza por filtrar las observaciones según su distancia a x.
- Los estimadores tipo núcleo suponen cierta suavidad sobre μ, de manera que si x_i está suficientemente cerca de x entonces esa relación se mantiene entre $\mu(x_i)$ y $\mu(x)$.

Regresión Local

- La estimación por Regresión Local de $\mu(x)$ se caracteriza por filtrar las observaciones según su distancia a x.
- Los estimadores tipo núcleo suponen cierta suavidad sobre μ, de manera que si x_i está suficientemente cerca de x entonces esa relación se mantiene entre $\mu(x_i)$ y $\mu(x)$.

Ejemplo

Se llama **estimador por medias locales** *de la función de regresión a*

$$\hat{\mu}_{loc}(x) = \frac{\sum_{|x_i - x| < h} Y_i}{n_h(x)}$$

Conceptos previos

- Variables aleatorias y conceptos asociados estudiados en el grado.

Conceptos previos

- Variables aleatorias y conceptos asociados estudiados en el grado.
- Ancho de banda (h), que determina la ventana de estimación $[x - h, x + h]$.

Conceptos previos

- Variables aleatorias y conceptos asociados estudiados en el grado.
- Ancho de banda (h), que determina la ventana de estimación $[x - h, x + h]$.
- Funciones núcleo:

Conceptos previos

- Variables aleatorias y conceptos asociados estudiados en el grado.
- Ancho de banda (h), que determina la ventana de estimación $[x - h, x + h]$.
- Funciones núcleo:

Definición

Se llama **función núcleo** *a cualquier función real K que verifique:*

N1. $K(x) \geq 0 \quad \forall x \in \mathbb{R}$

N2. $\displaystyle\int_{-\infty}^{\infty} K(x)dx = 1$

N3. $K(-x) = K(x) \quad \forall x \in \mathbb{R}$

N4. $\displaystyle\int_{-\infty}^{\infty} xK(x)dx = 0$

Índice

1. Motivación y objetivos
2. Preliminares
3. Estimadores tipo núcleo
4. Elección óptima de los parámetros
   - Núcleo óptimo
   - Ancho de banda óptimo
5. Aplicaciones de la Regresión Local tipo Núcleo
   - Detección de Outliers
   - Estimación del ruido de una Serie Temporal
   - Análisis de series temporales en tiempo real
6. Conclusiones

Estimador de Nadaraya-Watson

Definición

Se llama **estimador de Nadaraya-Watson** de la función de regresión al dado por

$$\hat{\mu}_{NW}(x) = \sum_{i=1}^{n} \frac{K_h\left(x - x_i\right) Y_i}{\sum_{i=1}^{n} K_h\left(x - x_i\right)}$$

donde $K_h(x) = \dfrac{1}{h} K(x/h)$ y K es una función núcleo.

Se puede expresar en forma de media ponderada como

$$\hat{\mu}_{NW}(x) = \sum_{i=1}^{n} W_{hi}(x) Y_i.$$

Convergencia del estimador

Condiciones:

(H1) X es continua con función de densidad $f(x)$,

(H2) $\lim\limits_{|u|\to\infty} uK(u) = 0$,

(H3) $\int |K(u)|du < \infty$,

(H4) $E(Y^2) < \infty$,

(H5) $n \longrightarrow \infty$, $h \longrightarrow 0$ y $nh \longrightarrow \infty$,

Proposición

Sea (X, Y) una variable aleatoria bidimensional, $\mu(x) = E(Y|X = x)$ y verificándose (H1)-(H5). Entonces en todo punto de continuidad de $\mu(x)$ y $f(x)$, con $f(x) > 0$ se cumple que:

$$\hat{\mu}_{NW}(x) \xrightarrow{p} \mu(x).$$

Convergencia del estimador

Teorema

Bajo las condiciones del resultado anterior, verificándose además que

(H6) $\mu, f \in \mathcal{C}^2$, (H7) x está en el interior del soporte de f,
y siendo

$$c_K = \int K^2(u)du, \quad d_k = \int u^2 K(u)du, \quad \sigma^2(x) = Var(Y|X=x)$$

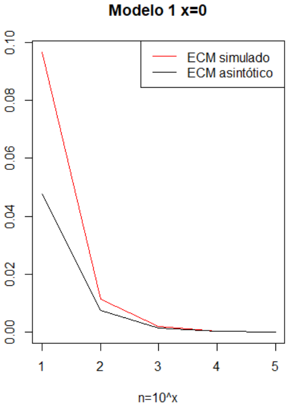
se verifica:

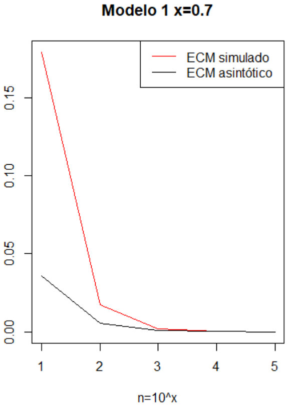
$$ECM(\hat{\mu}_{NW}(x)) = \frac{h^4}{4}\left(\mu^{''}(x) + 2\frac{\mu^{'}(x)f^{'}(x)}{f(x)}\right)d_K^2 + \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}c_K^2 + R(x)$$

donde $R(x)$ representa los términos de mayor orden.

Convergencia del estimador

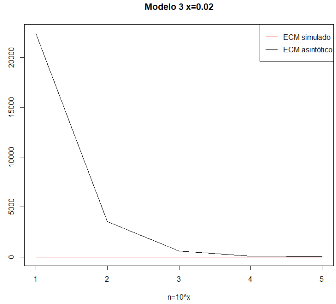
Función de regresión	Distribución de X_i	Errores
$\mu(x) = e^x$	$X_i \sim \mathcal{N}(0,1)$	$\epsilon_i \sim \mathcal{N}(0, 0.5)$





Convergencia del estimador

Función de regresión	Distribución de X_i	Errores
$\mu(x) = x^2 sen\left(\frac{1}{x}\right) I_{x\neq 0}$	$X_i \sim U(-1,1)$	$\epsilon_i \sim \mathcal{N}(0, 0.2)$



n	ECM simulado	Exp. asintótica
10	$1.253 \cdot 10^{-2}$	22401.16
10^2	$8.486 \cdot 10^{-4}$	3550.35
10^3	$1.132 \cdot 10^{-4}$	562.69
10^4	$1.567 \cdot 10^{-5}$	89.18
10^5	$2.920 \cdot 10^{-6}$	14.13

Distribución asintótica

Condiciones adicionales:

(H8) $\int |K(u)|^{2+\eta} < \infty$ para algún $\eta > 0$,

(H9) $h \sim n^{-1/5}$,

(H10) $x_1, ..., x_k$ son puntos de continuidad de $\sigma^2(x)$ y $E(|Y|^{2+\eta} \, |X = x)$

Teorema

Bajo las condiciones (H1)-(H10) se verifica:

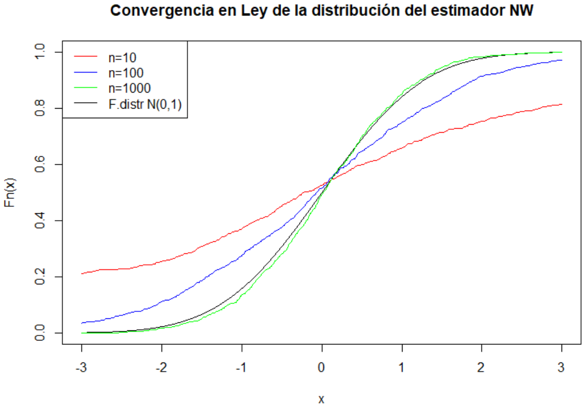
$$\left((nh)^{1/2} \left[\frac{\hat{\mu}_{NW}(x_j) - \mu(x_j)}{(\sigma^2 c_K / f(x_j))^{1/2}} \right] \right)_{j=1}^{k} \xrightarrow{L} \mathcal{N}_k(B, I_k)$$

donde

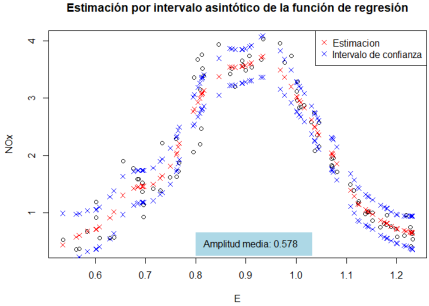
$$B = \left(d_K(\mu''(x_j) + 2\mu'(x_j)(f'(x_j)/f(x_j))) \right)_{j=1}^{k}.$$

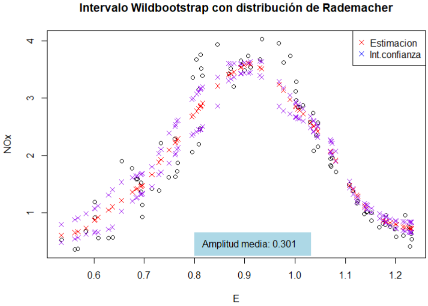
Distribución asintótica

Función de regresión	Distribución de X_i	Errores
$\mu(x) = x^3$	$X_i \sim \mathcal{U}(-1, 1)$	$\epsilon_i \sim \mathcal{N}(0, 0.5)$

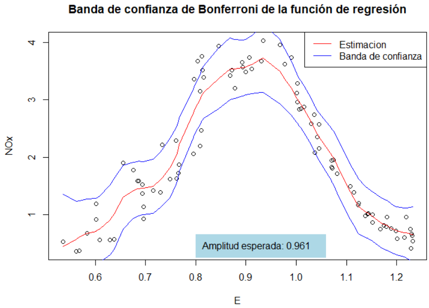


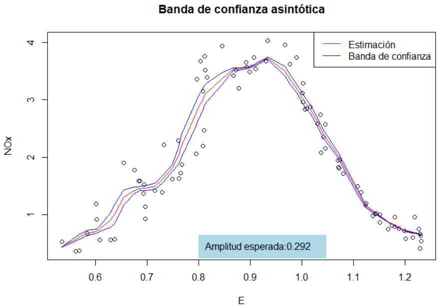
Estimación por intervalo: intervalos de confianza





Estimación por intervalo: bandas de confianza





Índice

1 Motivación y objetivos

2 Preliminares

3 Estimadores tipo núcleo

4 Elección óptima de los parámetros
- Núcleo óptimo
- Ancho de banda óptimo

5 Aplicaciones de la Regresión Local tipo Núcleo
- Detección de Outliers
- Estimación del ruido de una Serie Temporal
- Análisis de series temporales en tiempo real

6 Conclusiones

Núcleo óptimo

Problema

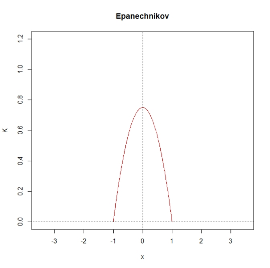
Encontrar la función núcleo K que minimiza el Error Cuadrático Medio:

$$ECM(\hat{\mu}_{NW}(x)) = \frac{h^4}{4}\left(\mu^{''}(x) + 2\frac{\mu^{'}(x)f^{'}(x)}{f(x)}\right) d_K^2 + \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}c_K^2$$

Solución

A través de los multiplicadores de Lagrange y aplicando la ecuación de Euler del cálculo variacional se obtiene como solución el Núcleo de Epanechnikov:

$$K(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}(x)$$



Núcleo óptimo

Efecto de la elección del núcleo sobre el ECM calculado sobre un modelo de datos simulados.

Núcleo	c_K	d_K	ECM(0)
Naïve	$1/2$	$1/3$	0.01533379
Gauss	$1/2\sqrt{\pi}$	1	0.01505352
Triangular	$2/3$	$1/6$	0.01462812
Epanechnikov	$3/5$	$1/5$	0.01446289
Biweight	$5/7$	$1/7$	0.01453384
Triweight	$350/429$	$1/9$	0.01461886
Coseno	$\pi^2/16$	$1 - 8/\pi^2$	0.01446924

Una elección no óptima del núcleo tiene un efecto muy escaso sobre el error.

Ancho de banda óptimo

- Elección subjetiva.

Ancho de banda óptimo

- Elección subjetiva.
- Elección según k-vecinos.

Ancho de banda óptimo

- Elección subjetiva.
- Elección según k-vecinos.
- Elección según AMSE (*Asymptotic Mean Squared Error*) o métodos plug-in:

$$AMSE(h) = \frac{h^4}{4}\left(\mu^{''}(x) + 2\frac{\mu^{'}(x)f^{'}(x)}{f(x)}\right) d_K^2 + \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}c_K^2.$$

Ancho de banda óptimo

- Elección subjetiva.
- Elección según k-vecinos.
- Elección según AMSE (*Asymptotic Mean Squared Error*) o métodos plug-in:

$$AMSE(h) = \frac{h^4}{4}\left(\mu^{''}(x) + 2\frac{\mu^{'}(x)f^{'}(x)}{f(x)}\right) d_K^2 + \frac{1}{nh}\frac{\sigma^2(x)}{f(x)} c_K^2.$$

- Elección según ASE (*Average Squared Error*) o métodos data-driven:

$$ASE(h) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_h(X_i) - \mu(X_i))^2.$$

Ancho de banda óptimo: métodos data-driven

Para minimizar la expresión $ASE(h)$ se propone un estimador que solo dependa de la muestra:

$$p(h) = \frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}_h(X_j) - Y_j)^2 w(X_j).$$

Problema

$p(h)$ no es un estimador insesgado ni asintóticamente insesgado de $ASE(h)$, y su sesgo depende de h.

Para corregir el sesgo de $p(h)$ se proponen tres alternativas que dan lugar a los métodos data-driven.

Ancho de banda óptimo: métodos data-driven

- Método de las funciones penalizadoras:

$$G(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{\mu}(X_j))^2 \mathit{\Xi}\left(\frac{1}{n}W_{h,j}(X_j)\right) w(X_j)$$

$$\mathit{\Xi}(u) = 1 + 2u + O(u^2) \quad , \ u \to 0$$

Ancho de banda óptimo: métodos data-driven

- Método de las funciones penalizadoras:

$$G(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{\mu}(X_j))^2 \Xi\left(\frac{1}{n}W_{h,j}(X_j)\right) w(X_j)$$

$$\Xi(u) = 1 + 2u + O(u^2) \quad , \ u \to 0$$

- Método *Leave-One-Out*:

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{\mu}_{h,-j}(X_j))^2 w(X_j)$$

$$\hat{\mu}_{h,-j}(x) = n^{-1}\sum_{i \neq j} W_{hi}(x) Y_i$$

Ancho de banda óptimo: métodos data-driven

- Método de las funciones penalizadoras:

$$G(h) = \frac{1}{n} \sum_{j=1}^{n} (Y_j - \hat{\mu}(X_j))^2 \Xi \left(\frac{1}{n} W_{h,j}(X_j) \right) w(X_j)$$

$$\Xi(u) = 1 + 2u + O(u^2) \quad , \; u \to 0$$

- Método *Leave-One-Out*:

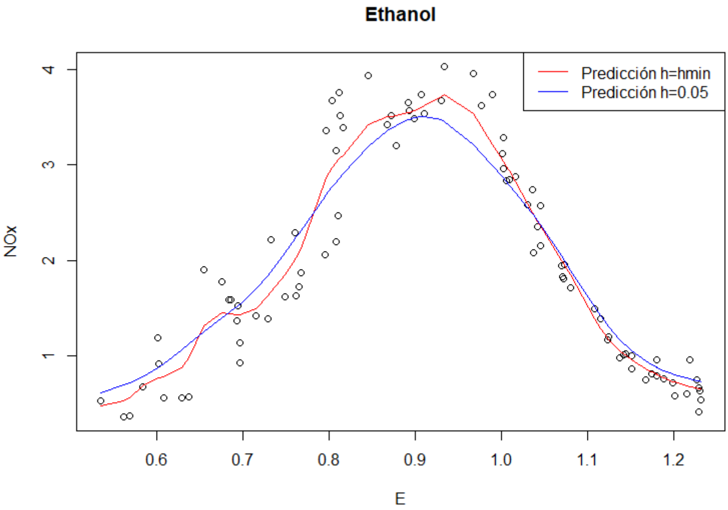
$$CV(h) = \frac{1}{n} \sum_{j=1}^{n} (Y_j - \hat{\mu}_{h,-j}(X_j))^2 w(X_j)$$

$$\hat{\mu}_{h,-j}(x) = n^{-1} \sum_{i \neq j} W_{hi}(x) Y_i$$

Proposición

El sesgo de $G(h)$ y $CV(h)$ como estimadores de $ASE(h)$ no depende de h.

Ancho de banda óptimo



Índice

1. Motivación y objetivos
2. Preliminares
3. Estimadores tipo núcleo
4. Elección óptima de los parámetros
   - Núcleo óptimo
   - Ancho de banda óptimo
5. Aplicaciones de la Regresión Local tipo Núcleo
   - Detección de Outliers
   - Estimación del ruido de una Serie Temporal
   - Análisis de series temporales en tiempo real
6. Conclusiones

Detección de Outliers

1. Se obtiene una muestra de datos no anómalos:
   entrenamiento $\{(x_i^e, y_i^e) : i = 1, ..., s\}$ y
   validación $\{(x_i^v, y_i^v) : i = 1, ..., q\}$.
2. Se obtiene una muestra de datos **test** $\{(x_i^t, y_i^t) : i = 1, ..., p\}$.
3. Se construye el estimador basado en los datos de entrenamiento

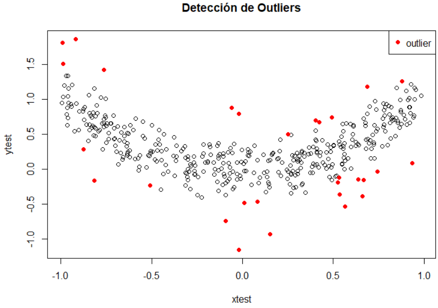
$$\hat{\mu}_{NW}(x) = \sum_{i=1}^{n} \frac{K_h\left(x - x_i^e\right) Y_i}{\sum_{i=1}^{n} K_h\left(x - x_i^e\right)}.$$

4. Se calcula el umbral de anomalía: umbral $= \max_i |\hat{\mu}_{NW}(x_i^v) - y_i^v|$.
5. Se consideran outliers los individuos (x_i^t, y_i^t) tales que

$$|\hat{\mu}_{NW}(x_i^t) - y_i^t| > \text{umbral}.$$

Detección de Outliers: grado de acierto

Se repite el método sobre diferentes muestras simuladas de una variables (X, Y) introduciendo anomalías aleatorias, para estimar la **sensibilidad** y la **especificidad**.



Sensibilidad	Especificidad
0.8836667	0.9361298

Estimación del ruido de una Serie Temporal

Para aplicar el modelo de regresión local a una serie temporal se consideran los valores de la misma como observaciones de la variable respuesta $\{y_t : t = 1, ..., n\}$ y el tiempo como variable predictora, siendo $\{x_t : t = 1, ..., n\}$ una cierta partición del intervalo observado. Se plantea así el modelo

$$y_t = \mu(x_t) + \epsilon_t \quad \text{ó} \quad y_t = \mu_t + \epsilon_t.$$

Estimación del ruido de una Serie Temporal

Para aplicar el modelo de regresión local a una serie temporal se consideran los valores de la misma como observaciones de la variable respuesta $\{y_t : t = 1, ..., n\}$ y el tiempo como variable predictora, siendo $\{x_t : t = 1, ..., n\}$ una cierta partición del intervalo observado. Se plantea así el modelo

$$y_t = \mu(x_t) + \epsilon_t \quad \text{ó} \quad y_t = \mu_t + \epsilon_t.$$

Se proponen dos métodos para la estimación del ruido:

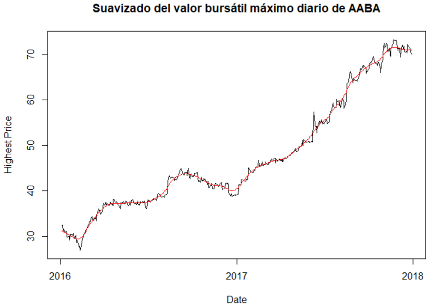
- Estimación del ruido a través del suavizado de la serie.
- Descomposición STL (*Seasonal-Trend Loess decomposition*).

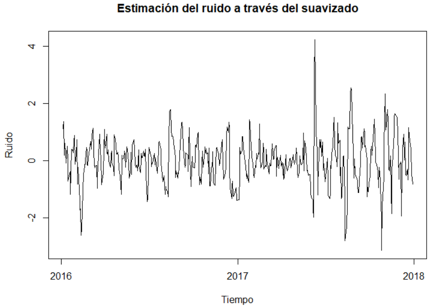
Estimación del ruido de una Serie Temporal: suavizado

- Se estima μ_t en cada instante observado:

$$\hat{\mu}_{NW}(x_t) = \sum_{i=1}^{n} \frac{K_h(x - x_t)\, y_t}{\sum_{i=1}^{n} K_h(x - x_t)}.$$

- Se estima el ruido en cada instante: $\hat{\epsilon}_t = \hat{\mu}_{NW}(x_t) - y_t$.





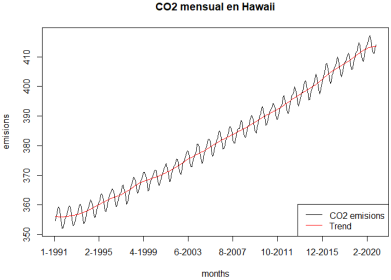
Estimación del ruido de una Serie Temporal: descomposición STL

El método STL (Cleveland et al.) se basa en la descomposición de μ_t en dos componentes: la tendencia y la estacionalidad.

$$Y_t = T_t + S_t + \epsilon_t.$$

Paso 1 Estimación de la tendencia:

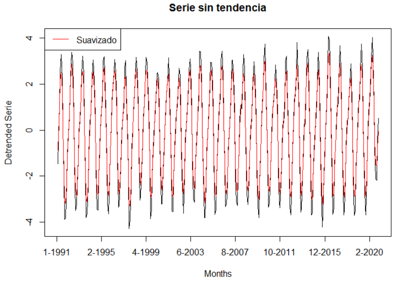
$$\hat{T}_t = \sum_{i=1}^{n} \frac{K_h\left(x_t - x_i\right) y_t}{\sum_{i=1}^{n} K_h\left(x_t - x_i\right)}$$



Estimación del ruido de una Serie Temporal: descomposición STL

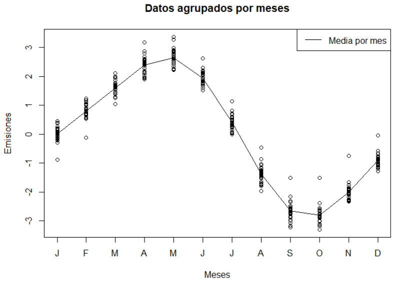
Paso 2 Eliminación de la tendencia, $D_t = y_t - \hat{T}_t \quad t = 1, ..., n$, y suavizado de la serie sin tendencia:

$$\hat{D}_t = \sum_{i=1}^{n} \frac{K_h\left(x_t - x_i\right) D_t}{\sum_{i=1}^{n} K_h\left(x_t - x_i\right)}.$$



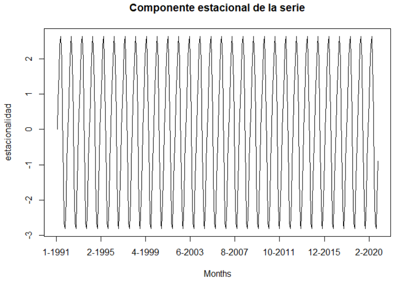
Estimación del ruido de una Serie Temporal: descomposición STL

Paso 3 Agrupación de datos según el ciclo y cálculo de su media.



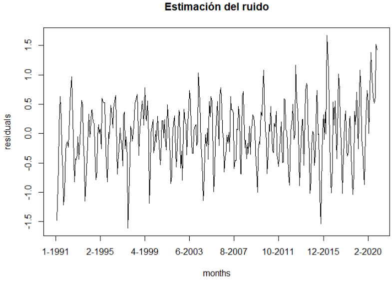
Estimación del ruido de una Serie Temporal: descomposición STL

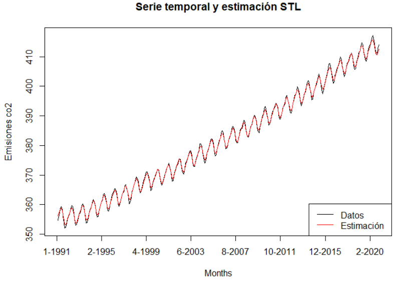
Paso 4 Estimación de la estacionalidad, $\hat{S}_t$ extendiendo las medias anteriores a todo el intervalo temporal.



Estimación del ruido de una Serie Temporal: descomposición STL

Paso 5 Estimación del ruido y de μ_t: $\hat{\epsilon}_t = y_t - \hat{T}_t - \hat{S}_t$, $\hat{\mu}_t = \hat{S}_t + \hat{T}_t$.





Análisis de series temporales en tiempo real

Se pueden modificar las ventanas de estimación y escoger $[x_t - h, x_t]$ y considerar así en cada instante de temporal únicamente los datos correspondientes a instantes anteriores. Se consigue de esta manera obtener componentes de interés (ruido, tendencia, etc.) de la serie temporal al mismo tiempo que se toman las medidas.

Ejemplo

Suavizado de series temporales a tiempo real:

$$\hat{y}_t = \sum_{i=1}^{t} \frac{K_h(x - x_i)\, y_t}{\sum_{i=1}^{t} K_h(x - x_i)}.$$

Índice

1 Motivación y objetivos

2 Preliminares

3 Estimadores tipo núcleo

4 Elección óptima de los parámetros
- Núcleo óptimo
- Ancho de banda óptimo

5 Aplicaciones de la Regresión Local tipo Núcleo
- Detección de Outliers
- Estimación del ruido de una Serie Temporal
- Análisis de series temporales en tiempo real

6 Conclusiones

Conclusiones

- Se ha abordado el problema de la regresión a través de los estimadores tipo núcleo.
- Se han estudiado las propiedades del estimador de Nadaraya-Watson como estimador de la función de regresión.
- Se han mostrado y justificado métodos que guían la elección del núcleo y el ancho de banda.
- Se han presentado ejemplos de problemas reales donde los métodos estudiados tienen aplicación.

Conclusiones

- Se ha abordado el problema de la regresión a través de los estimadores tipo núcleo.
- Se han estudiado las propiedades del estimador de Nadaraya-Watson como estimador de la función de regresión.
- Se han mostrado y justificado métodos que guían la elección del núcleo y el ancho de banda.
- Se han presentado ejemplos de problemas reales donde los métodos estudiados tienen aplicación.

Trabajos futuros

1. Extensión a variables multidimensionales.
2. Técnicas de regresión local más complejas: splines de suavizado, regresión local polinómica o estimación a través de series ortogonales.
3. Estudio de las derivadas de la función de regresión.

Referencias

[1] R. B. Cleveland, W. S. Cleveland, J. E. McRae e I. Terpenning. "STL: A Seasonal-Trend Decomposition Procedure Based on Loess". En: *Journal of Official Statistics* (1990).

[2] V. A. Epanechnikov. "Non-Parametric Estimation of a Multivariate Probability Density". En: *Theory of Probability & Its Applications* (1969).

[3] Wolfgang Härdle. *Applied Nonparametric Regression*. Cambridge University Press, 1992.

[4] G. Johnston. "Smooth Nonparametric Regression Analysis". Tesis doct. University of North Carolina at Chapel Hill, 1979.

[5] M. Köhler, A. Schindler y S. Sperlich. "A Review and Comparison of Bandwidth Selection Methods for Kernel Regression". En: *International Statistical Review* (2014).

[6] L. Wasserman. "Nonparametric Regression". En: *All of Nonparametric Statistics* (2006).