Ejercicio de regresión no paramétrica

La técnica conocida como LIDAR (light detection and ranging) usa la reflexión de luz de láser emitida para detectar compuestos químicos en la atmósfera. Esta técnica ha probado ser una herramienta muy eficiente para el monitoreo de la distribucion de diversos elementos polulantes en la atmósfera (Sigrist, 1994). En el archivo `lidar.txt` se encuentran datos medidos con a la técnica LIDAR. La variable range es la distancia recorrida antes de que la luz sea reflejada de regreso hacia su fuente. La variable logratio es el logaritmo del cociente de la luz recibida de dos fuentes de luz láser de distinta frecuencia.

1. A partir de los datos de lidar.txt realizar un diagrama de dispersión o scatter plot de range (eje x) vs. logratio (eje y). Describir

2. La función de R **ksmooth** computa el estimador de Nadaraya-Watson a partir de un conjunto de datos $(x_1, y_1), \ldots, (x_n, y_n)$ y lo evalúa en un conjunto de puntos intermedios. Mediante la función de R ksmooth estimar la funcion de regresión que relaciona a las variables range y logratio, tomando como variable explicativa a range, a partir de los datos dados usando el núcleo normal con ventana $h = 5$.

3. Graficar la función de regresión estimada. Repetir para valores de la ventana $h = 10, 30, 50$ y superponer en el mismo plot los puntos correspondientes a las observaciones y el valor estimado de la función de regresión obtenida para $h = 5, 10, 30, 50$. Comparar los resultados obtenidos con las 4 ventanas.

4. Repetir los dos ítems anteriores utilizando otro núcleo.

5. Para cada una de las estimaciones obtenidas computar el Error Cuadrático de Predicción Promediado (ECPP(h)). ¿Cual de las 4 ventanas consideradas da el menor ECPP(h)? ¿Como se puede justificar lo que esta ocurriendo?

6. Hallar mediante el criterio de Convalidación Cruzada $CV(h)$ la ventana optima. Realizar la búsqueda en una grilla para valores de h entre 3 y 165 con paso 1. Realizar un plot de h vs. $CV(h)$.