USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Análisis estadístico de las evidencias estadísticas del cambio climático

Alexander Suárez Soto

Curso 2023/2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Análisis estadístico de las evidencias estadísticas del cambio climático

Alexander Suárez Soto

Junio, 2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Agradecimientos

En primer lugar, me gustaría agradecer a mi tutor, Alberto Rodríguez Casal, todo el tiempo, esfuerzo y dedicación invertidos en este proyecto. Desde el comienzo se trató de un trabajo algo atípico, pero gracias a su paciencia y confianza hemos conseguido llevarlo a cabo. Muchas gracias, Alberto.

En segundo lugar, no puedo dejar de mencionar a mi familia, pareja, amigos y amigas, por su apoyo incondicional constante, tiempo invertido en leer el proyecto y sus críticas, siempre constructivas, que han contribuido a que este trabajo se acercara un poco más a la versión deseada.

Por último, me gustaría agradecer también a todos los profesores, profesoras y profesionales que me han acompañado a lo largo de mi etapa universitaria, ya que de una forma u otra, todos han aportado lo que ha estado en su mano para que llegara a este momento.

Trabajo propuesto

Área de Coñecemento: Área de Estadística e Investigación Operativa
Título: Análisis estadístico de las evidencias estadísticas del cambio climático
Breve descrición do contido
El estudio del cambio climático implica el análisis de la influencia que pueden tener determinadas variables, como puede ser la emisión de CO_2 en la evolución de la temperatura media. También aparece la posibilidad de contrastar si los datos sustentan el aumento de eventos climáticos extremos. En este trabajo se trata de analizar de forma empírica la consistencia de esas afirmaciones. Para eso se recogerán datos reales de variables que midan el estado del clima, así como del nivel de emisiones de gases que se sospecha tienen influencia en el clima. Después se analizarán mediante modelos estadísticos esos datos, extrayendo las conclusiones pertinentes.
Recomendacións
-
Outras observacións
-

Índice

Resumen

La regresión no paramétrica es una alternativa muy interesante para los modelos usuales de regresión lineal, dado que aporta una flexibilidad de la que el usuario se puede beneficiar enormemente en múltiples situaciones. Como todo, la regresión no paramétrica también tiene sus desventajas, y es que, como veremos, la flexibilidad de la que uno se beneficia, se termina pagando.

A lo largo del documento trabajaremos en la mencionada regresión no paramétrica, centrándonos especialmente en el estimador local lineal y en sus propiedades, así como en el efecto del parámetro ventana. Además, trataremos una herramienta tremendamente útil y conectada tanto a las propiedades del estimador local lineal como al efecto del parámetro ventana, llamada SiZer.

Para finalizar el trabajo, tomaremos unos datos de gran interés y actualidad, como son los correspondientes a variables climatológicas, y haciendo uso de todo lo visto a lo largo del estudio, trataremos de llevar a cabo un análisis sobre los mismos con el objetivo de apoyar la existencia del cambio climático, siempre teniendo presente que el estudio no deja de ser parte de un trabajo de fin de grado y un estudio riguroso del cambio climático excede por completo los objetivos del mismo.

Abstract

Nonparametric regression is a very interesting alternative to the usual linear regression models, since it provides a flexibility from which the user can benefit enormously in multiple situations. As with everything, nonparametric regression also has its disadvantages, and, as we shall see, the flexibility that one benefits from ends up paying for it.

Throughout the paper we will work on the aforementioned nonparametric regression, focusing especially on the local linear estimator and its properties, as well as on the effect of the bandwidth parameter. In addition, we will discuss a tremendously useful tool connected to both the properties of the local linear

estimator and the effect of the bandwidth parameter, called SiZer.

To conclude the work, we will take data of great interest and topicality, such as those corresponding to climatological variables, and making use of everything seen throughout the study, we will try to carry out an analysis on them with the aim of supporting the existence of climate change, always bearing in mind that the study is still part of a final degree project and a rigorous study of climate change completely exceeds the objectives of the same.

Introducción

Titulares alarmantes, pánico generalizado, perspectivas apocalípticas (Figura 1 [1], obtenida de [5]). Negacionismo, escepticismo, dudas. En resumidas cuentas, una sociedad dividida en torno a un posible problema de gran actualidad e impacto. El cambio climático. Se trata de un tema que ha llegado para quedarse y está en boca de todos, pero de la boca de cada uno sale algo diferente. ¿Se trata de un problema de urgente abordaje? ¿existe peligro de calentamiento global excesivo o, por el contrario, estamos más próximos a una glaciación (Figura 2 [2], obtenida de [14])? ¿se trata de un problema originado por nosotros, los seres humanos? ¿se trata siquiera de un problema? Cada cuestión es más ambiciosa que la anterior, pero para comenzar lo correcto sería hacernos la siguiente pregunta: ¿qué es el cambio climático?

Figura 1: Emisiones efecto invernadero.

Figura 2: Glaciación.

Normalmente el cambio climático se asocia directamente con una variación en la temperatura. Más concretamente se suele pensar en una subida progresiva de la misma pero, ¿se queda ahí la cuestión? La respuesta es no. Cuando se habla de cambio climático, en realidad se está hablando de un cambio que no tiene por qué estar enfocado únicamente en la temperatura y que además afecta a todas las componentes

[1] Licencia: https://creativecommons.org/licenses/by/3.0/

[2] Licencia: https://creativecommons.org/licenses/by-nc-sa/2.0/

del conocido como *Sistema Climático*, que es un sistema compuesto por diferentes subsistemas, a saber: la atmósfera, que es una capa homogénea de gases concentrada alrededor de un planeta o astro celeste y que se mantiene en su posición por acción de la gravedad, la hidrosfera, que puede entenderse como la capa de agua que rodea la tierra (donde, por supuesto, predomina la componente oceánica), la superficie sólida, la biosfera, que es la suma de todos los ecosistemas y la criosfera, que es el término que describe las partes de la tierra cubieras por hielo. Pero además, tenemos que asegurarnos de que los cambios sean a escalas prolongadas de tiempo, para poder empezar a hablar del cambio climático. Es decir, si durante un periodo lo suficientemente grande de tiempo (de décadas por lo menos) y a una escala global (continental, por ejemplo) se producen cambios que sean significativos estadísticamente hablando en los valores de ciertos parámetros como la media, mediana y/o varianza de la distribución de frecuencias que siga una cierta variable climática, podríamos empezar a sacar a relucir el término cambio climático, con efectos sobre todo el sistema climático.

En el párrafo anterior hacemos referencia a que el cambio observado ha de ser por lo menos decadal. En este punto, conviene diferenciar los términos clima y tiempo meteorológico, ya que tienden a mezclarse, y es que el tiempo meteorológico se refiere a las condiciones climáticas dadas en una región concreta durante un periodo breve de tiempo, mientras que el clima abarca una superficie más extensa, así como un intervalo temporal mucho mayor. Es importante marcar bien la diferencia entre ambas, ya que el tiempo meteorológico está en constante cambio y no es de especial preocupación, mientras que si el clima cambia, podría suponer un problema de gran calibre.

Hablemos ahora del *Efecto invernadero*. No se trata de un término nuevo para nadie, ya que todos hemos tenido ocasión de oírlo alguna vez en nuestra vida, pero ¿qué es en realidad el efecto invernadero y por qué nos interesa saberlo? Empecemos por el principio. Para visualizar mejor la atmósfera, podríamos verla como una capa protectora compuesta de cinco capas intermedias, que filtra los rayos y radiaciones procedentes del espacio (Figura 3[3], obtenida de [9]).

[3] Licencias:
https://creativecommons.org/licenses/by-sa/4.0/
https://creativecommons.org/licenses/by-sa/3.0/
https://creativecommons.org/licenses/by-sa/2.5/
https://creativecommons.org/licenses/by-sa/2.0/
https://creativecommons.org/licenses/by-sa/1.0/

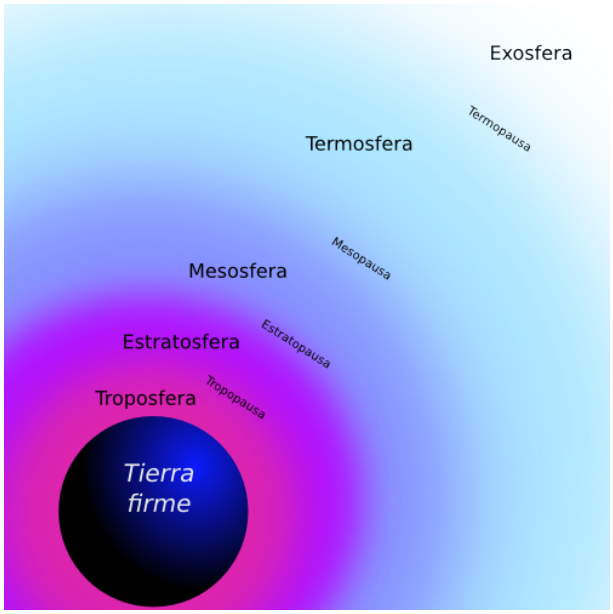

Figura 3: Esquema de las capas de la atmósfera.

Para completar la explicación, podríamos tratar de imaginar que dicha capa protectora contiene una serie de gases, y estos gases son capaces de absorber la emisión terrestre, es decir, la proveniente de la superficie. Una vez tengamos una imagen mental de lo que sería la atmósfera (o lo que nos interesa ahora mismo de ella), pensemos en qué ocurriría cuando tras la absorción ya mencionada, hubiese una reemisión hacia la tierra. Se llega a una posible conclusión: la temperatura aumenta. Pues bien, a este suceso se le llama el efecto invernadero. Entonces, ¿qué es exactamente lo que favorece el aumento de este fenómeno? Simplificando enormemente lo que podemos ver en *Causas del cambio climático* ([4]) para centrarnos en lo que será mas relevante a lo largo del trabajo, la respuesta es: el aumento de la concentración atmosférica de determinados gases, denominados precisamente de efecto invernadero, que son, entre otros:

- Dióxido de carbono, CO_2.
- Metano, CH_4.

Esta es la motivación para trabajar la concentración atmosférica de dos de esos dos gases más adelante.

En otro orden de cosas, el cambio climático, a su vez tiene una gran influencia en el subsistema del sistema climático correspondiente al océano, al que teóricamente afecta de forma perjudicial. Estaríamos hablando de un aumento de la temperatura oceánica, por no hablar de la acidificación del mismo. Por otra parte, pero siguiendo en esta línea, el cambio climático también afecta a la disminución de la capa de hielo, o más técnicamente, afecta al subsistema del sistema climático correspondiente a la criosfera. Pero es que el hecho de que la capa de hielo terrestre se derrita, a su vez repercute en el nivel del mar,

haciendo que aumente. En resumen, un análisis global de los efectos y/o causas del cambio climático es muy complejo, y queda fuera de los objetivos de un Trabajo de Fin de Grado. Nosotros nos centraremos en una parte pequeña y necesariamente restringida de toda esta problemática.

En este punto de la introducción, ya podemos empezar a notar la importancia de dar la correcta definición de los términos más importantes, como la de sistema climático, porque podemos empezar a atisbar la interrelación que hay entre ellos. A lo largo de este estudio, vamos a trabajar con los datos correspondientes a ciertas variables climatológicas para tratar de ver si mediante los resultados que obtengamos se puede apoyar la hipótesis de que efectivamente existe un cambio en el clima o por el contrario nos veríamos obligados a buscar las evidencias siguiendo otros caminos. A su vez, el lector podrá observar que se trabaja con ciertos datos correspondientes a medias (temperatura media, por ejemplo). Esto se debe a que no sería significativo ni tan siquiera alarmante que hubiese algún evento meteorológico extremo, ya que estos son esperables dentro de un comporamiento normal del clima. Estos eventos extremos reciben mucha atención mediática, pero sólo sería relevante un cambio en su distribución o probabilidad de ocurrencia.

Capítulo 1

Base de datos y metodología estadística

En este primer capítulo presentaremos la base de datos que vamos a utilizar para llevar a cabo nuestro estudio. Nuestras variables representarán ciertos factores que en la introducción ya mencionamos como posibles causas o consecuencias del cambio climático. A su vez, a la hora de presentar cada conjunto de datos, propocionaremos la procedencia del mismo y comentaremos qué haremos con ellos y cuál será la función que cumplirá cada conjunto en nuestro estudio. Por otra parte, introduciremos las técnicas estadísticas que se llevarán a cabo, ilustrando la estructura del trabajo. Pasemos entonces con la presentación de los datos.

1.1. Base de datos

En esta sección nos encargaremos de describir los datos, mencionando además cómo se obtuvieron y finalmente explicar para qué los usaremos. El primer conjunto que presentaremos será el de los datos correspondientes a la temperatura. Estos no son exactamente datos correspondientes a temperaturas, sino índices de la misma. Es decir, los datos que tenemos son de los índices de las temperaturas medias anuales desde el año 1880 hasta 2022, medidos en ºC. Esto significa que se tomó una temperatura que desconocemos como temperatura de referencia, y a partir de ahí solamente se midió cuánto aumentó o disminuyó la temperatura media anualmente con respecto a esta temperatura de referencia. Veamos cuáles son los datos en cuestión:

Año	Índice Temperatura
1880	-0,17
1881	-0,09
1882	-0,11
1883	-0,17
1884	-0,28
1885	-0,33
1886	-0,32
1887	-0,36
1888	-0,18
1889	-0,11
1890	-0,35
1891	-0,23
1892	-0,27
1893	-0,31
1894	-0,31
1895	-0,23
1896	-0,11
1897	-0,11
1898	-0,27
1899	-0,18
1900	-0,08
1901	-0,15
1902	-0,28
1903	-0,37
1904	-0,47
1905	-0,26
1906	-0,22
1907	-0,39
1908	-0,43
1909	-0,49
1910	-0,44
1911	-0,44
1912	-0,36
1913	-0,35
1914	-0,15
1915	-0,14
1916	-0,36
1917	-0,46
1918	-0,30
1919	-0,28
1920	-0,27
1921	-0,19
1922	-0,28
1923	-0,26
1924	-0,27
1925	-0,22
1926	-0,11
1927	-0,22
1928	-0,20
1929	-0,36
1930	-0,16
1931	-0,09
1932	-0,16
1933	-0,28
1934	-0,12
1935	-0,20
1936	-0,14
1937	-0,03
1938	0,00
1939	-0,02
1940	0,13
1941	0,19
1942	0,07
1943	0,09
1944	0,21
1945	0,10
1946	-0,07
1947	-0,02
1948	-0,10
1949	-0,11
1950	-0,17
1951	-0,07
1952	0,01
1953	0,08
1954	-0,13
1955	-0,14
1956	-0,19
1957	0,05
1958	0,06
1959	0,03
1960	-0,02
1961	0,06
1962	0,03
1963	0,05
1964	-0,20
1965	-0,11
1966	-0,06
1967	-0,02
1968	-0,08
1969	0,05
1970	0,03
1971	-0,08
1972	0,01
1973	0,16
1974	-0,07
1975	-0,01
1976	-0,10
1977	0,18
1978	0,07
1979	0,16
1980	0,26
1981	0,32
1982	0,14
1983	0,31
1984	0,16
1985	0,12
1986	0,18
1987	0,32
1988	0,39
1989	0,27
1990	0,45
1991	0,41
1992	0,22
1993	0,23
1994	0,31
1995	0,45
1996	0,33
1997	0,46
1998	0,61
1999	0,38
2000	0,39
2001	0,54
2002	0,63
2003	0,62
2004	0,53
2005	0,68
2006	0,64
2007	0,66
2008	0,54
2009	0,66
2010	0,72
2011	0,61
2012	0,65
2013	0,68
2014	0,75
2015	0,90
2016	1,02
2017	0,92
2018	0,85
2019	0,98
2020	1,02
2021	0,85
2022	0,89

Tabla 1.1: Índice de las temperaturas medias anuales (en ºC) desde el año 1880 hasta 2022.

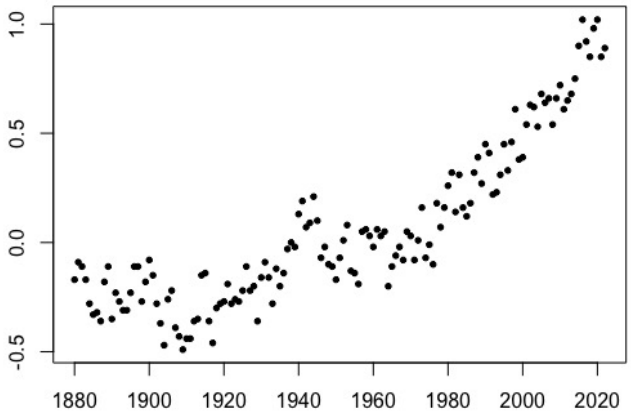

Figura 1.1: Gráfica de los índices de la temperatura media anual (ºC) frente al año en que fueron tomados, desde el año 1880 hasta 2022.

En la Tabla 1.1 podemos ver los datos correspondientes a los índices de la temperatura media anual, los cuales están representados gráficamente en la Figura 1.1. Presentaremos a continuación los datos de otra variable mencionada en la introducción: la concentración de CO_2 atmosférico, junto con una representación gráfica de dicha concentración.

Año	**Concentración** CO_2
1958	315,70
1959	315,58
1960	316,43
1961	316,89
1962	317,94
1963	318,74
1964	319,57
1965	319,44
1966	320,62
1967	322,33
1968	322,57
1969	324,00
1970	325,06
1971	326,17
1972	326,77
1973	328,55
1974	329,35
1975	330,73
1976	331,56
1977	332,68
1978	334,94
1979	336,14
1980	337,90
1981	339,29
1982	340,93
1983	341,57
1984	344,21
1985	345,48
1986	346,78
1987	348,73
1988	350,51
1989	353,07
1990	353,86
1991	354,93
1992	356,34
1993	357,10
1994	358,36
1995	360,04
1996	362,20
1997	363,24
1998	365,39
1999	368,35
2000	369,45
2001	370,76
2002	372,70
2003	375,07
2004	377,17
2005	378,63
2006	381,58
2007	383,10
2008	385,78
2009	387,17
2010	388,91
2011	391,50
2012	393,31
2013	395,78
2014	398,04
2015	400,18
2016	402,73
2017	406,36
2018	408,15
2019	411,03
2020	413,59
2021	415,49
2022	418,13
2023	419,48

Tabla 1.2: Concentraciones atmosféricas de CO_2 (en ppm) desde el año 1958 hasta 2023.

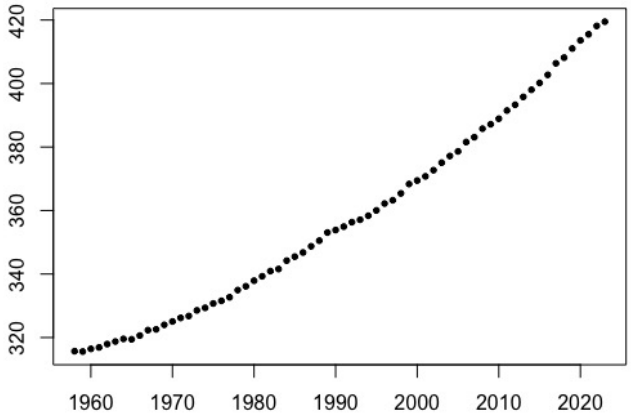

Figura 1.2: Gráfica de las concentraciones atmosféricas de CO_2, medidas en partes por millón (ppm), por año.

En la Tabla 1.2 podemos encontrar las medidas de las concentraciones atmosféricas de CO_2 medidas en partes por millón desde el año 1958 hasta el 2023. Además, en la Figura 1.2, se puede ver una representación gráfica del contenido de la susodicha tabla. Mostremos ahora los datos correspondientes a dos variables adicionales, con sus respectivas representaciones gráficas. Primero, tratemos los datos del CH_4.

Año	**Concentración** CH_4
1984	1644,85
1985	1657,29
1986	1670,09
1987	1682,70
1988	1693,16
1989	1704,53
1990	1714,43
1991	1724,82
1992	1735,47

1993	1736,50
1994	1742,07
1995	1748,88
1996	1751,28
1997	1754,53
1998	1765,54
1999	1772,34
2000	1773,33
2001	1771,22
2002	1772,66

2003	1777,33
2004	1777,05
2005	1774,16
2006	1774,95
2007	1781,37
2008	1787,01
2009	1793,53
2010	1798,93
2011	1803,14
2012	1808,12

2013	1813,41
2014	1822,57
2015	1834,26
2016	1843,12
2017	1849,58
2018	1857,33
2019	1866,58
2020	1878,93
2021	1895,28
2022	1911,82

Tabla 1.3: Concentraciones atmosféricas de CH_4 (en ppb) desde el año 1984 hasta 2022.

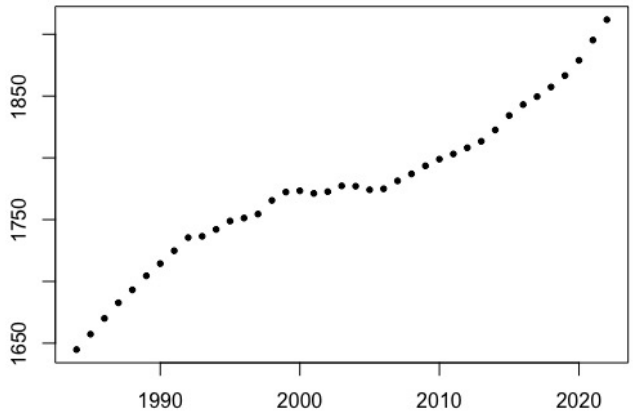

Figura 1.3: Gráfica de las concentraciones atmosféricas de CH_4, medidas en partes por billón (ppb), por año, desde el año 1984 hasta 2022.

En la Tabla 1.3 tenemos las concentraciones atmosféricas de CH_4 medidas en partes por billón (ppb) y en la Figura 1.3 se puede encontrar la representación gráfica de los datos de la susodicha tabla. Presentaremos a continuación el último conjunto de datos que trabajaremos: el correspondiente a las mediciones del crecimiento/decrecimiento del nivel del mar.

Año	Nivel del mar
1880	0,00
1881	0,56
1882	-1,12
1883	-0,59
1884	1,50
1885	1,35
1886	1,11
1887	0,55
1888	0,76
1889	0,92
1890	1,12
1891	0,95
1892	1,27
1893	1,74
1894	0,77
1895	1,95
1896	1,19
1897	1,71
1898	2,65
1899	3,40
1900	2,86
1901	2,82
1902	3,28
1903	4,08
1904	3,05
1905	2,50
1906	3,18
1907	3,04
1908	2,79
1909	3,24
1910	3,23
1911	4,06
1912	3,75
1913	3,93
1914	4,56
1915	5,35
1916	5,16
1917	4,71
1918	4,55
1919	4,71
1920	4,84
1921	5,05
1922	4,96
1923	5,08
1924	4,35
1925	4,55
1926	5,20
1927	5,09
1928	4,70
1929	4,84
1930	5,24
1931	5,20
1932	5,77
1933	6,20
1934	5,66
1935	6,22
1936	5,83
1937	6,40
1938	6,66
1939	7,18
1940	6,65
1941	7,87
1942	7,87
1943	7,87
1944	7,23
1945	7,51
1946	8,26
1947	8,57
1948	9,05
1949	8,92
1950	9,14
1951	10,09
1952	9,83
1953	10,27
1954	9,98
1955	10,07
1956	9,56
1957	10,90
1958	11,04
1959	11,07
1960	11,44
1961	12,06
1962	11,54
1963	11,38
1964	10,59
1965	11,71
1966	11,17
1967	11,31
1968	11,39
1969	12,07
1970	11,88
1971	12,40
1972	13,31
1973	12,71
1974	13,90
1975	13,74
1976	13,64
1977	13,47
1978	14,11
1979	13,62
1980	14,22
1981	15,46
1982	14,88
1983	15,72
1984	15,63
1985	14,60
1986	14,66
1987	14,72
1988	15,19
1989	15,64
1990	15,87
1991	16,12
1992	16,19
1993	16,01
1994	16,53
1995	16,82
1996	17,23
1997	17,93
1998	16,94
1999	17,79
2000	17,92
2001	18,47
2002	18,71
2003	19,63
2004	19,59
2005	19,60
2006	20,03
2007	20,22
2008	21,09
2009	21,67
2010	22,44
2011	22,60
2012	23,48
2013	22,64
2014	21,79
2015	22,67
2016	22,98
2017	23,14
2018	23,46
2019	24,08
2020	24,36
2021	24,88

Tabla 1.4: Crecimiento y decrecimiento anual del nivel del mar (en cm) desde el año 1880 hasta 2021.

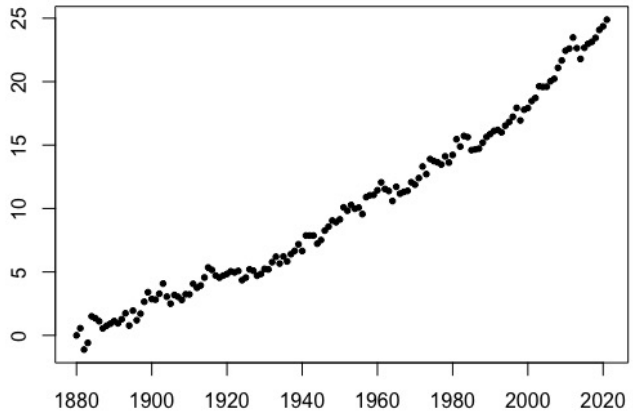

Figura 1.4: Gráfica del crecimiento o decrecimiento relativo anual del nivel del mar frente al año en que fueron tomados, medidos en cm.

En la Tabla 1.4 podemos encontrar los datos correspondientes al crecimiento o decrecimiento anual del nivel del mar (medido en cm) con respecto a una medida de referencia que desconocemos, que usaremos para nuestro estudio. Estos datos podemos encontrarlos representados en la Figura 1.4.

Una vez presentados todos los datos que utilizaremos para nuestro estudio, consideramos importante mencionar de dónde los obtuvimos. El proceso de búsqueda fue largo y requirió la ayuda de ciertos expertos que pasaré a mencionar después. Lo primero fue escribir a la *NOAA (National Oceanic and Atmospheric Administration)*, concretamente a un miembro de la mencionada agencia llamado *Theo Stein (theo.stein@noaa.gov)*, para ver si era posible obtener los datos climatológicos que ellos poseían. Me contestó de forma muy favorable, enviándome dos direcciones donde podría encontrar los datos que pedía (se pueden consultar en [11] y [12]). Por otra parte, me dijo que él no tenía los datos de las temperaturas, pero muy amablemente me proporcionó otra dirección de correo electrónico a la que podía escribir para pedírselo. Dicha dirección de correo correspondía a un científico llamado *Scott Stephens (scott.stephens@noaa.gov)*, de quien una vez más obtuve una respuesta más que amable y satisfactoria, ya que me contestó al correo con una dirección en la que estaban los datos que yo le pedía (se puede ver en [13]). Aun así, traté de recopilar más datos, para poder compararlos y contrastarlos con los que ya tenía. Esta búsqueda me llevó a muchas páginas oficiales y direcciones de las que no pude obtener nada aprovechable, hasta que encontré lo que buscaba en una sección de la página oficial de la NASA (se puede encontrar en [10]). En ella pude comprobar que había datos muy útiles y sencillos de interpretar, pero aún así continué la búsqueda para ver si era capaz de encontrar, ya no datos diferentes, sino más cantidad, para que el estudio fuese lo más completo posible. Lo conseguí cuando di con la *EPA (United States*

Environmental Protection Agency), donde encontré gran cantidad de datos que me resultaron realmente útiles (disponibles en [6]).

En esta sección, además de presentar los datos y decir de dónde los obtuvimos, explicaremos la manera en la que los conectaremos a lo largo del estudio. Primero, trataremos de relacionar el índice de la temperatura media anual con la concentración atmosférica de CO_2, como no podía ser de otra forma, ya que es una de las relaciones que más se estudia cuando se trata de calentamiento global. Nuestra variable explicativa será la concentración atmosférica de CO_2, medida en partes por millón, que tratará de explicar la variable respuesta, que será el índice de la temperatura media anual, medida en ºC. A su vez, trataremos de probar la existencia de una relación entre el índice de la temperatura media anual y la concentración atmosférica de CH_4. Como hemos comentado en la introducción, el CH_4 es un gas de efecto invernadero, luego por el mismo motivo que el que teníamos para el CO_2, se trata de un estudio de gran interés. En este caso, la variable explicativa que tratará de explicar nuestra variable respuesta será la concentración atmosférica de CH_4, medida en partes por billón, mientras que la segunda será el índice de la temperatura media anual, medida en ºC. Por último, trataremos una última relación entre las variables. Para esto, realizaremos una hipótesis algo más compleja que las dos anteriores, y es que trataremos de relacionar el índice de la temperatura media anual, medida en ºC, con el aumento o disminución del nivel del mar, medido en centímetros. En este caso, trataremos de explicar el aumento del nivel del mar mediante el aumento de la temperatura media anual. La motivación detrás de este análisis es que se puede pensar que el aumento de la temperatura media anual provoca que la criosfera se derrita, causando de este modo un aumento del nivel del mar. A continuación, en la sección siguiente, veremos cómo estudiaremos las mencionadas relaciones entre las variables.

Observación 1.1. Destacamos también que el número de datos de cada variable es distinto (salvo temperatura y nivel del mar). Es por esto que cuando comparemos dos variables, tendremos que asegurarnos de emparejar correctamente los años de la muestra, tomando únicamente los que ambas posean.

1.2. Procedimiento en el trabajo y análisis de datos

En esta sección trataremos el esquema que seguiremos a lo largo del trabajo, así como las técnicas estadísticas que utilizaremos para analizar e interpretar los datos vistos en la sección anterior. Este trabajo constará de tres capítulos, contando el actual. En este primero, el objetivo principal es presentar los datos y técnicas que usaremos en los dos restantes, así como comenzar a ilustrar los datos, para que el lector se vaya familiarizando tanto con el tema como con los datos. A continuación, en el Capítulo 2, nos encontraremos el bloque más teórico del trabajo. Se presentará la regresión no paramétrica, y poco a poco se irá construyendo el trasfondo necesario hasta llegar a presentar el estimador local lineal, que será el núcleo del capítulo dos. Se calculará su expresión analítica y también su expresión matricial. Analizaremos sus propiedades estadísticas básicas como son su sesgo y varianza. Después de esto, en

el mismo capítulo hablaremos del parámetro ventana h, crucial a la hora de determinar la suavidad de nuestra estimación, y proporcionaremos dos formas de encontrar un valor óptimo para el mismo, una de ellas más teórica y la otra más práctica. Finalizaremos el capítulo hablando del SiZer, explicando lo que es, sus características y finalmente ilustrando la teoría con dos ejemplos.

Continuamos con el Capítulo 3. En este podremos encontrar el análisis de los datos que presentamos en la sección anterior, mediante las técnicas que encontraremos en el Capítulo 2. Dividiremos el tercer capítulo en tres secciones, y cada una corresponderá al estudio de uno de los emparejamientos ya mencionados. La primera sección tratará de analizar el índice de la temperatura media anual (medida en ºC) como variable respuesta y la concentración atmosférica de CO_2 (medida en ppm) como variable explicativa. Para este análisis representaremos los datos conjuntos, y estimaremos mediante el estimador local lineal la curva correspondiente, para cuatro valores distintos del mencionado parámetro ventana h. Analizaremos las curvas obtenidas y además se realizará un contraste de linealidad y dependiendo del resultado de dicho contraste, el estudio se enfocará de una forma o de otra totalmente diferente. La segunda sección del capítulo irá sobre el estudio de las variables concentración atmosférica de CH_4 (medida en ppb) como variable explicativa y el índice de la temperatura media anual (medida en ºC) como variable respuesta. En esta sección representaremos los datos, aplicaremos una vez más un test de linealidad y realizaremos las distintas estimaciones para los valores de h, interpretando a continuación los resultados. Después se realizará un pequeño estudio de la variabilidad del modelo. Se completará el análisis de este conjunto de datos mediante la representación del SiZer correspondiente, que nos permitira realizar inferencia sobre el crecimiento y/o decrecimiento de las distintas curvas estimadas en el inicio de la sección. Finalmente, concluiremos el capítulo con la tercera sección, que tratará de analizar el aumento o disminución del nivel del mar (medida en cm.) como variable respuesta y el índice de la temperatura media anual (medida en ºC) como variable explicativa. En este estudio, después de representar los datos y realizar un contraste de linealidad, proporcionando su gráfica asociada, se llevarán a cabo las estimaciones para distintos valores de h, acompañándolas de la correspondiente interpretación de los resultados y después de la cual se procederá con un análisis de la variabilidad. Además, finalizaremos el estudio de la misma forma que en la sección previa, estudiando el SiZer correspondiente a estos datos, ya que resulta realmente útil para poder extraer conclusiones acerca de las características que aparecen en nuestras estimaciones, tal y como veremos cuando hablemos del mismo en el Capítulo 2. Con esto podemos concluir el presente capítulo, dejando paso al siguiente.

Capítulo 2

El estimador local lineal

Damos inicio al segundo capítulo del trabajo explicando brevemente el por qué y qué haremos a lo largo del mismo. Como bien sabemos, cuando en el Capítulo 3 llevemos a cabo el estudio relativo a los datos presentados en el Capítulo 1, aplicaremos la regresión no paramétrica (dado que a priori no podemos encasillarlos en ningún modelo predeterminado). Por este motivo, comenzaremos explicando en qué consiste este tipo de regresión. Siguiendo esta explicación paso por paso, finalmente necesitaremos mencionar el estimador que utilizaremos, que se trata del estimador local lineal. Trataremos sus expresiones y hablaremos de sus propiedades estadísticas, como el sesgo y la varianza. Esto último nos propiciará ahondar más en el parámetro ventana y los efectos del mismo, así como distintas formas de llegar un valor óptimo para este. Cerraremos el capítulo estudiando lo que es un SiZer, sus propiedades y utilidades. Esta herramienta permite localizar zonas de crecimiento y/o decrecimiento significativo en el modelo de regresión. En el problema que analizamos en este TFG, establecer esta relación creciente entre variable explicativa, X, y dependiente Y, es especialmente relevante.

Por otra parte, aprovechamos esta breve introducción para mencionar las fuentes principales de las que obtuvimos la inmensa mayoría de la información presente en el capítulo. Para el desarrollo del mismo, se tuvieron muy en cuenta las siguientes referencias: Capítulo 1 de *Introduction to Nonparametric Estimation* ([16]), Capítulo 5 de *All of Nonparametric Statistics* ([18]), Capítulo 5 de *Kernel Smoothing* ([17]), Capítulos 3, 4 y 5 de *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations* ([1]), Capítulo 3 de *Generalized Additive Models* ([8]) y *SiZer for Exploration of Structures in Curves* ([3]).

2.1. Definición y propiedades

En ocasiones, nos interesa medir la relación que guardan dos variables, llamémoslas X e Y. Mediante un modelo de regresión somos capaces de obtener en qué medida la variable Y, que recibe el nombre de variable dependiente o variable respuesta viene explicada a partir de la X, denominada variable independiente o explicativa. Entendemos entonces que podríamos verlo como una posible dependencia de la variable Y sobre la X, y con esta motivación introducimos la **función** m que modela esa dependencia, a la cual no le vamos a pedir nada adicional por ahora, dado que nuestra única pretensión por el momento es la de presentarla. El motivo principal es que mediante esta función podemos manifestar la potencial dependencia de Y sobre X de la forma:

$$Y = m(X).$$

Ilustremos todo lo mencionado con un pequeño ejemplo sobre el que volveremos con más detalle más adelante, en el Capítulo 3. Podríamos considerar como Y **el índice de temperatura media anual**, que es, tal y como hemos adelantado en el Capítulo 1, una serie de datos que muestran cuánto ha variado la temperatura media anualmente con respecto a un año de referencia. Por otra parte, podríamos tomar como nuestra variable X **la concentración atmosférica de** CO_2. En este ejemplo, nos interesa determinar si el aumento de la concentración del gas de efecto invernadero CO_2 en la atmósfera presenta relación con el aumento de la temperatura media anual. Los resultados obtenidos de esta regresión podrían llegar a ser muy beneficiosos, ya que si se llega a que, efectivamente, la variable Y se ve explicada por la variable X, sería un indicio de que tendríamos que comenzar a vigilar la cantidad de CO_2 que se emite a la atmósfera. La naturaleza de esta relación no está del todo clara a priori y por tanto la forma de m. Aunque se espera que el CO_2 tenga un efecto invernadero, este efecto puede notarse, por ejemplo, solo a partir de ciertas concentraciones atmosféricas.

El primer paso para llevar a cabo nuestra regresión y ya dejando atrás el pequeño ejemplo ilustrativo que hemos planteado, sería hacernos con un conjunto finito, fiable y significativo de datos y estimar nuestra función m a partir del mismo. De este modo tendríamos una muestra de datos del vector bidimensional (X, Y). A su vez, sobre el modelo contemplado anteriormente, habría que añadir un término de error que hiciese, por así decirlo, las veces de representante de la variabilidad natural de cada medición Y_i, y por tanto, de la precisión o calidad del modelo funcional m. Así, tendríamos por observaciones:

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, ...n,$$

siendo n el número de datos. En el modelo anterior estamos expresando el valor de Y, Y_i, como aleatorio, y el de X, x_i, como fijo, ya que el objetivo del estudio es modelar la variabilidad de la variable respuesta y su dependencia de los datos $x_1, ..., x_n$ observados. A nivel poblacional podríamos considerar el siguiente modelo:

$$Y = m(X) + \varepsilon,$$

donde ε representa esa variabilidad natural de Y no explicada por X. Pero ¿qué es lo que representa la función m exactamente? De forma precisa, esta función representa la mejor aproximación de Y como función de X, en términos de error cuadrático, lo que matemáticamente se representa como:

$$m(x) = \mathbb{E}[Y|X = x].$$

Esto formalmente podría ponerse como:

$$\mathbb{E}[Y - m(X)]^2 = \min_g \mathbb{E}[Y - g(X)]^2.$$

Tal y como hemos indicado en la introducción, la idea del estudio es llevar a cabo una regresión no paramétrica o flexible. El motivo es que vistos los datos con los que contamos, a priori no podríamos encasillarlos en ninguna distribución conocida para, de este modo, poder optar a un enfoque paramétrico. Las ventajas del enfoque no paramétrico son que no tenemos por qué imponer ninguna condicion previa a nuestra función m, lo cual evidentemente le otorga una mayor flexibilidad y capacidad de adaptación a los datos. Por supuesto, existe un precio a pagar por dicha flexibilidad, en términos de tasa de error en la estimación.

Para nuestro estudio, se ha escogido utilizar un modelo polinómico local, en particular de grado uno. Es por esto que introducimos ahora lo que es el **Modelo local lineal**. La idea del mismo es muy sencilla, y su nombre es increíblemente descriptivo, ya que consiste en tratar de buscar una recta que ajuste bien localmente, en un entorno de x, para cada punto x. Es decir, solamente nos van a preocupar los puntos que estén en un entorno del mismo. La siguiente pregunta que uno podría hacerse es, cuán local es el método, cuán grande ha de ser el entorno que tomemos para x. Con el fin de responder a esto, sacamos a colación el parámetro ventana. El parámetro h, conocido como ventana, es un parámetro de escala que cuenta con una gran influencia en la estimación resultante, en particular en el nivel de influencia de los datos que están en el entorno de x, y determina en cierta medida el tamaño de dicho entorno. La idea de introducir el parámetro ventana ahora es que es crucial a la hora de definir nuestro modelo local lineal, ya que supondremos que éste sea aproximadamente válido únicamente en el intervalo $(x - h, x + h)$. Por otra parte, no podemos perder de vista el carácter lineal. Así supondremos que el modelo tendrá, aproximadamente la siguiente forma en el entorno $(x - h, x + h)$:

$$Y_i = a(x) + b(x)(x_i - x) + \varepsilon_i, \quad x_i \in (x - h, x + h).$$

Para definir un poco mejor el carácter local será útil la siguiente definición.

Definición 2.1 (Kernel)**.** Un kernel es una función de variable real $K\colon \mathbb{R} \longrightarrow \mathbb{R}$ que cumple las siguientes propiedades:

- No negativa: $K(x) \geq 0, \quad \forall x.$

- Soporte en $[-1,1]$: $K(x) = 0, \quad x \notin [-1,1]$.
- Simétrica: $K(x) = K(-x)$.
- $\int_{-1}^{1} K(x)dx = 1$.
- $\sigma^2 = \int_{-1}^{1} x^2 K(x)dx \in \mathbb{R}^+$, donde σ^2 representa su varianza.

Observación 2.2. Estamos suponiendo que K es una densidad unimodal simétrica alrededor del cero con soporte en el intervalo $[-1,1]$. Como consecuencia de su simetría, podemos añadir que $\int_{-1}^{1} xK(x)dx = 0$.

La hipótesis dada en relación al soporte es meramente de tipo técnico, para simplificar ciertos razonamientos que siguen, pero que se puede reemplazar por hipótesis sobre la existencia y finitud de ciertos momentos $\mu_r(K)$, donde $\mu_r(K) = \int z^r K(z)dz$. Se trata de controlar que K no puede tomar con probabilidad demasiado alta valores grandes. Comentar que, cuando presentemos los distintos kernels, el lector podrá observar que el **kernel Gaussiano** no se adapta a algunas de las características dadas en la definición, y es por esto que lo consideraremos como una **excepción** a la misma. En este caso se puede incluir, aunque no tenga soporte compacto, ya que tiene momentos finitos de cualquier orden r. Existen múltiples tipos de kernel. Entre los más conocidos se encuentran los que definiremos a continuación, que además pueden encontrarse representados gráficamente en la Figura 2.1:

Ejemplo 2.3. Tipos de kernel.

- "**Kernel Rectangular**"

  $K(x) = \frac{1}{2}\, I_{[-1,1]}(x), \quad x \in \mathbb{R}$, donde

  $$I_{[-1,1]}(x) = \begin{cases} 1, & \text{si} \quad |x| \leq 1. \\ 0, & \text{si} \quad |x| > 1. \end{cases}$$

- "**Kernel Gaussiano**"

  $K(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}$.

- "**Kernel Epanechnikov**"

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

- "**Kernel Tricubo**"

  $K(x) = \frac{70}{81}\,(1 - |x|^3)^3\, I_{[-1,1]}(x), \quad x \in \mathbb{R}$.

- "**Kernel Bipeso**"

  $K(x) = \frac{15}{16}\,(1 - x^2)^2\, I_{[-1,1]}(x), \quad x \in \mathbb{R}$.

- "**Kernel Triangular**"

  $K(x) = (1 - |x|)\, I_{[-1,1]}(x), \quad x \in \mathbb{R}$.

- "**Kernel Coseno**"

  $K(x) = \frac{1}{2}(\cos(\pi x) + 1)\, I_{[-1,1]}(x), \quad x \in \mathbb{R}$.

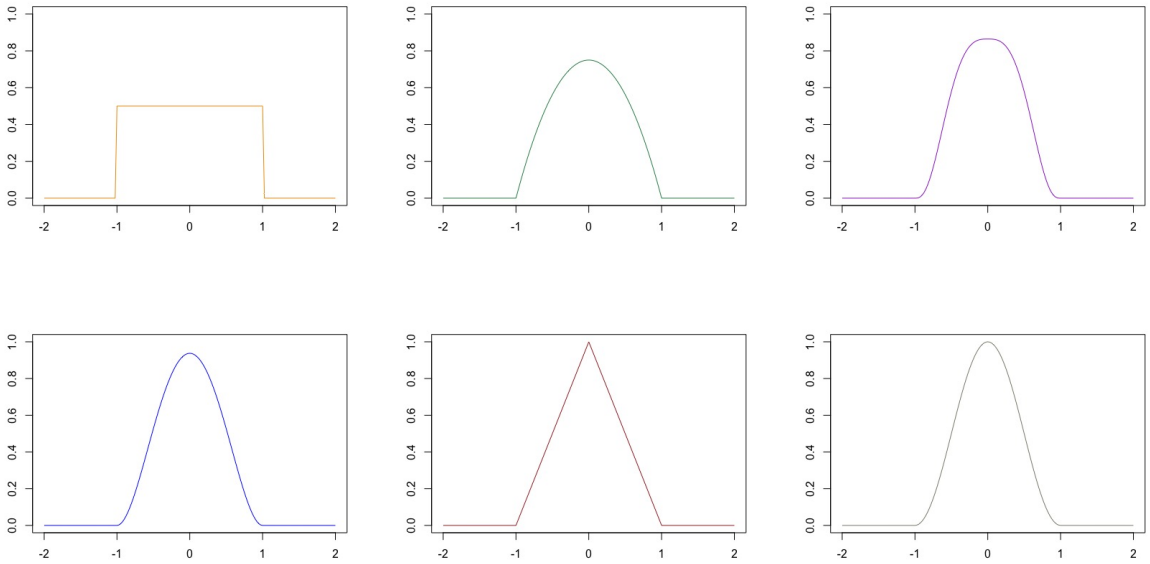

Figura 2.1: De izquierda a derecha y de arriba a abajo: Kernel Rectangular, Kernel Epanechnikov, Kernel Tricubo, Kernel Bipeso, Kernel Triangular y Kernel Coseno.

Observación 2.4. Desde este punto del documento cada vez que hablemos de K, asumiremos que está definido en el intervalo en el cual es positivo. Recordemos que este intervalo es por hipótesis el $[-1, 1]$, aunque éste puede modificarse por un parámetro tal y como veremos en la definición siguiente. A su vez, en algunas partes del documento, se pedirá que K sea estrictamente suave.

Tal y como hemos mencionado antes, el parámetro h juega un papel de escala. Para eso se define el kernel reescalado de la forma siguiente.

Definición 2.5 (Escalado de kernels)**.** Sea $h \in \mathbb{R}^+$ el parámetro ventana. Definimos el kernel reescalado por h como sigue:

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right), \quad x \in \mathbb{R}.$$

Observación 2.6. Si K es una densidad que toma valores en $[-1, 1]$, entonces K_h es una densidad que toma valores en $[-h, h]$. Además, si Z tiene densidad K entonces hZ tendrá densidad K_h.

El método local lineal trata de ajustar un polinomio de grado 1, más concretamente los parámetros del mismo, por el método de **mínimos cuadrados ponderados**, usando únicamente los puntos del entorno $(x - h, x + h)$. Una primera idea sería minimizar localmente la suma de cuadrados:

$$\sum_{i=1}^{n} (Y_i - a(x) - b(x)(x_i - x))^2 I\left[x_i \in [x - h, x + h]\right].$$

El problema de la suma de cuadrados anterior es que le da la misma importancia a todos los datos del intervalo $(x-h, x+h)$, independientemente de su distancia al punto x. El método local lineal propone que los puntos más cercanos a x tendrán un peso mayor en la estimación que los que, estando dentro del entorno $[x-h, x+h]$, estén más alejados. Es decir, la forma que tiene la expresión a minimizar es

$$\sum_{i=1}^{n}(Y_i - a(x) - b(x)(x_i - x))^2 K_h(x - x_i), \tag{2.1}$$

donde el factor $K_h(x - x_i)$ representa las mencionadas ponderaciones del método.

Nuestro objetivo actualmente es obtener una expresión de la forma:

$$\hat{m}(x) = \alpha(x),$$

donde $\alpha(x)$ y $\beta(x)$ representan los valores de los coeficientes $a(x)$ y $b(x)$ que minimizan la suma de cuadrados mostrada en la ecuación (2.1), respectivamente.

Antes de continuar con el teorema que nos dará la expresión del estimador local lineal, aplicaremos otro pequeño cambio en (2.1):

$$\min_{a,b} \sum_{i=1}^{n}(Y_i - a - b(x_i - x))^2 \hat{K}_{ih}, \tag{2.2}$$

donde hemos sustituido $K_h(x - x_i)$ por $\hat{K}_{ih}$, que representa:

$$\hat{K}_{ih} = \frac{K_h(x - x_i)}{\sum_{k=1}^{n} K_h(x - x_k)}.$$

El motivo fundamental de este cambio es conseguir que los pesos de la suma de cuadrados ponderada que aparece en (2.2) sumen uno. A lo largo de la demostración del próximo teorema se podrá apreciar la utilidad del mismo. Antes de presentar su enunciado daremos una breve idea de lo que haremos. La idea del teorema es obtener la expresión del estimador local lineal. En la demostración, primero se tratará de probar que la expresión analítica tiene los coeficientes que dice el enunciado del teorema, y después se obtendrá la expresión matricial, que será la que más nos sirva para la práctica y para el análisis teórico.

Teorema 2.7. *Los coeficientes del **Estimador Local Lineal** $\hat{m}(x) = \alpha$ vendrán dados explícitamente por las siguientes expresiones:*

$$\alpha = \xi_{1,0} - \frac{\xi_{1,1} - \xi_{1,0}\xi_{0,1}}{\xi_{0,2} - (\xi_{0,1})^2}\xi_{0,1},$$

$$\beta = \frac{\xi_{1,1} - \xi_{1,0}\xi_{0,1}}{\xi_{0,2} - (\xi_{0,1})^2},$$

supuesto que $\xi_{0,2} - (\xi_{0,1})^2 > 0$*, donde* $\xi_{i,j} = \sum_{i=1}^{n} \hat{K}_{ih} Y_i^i (x_i - x)^j$ *y donde* α *y* β *representan los valores de los coeficientes* a *y* b *que minimizan la suma de cuadrados mostrada en la ecuación* (2.1)*, respectivamente.*

Por otra parte, su expresión matricial es la que sigue:

$$\hat{m}(x) = \mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x Y}. \tag{2.3}$$

En esta última, tenemos que $\boldsymbol{X}$ *es la matriz de diseño,* $\boldsymbol{X^T}$ *es su traspuesta,* $\boldsymbol{P_x}$ *es la matriz de pesos e* $\boldsymbol{Y}$ *es un vector columna que contiene las componentes de la variable* $\boldsymbol{Y}$ *y* $\boldsymbol{e_1}$ *es un vector columna cuya primera componente es igual a 1 y la segunda 0. Tienen la siguiente expresión explícita:*

$$\boldsymbol{P_x} = \begin{bmatrix} \boldsymbol{K_h(x - x_1)} & . & . & . & \mathbf{0} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \mathbf{0} & . & . & . & \boldsymbol{K_h(x - x_n)} \end{bmatrix}_{\boldsymbol{n \times n}}, \quad \boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{(x_1 - x)} \\ . & . \\ . & . \\ . & . \\ \mathbf{1} & \boldsymbol{(x_n - x)} \end{bmatrix}_{\boldsymbol{n \times 2}}, \quad \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y_1} \\ . \\ . \\ . \\ \boldsymbol{Y_n} \end{bmatrix}_{\boldsymbol{n \times 1}}, \quad \boldsymbol{e_1} = \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}_{\mathbf{2 \times 1}}.$$

Observación 2.8. En la demostración que sigue, observaremos que en dos puntos consideramos la inversa de una matriz. Justificamos en esta observación que podemos suponer la existencia de dicha inversa. Para verlo consideremos la matriz en cuestión:

$$\begin{bmatrix} 1 & \xi_{0,1} \\ \xi_{0,1} & \xi_{0,2} \end{bmatrix}_{2 \times 2}.$$

Vemos que su determinante es $\xi_{0,2} - \xi_{0,1}^2$, que se trata de la diferencia entre el momento de orden dos y el de orden uno elevado al cuadrado, es decir, una varianza. Luego $\xi_{0,2} - \xi_{0,1}^2 > 0$, como es una varianza, es equivalente a que exista $i / x_i \neq x$ con $0 < \hat{K}_{ih} < 1$, lo que implica una distribución no degenerada de los puntos. Dicho índice i existirá siempre que haya suficientes datos x_i y h no sea un valor demasiado pequeño.

Demostración. Derivemos (2.2), primero con respecto de a y después con respecto de b para después igualar a cero ambas expresiones. Derivando con respecto de la variable a obtenemos:

$$\sum_{i=1}^{n} 2(Y_i - a - b(x_i - x))(-1)\hat{K}_{ih} = 0 \Leftrightarrow \sum_{i=1}^{n} Y_i \hat{K}_{ih} = a \sum_{i=1}^{n} \hat{K}_{ih} + b \sum_{i=1}^{n} (x_i - x)\hat{K}_{ih}$$

en este punto es cuando nos es de utilidad el cambio que hemos hecho en el factor de los pesos, ya que $\sum_{i=1}^{n} \hat{K}_{ih} = 1$ y por lo tanto:

$$\sum_{i=1}^{n} Y_i \hat{K}_{ih} = a + b \sum_{i=1}^{n} (x_i - x)\hat{K}_{ih},$$

que usando la notación de los $\xi_{i,j}$ nos quedaría como sigue:

$$\boxed{\xi_{1,0} = a + b \cdot \xi_{0,1}} \tag{2.4}$$

Continuemos con el proceso, esta vez derivando con respecto de b.

$$\sum_{i=1}^{n} 2\left(Y_i - a - b(x_i - x)\right)\left(-(x_i - x)\right)\hat{K}_{ih} = 0 \Leftrightarrow \sum_{i=1}^{n} Y_i(x_i - x)\hat{K}_{ih} = a\sum_{i=1}^{n}(x_i - x)\hat{K}_{ih} + b\sum_{i=1}^{n}(x_i - x)^2\hat{K}_{ih}$$

luego:

$$\boxed{\xi_{1,1} = a \cdot \xi_{0,1} + b \cdot \xi_{0,2}} \tag{2.5}$$

Ahora, combinando (2.4) y (2.5) llegamos a un sistema de la forma siguiente:

$$\begin{bmatrix}\xi_{1,0}\\ \xi_{1,1}\end{bmatrix}_{2\times 1} = \begin{bmatrix}1 & \xi_{0,1}\\ \xi_{0,1} & \xi_{0,2}\end{bmatrix}_{2\times 2} \cdot \begin{bmatrix}a\\ b\end{bmatrix}_{2\times 1} \Rightarrow \begin{bmatrix}1 & \xi_{0,1}\\ \xi_{0,1} & \xi_{0,2}\end{bmatrix}^{-1}_{2\times 2} \cdot \begin{bmatrix}\xi_{1,0}\\ \xi_{1,1}\end{bmatrix}_{2\times 1} = \begin{bmatrix}a\\ b\end{bmatrix}_{2\times 1}$$

Ahora, calculando la inversa:

$$\frac{1}{\xi_{0,2} - \xi_{0,1}^2} \cdot \begin{bmatrix}\xi_{0,2} & -\xi_{0,1}\\ -\xi_{0,1} & 1\end{bmatrix}_{2\times 2}$$

e implementándola en la igualdad se termina llegando a:

$$\frac{1}{\xi_{0,2} - \xi_{0,1}^2} \cdot \begin{bmatrix}\xi_{1,0}\xi_{0,2} - \xi_{1,1}\xi_{0,1}\\ \xi_{1,1} - \xi_{1,0}\xi_{0,1}\end{bmatrix}_{2\times 1} = \begin{bmatrix}a\\ b\end{bmatrix}_{2\times 1}$$

concluyendo así el resultado de la expresión analítica.

El procedimiento para la obtención de la forma matricial del estimador local lineal es realmente más sencillo, si se usa cálculo vectorial, y lo veremos brevemente para completar la demostración. Lo primero sería considerar la expresión de partida en forma matricial:

$$\min_{\mathbf{\Lambda}}(\mathbf{Y} - \mathbf{X}\mathbf{\Lambda})^T\mathbf{P_x}(\mathbf{Y} - \mathbf{X}\mathbf{\Lambda}),$$

donde mantenemos la notación presentada en el enunciado del teorema, añadiendo únicamente:

$$\mathbf{\Lambda} = \begin{bmatrix}a\\ b\end{bmatrix}_{2\times 1}$$

que es la matriz que contiene los coeficientes del estimador local lineal. Ahora basta con diferenciar la expresión e igualarla a cero para obtener:

$$\mathbf{X^T P_x X \Lambda} = \mathbf{X^T P_x Y}.$$

Despejando $\mathbf{\Lambda}$ de esta expresión:

$$\hat{\mathbf{\Lambda}} = (\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x Y},$$

y por último, dado que el estimador buscado es el intercepto, tenemos que:

$$\hat{m}(x) = \mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x Y}. \tag{2.6}$$

□

2.2. Sesgo y varianza. El efecto de la ventana h

Continuaremos el trabajo con el análisis de las propiedades estadísticas del estimador local lineal. En concreto nos centraremos en el análisis de su sesgo y de su varianza que, tal y como veremos a lo largo de la sección, están tremendamente relacionados, lo que le confiere un interés añadido. Por otra parte, como ya mencionamos al comienzo del capítulo supondremos un diseño fijo. Además, en esta sección profundizaremos más en la elección del parámetro ventana h que, a su vez, está más relacionado con el sesgo y la varianza de lo que puede parecer a priori. Dado que se trata de una sección clave para comprender teóricamente la forma en la que funciona el estimador local lineal y el trasfondo que habrá tras el análisis de datos, acompañaremos la teoría con imágenes y gráficas que ayuden a comprender la misma. Para concluir esta pequeña introducción, añadir únicamente que a lo largo de esta sección asumiremos estas cinco hipótesis como ciertas:

1. La función m está definida en el intervalo $[0, 1]$. Existen las derivadas segunda y tercera de la función m, m'' y m''', y son continuas en el intervalo $[0, 1]$.

2. El kernel es simétrico en torno al cero y su soporte es el intervalo $[-1, 1]$.

3. El parámetro ventana, h, depende únicamente de n, h_n, y cumple $h_n \to 0$ y $n \cdot h_n \to \infty$ cuando $n \to \infty$. Por comodidad en la notación seguiremos denotando el parámetro ventana como h, omitiendo su dependencia de n.

4. El análisis se hará para $x \in (0, 1)$. Además se supondrá que n es suficientemente grande para que se cumpla que $h < x < 1 - h$.

5. Existe la primera derivada de K, K', y está acotada.

Observemos que gracias a las condiciones 2) y 5), podemos suponer a lo largo de la sección que ambas funciones K y K' están acotadas, y en particular, que existen. A priori, no parece algo a lo que pueda sacársele demasiado partido, en cambio, veremos como resulta de gran interés. A su vez, que el kernel sea simétrico en torno al cero, facilita enormemente el trabajo que realizaremos con el mismo. Por otra

parte, la hipótesis 1) nos está proporcionando una condición para aproximar linealmente m por puntos próximos, lo cual será esencial para el análisis que realizaremos en esta sección, así como la 4), de la que también nos beneficiaremos cuando hablemos del sesgo asintótico. También es importante tener presente en todo momento la hipótesis 3), ya que nos recuerda que, a pesar de que denotemos al parámetro ventana por h, se trata en realidad de una sucesión que depende de n, que es como tenemos que verlo.

Comenzaremos hablando del sesgo. El sesgo en una estimación es, por así decirlo, cuanto difiere en promedio ésta de la curva real. Dicho de otra forma, podría verse como una forma de cuantificar lo bien o mal que se ajusta nuestra estimación a la curva real. Analíticamente, el sesgo es la diferencia entre el promedio del estimador y el valor real, y para tratarlo seguiremos un camino que nos irá llevando de forma natural hasta la expresión asintótica del sesgo del estimador local lineal. Comenzaremos con el teorema que viene a continuación, el cual nos dará la expresión de la esperanza del estimador local lineal. Resultará útil para comenzar a observar características y empezar a formarnos hipótesis que trataremos de fundamentar más adelante. Sin ir más lejos, el teorema que proporcionaremos ahora, nos será de gran ayuda para comenzar a graficar el sesgo en ejemplos concretos y de este modo dar el primer paso del mencionado camino para el análisis asintótico del sesgo.

Teorema 2.9. *La esperanza del estimador local lineal tiene la siguiente expresión:*

$$\mathbb{E}[\hat{m}(x)] = \mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x M}, \quad \textit{donde} \quad \boldsymbol{M} = (m(x_1), ..., m(x_n))^T. \tag{2.7}$$

Demostración. Por el Teorema 2.7, ver expresión (2.3), podemos deducir la esperanza usando que estamos suponiendo un diseño fijo, es decir, los valores $x_1, ..., x_n$ son cantidades fijas. Así:

$$\mathbb{E}[\hat{m}(x)] = \mathbb{E}\left[\mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x Y}\right] = \mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x}\mathbb{E}[\mathbf{Y}],$$

y por lo tanto se tiene que:

$$\mathbb{E}[\hat{m}(x)] = \mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x M}.$$

□

Gracias a esta fórmula que acabamos de obtener, somos capaces de crear representaciones gráficas para analizar el sesgo y su dependencia de la ventana h y de la forma de la curva m. Es por esto que realizaremos y analizaremos dos mapas de calor o *heatmaps* a continuación.

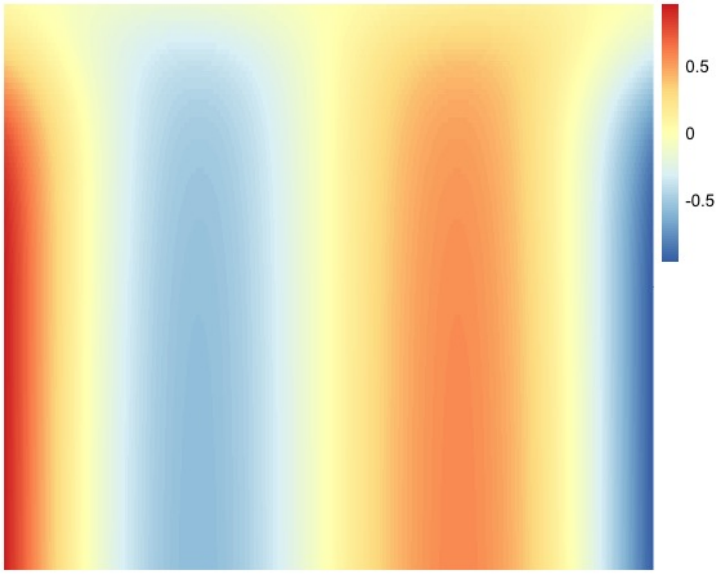

Figura 2.2: En esta gráfica se puede apreciar un *heatmap* en el que, partiendo de una muestra $x_1, ..., x_n$ de datos equiespaciados que van desde cero hasta uno con longitud $n = 100$ se ha calculado el sesgo del estimador local lineal con la función $m(x) = \sin(2\pi x)$ en una rejilla 100×1000, con $Y_i = m(x_i) + \varepsilon_i$, donde $\mathbb{E}(\varepsilon_i) = 0$. En el eje X aparece el sesgo calculado en un vector $r_1, ..., r_M$ de datos equiespaciados que van desde cero hasta uno con longitud $M = 1000$ para un valor fijo de h, y en el Y aparecen distintos valores del parámetro ventana, que aumentan desde la primera fila, es decir, la superior, en la que se tomó $h = 0{,}1$ y llegan de forma creciente y equiespaciada a la última fila, donde $h = 1$.

La Figura 2.2, correspondiente al mapa de calor, resulta muy interesante ya que nos muestra muchas cosas. Comencemos dejando claro lo que estamos representando mediante este mapa de calor. En él, estamos midiendo el sesgo, es decir, por dar una primera idea antes de definir el concepto formalmente, diríamos que es la bondad con la que se ajusta (en promedio) nuestra estimación a la curva real. Por otra parte, tal y como se indica en el pie de la susodicha figura, hemos tomado la función $m(x) = \sin(2\pi x)$, para $x \in [0, 1]$, y partiendo de los valores contenidos en un vector al que hemos llamado x, que representa los valores entre cero y uno, ambos incluidos, equiespaciados y con longitud igual a cien, se ha calculado el sesgo del estimador local lineal en un vector $r_1, ..., r_M$ de datos equiespaciados que van desde cero hasta uno con longitud $M = 1000$. Por otra parte, hemos repetido este proceso para cien valores de h comenzando en $0{,}1$, fila superior, y avanzando de forma equiespaciada hasta llegar al valor $h = 1$ que se encuentra en la última fila, es decir, en la fila inferior. En resumen, en la casilla que se encuentra en la fila i-ésima y en la columna j-ésima de nuestra rejilla 100×1000 podremos encontrar el sesgo de estimar mediante el estimador local lineal la función $m(x) = \sin(2\pi x)$ en r_j, para el valor i-ésimo de h cuando disponemos de una muestra equiespaciada de $n = 100$ datos x_i. Las casillas de la rejilla que tengan un color rojo más fuerte representan un valor más positivo del sesgo, mientras que las que tengan un color azul más fuerte representan un valor más negativo del mismo. Dicho esto, observemos ahora el mapa de

calor. Lo primero que nos llama la atención al mirarlo son las columnas laterales, donde el color rojo y azul es más fuerte, lo que nos indica que ahí el sesgo es más alto y más bajo, respectivamente. Además, se puede apreciar una clara oscilación entre valores positivos y negativos del sesgo. Podríamos achacar este fenómeno a la oscilación de la propia función entre sus máximos y mínimos relativos. Si miramos el *heatmap* de izquierda a derecha, vemos que comienza en valores muy positivos, para después ir poco a poco bajando hasta tocar valores negativos, para seguir después con un aumento gradual hasta valores positivos y finalmente bajar hasta los valores más negativos. El hecho de que en las columnas iniciales se encuentre el sesgo más positivo, mientras que en las finales se encuentre el más negativo tiene que ver con el denominado efecto frontera, y motiva que pidamos en el análisis asintótico que $h < x < 1 - h$.

Una vez analizadas las columnas, centrémonos en las filas. Observamos en el mapa de calor que a medida que bajamos en las filas, da igual en la columna que nos encontremos, el color se hace más fuerte. En las columnas en las que el sesgo es positivo, este se hace aún más positivo al bajar por las filas y aquellas en las que el sesgo era negativo, se hace aún más negativo al bajar por ellas. En resumidas cuentas, al bajar por las filas el **sesgo aumenta**. Si pensamos en qué estamos haciendo cuando bajamos por las filas, recordaremos que este proceso no es más que un aumento progresivo del valor del parámetro ventana. Es decir, este mapa de calor nos está mostrando gráficamente que el sesgo aumenta a medida que lo hace el valor del parámetro h. O, equivalentemente, disminuye al tomar valores pequeños de h. Veamos un ejemplo más antes de continuar.

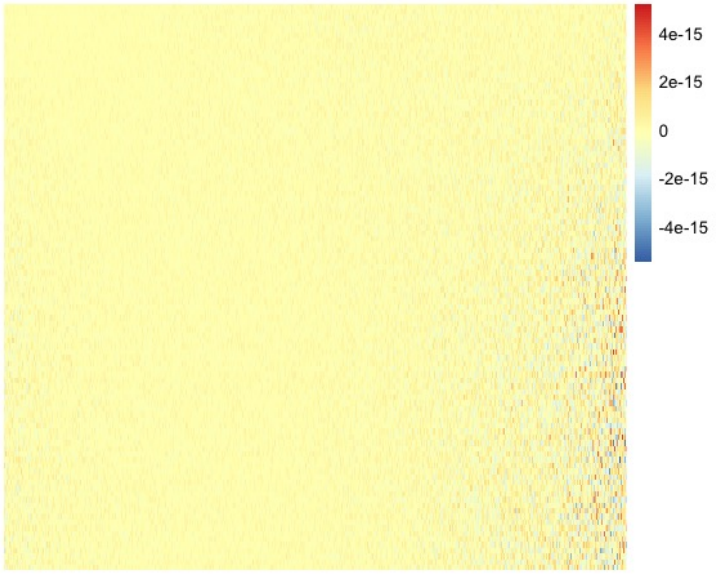

Figura 2.3: En esta gráfica se puede apreciar un *heatmap* en el que, partiendo de una muestra $x_1, ..., x_n$ de datos equiespaciados que van desde cero hasta uno con longitud $n = 100$ se ha calculado el sesgo del estimador local lineal con la función $m(x) = x$ en una rejilla 100×1000, con $Y_i = m(x_i) + \varepsilon_i$, donde $\mathbb{E}(\varepsilon_i) = 0$. En el eje X aparece el sesgo en un vector $r_1, ..., r_M$ de datos equiespaciados que van desde cero hasta uno con longitud $M = 1000$ para un valor fijo de h, y en el Y aparecen distintos valores del parámetro ventana, que aumentan desde la primera fila, en la que $h = 0{,}1$ y llegan de forma creciente y equiespaciada a la última fila, donde $h = 1$.

En la Figura 2.3 nos encontramos con otro mapa de calor. Aunque éste es sustancialmente diferente al anterior, el proceso de obtención es análogo al que explicamos en el párrafo anterior, solo que esta vez la función que implementamos no fue el $m = \sin(2\pi x)$, si no $m(x) = x$. Ahora el *heatmap* es mucho más monótono, ya que solo cuenta con un color principal, a diferencia del anterior. Fijémonos ahora en la escala. Los puntos en los que el sesgo no es cero exactamente, lo es prácticamente, y al igual que antes, estos puntos en los que el sesgo no es cero, se encuentran en la parte más baja del mapa de calor, es decir, en las últimas filas, donde el valor de la h es más alto y próximo a la frontera del intervalo $[0, 1]$.

En este punto, podríamos formarnos una hipótesis, hablando únicamente a través de los resultados obtenidos gráficamente, acerca de la disminución del sesgo al reducir h, pero aún no contamos con ningún resultado que nos lo pueda confirmar ya que la expresión del sesgo depende de forma compleja de h. Por otro lado, así como antes la oscilación en los colores del *heatmap* podría uno achacárselos a la periodicidad de la función seno, dado que podría relacionarse con las zonas de máximos y mínimos de m, ¿podríamos atribuir a la forma de la función $m(x) = x$ el sesgo nulo visto en la Figura 2.3? ¿quizás a la linealidad de dicha función?

Estos mapas de calor nos permiten extraer muchas conclusiones, pero el Teorema 2.9 no nos proporciona la información suficiente para comprender el por qué de las características que observamos. Es por esto que introducimos ahora el siguiente teorema, que analiza el comportamiento asintótico, es decir, para muestras grandes, del sesgo.

Teorema 2.10. *Por las condiciones 1) - 5) mencionadas al inicio de la subsección, el estimador local lineal tiene un sesgo asintótico de la forma siguiente:*

$$\mathbb{E}[\hat{m}(x)] - m(x) = \frac{1}{2}h^2 m''(x) \int_{-1}^{1} z^2 K(z)dz + o(h^2)$$

Observación 2.11. Además, por simplicidad en los cálculos que siguen supondremos que $x_i = \frac{i}{n}$, con i=1,...,n. De este modo imponemos un diseño fijo y equiespaciado.

Demostración. Si ahora consideramos el teorema de Taylor de orden dos, para cualquier $x \in (0, 1)$:

$$m(x_i) = m(x) + (x_i - x)m'(x) + \frac{1}{2}(x_i - x)^2 m''(x) + \frac{1}{6}(x_i - x)^3 m'''(y_i),$$

donde y_i es un número real entre x_i y x. Matricialmente:

$$\begin{bmatrix} m(x_1) \\ . \\ . \\ . \\ m(x_n) \end{bmatrix}_{n\times 1} = M = \mathbf{X} \begin{bmatrix} m(x) \\ m'(x) \end{bmatrix}_{2\times 1} + \frac{1}{2}m''(x) \begin{bmatrix} (x_1 - x)^2 \\ . \\ . \\ . \\ (x_n - x)^2 \end{bmatrix}_{n\times 1} + \begin{bmatrix} R_2(x_1) \\ . \\ . \\ . \\ R_2(x_n) \end{bmatrix}_{n\times 1}, \tag{2.8}$$

siendo $R_2(x_i)$ el i-ésimo término del resto.

Observación 2.12. Observamos que si $m''(x) = 0$ (m es lineal) el sesgo sería cero (ver la ecuación (2.7) del Teorema 2.9). Esto explica lo que nos ocurría en el mapa de calor de la Figura 2.3. Al ser la función $m(x) = x$ lineal, su derivada segunda es cero y si lo implementamos en la fórmula proporcionada por la ecuación (2.8) obtendríamos como resultado lo que queda plasmado en el mapa de calor mencionado.

Ahora tomaremos la ecuación (2.7) del Teorema 2.9 para obtener lo siguiente:

$$\mathbb{E}[\hat{m}(x)] = \mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x X} \begin{bmatrix} m(x) \\ m'(x) \end{bmatrix}_{2\times 1} +$$

$$+\frac{1}{2}m''(x)\mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x} \begin{bmatrix} (x_1 - x)^2 \\ . \\ . \\ . \\ (x_n - x)^2 \end{bmatrix}_{n\times 1} + \mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x} \begin{bmatrix} R_2(x_1) \\ . \\ . \\ . \\ R_2(x_n) \end{bmatrix}_{n\times 1},$$

donde $R_2(x_i) = \frac{1}{6}m'''(y_i)(x_i - x)^3$. Analizaremos ahora la parte principal del desarrollo de Taylor. Más adelante analizaremos la parte del resto. Se tiene que, dado que:

$$\mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x X}\begin{bmatrix} m(x) \\ m'(x) \end{bmatrix}_{2\times 1} = m(x),$$

la parte principal del sesgo (omitido el término del resto) quedaría:

$$\mathbb{E}[\hat{m}(x)] - m(x) \cong \frac{1}{2}m''(x)\mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x}\begin{bmatrix} (x_1 - x)^2 \\ . \\ . \\ . \\ (x_n - x)^2 \end{bmatrix}_{n\times 1} . \tag{2.9}$$

Consideremos ahora el factor $\boldsymbol{X^T P_x X}$. Realizando el producto matricial, terminaremos llegando a una matrix de orden 2x2, cuyas componentes están compuestas por:

$$\begin{bmatrix} \sum_{i=1}^n K_h(x_i - x) & \sum_{i=1}^n (x_i - x)K_h(x_i - x) \\ \sum_{i=1}^n (x_i - x)K_h(x_i - x) & \sum_{i=1}^n (x_i - x)^2 K_h(x_i - x) \end{bmatrix}_{2\times 2} .$$

Por otra parte:

$$\mathbf{X^T P_x}\begin{bmatrix} (x_1 - x)^2 \\ . \\ . \\ . \\ (x_n - x)^2 \end{bmatrix}_{n\times 1} = \begin{bmatrix} \sum_{i=1}^n (x_i - x)^2 K_h(x_i - x) \\ \sum_{i=1}^n (x_i - x)^3 K_h(x_i - x) \end{bmatrix}_{2\times 1} .$$

En este punto de la demostración vamos a realizar un pequeño ajuste, para simplificar los cálculos que siguen. Introduzcamos la siguiente notación:

$$s_r(x) = \frac{1}{n}\sum_{i=1}^n (x_i - x)^r K_h(x_i - x), \quad r = 0, 1, 2, ...$$

y en consecuencia:

$$\frac{1}{n}\mathbf{X^T P_x X} = \begin{bmatrix} s_0(x) & s_1(x) \\ s_1(x) & s_2(x) \end{bmatrix}_{2\times 2} .$$

$$\frac{1}{n}\mathbf{X^T P_x}\begin{bmatrix} (x_1 - x)^2 \\ . \\ . \\ . \\ (x_n - x)^2 \end{bmatrix}_{n\times 1} = \begin{bmatrix} s_2(x) \\ s_3(x) \end{bmatrix}_{2\times 1} .$$

El próximo lema permite aproximar asintóticamente los elementos $s_r(x)$ mediante una representación integral.

Lema 2.13. *Por las condiciones (2) hasta la (4) mencionadas al inicio de la subsección, se cumple que:*

$$\int_0^1 (y-x)^r K_h(y-x)dy = \frac{1}{n}\sum_{i=1}^{n}(x_i-x)^r K_h(x_i-x) + o(h^r). \tag{2.10}$$

Demostración. Iniciaremos con la siguiente igualdad:

$$\int_0^1 (y-x)^r K_h(y-x)dy = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} (y-x)^r K_h(y-x)dy.$$

Si ahora tomamos la función $G(z) = \int_{x_i}^{z}(y-x)^r K_h(y-x)dy$ y, fijado i, $i \in \{1, ..., n-1\}$, realizamos su expansión en serie de Taylor alrededor del punto x_i obtenemos un polinomio para $z \in [x_i, x_{i+1}]$ de la forma:

$$G(z) = G(x_i) + G'(x_i)(z-x_i) + G''(t_i)\frac{(z-x_i)^2}{2},$$

siendo t_i un punto en $[x, z]$. En particular verifica $t_i \in [x_i, x_{i+1}]$, para cualquier $z \in [x_i, x_{i+1}]$. Por un lado, tenemos que:

$$G(x_i) = \int_{x_i}^{x_i} (y-x)^r K_h(y-x)dy = 0,$$

$$G'(x_i) = (x_i-x)^r K_h(x_i-x),$$

$$G''(t_i) = r(t_i-x)^{r-1}K_h(t_i-x) + (t_i-x)^r \frac{1}{h^2}K'\left(\frac{t_i-x}{h}\right).$$

En consecuencia, si evaluamos la función G en x_{i+1}, recordando que $x_{i+1} - x_i = \frac{1}{n}$:

$$\int_{x_i}^{x_{i+1}} K_h(y-x)dy = G(x_{i+1})$$

$$= (x_i-x)^r K_h(x_i-x)(x_{i+1}-x_i) + \frac{r}{h}(t_i-x)^{r-1}K\left(\frac{t_i-x}{h}\right)\frac{(x_{i+1}-x_i)^2}{2}$$

$$+\frac{1}{h^2}(t_i-x)^r K'\left(\frac{t_i-x}{h}\right)\frac{(x_{i+1}-x_i)^2}{2}$$

$$= \frac{1}{n}(x_i-x)^r K_h(x_i-x) + \frac{r}{2n^2h}(t_i-x)^{r-1}K\left(\frac{t_i-x}{h}\right) + \frac{1}{2n^2h^2}(t_i-x)^r K'\left(\frac{t_i-x}{h}\right).$$

Y por lo tanto:

$$\int_0^1 K_h(y-x)dy = \sum_{i=0}^{n-1} G(x_{i+1})$$

$$= \frac{1}{n}\sum_{i=0}^{n-1}(x_i-x)^r K_h(x_i-x) + \frac{r}{2n^2h}\sum_{i=0}^{n-1}(t_i-x)^{r-1}K\left(\frac{t_i-x}{h}\right) + \frac{1}{2n^2h^2}\sum_{i=0}^{n-1}(t_i-x)^r K'\left(\frac{t_i-x}{h}\right)$$

donde $t_i \in [x_i, x_{i+1}]$. Trabajando ahora sobre el segundo y tercer sumando, y teniendo en cuenta que K_h sólo toma valores no nulos en $[x-h, x+h]$:

- Si $x_i > x + h \Rightarrow t_i > x_i > x + h \Rightarrow K\left(\frac{t_i - x}{h}\right) = K'\left(\frac{t_i - x}{h}\right) = 0$
- Si $x_{i+1} < x - h \Rightarrow t_i \leq x_{i+1} < x - h \Rightarrow K\left(\frac{t_i - x}{h}\right) = K'\left(\frac{t_i - x}{h}\right) = 0$

se tendrá que nos deberíamos limitar a sumar los elementos pertenecientes a B_x, donde B_x representa el conjunto donde podría ocurrir que $K_h(t_i - x) \neq 0$ y $K'_h(t_i - x) \neq 0$, es decir,

$$B_x : \begin{cases} x_{i+1} \geq x - h, \\ x_i \leq x + h. \end{cases}$$

Por la definición de x_i, se tiene que:

$$\left.\begin{aligned} \frac{i+1}{n} &\geq x - h, \\ \frac{i}{n} &\leq x + h, \end{aligned}\right\} \qquad \Longrightarrow \left.\begin{aligned} i &\geq nx - nh - 1, \\ i &\leq nx + nh. \end{aligned}\right\} \quad .$$

De esto último, deducimos lo siguiente:

$$nx - nh - 1 \leq i \leq nx + nh \Rightarrow \#B_x \leq nx + nh - (nx - nh - 1) + 1 = 2nh + 2,$$

donde el 1 final se ha sumado debido al ajuste de los menores y mayores o iguales. Por lo tanto, existen a lo sumo $2(nh+1)$ sumandos distintos de 0. Teniendo en cuenta todo esto, usamos que K' está acotada por una constante que denotamos por C. A su vez podemos suponer que K está acotada por la misma constante C, y usando también que $nh \to \infty$, podríamos realizar la siguiente acotación:

$$\left| \frac{r}{2n^2h} \sum_{i \in B_x} (t_i - x)^{r-1} K\left(\frac{t_i - x}{h}\right) + \frac{1}{2n^2h^2} \sum_{i \in B_x} (t_i - x)^r K'\left(\frac{t_i - x}{h}\right) \right|$$

$$\leq \left| \frac{r}{2n^2h} \sum_{i \in B_x} (t_i - x)^{r-1} K\left(\frac{t_i - x}{h}\right) \right| + \left| \frac{1}{2n^2h^2} \sum_{i \in B_x} (t_i - x)^r K'\left(\frac{t_i - x}{h}\right) \right|$$

pero teniendo en cuenta que $|t_i - x| \leq h$ y recordando que $nh \to \infty$, podríamos acotar ahora por

$$\leq \frac{r}{2n^2h} \cdot (2nh + 2) \cdot h^{r-1} \cdot C + \frac{1}{2n^2h^2} \cdot (2nh + 2) \cdot h^r \cdot C$$

$$= o(1) \cdot h^r = o(h^r)$$

Por lo tanto, podemos concluir que:

$$\int_0^1 (y - x)^r K_h(y - x) dy = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - x)^r K_h(x_i - x) + o(h^r).$$

Por otra parte, dado que la suma vista en la ecuación (2.10) del Lema 2.13, vale cero para $x_i = x_0 = 0$ y $x_i = x_n$, es decir:

$$(x_0 - x)^r K_h(x_0 - x) = (x_n - x)^r K_h(x_n - x) = 0,$$

tendremos que:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-x)^r K_h(x_i-x) = \frac{1}{n}\sum_{i=0}^{n-1}(x_i-x)^r K_h(x_i-x), \text{donde } \; x_0=0,$$

y en consecuencia:

$$\boxed{\int_0^1 (y-x)^r K_h(y-x)dy = s_r + o(h^r).}$$

□

Ahora, hagamos en la expresión integral el cambio de variable $u = \frac{y-x}{h}$. De este modo, obtenemos:

$$\int_0^1 (y-x)^r K_h(y-x)dy = h^r \int_{\frac{-x}{h}}^{\frac{1-x}{h}} u^r K(u)du + o\left(h^r\right).$$

Dado que cuando $h \to 0$ los límites de integración tenderán a $+\infty$ y $-\infty$, si recordamos que la función K está definida en el intervalo $[-1,1]$ se puede realizar el siguiente ajuste:

$$= h^r \int_{-1}^{1} u^r K(u)du + o\left(h^r\right).$$

Por la simetría respecto al origen de K, sus momentos impares desaparecen y nos quedamos con las siguientes entradas dentro de nuestras matrices, apoyándonos a su vez en las condiciones 4 y 5 de la Definición 2.1 y en el Lema 2.13:

$$\frac{1}{n}\mathbf{X^T P_x X} = \begin{bmatrix} 1+o(1) & o(h) \\ o(h) & h^2\int_{-1}^{1} z^2K(z)dz + o(h^2) \end{bmatrix}_{2\times 2}$$

y

$$\frac{1}{n}\mathbf{X^T P_x}\begin{bmatrix} (x_1-x)^2 \\ . \\ . \\ . \\ (x_n-x)^2 \end{bmatrix}_{n\times 1} = \begin{bmatrix} h^2\int_{-1}^{1} z^2K(z)dz + o(h^2) \\ o(h^3) \end{bmatrix}_{2\times 1}$$

Por último, introduciendo en la fórmula matricial del estimador los términos obtenidos y realizando un simple cálculo matricial, podríamos llegar a una expresión explícita para el término principal del sesgo (recordemos la ecuación (2.9), que es la expresión del mismo que habíamos visto previamente). Veamos los mencionados cálculos para concluir con el resultado.

Multiplicando la inversa de $\frac{1}{n}\boldsymbol{X^T P_x X}$ por

$$\begin{bmatrix} h^2\int_{-1}^{1} z^2K(z)dz + o(h^2) \\ o(h^3) \end{bmatrix}_{2\times 1}$$

y quedándonos con la primera componente, ya que recordemos que todo este producto va multiplicado también por $\boldsymbol{e_1}$, llegamos a un cociente de la forma siguiente:

$$\frac{h^4(\int_{-1}^{1} z^2K(z)dz)^2 + o(h^4)}{h^2\int_{-1}^{1} z^2K(z)dz + o(h^2)}.$$

Extrayendo factor común del numerador nos quedaría:

$$h^2\int_{-1}^{1} z^2K(z)dz\left(\frac{h^2\int_{-1}^{1} z^2K(z)dz + o(h^2)}{h^2\int_{-1}^{1} z^2K(z)dz + o(h^2)}\right) = h^2\int_{-1}^{1} z^2K(z)dz\cdot(1+o(1)).$$

Realizando el producto:

$$h^2\int_{-1}^{1} z^2K(z)dz\,(1+o(1)) = h^2\int_{-1}^{1} z^2K(z)dz + o(h^2).$$

Incorporando esto último al resto de la fórmula principal del sesgo, llegaríamos a una expresión para el primer término del sesgo:

$$\boxed{\mathbb{E}[\hat{m}(x)] - m(x) = \frac{1}{2}h^2m''(x)\int_{-1}^{1} z^2K(z)dz + o(h^2)}$$

Finalmente, volvemos sobre el término del resto que involucraba la derivada tercera de m. Si nos fijamos en el proceso que hemos llevado a cabo con la derivada segunda, podemos deducir que si ahora realizásemos un procedimiento análogo con la derivada tercera de m, m''', obtendríamos un término de error del orden de h^3, que podría ser absorbido por el término $o(h^2)$ que tenemos de la parte principal analizada anteriormente. Es por esto que el resto no alterará el comportamiento asintótico del sesgo. □

Observación 2.14. Se puede encontrar un resultado más general en *Kernel Smoothing* ([17]), concretamente en el Capítulo 5, Sección 5.3 del mismo.

Resulta claro después de ver este teorema que el sesgo dependerá del valor de h. Cuanto mayor sea este, mayor será el sesgo y viceversa. Por este motivo es por el que hemos podido ver un aumento en el sesgo a medida que bajábamos en las filas de nuestros mapas de calor. Además, observemos que en la fórmula proporcionada por el teorema aparece el factor $m''(x)$, el cual puede cambiar de signo, de donde deducimos que el sesgo puede ser tanto positivo como negativo. Pensemos en esto y recordemos que cuando una función tiene un máximo, su derivada segunda es negativa en el mismo, mientras que en el mínimo de la función, su derivada segunda será positiva. Gracias a esto podemos inferir que el sesgo será negativo en los máximos de la función m y positivo en los mínimos de la susodicha función. Dicho de otra forma, este teorema nos proporciona las herramientas necesarias para justificar y entender lo que hemos podido intuir gracias a la Figura 2.2 y a la Figura 2.3.

Recordemos la oscilación entre valores positivos y negativos de la Figura 2.2, ya que ahora estamos en condiciones de explicarla. Sabemos que el seno es una función sinusoidal, es decir, que oscila creando

máximos y mínimos locales de forma periódica. Tal y como explicamos antes, cuando tenemos un máximo la derivada segunda será negativa, creando un sesgo negativo mientras que en los mínimos esta será positiva, haciendo que el sesgo sea positivo. Esto se puede ver gráficamente, y para ello nos ayudaremos de la siguiente gráfica:

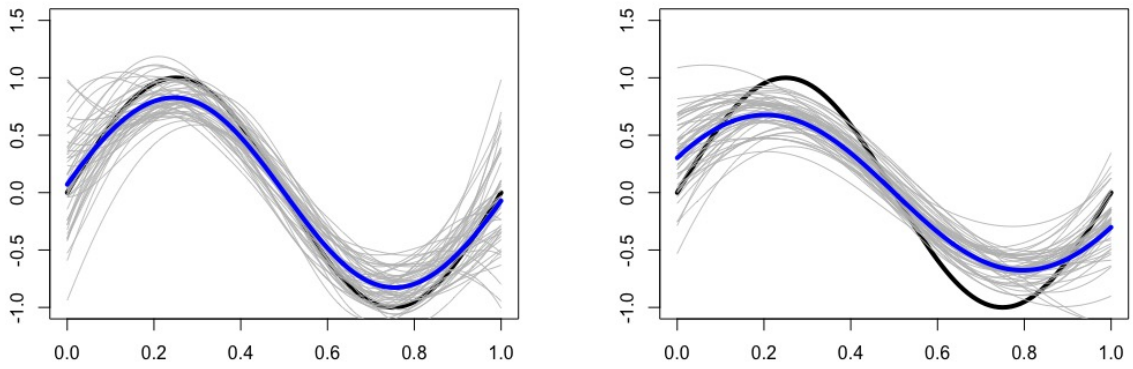

Figura 2.4: En esta gráfica se puede apreciar una representación del $\sin(2\pi x)$ para un vector x de longitud $n = 100$ de valores equiespaciados que van desde cero hasta uno (curva negra gruesa) y una representación de $\mathbb{E}[\hat{m}(x)]$ (línea azul gruesa), calculada partiendo del mencionado vector x en otro vector $r_1, ..., r_M$ de datos equiespaciados que van desde cero hasta uno con longitud $M = 1000$. Las líneas finas de color gris representan $L = 50$ estimaciones distintas calculadas mediante el estimador local lineal en el mencionado vector $r_1, ..., r_M$ para observaciones de la forma $\{Y_{ij} = m(x_j) + \varepsilon_j$, $j = 1, ..., n\}$, donde $\varepsilon_j \in N(0, 1)$, con $i = 1, ..., L$. Además, en la imagen de la izquierda se ha fijado un valor de $h = 0{,}1$ y en la imagen de la derecha $h = 0{,}175$.

Observemos ambas imágenes de la Figura 2.4. En el máximo de la curva negra representada en las dos imágenes, $\mathbb{E}[\hat{m}(x)]$ queda por debajo de la misma, así como la mayoría de las estimaciones para las distintas observaciones en la imagen de la izquierda, y todas en la de la derecha, lo que resultaría en un sesgo negativo, tal y como hemos dicho antes basándonos en el signo de la derivada segunda. Por otra parte, en el mínimo de la curva negra que llega a continuación, $\mathbb{E}[\hat{m}(x)]$ queda por encima, y casi todas las estimaciones para las distintas observaciones en la imagen de la izquierda y una vez más todas las de la imagen de la derecha quedan también por encima de la curva negra, siendo así el sesgo positivo, apoyando una vez más la teoría del signo de la derivada segunda. Además, resulta muy interesante comparar ambas imágenes, ya que al tener un distinto valor del parámetro ventana, tendrán también una forma diferente. Vemos que al aumentar el valor de h, tanto la curva azul, como las curvas grises se aplanan, asemejándose de este modo más a una recta, lo cual era de esperar, ya que a estas alturas, el Teorema 2.10 nos ha proporcionado la información necesaria para saber que el sesgo aumenta

con el valor del parámetro ventana. Por este motivo, vemos que los máximos que trazan tanto la curva azul como las curvas grises, cada vez son menos pronunciados, y de este modo ya no hay ni una sola estimación de ninguna observación que quede por encima del máximo, o por debajo del mínimo, ya que se agolpan cada vez más en torno a la curva azul, que a su vez cada vez es más recta. La Figura 2.4 resulta muy interesante para cerrar el estudio del sesgo, ya que ofrece de una forma muy visual y clara una explicación que apoya la teoría vista en el Teorema 2.10, que por su parte estaba aportando una explicación fundamentada para las conjeturas formadas en la Figura 2.2 y la Figura 2.3.

Observación 2.15. Es importante recalcar que a lo largo del capítulo hemos trabajado con datos equiespaciados, pero en la página 80 de *All of Nonparametric Statistics* ([18]) se nos confirma que el diseño de los puntos no afectaría al sesgo. Es decir, la fórmula asintótica que hemos proporcionado se conservaría intacta tomando un diseño distinto al trabajado.

Una vez concluido el estudio del sesgo, pasamos a tratar con la varianza del estimador local lineal. Esta mide qué tan dispersos están los datos alrededor de su media (recordar las curvas gruesas de la Figura 2.4), por lo que es lógico pensar que nos interesará tenerla controlada. Al igual que antes, el esquema que seguiremos a lo largo del estudio de la varianza estará determinado por las necesidades que nos van surgiendo a medida que vamos avanzando. Comenzamos con el siguiente teorema, pero antes, recalcamos, ya que será importante para lo que sigue, que el modelo supuesto es un modelo **homocedástico**. Dado que estamos presentando un concepto nuevo, haremos una breve digresión para definir el concepto.

Definición 2.16. Se dice que un modelo predictivo presenta *homocedasticidad* cuando la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones. Matemáticamente, $\sigma_i^2(x) = \sigma^2$.

Teorema 2.17. *La expresión exacta de la varianza del estimador local lineal es de la forma siguiente:*

$$\mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x \Omega P_x X}(\mathbf{X^T P_x X})^{-1}\mathbf{e_1}$$

donde en este caso, $\mathbf{\Omega}$ *representa una matriz diagonal n×n en cuya componente i-ésima se encuentra var*$(Y_i) = \sigma_i^2(x) = \sigma^2$.

Demostración. Usando la expresión (2.3) del Teorema 2.7, se tiene que

$$\text{var}(\hat{m}(x)) = \text{var}\left(\mathbf{e_1^T}(\mathbf{X^T P_x X})^{-1}\mathbf{X^T P_x Y}\right).$$

Aplicando a la susodicha expresión la propiedad siguiente:

$$\text{var}(\mathbf{A}X) = \mathbf{A}\text{var}(X)\mathbf{A^T},$$

donde $\mathbf{A} \in \mathbb{M}_{mxn}$, X es un vector aleatorio de dimensión n y var$(\cdot)$ es un operador que, al aplicarle el vector aleatorio X, var(X), nos devolverá la matriz de varianzas y covarianzas del susodicho vector,

obtenemos:

$$\mathbf{e_1^T(X^T P_x X)^{-1} X^T P_x} \text{var(Y)} \mathbf{P_x X(X^T P_x X)^{-1} e_1}$$
$$= \mathbf{e_1^T(X^T P_x X)^{-1} X^T P_x \Omega P_x X(X^T P_x X)^{-1} e_1}.$$

□

La fórmula proporcionada por el Teorema 2.17 nos permite crear gráficas que nos ofrecen la oportunidad de observar la varianza en distintos puntos y para distintos valores del parámetro ventana. Para este fin, vamos a volver a optar por el *heatmap* que tan buen resultado nos dio con el sesgo. Podemos encontrarlo a continuación.

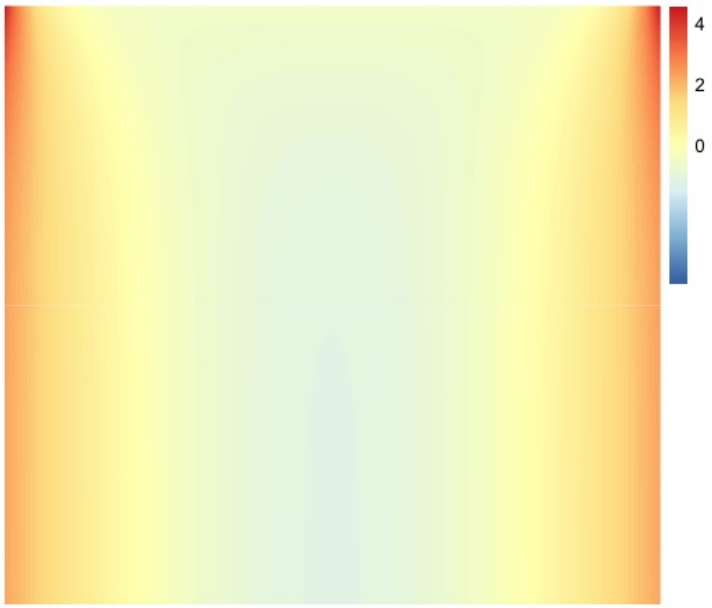

Figura 2.5: En esta gráfica se puede apreciar un *heatmap* en el que, partiendo de una muestra $x_1, ..., x_n$ de datos equiespaciados que van desde cero hasta uno con longitud $n = 100$ se ha calculado la varianza del estimador local lineal con un valor fijo de $\sigma^2 = 1$ en una rejilla 1000×1000, con $Y_i = m(x_i) + \varepsilon_i$, donde $\mathbb{E}(\varepsilon_i) = 0$. En el eje X aparece la varianza en un vector $r_1, ..., r_M$ de datos equiespaciados que van desde cero hasta uno con longitud $M = 1000$ para un valor fijo de h, y en el Y aparecen distintos valores del parámetro ventana, que aumentan desde la primera fila, en la que $h = 0{,}1$ y llegan de forma creciente y equiespaciada a la última fila, donde $h = 1$.

En la Figura 2.5 podemos encontrar el *heatmap* asociado a la varianza. Primero, explicaremos más en profundidad en qué consiste esta gráfica, para después continuar con un breve análisis de la misma. Para la realización del mapa de calor, hemos partido de una muestra $x_1, ..., x_n$ de datos equiespaciados que van desde cero hasta uno con longitud $n = 100$, hemos tomado un valor fijo $\sigma^2 = 1$, lo que convierte a la matriz Ω vista en el Teorema 2.17 en la matriz identidad, y hemos calculado la varianza del estimador

local lineal en un vector de puntos r de longitud $M = 1000$. Además, hemos repetido este proceso para $M = 1000$ valores de h y los hemos representado en las distintas filas, comenzando en la primera con un valor h tomado arbitrariamente y distinto de cero, $h = 0{,}1$, y avanzando de forma creciente y equiespaciada hasta llegar al valor $h = 1$, que se encuentra en la última fila. En resumen, en la casilla que se encuentra en la fila i-ésima y en la columna j-ésima de nuestra rejilla 1000×1000 podremos encontrar la varianza calculada en r_j para el valor i-ésimo de h. Dado que la varianza es siempre positiva, a diferencia de lo que hemos visto en el caso del sesgo, las casillas de la rejilla que tengan un color rojo más fuerte representan un valor mayor de la varianza, mientras que las que tengan un color más azulado representan un valor menor de la misma (observar la escala). Si ahora observamos el mapa de calor con la intención de analizarlo, podremos apreciar que la mayor varianza se encuentra en las casillas de las filas superiores, en las columnas de los extremos, y a medida que avanzamos hacia las columnas centrales, podemos ver que esta se va moderando. Una vez más, somos testigos del ya mencionado efecto frontera. Por otra parte, nos fijamos que esta vez al bajar por las filas de nuestra rejilla existe una disminución de la varianza, al revés de lo que nos ocurría para los mapas de calor del sesgo. Esto lo que nos dice es que con el aumento del valor del parámetro ventana se produce una disminución de la varianza. Presentemos entonces un teorema que nos permitirá confirmar nuestras conjeturas, pero antes, cabe mencionar que el mapa de calor de la Figura 2.5 ha sido reescalado para que el lector pueda apreciar de forma más nítida esta disminución de la varianza con el valor de h. Esto quiere decir que los valores numéricos de la varianza de cada casilla del mapa de calor no representarían la varianza real. El motivo es que lo nos interesa en este momento no es el valor exacto, sino la propiedad cualitativa que nos ofrece el *heatmap* y, en consecuencia, se han modificado los valores del mismo para que todos estén dentro de la misma escala, conservando los colores más rojos para los valores más altos y los azulados para los más bajos. De no haberse llevado a cabo esta pequeña modificación, resultaría muy complicado discernir los colores dentro del mapa de calor, impidiéndonos apreciar la disminución de la varianza a través de los colores.

Teorema 2.18. *Supongamos un modelo homocedástico y un diseño arbitrario, no necesariamente equiespaciado, que viene dado por la función $f(x)$, que es la función de densidad de los puntos observados y a la cual le pediremos que $f(x) > 0$. Por las condiciones 1) - 5) mencionadas al inicio de la subsección, el estimador local lineal tiene una varianza asintótica de la forma siguiente:*

$$var[\hat{m}(x)] = \frac{\sigma^2}{f(x)nh} \int_{-1}^{1} K(z)^2 dz + o\left(\frac{1}{nh}\right)$$

Observación 2.19. La demostración, que es relativamente laboriosa, incluye desarrollos similares a los empleados en la demostración del Teorema 2.10 y por eso se omite. Se puede encontrar la demostración de este teorema en *Kernel Smoothing* ([17]), concretamente en el Capítulo 5, Sección 5.3 del mismo.

Observación 2.20. Observamos que en la expresión presentada por el teorema, aparece $f(x)$. Esto se debe a que estamos enunciando el teorema para un diseño arbitrario, no necesariamente equiespaciado. Cabe destacar que, así como en el caso del sesgo el hecho de que el diseño sea equiespaciado o general no influye en la fórmula del mismo, en el caso de la varianza la fórmula sí que se vería ligeramente modificada al cambiar de un diseño general a uno equiespaciado.

A la vista del Teorema 2.18, resulta evidente que el parámetro ventana h también guarda una gran relación con la varianza de la estimación, pero en este caso, cuando aumenta el valor del parámetro ventana, la varianza disminuye y viceversa, al contrario de lo que nos pasaba con el sesgo. En conclusión, este teorema sí que nos permite fundamentar los cambios que se estaban produciendo en nuestro mapa de calor. Por otra parte, esto nos plantea un conflicto: dado que nos interesa mantener tanto el sesgo como la varianza lo más controlados posible, ¿optaremos por un valor de h más alto o más bajo? Para responder a esta cuestión, cerraremos la sección hablando del efecto de la ventana h.

La elección del parámetro ventana marcará una diferencia notable desde el primer momento en nuestra estimación, ya que influirá de manera directa en el sesgo y la varianza de la misma. Por este motivo, nos interesa encontrar una respuesta para el conflicto planteado en el párrafo anterior. Uno podría plantearse si existe un valor óptimo para el parámetro h que nos permitiese paliar este problema. La respuesta es que sí. Para poder hallar un valor óptimo del susodicho parámetro, tendríamos que ser capaces de llegar a un punto de encuentro entre el sesgo y la varianza. Para ello, una estrategia simple pero efectiva es definir el **Error Cuadrático Medio** en cada punto x:

$$\text{ECM}(x) = \mathbb{E}[\hat{m}(x) - m(x)]^2$$

que, como es conocido, es la suma de los términos del sesgo al cuadrado y la varianza. El error cuadrático medio es el segundo momento, sobre el orígen, del error y en consecuencia incorpora tanto el sesgo como la varianza del estimador, lo que lo convierte en una buena forma de medir la calidad del mismo. Tal como está definido dependerá del punto x, donde estamos realizando la estimación. Para crear una medida de rendimiento que represente el comportamiento global del estimador, los errores cuadráticos medios se pueden "sumar" sobre los puntos de diseño observados. Para allanar el camino para el análisis asintótico, es natural considerar el **Error Cuadrático Medio Integrado**, cuyo nombre en inglés nos proporciona las siglas que representan a la expresión en cuestión, MISE (Mean Integrated Squared Error):

$$\text{MISE} = \int \mathbb{E}[\hat{m}(x) - m(x)]^2 f(x) dx,$$

donde $f(x)$ representará la densidad de los puntos de diseño observados. MISE es una función que depende del parámetro h, y por ende es lógico pensar que para encontrar un valor óptimo del mismo, deberíamos encontrar el valor de h que minimice la función. Ahora proporcionaremos un teorema que nos dará el valor asintóticamente óptimo del parámetro ventana, junto con una demostración "heurística" del mismo. En la demostración que veremos a continuación, no trabajaremos con el sesgo y varianza explícitos, si no con sus aproximaciones asintóticas. Las aproximaciones en cuestión serían:

$$\mathbb{E}[\hat{m}(x)] \cong m(x) + \frac{1}{2} h^2 m''(x) \int z^2 K(z) dz, \tag{2.11}$$

$$\text{var}[\hat{m}(x)] \cong \frac{\sigma^2(x)}{f(x) n h} \int K(z)^2 dz. \tag{2.12}$$

Por este motivo introducimos AMISE. AMISE representa las siglas *Asymptotic Mean Integrated Squared Error*, es decir, *Asymptotic MISE* y sería, esencialmente, MISE sin contar con los términos más pequeños para muestras grandes. MISE = AMISE + error. Una vez visto esto, estamos en condiciones de presentar el teorema siguiente:

Teorema 2.21. *Supongamos un modelo no necesariamente homocedástico y con diseño general. Usando las aproximaciones asintóticas dadas por las ecuaciones* (2.11) *y* (2.12) *en la evaluación del AMISE, llegamos a un valor óptimo para el parámetro ventana h que puede expresarse de la forma siguiente:*

$$h_{OAMISE} = \left[\frac{\int \sigma^2(x)dx \frac{\int K^2(z)dz}{\left(\int z^2 K(z)dz\right)^2}}{n \int f(x)(m''(x))^2 dx} \right]^{\frac{1}{5}},$$

donde h_{OAMISE} representa el valor óptimo de h que viene dado por AMISE.

Observación 2.22. El lector encontrará una demostración más completa del resultado en el Teorema 2 de *Design-adaptive Nonparametric Regression* ([7]).

Demostración. Combinando (2.11) y (2.12) y teniendo en cuenta que el Error Cuadrático Medio se puede expresar como sesgo^2 + varianza, podemos reescribir AMISE como:

$$AMISE = \int \left(\frac{h^4}{4} \left(\int z^2 K(z)dz \right)^2 (m''(x))^2 + \frac{\sigma^2(x)}{nh} \frac{\int K^2(z)dz}{f(x)} \right) f(x)dx$$

$$= \int \frac{h^4}{4} f(x) \left(\int z^2 K(z)dz \right)^2 (m''(x))^2 dx + \int \frac{\sigma^2(x)}{nh} \int K^2(z)dzdx.$$

Extraigamos de las integrales los valores que no dependen de x y derivemos con respecto de h para despues igualar a cero y de este modo obtener el valor mínimo:

$$\frac{d}{dh}\left[\frac{h^4}{4}\right] \left(\int z^2 K(z)dz \right)^2 \int f(x)(m''(x))^2 dx + \frac{d}{dh}\left[\frac{1}{h}\right] \frac{\int K^2(z)dz}{n} \int \sigma^2(x)dx$$

$$= h^3 \left(\int z^2 K(z)dz \right)^2 \int f(x)(m''(x))^2 dx - \frac{1}{h^2} \frac{\int K^2(z)dz}{n} \int \sigma^2(x)dx = 0$$

$$\Leftrightarrow h^3 \left(\int z^2 K(z)dz \right)^2 \int f(x)(m''(x))^2 dx = \frac{1}{h^2} \frac{\int K^2(z)dz}{n} \int \sigma^2(x)dx$$

$$\Leftrightarrow h^5 = \frac{\int \sigma^2(x)dx \frac{\int K^2(z)dz}{\left(\int z^2 K(z)dz\right)^2}}{n \int f(x)(m''(x))^2 dx}$$

$$\Leftrightarrow \boxed{h_{OAMISE} = \left[\frac{\int \sigma^2(x)dx \frac{\int K^2(z)dz}{\left(\int z^2 K(z)dz\right)^2}}{n \int f(x)(m''(x))^2 dx} \right]^{\frac{1}{5}}} \tag{2.13}$$

□

Observemos el valor que acabamos de obtener. Para empezar, nos fijamos en la n que hay en el denominador de la fracción, ¿qué ocurría con h_{OAMISE} si vamos cambiando sus valores? Supongamos que n fuese muy grande, es decir, supongamos que $n \to \infty$. En este caso, tendríamos que h_{OAMISE} sería cada vez más pequeña, mientras que si, por el contrario, n se hiciese cada vez más pequeña, el valor de h_{OAMISE} aumentaría. Esto es, el valor óptimo de nuestro parámetro ventana tenderá a ser mayor en regresiones con menor número de observaciones que en aquellas que cuentan con un tamaño muestral mayor, para las que h_{OAMISE} será más pequeño. Dirijamos el foco ahora hacia la derivada segunda de m que se encuentra dentro de la integral, en el denominador, y preguntémonos qué ocurriría si m fuese una función lineal. Tal y como hemos visto antes, si m fuese una función lineal, su sesgo sería nulo y por este motivo no tendríamos el conflicto que comentamos previamente de tratar de minimizar el sesgo y la varianza simultáneamente. Ahora únicamente nos preocupará la varianza, luego tendremos que tomar el valor del parámetro ventana lo más alto posible para minimizarla. Pensemos en esto, si tuviésemos una función lineal, quizá no nos interesaría un enfoque local lineal, sino uno global, y esta es la idea que hay tras esta selección de la h óptima para funciones lineales. La h, al final, lo que determina es cuán grande es el entorno en el que estamos trabajando, y si tenemos una función lineal, nos interesará un valor de h lo más grande posible, y por ende un entorno del punto lo más grande posible, optando por un enfoque lineal global. Por otra parte, así como la derivada segunda nos aportaba un signo en la fórmula del sesgo, podemos ver que en esta se encuentra elevada al cuadrado, por lo que no contará el signo de la misma para modificar el de h_{OAMISE}, lo que a su vez tiene sentido, ya que recordemos que el parámetro ventana es siempre positivo.

Uno puede observar que la h_{OAMISE} que acabamos de obtener en la ecuación (2.13) no resulta demasiado útil para la práctica, ya que depende de la curva desconocida $m(x)$, en particular de su curvatura, además de la densidad del diseño. Es por esto que tendremos que buscar una alternativa para aplicarla en el apartado de análisis de datos, cuando calculemos el valor óptimo del parámetro ventana para nuestros datos. Lo que haremos será aplicar el método de la **validación cruzada** o *cross validation* en inglés.

Este método se basa en construir una estimación empírica del MISE y minimizarla sobre h. La filosofía que aplica es tratar de predecir cada valor respuesta, Y_i, a partir del resto de los datos. Para el valor Y_i, podemos denotar a la función que se encargará de su predicción por $\hat{m}_{-i}(x_i)$, donde el subíndice " -i " indica que la observación (x_i, Y_i) ha sido omitida a la hora de calcular el estimador local lineal. Por ende, podemos definir la función de validación cruzada de la siguiente manera:

$$\text{VC}(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{-i}(x_i)\}^2.$$

Seleccionaremos h minimizando la función previa, VC(h). Terminaremos la sección con el siguiente lema. El interés del mismo reside en el hecho de que nos va a permitir conectar la validación cruzada con el MISE estudiado anteriormente.

Lema 2.23. *Supongamos un modelo homocedástico. Entonces:*

$$\mathbb{E}\left(VC(h)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{\hat{m}_{-i}(x_i) - m(x_i)\}^2 + \sigma^2,$$

donde σ^2 representa la varianza.

Demostración. $\mathbb{E}\left(\mathrm{VC}(h)\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{-i}(x_i)\}^2\right)$

$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(x_i) + m(x_i) - \hat{m}_{-i}(x_i)\}^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{\hat{m}_{-i}(x_i) - m(x_i)\}^2 + \sigma^2.$ □

Observación 2.24. Podemos apreciar que el término $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{\hat{m}_{-i}(x_i) - m(x_i)\}^2$ aproxima numéricamente la integral vista en el MISE, a saber: $\int \mathbb{E}[\hat{m}(x) - m(x)]^2 f(x)dx$. Además, se le añade una constante σ^2, la cual no tiene importancia en la búsqueda del mínimo en h, ya que no depende de esta variable. Es por esto que el método de la validación cruzada provee un estimador razonable de MISE.

2.3. SiZer

En la sección previa hemos hecho hincapié en que el parámetro ventana regula la suavidad de la estimación. Recordemos que cuanto menor era el valor del susodicho parámetro, más rugosa sería nuestra estimación, debido a un sobreajuste que podría derivar en la aparición de modas que, realmente, no existirían con una elección más moderada del parámetro ventana. Por el contrario, tomando un valor grande de h, la estimación se suaviza sobremanera, pero ahora nos ocurre lo contrario que nos pasaba para un valor pequeño, que es que podríamos terminar obviando elementos de la estimación que sí son significativos. En este caso, ¿qué valores de h nos permiten fiarnos de las características que aparezcan en nuestra estimación, para no caer en alguna falsa o no saltarnos ninguna real? Una forma sería volver a hablar del valor óptimo de h, pero en su lugar, introduciremos el SiZer, ya que este nos ofrece una clara ventaja con respecto a otros métodos, y es que mediante el mismo seremos capaces de comparar una larga serie de valores de h simultáneamente, evitando así el momento de elegir un solo valor para el susodicho parámetro.

Un SiZer es una representación de un mapa de colores que muestra mediante los mismos en qué puntos la curva promedio crece y en cuáles decrece significativamente. Cada SiZer contará con un máximo de cuatro colores posibles, uno de los cuales representará un crecimiento significativo, otro un decrecimiento significativo, otro que no hay ni un crecimiento ni decrecimiento significativo y el último representará la ausencia de datos suficientes para llegar a una conclusión. Una vez presentada la motivación y la idea subyacente del SiZer, tratémoslo matemáticamente.

Consideremos el estimador con el que veníamos trabajando hasta el momento, el visto en el Teorema

2.7. La idea principal consiste en buscar intervalos de confianza para $\hat{m}'(x)$, que adoptará la siguiente expresión:

$$\hat{m}'(x) = \arg\min_b \sum_{i=1}^{n} (Y_i - a - b(x_i - x))^2 K_h(x_i - x),$$

donde el mínimo, al igual que en el Teorema 2.7, lo calculamos sobre a y b para después quedarnos con b, cuya expresión del mínimo, β, se encuentra explicitada en el Teorema 2.7. Recordemos que en este teorema afirmamos que $\hat{m}(x) = \alpha$. Ahora, en esta sección afirmamos que $\hat{m}'(x) = \beta$, donde β es el argumento donde b alcanza su mínimo. Basaremos el estudio visual de las características significativas de nuestro modelo para una familia de funciones $\{\hat{m}(x), h \in [h_{min}, h_{max}]\}$ en los límites de confianza para la derivada $\hat{m}'(x)$. Antes de continuar, introducimos el concepto de *plano espacio - escala*, que no es más que un plano cuyo eje X viene determinado por los puntos x correspondientes a la covariable y cuyo eje Y contiene los valores del parámetro ventana h para los cuales vamos a realizar el contraste. Dicho esto, el comportamiento en un punto arbitrario (x, h) del plano espacio - escala se presentará mediante un color determinado en el mapa de color que nos ofrece el SiZer, donde repetimos, existen cuatro colores en función de si $\hat{m}'(x)$ es significativamente positiva, negativa, ninguna de las anteriores o no hay suficientes datos. Los límites de confianza para la susodicha derivada son:

$$\hat{m}'(x) + q \cdot \hat{SD}(\hat{m}'(x)), \tag{2.14}$$

$$\hat{m}'(x) - q \cdot \hat{SD}(\hat{m}'(x)), \tag{2.15}$$

donde q representa un cuantil apropiado, que puede ser calculado de diversas formas. Una forma habitual es usar la aproximación normal, la cual, para completar la información, añadimos que se puede aplicar puntualmente o por bloques de puntos. Una localización (x, h) en el plano espacio - escala se dice que es significativamente creciente, decreciente o no significativo cuando el cero está bajo, sobre o dentro de estos límites de confianza, respectivamente.

Ilustramos lo visto en la sección con dos sencillos ejemplos que podremos encontrar en la Figura 2.6 y en la Figura 2.7. En la primera podemos ver un mapa de colores SiZer para los datos de concentración de CO_2 e índice de temperatura media que corresponden a las variables X e Y, respectivamente. Dado que este primer gráfico quizás no nos resulte del todo completo dada la forma de los datos, hemos realizado otro gráfico SiZer, cuya variable X será un conjunto de datos equiespaciados que van desde cero hasta uno con longitud $n = 100$, y cuya variable Y corresponderá a los datos resultantes de aplicarle la función $f(x) = \sin(2\pi x)$ a nuestra variable X. Además, vamos a comparar la gráfica correspondiente a este último SiZer con otra en la que simplemente representaremos las variables X e Y en conjunto y que también podremos encontrar en la Figura 2.7. Para obtener los mencionados mapas de calor utilizamos una librería de R, llamada *SiZer*, en la cual no profundizaremos. Si el lector deseara más información sobre la misma, puede consultar [15]. Una vez presentada la librería que nos permitirá la creación de los gráficos, veamos el primer resultado, pero antes, a modo de ejemplo, especificaremos el código necesario para obtener los gráficos presentes en la Figura 2.7:

```
> n=100
> x<-seq(0, 1, len=n)
> f<-function(x) sin(2*pi*x)
> y<-f(x)
> plot(x,y,pch = 19, col = "black", cex = 0.6, xlab = "", ylab = "")
> SiZer.1 <- SiZer(x, y, h=c(0.05,0.2), degree=1, derv=1, grid.length
    =100, quiet = FALSE)
> plot(SiZer.1)
> plot(SiZer.1, ggplot2=TRUE)
> ggplot_SiZer(SiZer.1)
```

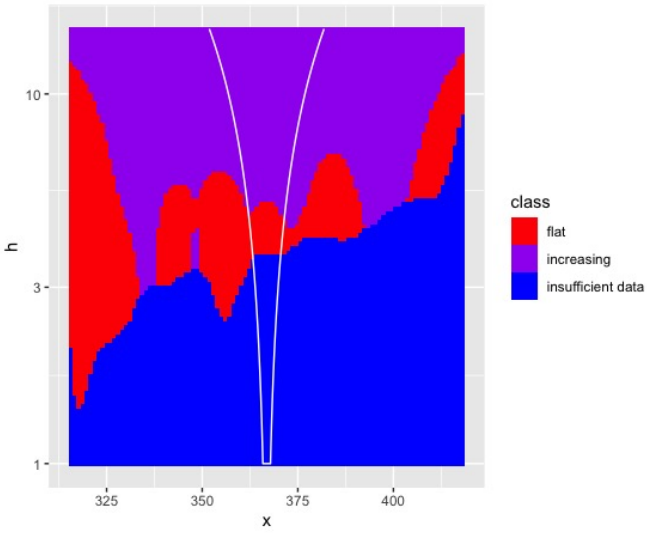

Figura 2.6: SiZer para los datos Concentración CO_2 - Temperatura.

Antes de comenzar a comentar el SiZer, cabe destacar algo que ya dijimos en un párrafo previo de la sección, y es que podemos ver que, dado que se trata de un plano espacio - escala, en el eje X se encuentran los datos correspondientes a la concentración de CO_2 atmosférico (nuestra variable X) y en el eje Y se encuentran los distintos valores del parámetro h sobre los que se realizará el estudio. Observamos que para valores de h cercanos a 1, los datos son insuficientes para cualquier valor de x, mientras que para los valores superiores a 10, existe un crecimiento significativo para la curva promedio, para casi todo valor de x. Aquí es necesario recordar que para valores grandes de h puede haber un gran sesgo, y el promedio estar muy alejado de la curva de regresión. Por lo demás, el gráfico nos muestra claramente los puntos (x, h) del plano espacio - escala en los que hay crecimiento significativo (morado), datos insuficientes (azul) o ni crecimiento ni decrecimiento significativo (rojo). En otro orden de cosas, una observación a tener en cuenta es que el valor óptimo de h para estos datos debería ser, aproximadamente, mayor que 3.5, dado que a partir de este valor parece que podemos empezar a observar datos suficientes

para extraer conclusiones. Esto último lo comprobaremos más adelante, en el próximo capítulo.

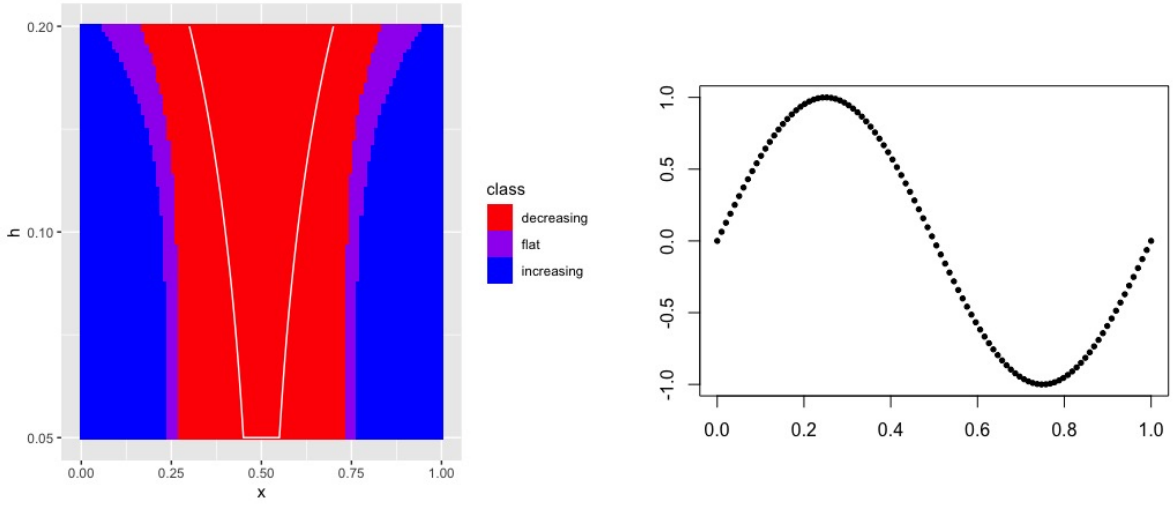

Figura 2.7: De izquierda a derecha: SiZer para los datos correspondientes a la función $f(x) = \sin(2\pi x)$ y gráfica de las variables X e Y para las cuales hemos realizado el SiZer.

En este caso, en el eje X de nuestro plano espacio - escala se encuentra el conjunto de datos equiespaciados que van desde cero hasta uno con longitud $n = 100$, mientras que en el eje Y se encuentran una vez más los valores del parámetro h sobre los que realizaremos el estudio mediante el SiZer. Comentemos brevemente la Figura 2.7. Vemos que en ella, además del SiZer, hay una gráfica que representa la función $f(x) = \sin(2\pi x)$ para el vector de puntos previamente mencionado. El interés de aportar dicha gráfica reside en poder comparar ambas imágenes, con el fin de observar si somos capaces de ver una relación entre las mismas. Observemos los valores de h más pequeños, en los que el sesgo será menor y existirá un mayor ajuste de la estimación a los puntos (presentes en la gráfica de la derecha). En estos valores de h, vemos que entre 0 y $0{,}2$ y entre $0{,}8$ y 1 existe un crecimiento significativo, lo que se ve respaldado por la gráfica de la derecha. Por otra parte, entre $0{,}3$ y $0{,}7$ vemos un decrecimiento significativo, una vez más apoyado por la gráfica de los puntos. Finalmente, en los intervalos restantes, no se aprecia ni un crecimiento ni un decrecimiento significativo, algo que también se ve respaldado por la gráfica de la derecha, ya que observemos que estos intervalos se corresponden, aproximadamente, con los máximos y los mínimos que traza la susodicha función. Por otra parte, a medida que aumentamos el valor de h, aumentamos también el sesgo y por lo tanto estos puntos en los que no hay ni crecimiento ni decrecimiento significativo serán más abundantes, ya que cada vez existirá un menor ajuste a los puntos y las zonas de crecimiento y decrecimiento serán menos pronunciadas. Con este vistazo al SiZer, cerramos el presente capítulo, para dar paso al siguiente.

Capítulo 3

Análisis de datos

El análisis de los datos presentados en el Capítulo 1 lo dividiremos en tres secciones, donde cada una de ellas representará el estudio de un emparejamiento distinto de los datos, que dará nombre a su respectiva sección. Los estudios que se verán a continuación se llevarán a cabo usando el lenguaje de programación R.

Antes de comenzar, faltan por añadir dos detalles técnicos. El primero es que el kernel que usaremos a lo largo del capítulo será el **kernel Gaussiano**, visto en el Capítulo 2. Por otra parte, el segundo, es que el estimador que utilizaremos sera el **estimador local lineal**. Una vez aclarados estos puntos, procedamos con el análisis de los datos.

3.1. Asociación entre temperatura y niveles de CO_2

El primer conjunto que analizaremos será el correspondiente al CO_2, medido en partes por millón (ppm), e índice de temperatura media anual, medida en ºC. La primera será nuestra variable explicativa X y la segunda la variable respuesta Y. Abrimos la sección con una representación conjunta de los datos del CO_2 y temperatura emparejados por año. Además, añadimos otra igual a la que le superponemos la recta de mínimos cuadrados. Todo esto se puede apreciar en la Figura 3.1.

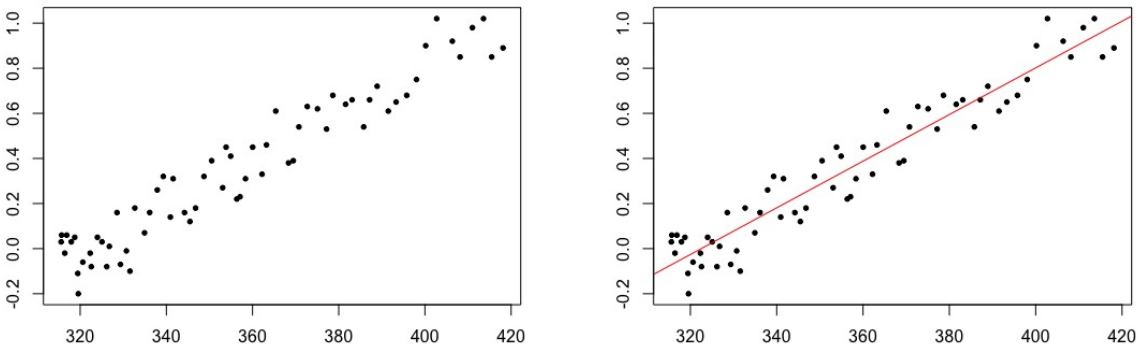

Figura 3.1: En la gráfica de la izquierda podemos ver una representación de los datos. En la gráfica de la derecha podemos ver una representación de los datos a la que se le ha añadido la recta de mínimos cuadrados. En el eje X la concentración de CO_2 está medida en ppm y en el eje Y la temperatura se mide en ºC.

A continuación, calcularemos y representaremos la curva resultante del estimador local lineal con el kernel Gaussiano para los datos que acabamos de ver. Es interesante poder comparar las estimaciones para los distintos valores de h. Es por esto que realizaremos la respresentación gráfica para cuatro valores distintos de h, a saber: $0{,}3$, 1, 10 y 100. Hemos elegido estos valores ya que nos permiten observar la diferencia entre las estimaciones de forma nítida. A su vez, el motivo de fijar un valor de h tan alto como 100, es que se pueda apreciar la similitud entre la estimación resultante y la recta de mínimos cuadrados, plasmada en la gráfica de la derecha de la Figura 3.1. Las cuatro representaciones mencionadas pueden verse en la Figura 3.2.

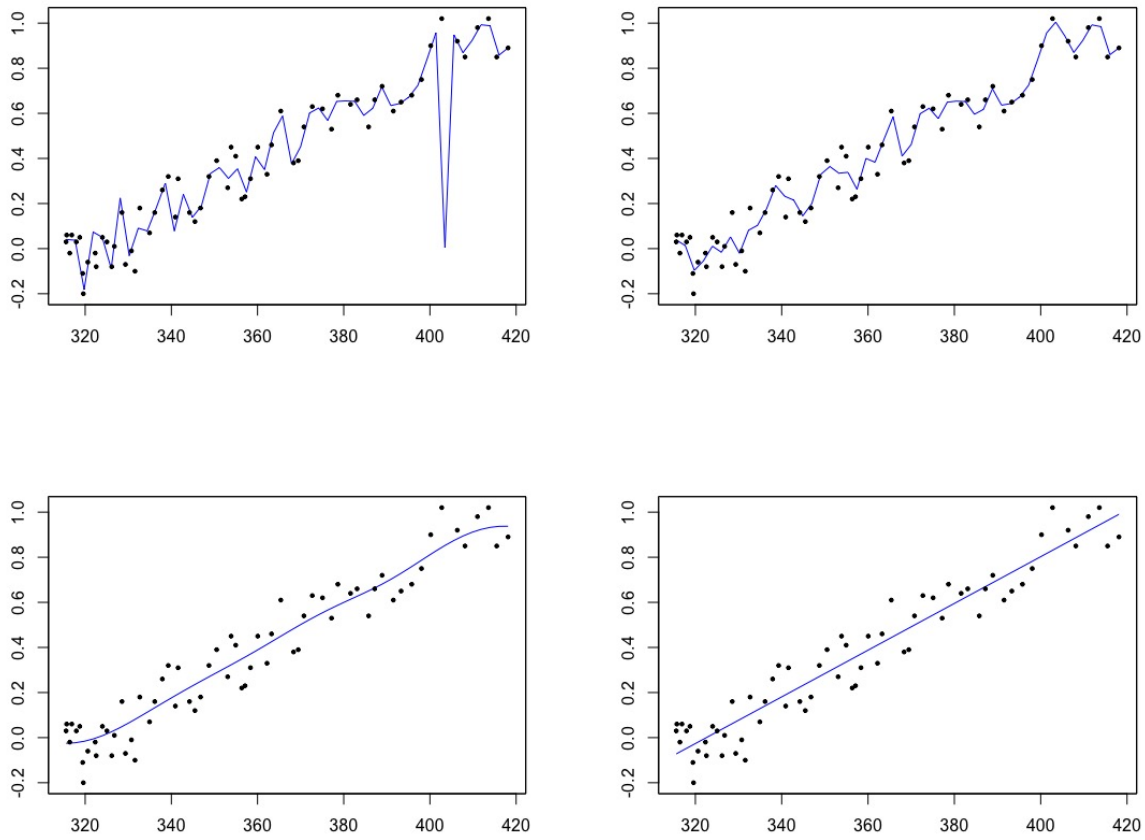

Figura 3.2: De izquierda a derecha y de arriba a abajo: estimación de la curva con $h = 0{,}3$, estimación de la curva con h = 1, estimación de la curva con $h = 10$ y estimación de la curva con $h = 100$.

Se aprecia de una forma muy clara cómo la estimación es cada vez más suave hasta llegar al valor 100. Si observamos la gráfica correspondiente a dicho valor y la comparamos con la proporcionada al principio que contenía la recta de mínimos cuadrados, no podremos apreciar una gran diferencia. En otro orden de cosas, en un punto del Capítulo 2, mencionamos la utilidad del **método de la validación cruzada** para obtener el valor óptimo del parámetro ventana en la práctica. Por lo tanto, calculemos dicho valor para nuestros datos mediante el mencionado método y hagamos una estimación más tomando el valor de h en cuestión. Podemos observar el resultado en la Figura 3.3.

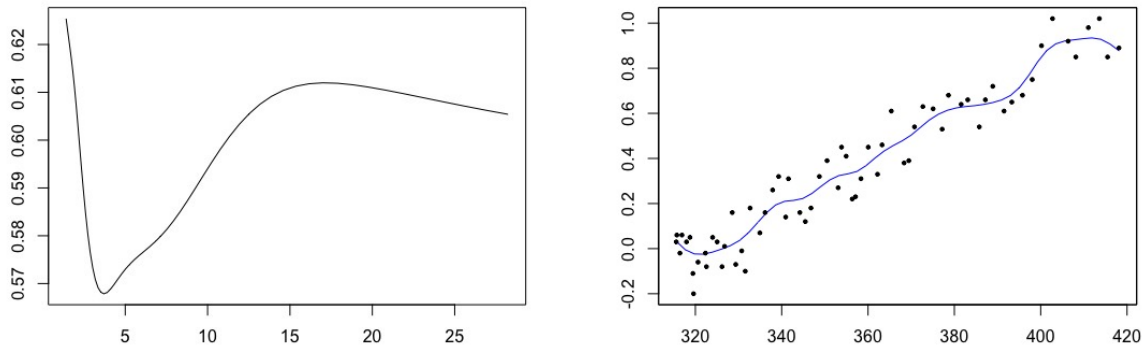

Figura 3.3: De izquierda a derecha: gráfica obtenida a partir del método de la validación cruzada que nos permite obtener el valor óptimo de h y gráfica de una nueva estimación utilizando el valor de h óptimo, obtenido mediante validación cruzada.

Como vemos en la gráfica de la izquierda de la Figura 3.3, el valor de la ventana de validación cruzada es $h = 3{,}71$, que es el valor empleado en la estimación de la derecha. Para finalizar el análisis de este conjunto de datos, vamos a volver sobre los dos primeros gráficos de la sección, es decir, a la Figura 3.1. Deteniéndonos a observar el patrón que siguen los puntos en el espacio, podría considerarse de interés realizar un contraste de linealidad, ya que gráficamente tienen un aspecto sugerente. Por este motivo, implementaremos en R un test cuya idea general se basa en medir diferencias entre el estimador lineal paramétrico y otro no paramétrico. No entraremos en detalles más específicos acerca del mismo, pero antes de proceder a su utilización, añadiremos que pertenece a la librería *sm* (ver [2]). Aplicando el susodicho test a nuestros datos, obtenemos un p-valor de $0{,}202$, valor muy alto, por lo que no existirían evidencias estadísticamente significativas para anteponer un modelo no paramétrico a uno lineal. En consecuencia, para estos datos podemos aceptar como válido el modelo lineal. A pesar de ello, nos han sido realmente útiles para ilustrar todo lo visto hasta el momento.

A la vista de esta nueva característica de los datos, podemos calcular ciertas medidas adicionales, como puede ser la correlación muestral entre ambas variables (denotada por R) o el coeficiente de determinación (denotado por R^2), que nos permitirá ver la proporción de variación de los datos que puede explicarse mediante el modelo y a su vez determinar la calidad del modelo para replicar resultados. Si calculamos el primero, obtenemos que R = *0.9596*, que es un valor realmente alto. La correlación entre dos variables es un valor que se encuentra entre -1 y 1, siendo los valores más proximos a 1 los que representan una mayor relación lineal directa entre las variables, los más próximos a -1 los que representan una mayor relación lineal inversa y los más próximos a 0 los que marcan que las variables no guardan una relación lineal. Nuestra correlación para las variables X e Y está realmente próxima a 1, lo que indica

una gran relación lineal directa entre ambas. Centrémonos ahora en el coeficiente de determinación, que resulta muy sencillo de obtener, dado que se trata del cuadrado del coeficiente de correlación. Este se mueve entre 0 y 1, siendo los valores más próximos a 1 los que representan una mayor bondad de ajuste, ya que se explicará una mayor proporción de la variación de los datos mediante el modelo. Si realizamos el cálculo, llegamos a un valor de R^2 = *0.9208*. Tal y como era de esperar, se trata de un valor realmente alto, por lo que podemos concluir que nuestro modelo es de alta calidad y que una gran proporción de la variación de los datos se puede explicar mediante el mismo.

En conclusión, nuestro modelo nos permite ver que, tal y como hemos llevado a cabo el estudio, el índice de la temperatura media anual y la concentración de CO_2 atmosférico guardan una gran relación directa lineal.

Observación 3.1. Es necesario hacer hincapié en que las conclusiones a las que hemos llegado son fruto únicamente de los procesos que hemos seguido. Por supuesto que puede haber alguna variable confusora que esté alterando los resultados, así como factores que no hayan sido tenidos en cuenta. Es por esto que remarcamos que los resultados que se concluyen en esta sección y las que vienen, hay que entenderlos dentro del contexto de un TFG.

3.2. Asociación entre temperatura y niveles de CH_4

A continuación trataremos los datos correspondientes a otro de los gases de efecto invernadero de más importancia en los estudios, el metano (CH_4). Esta será nuestra variable X, medida en partes por billón (ppb) y mantendremos nuestra variable respuesta Y, el índice de la temperatura media anual, medida en ºC. Dado que tenemos entre manos un nuevo conjunto de datos, comenzamos el estudio con la representación gráfica de los mismos. Se puede observar que esta vez la linealidad no es tan evidente, y por lo tanto vamos a empezar llevando a cabo el contraste de modelo lineal frente a modelo no paramétrico visto en la sección anterior y representando el gráfico resultante. Podemos verlo en la Figura 3.4.

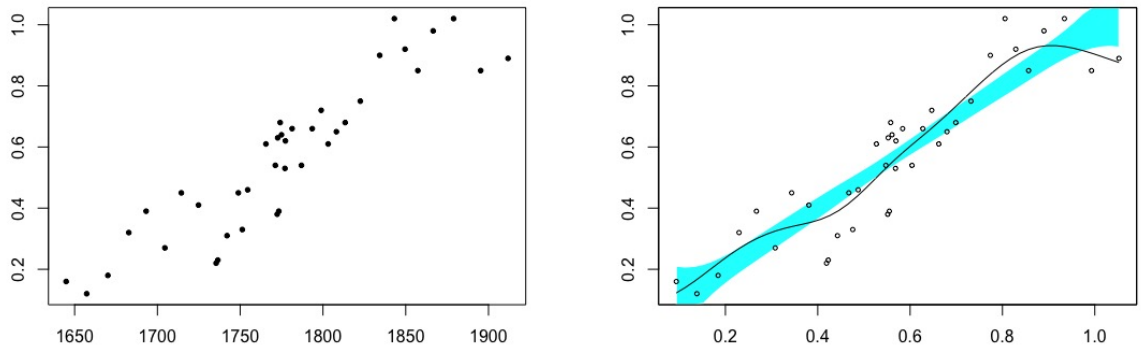

Figura 3.4: De izquierda a derecha: gráfico de los datos iniciales y gráfico correspondiente al contraste de linealidad de los datos.

En la imagen de la derecha de la Figura 3.4 podemos ver el mencionado gráfico que nos proporciona el test. En él se puede apreciar una superficie de color azul. Recordemos que este test está comprobando si nuestros datos se ajustan a un modelo lineal, y eso es justo lo que representa la zona azul, unas bandas de referencia que nos permitirán ver gráficamente si contienen a una recta. En cuanto a los resultados analíticos, tal y como augurábamos, el p-valor asociado al mencionado contraste de modelo lineal frente a modelo no paramétrico es de *0.008*. Por lo tanto, podemos concluir que existen evidencias estadísticamente significativas para optar por un modelo no paramétrico frente a uno lineal, ya que se rechaza linealidad para los valores usuales del 1 %, 5 % y 10 %.

Continuamos el estudio representando la curva trazada por nuestro estimador para nuestros nuevos datos con tres valores distintos de h. A saber: $1{,}75$, 10 y el valor óptimo de h, seleccionado por validación cruzada. En esta sección nos topamos con una peculiaridad, y es que realizando la validación cruzada tal y como veníamos haciendo hasta el momento, llegamos a un valor de $h_{opt} = 3{,}49$, que proveería una estimación demasiado rugosa. Es por esto que hemos optado por una validación cruzada local para valores de h entre 5 y 40 y de esta forma hemos obtenido $h_{opt} = 5{,}78$, que proporciona una estimación más satisfactoria. El resto de valores han sido elegidos de tal forma que pueda compararse la estimación correspondiente a h_{opt} con otra más lisa (h más grande) y otra más rugosa (h más pequeña). Podemos encontrar los gráficos asociados tanto a las distintas regresiones, como a la curva de h resultante de la validación cruzada local en la Figura 3.5.

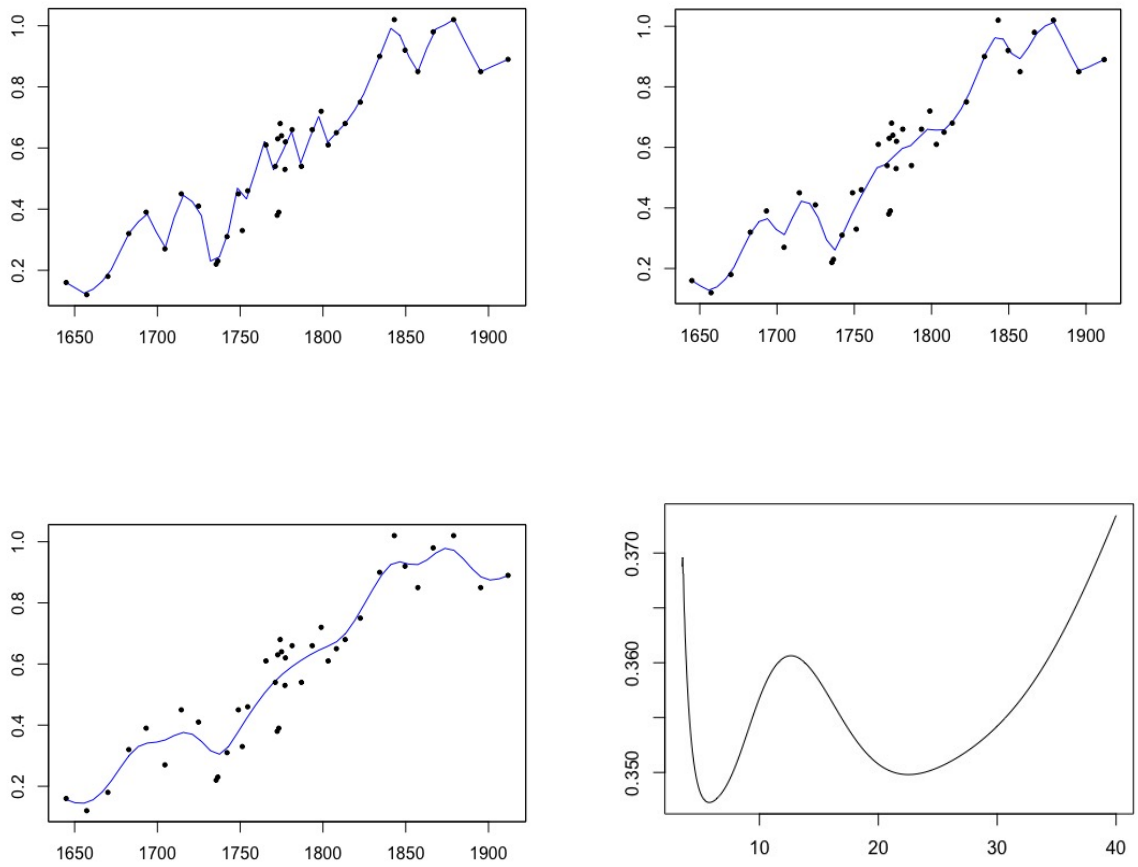

Figura 3.5: De izquierda a derecha y de arriba a abajo: estimación con $h = 1{,}75$, estimación con $h = h_{opt}$, estimación con $h = 10$ y gráfica asociada al método de validación cruzada local.

Podemos apreciar que a pesar de no ser lineal, existe una tendencia creciente a lo largo de la curva estimada, lo que querría decir que existe una cierta relación directa entre ambas variables. Es decir, que a medida que aumenta la concentración atmosférica de CH_4, también lo haría el índice de la temperatura media anual. Sin embargo, como ya hemos señalado antes y podemos reiterar una vez vistas las estimaciones de la curva, no se trata de una curva lineal, y por tanto no podríamos hablar de una relación lineal simple o relación lineal directa, como hemos hecho en la sección anterior. En lo que a la suavidad se refiere, podemos apreciar un nivel de suavidad satisfactorio en las estimaciones realizadas con $h = h_{opt}$ y $h = 10$, pero podría afirmarse que para el valor de $h = 1{,}75$ nuestra curva presenta ciertos cambios bruscos en la monotonía, incluso algunos picos prominentes, que nos permitirían inferir que nuestra estimación para este valor de h es ciertamente rugosa. Por lo demás, la estimación de la curva se ajusta satisfactoriamente a los puntos, es decir, captura la variabilidad de manera adecuada.

Continuando por este camino, pasaremos a estudiar la variabilidad del modelo. Para ello, nos valdremos de un recurso que no presentamos en la sección teórica, pero nos tomaremos ahora la libertad de explicarlo brevemente para posteriormente utilizarlo en el estudio. Hablamos de las bandas de variabilidad. Estas indican el nivel de variabilidad con el que cuenta un estimador de regresión no paramétrica, sin tratar de ajustar el sesgo existente. Las bandas de este tipo no tienen una construcción compleja, pero han de ser interpretadas con cuidado. Para empezar, hay que remarcar su carácter **puntual**, es decir, no actúan globalmente, y lo que hacen es indicar los intervalos de confianza puntuales para $\mathbb{E}[\hat{m}(x)]$, no para $m(x)$. Recordemos que $\hat{m}(x)$ tiene sesgo, que puede ser importante para valores grandes de h y modelos no lineales. Para calcularlas, el único requerimiento es una estimación de la varianza de $\hat{m}(x)$, y serán construidas introduciendo el doble del valor de la desviación típica estimada tanto por encima como por debajo del estimador, suponiendo un comportamiento asintóticamente normal. A continuación representaremos las bandas de variabilidad para los valores de h que utilizamos para la estimación de las curvas. A su vez, traeremos de vuelta el SiZer que presentamos en la Sección 2, con el fin de ver si nos da alguna información adicional que nos ayude a enriquecer el estudio. Todo esto podremos encontrarlo en la Figura 3.6.

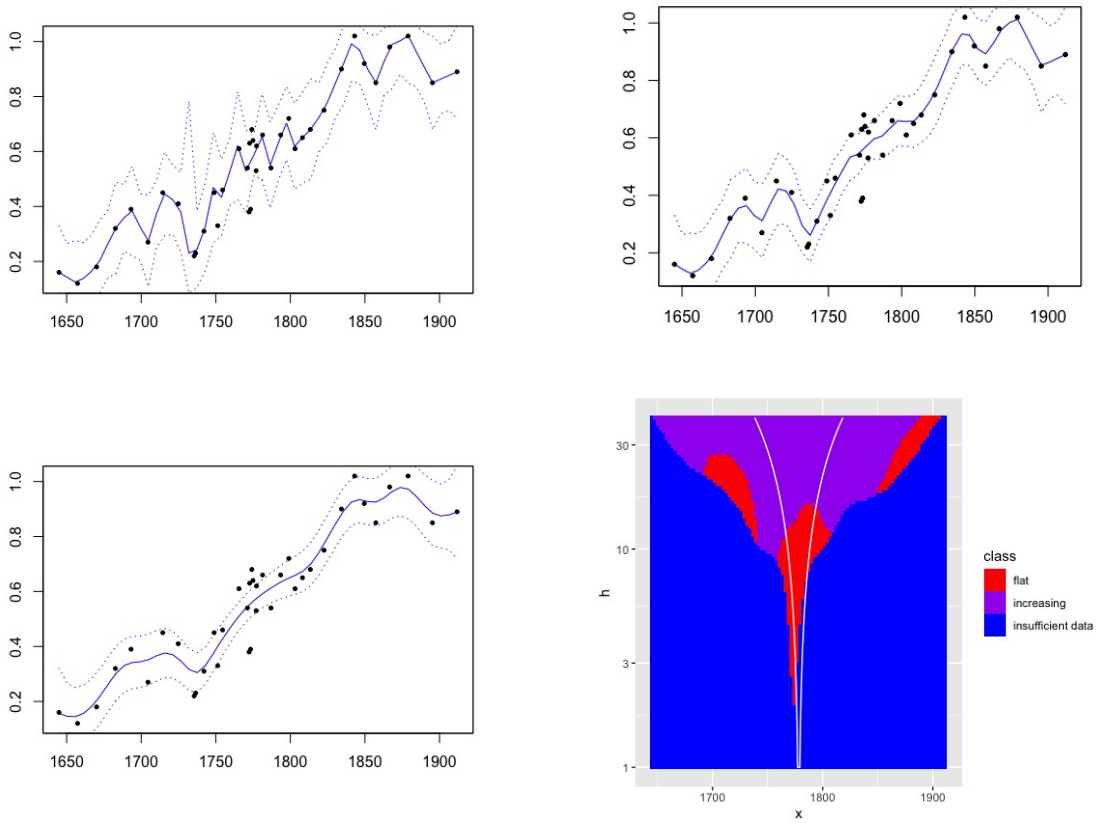

Figura 3.6: De izquierda a derecha y de arriba a abajo: bandas de variabilidad con $h = 1{,}75$, bandas de variabilidad con $h = h_{opt}$, bandas de variabilidad con $h = 10$ y SiZer correspondiente a estos datos.

Podemos observar la diferencia que existe entre la anchura de las bandas de las dos últimas estimaciones y la correspondiente a $h = 1{,}75$. Esto se debe a que, tal y como hemos mencionado más arriba, la anchura de las bandas de variabilidad vendrá determinada por la varianza de $\hat{m}(x)$, y recordemos que esta se reducía con el aumento del valor del parámetro ventana. En relación al SiZer, podemos afirmar que los valores de h inferiores a 10 no nos van a proporcionar ninguna información de utilidad. Si nos vamos a valores más altos, a saber: desde $h = 20$ en adelante, ya empezamos a ver resultados concluyentes, esto es, se puede apreciar que existe un crecimiento significativo en la mayoría de los puntos. A su vez, vemos que la baja vista en la gráfica correspondiente a la estimación con $h = 10$ de la Figura 3.5 no sería significativa, tal y como cabía esperar. Es más, no se observa ningún punto en el cual se produzca un decrecimiento significativo en todo mapa de colores.

En conclusión, hemos visto que las variables no guardan una relación lineal, pero dada la tendencia creciente de las curvas, sin importar el valor de h, si se podría confirmar una relación directa entre ambas. Es por este motivo que basándonos en el proceso seguido hasta el momento y siempre teniendo en cuenta que no contábamos con una gran cantidad de datos, diríamos que la concentración de CH_4 sí presenta una asociación creciente con el índice de temperatura media anual. Hacemos especial hincapie en que estas conclusiones las estamos extrayendo de los procesos vistos a lo largo de la sección. Cabe la posibilidad de que haya alguna variable confusora, o de que exista la influencia de otros factores no tenidos en cuenta. Es un problema de naturaleza multifactorial compleja cuyo análisis está fuera de los objetivos de un TFG.

3.3. Asociación entre nivel del mar y temperatura

En esta sección, nuestra covariable, X, será la que hasta ahora ha representado a nuestra variable Y, el índice de la temperatura media (ºC), y nuestra variable respuesta Y será el nivel del mar (medida en cm). El motivo de tratar de relacionar estos dos conjuntos de datos es que podemos hipotetizar que al aumentar la temperatura, se funde la capa de hielo del planeta, transformándola en agua y haciendo así que aumente el nivel del mar. El esquema que seguiremos será similar a los dos anteriores. Dicho esto, comencemos plasmando el gráfico de los datos para formarnos una idea visual de los mismos. Para este nuevo conjunto, resulta de gran interés comprobar la conveniencia de optar por un modelo no paramétrico, con el fin de ver si el estudio de estos datos será ilustrativo para las técnicas vistas en el Capítulo 2. Para ello, recurriremos una vez más al test de modelo lineal frente a modelo no paramétrico, aunque un primer juicio visual nos adelanta que quizá sean los datos que con más motivo requieren optar por un modelo no paramétrico. Tanto el gráfico de los datos como la gráfica, vista previamente, asociada al mencionado contraste se pueden encontrar en la Figura 3.7.

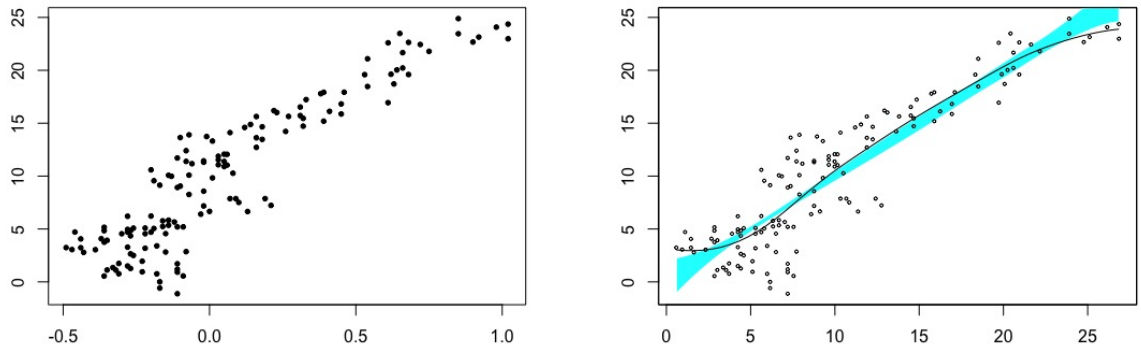

Figura 3.7: De izquierda a derecha: gráfico de los datos iniciales y gráfico correspondiente al contraste de linealidad de los datos.

Con un p-valor de *0.003*, el contraste hace aún más evidente la preferencia de un modelo no paramétrico sobre uno lineal, dadas las claras evidencias estadísticas que existen a favor del primero, en particular, para los niveles de significación habituales, esto es: 1 %, 5 % y 10 %. Para verlo de forma gráfica, basta con observar cómo la estimación rebasa los límites de las bandas de referencia en la imagen de la derecha de la Figura 3.7. El siguiente objetivo será representar la estimación de la curva mediante el estimador local lineal. Como hemos comprobado en las secciones previas, resulta de gran interés comparar la suavidad de las estimaciones con distintos valores de h. Por este motivo, aplicaremos nuevamente el método de la validación cruzada para obtener el valor óptimo del parámetro ventana y de este modo poder seleccionar un valor más alto y otro más bajo, enriqueciendo la comparativa. Esta vez, llegamos a que $h_{opt} = 0{,}11$. A continuación representaremos las estimaciones con $h = 0{,}01$, $h = h_{opt}$ y $h = 0{,}2$ conjuntamente con la gráfica asociada al método de la validación cruzada. Podremos encontrarlo en la Figura 3.8.

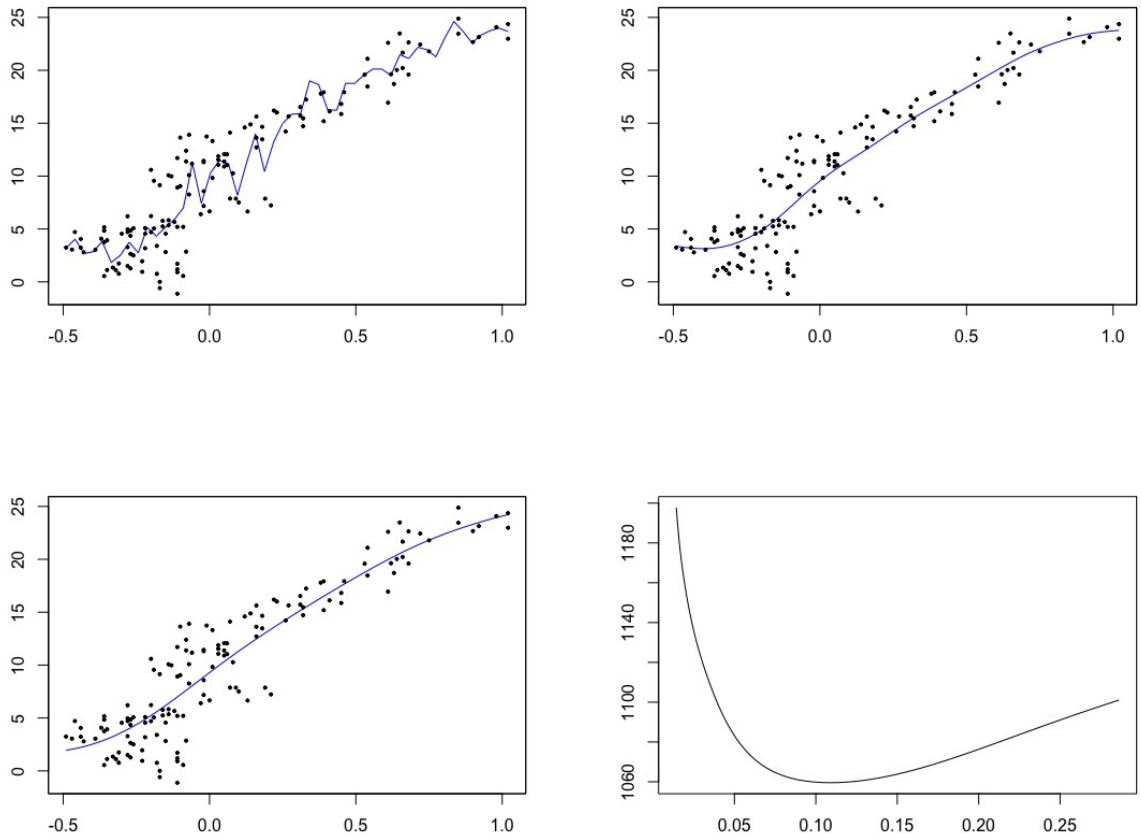

Figura 3.8: De izquierda a derecha y de arriba a abajo: estimación con $h = 0.01$, estimación con $h = h_{opt}$, estimación con $h = 0.2$ y gráfica asociada al método de validación cruzada.

Vemos que en la estimación correspondiente a $h = 0{,}01$ surgen ciertos picos entre los 0 ºC y $0{,}5$ ºC de temperatura, pero después pasan a desaparecer por completo cuando observamos la estimación correspondiente a h_{opt}, lo que hace que resulte especialmente interesante consultar el SiZer asociado a estos datos, ya que recordemos que este, teóricamente, nos permite inferir cuando un crecimiento y/o decrecimiento es significativo. A pesar de esto último, el nivel de suavidad general es satisfactorio, lo que indicaría una relación estable entre ambas variables. En otro orden de cosas, lo que no cambia en ninguna de las tres estimaciones es la tendencia creciente de la curva, lo que nos dice que existe una cierta relación directa entre ambas variables. A su vez, no parece que haya ningún punto de inflexión relevante en las curvas estimadas, lo que nos indica que no hay cambios de relación aparentes entre las variables, manteniéndose la relación directa como la predominante en las estimaciones.

Al igual que antes, estamos interesados en estudiar la variabilidad del modelo. Por este motivo traeremos

de vuelta las bandas de variabilidad, pero esta vez solamente se llevarán a cabo para el valor óptimo de h. Además, como ya adelantamos previamente, realizaremos el SiZer correspondiente a estos datos. Podemos encontrar todo esto en la Figura 3.9.

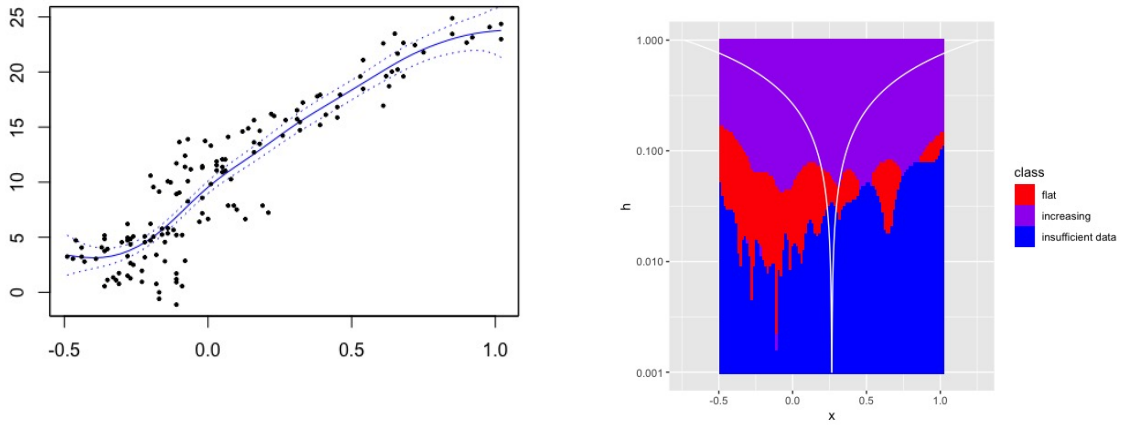

Figura 3.9: De izquierda a derecha: bandas de variabilidad con $h = h_{opt}$ y SiZer correspondiente a estos datos.

En la imagen de la izquierda, la correspondiente a las bandas, se puede apreciar muy poca variabilidad, ya que las dos bandas que marcan los límites en torno a la estimación se encuentran muy próximas a la misma, salvo en las fronteras. En lo que al SiZer se refiere, vemos que para obtener resultados significativos deberíamos irnos a valores de h superiores a $0{,}1$, que curiosamente, es un valor muy próximo a h_{opt}. Esto, en cierto modo apoya que la elección óptima de h sea la que es. Observamos que para los mencionados valores de h superiores a $h = 0{,}1$ hay un crecimiento significativo. Por otra parte, no existe ningún punto en todo el SiZer en el que se produzca un decrecimiento significativo.

Cerramos el estudio, el capítulo y el trabajo con una breve conclusión de lo visto en esta sección. Hemos visto que no podemos relacionar linealmente el aumento de la temperatura media anual con el aumento del nivel del mar, pero tampoco era nuestra pretensión. En cambio, sí que podemos relacionarlas directamente, que era lo que nos interesaba ver, ya que de este modo, a través de los procesos seguidos y de los resultados obtenidos, se podría inferir que la temperatura media anual sí que estaría influyendo de forma directa en el nivel del mar. Dado que pueden existir muchos factores que no hemos tenido en cuenta, especialmente en esta sección, en la cual la hipótesis que hemos trazado antes del estudio era más compleja, y sobretodo, dados los medios con los que hemos contado para la realización del mismo, sería muy atrevido afirmar una conclusión de este calibre con rotundidad, dada su naturaleza multifactorial.

Bibliografía

[1] A. W. Bowman y A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*. Vol. 18. OUP Oxford, 1997.

[2] A. W. Bowman y A. Azzalini. *R package sm: nonparametric smoothing methods (version 2.2-5.7)*. University of Glasgow, UK y Università di Padova, Italia, 2021. URL: http://www.stats.gla.ac.uk/~adrian/sm/.

[3] P. Chaudhuri y J. S. Marron. «SiZer for exploration of structures in curves». En: *Journal of the American Statistical Association* 94.447 (1999), págs. 807-823.

[4] Comisión-Europea. *Causas del Cambio Climático*. URL: https://climate.ec.europa.eu/climate-change/causes-climate-change_es (visitado 2024).

[5] R. S. Donovan. *Una central eléctrica de carbón en el río Ohio, al oeste de Cincinnati*. 2013. URL: https://tinyurl.com/yc796hp5 (visitado 2023).

[6] EPA. *U.S. Environmental Protection Agency*. URL: https://www.epa.gov (visitado 2023).

[7] J. Fan. «Design-adaptive nonparametric regression». En: *Journal of the American statistical Association* 87.420 (1992), págs. 998-1004.

[8] T. J. Hastie y R. J. Tibshirani. *Generalized Additive Models*. CRC Press, 1990.

[9] Josell7. *Atmosferaterra*. 2010. URL: https://upload.wikimedia.org/wikipedia/commons/0/03/Atmosferaterra.png (visitado 2023).

[10] NASA. *Global Climate Change - Vital Signs of the Planet*. URL: https://climate.nasa.gov/en-espanol/signos-vitales/temperatura-global/ (visitado 2023).

[11] NOAA. *Global Monitoring Laboratory*. URL: https://gml.noaa.gov/ccgg/data/ (visitado 2023).

[12] NOAA. *Global Monitoring Laboratory*. URL: https://gml.noaa.gov/hats/data.html (visitado 2023).

[13] NOAA. *Global Monitoring Laboratory*. URL: https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-monthly (visitado 2023).

[14] Raiden.tk. *Glaciación*. 2007. URL: https://www.flickr.com/photos/raiden/2122781678/ (visitado 2023).

[15] D. Sonderegger. *SiZer: Significant Zero Crossings*. R package version 0.1-8. 2022. URL: https://CRAN.R-project.org/package=SiZer.

[16] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, NY, 2008.

[17] M. P. Wand y M. C. Jones. *Kernel smoothing*. CRC press, 1994.

[18] L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.