Trabajo Práctico 3
Regresión Local Pesada Multivariada

Métodos Numéricos

Integrante	LU	Correo Electrónico
Teo Kohan	385/20	teo.kohan@gmail.com
Martin Santesteban	397/20	martin.p.santesteban@gmail.com

Abstract

Se desarrolla e implementa *loess* : regresión local pesada. Con un conjunto de muestras formadas por variables aleatorias, se usa para estudiar la relación de dependencia de una variable en función del resto, trazando una superficie de regresión. Se discute sus aplicaciones y se experimenta con los elementos fundamentales de la metodología. También se experimenta con conjuntos de datos sintéticos y se ve el rendimiento al utilizar distintos tipos de regresiones locales.

Keywords— Regresión lineal, Prediccion, *LOESS*, Cuadrados mínimos lineales

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
https://exactas.uba.ar

Índice

1 Introducción

1.1 Motivación

Hay muchas motivaciones para implementar un buen algoritmo de regresión lineal. Dado un conjunto de muestras, realiza predicciones y previsiones con respecto a las variables asociadas. Sin embargo, *loess* es un método que permite realizar regresiones lineales locales, y que sigue otorgando la flexibilidad de las regresiones no lineales. En este trabajo se experimenta con los parámetros de mayor importancia, y se exponen las fortalezas y debilidades del método.

1.2 Regresión

La regresión es un proceso estadístico, mediante el que se estudia la relación entre p variables independientes y una variable dependiente (target, objetivo). El objetivo es ver como cambia la variable dependiente en función de las independientes, encontrando una función de regresión capaz de estimar la relación de dependencia, a partir de un conjunto base de muestras. Esta regresión, de ser la correcta, nos permitirá realizar predicciones con respecto a la variable objetivo.

1.3 Regresión Lineal Simple y Múltiple

La regresión lineal busca, dado un conjunto de puntos $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$, encontrar la función de regresión de grado menor o igual a uno que minimiza alguna distancia. Cuando se trata del caso en que hay un única variable independiente se hablará de regresión simple, mientras que en caso contrario utilizaremos la regresión múltiple. En la regresión lineal se asume que las variables independientes (también llamados predictores) se relacionan con la variable dependiente

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon$$

donde β_0 y β_1 son coeficientes desconocidos y ϵ es una variable independiente con distribución normal con media igual a cero y desviación estándar σ^2. Para poder hacer una regresión lineal eficientemente, necesitamos que las variables estén relacionadas linealmente. Uno de los métodos de regresión lineal más utilizado es el método de Cuadrados Mínimos. Este despeja a los coeficientes de regresión buscando la solución al siguiente problema de minimización

$$\min \|y - Ax\|_2$$

Que se resuelve con las siguientes ecuaciones normales

$$A^t \cdot A \cdot x = A^t \cdot y$$

donde A es una matriz en $\mathbb{R}^{n \times p+1}$ (si $p = 1$ estaríamos hablando de regresión simple y si p fuera mayor de múltiple) que tiene unos en la primer columna y los datos x_j completando la fila (al utilizar cuadrádos mínimos lineales). El vector $y^t = (y_1, y_2, \ldots, y_n)$ y el vector incógnita x tendrá a los coeficientes de regresión. Sin embargo, hay que notar que este método hace la regresión en función a todos los puntos del conjunto, de forma global. A nosotros nos interesará usar cuadrados mínimos para realizar regresiones locales.

1.4 Regresión Local Pesada: Una Introducción a *loess*.

La regresión local encuentra la superficie de regresión en función a cuadrados mínimos lineales calculados en subconjuntos de puntos agrupados, o vecindarios. Por cada punto, se calculará Cuadrados Mínimos teniendo en cuenta a todos los puntos, pero dándole más importancia a los más cercanos utilizando una función de peso. Esta función de peso se encargará de hacer que las muestras por fuera del vecindario no aporten nada al cálculo de la recta de cuadrados mínimos, en relación a la muestra donde se ha centrado el cálculo; de ahí el nombre, regresión local pesada.

1.4.1 Función de Peso

Para poder llevar a cabo *loess*, se necesita de cualquier función de peso $W : \mathbb{R}^p \to \mathbb{R}$ que cumpla las siguientes propiedades,

1. $W(x) > 0$ para $|x| < 1$
2. $W(-x) = W(x)$ función par.
3. $W(x)$ es monótona decreciente para x ≥ 0.
4. $W(x) = 0$ para todo $|x| \geq 1$.

Esta función, será utilizada para asignarle peso a las muestras al calcular Cuadrados Mínimos lineales centrado en una muestra $x^{(j)}$. Para poder hacer eso, sin embargo, deberemos definir una segunda función que evalúe la función de peso centrándola en $x^{(i)}$ que llamaremos $\omega : \mathbb{R}^p \to \mathbb{R}$.

$$\omega_i(x) = W(\|x^{(i)}, x\|_2 / d(x^{(i)}))$$

donde $d(x^{(i)})$ es la distancia entre $x^{(i)}$ y el q-ésimo vecino. La cantidad de vecinos de $x^{(i)}$ a considerar es q, donde $q = f \cdot n$, f es un real entre 0 y 1 y n es la cantidad de muestras.

A menos que se especifique lo contrario, a lo largo del trabajo se utilizará como función de peso la función tri-cúbica

$$W(x) = \begin{cases} (1 - x^3)^3 & 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

1.5 Desarrollo

Se cuenta con un conjunto de $n \in \mathbb{N}$ muestras de p variables aleatorias independientes, $X_1, X_2, \ldots, X_p$ junto con sus respectivas mediciones de la variable aleatoria dependiente Y, $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$. Para conseguir la superficie de regresión, se debe aplicar una regresión local pesada sobre cada muestra $(x^{(i)}, y^{(i)})$ del *dataset* inicial.

Sea $x^\star$ una muestra del *dataset* con su respectivo $y^\star$, para realizar la regresión local pesada en ese punto debemos utilizar Cuadrados Mínimos, utilizando la función de peso introducida en 1.4.1. Resolvemos las ecuaciones normales pero agregando una matriz de pesos centrados en nuestra muestra $x^\star$. Para poder asignar los pesos, necesitamos un parámetro q que haga referencia a la

cantidad de vecinos que se tendrán en cuenta al calcular los pesos de las muestras; este se podrá descomponer como $f \cdot n$, donde f (el parámetro más importante de *loess*) es un real entre 0 y 1. A partir del q-ésimo vecino más cercano de $x^\star$, el peso que corresponderá a todas las muestras será 0.

Dado el vecindario de $x^\star$, podemos calcular los pesos de cada vecino, lo que resta es calcular los coeficientes de regresión β vistos anteriormente, solución de las ecuaciones normales. Debemos obtener los pesos de cada muestra respecto a $x^\star$, para eso, definimos la matriz de pesos $W_\star \in \mathbb{R}^{n\times n}$, tal que es diagonal y cumple que $(W_\star)_{ii} = \omega_\star(x^{(i)})$, para todo $1 \leq i \leq n$ donde $\omega_\star(x^{(i)})$ devuelve el peso de $x^{(i)}$ en función de su distancia con $x^\star$. Para encontrar los coeficientes de regresión, resolvemos las ecuaciones normales

$$A^t W A x = A^t y$$

Una vez calculados los coeficientes de regresión de $x^\star$, podemos calcular el ajuste de su variable dependiente $\hat{y}^\star$ (fitted value) dependiendo de si estamos usando *loess* con regresión lineal o cuadrática.

Regresión Lineal y Cuadrática

Al utilizar *loess*, podemos elegir utilizar regresión local lineal o cuadrática. Al utilizar cuadrados mínimos lineales, se busca la recta que minimice la distancia con todos los puntos. En cambio, cuadrados mínimos cuadráticos permite encontrar, dado un conjunto de puntos en el plano, la parábola que tenga una distancia mínima con todos los puntos. Utilizando esta variación de Cuadrados Mínimos, A cambiará de forma. A continuación podremos ver la matriz A al utilizar *loess* con regresión lineal y cuadrática, con n muestras independientes tridimensionales ($p = 3$).

$$A = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & x_3^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & x_3^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & x_3^{(3)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & x_3^{(n)} \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & (x_1^{(1)})^2 & x_1^{(1)} \cdot x_2^{(1)} & x_1^{(1)} \cdot x_3^{(1)} & (x_2^{(1)})^2 & x_2^{(1)} \cdot x_3^{(1)} & (x_3^{(1)})^2 \\ 1 & x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & (x_1^{(2)})^2 & x_1^{(2)} \cdot x_2^{(2)} & x_1^{(2)} \cdot x_3^{(2)} & (x_2^{(2)})^2 & x_2^{(2)} \cdot x_3^{(2)} & (x_3^{(2)})^2 \\ 1 & x_1^{(3)} & x_2^{(3)} & x_3^{(3)} & (x_1^{(3)})^2 & x_1^{(3)} \cdot x_2^{(3)} & x_1^{(3)} \cdot x_3^{(3)} & (x_2^{(3)})^2 & x_2^{(3)} \cdot x_3^{(3)} & (x_3^{(3)})^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & x_3^{(n)} & (x_1^{(n)})^2 & x_1^{(n)} \cdot x_2^{(n)} & x_1^{(n)} \cdot x_3^{(n)} & (x_2^{(n)})^2 & x_2^{(n)} \cdot x_3^{(n)} & (x_3^{(n)})^2 \end{bmatrix}$$

Luego de calcular la solución de las ecuaciones normales 1.5, el valor ajustado de la variable independiente en el caso lineal será, para todo $x^{(i)} \in \mathbb{R}^3$,

$$\hat{y}^{(i)} = \beta_0 + \beta_1 \cdot x_1^{(i)} + \beta_2 \cdot x_2^{(i)} + \beta_3 \cdot x_3^{(i)}$$

mientras que en el caso cuadrático

$$\hat{y^{(i)}} = \beta_0 + \beta_1 \cdot x_1^{(i)} + \beta_2 \cdot x_2^{(i)} + \beta_3 \cdot x_3^{(i)} + \beta_4 \cdot (x_1^{(i)})^2 + \beta_5 \cdot x_1^{(i)} \cdot x_2^{(i)} + \beta_6 \cdot x_1^{(i)} \cdot x_3^{(i)} + \beta_7 \cdot (x_2^{(i)})^2 + \beta_8 \cdot x_2^{(i)} \cdot x_3^{(i)} + \beta_9 \cdot (x_3^{(i)})^2$$

1.6 Implementación del algoritmo

Para ejecutar el algoritmo, primero necesitamos tener los siguientes parámetros,

- Una conjunto de n muestras constituidas por p-variables independientes y una variable dependiente.
- Una función de peso.
- Un número $f \in \mathbb{R}$, $0 \leq f \leq 1$, de tal forma que la función de peso sea positiva para vecindarios de tamaño $f \cdot n$.

Luego, en base a estos parámetros, debemos

- Construir una matriz $A \in \mathbb{R}^{n \times (p+1)}$, que depende de si se está usando Cuadrados Mínimos lineales o cuadráticos.
- Para toda muestra $(x^{(i)}, y^{(i)})$ del conjunto base de datos, se debe
  - Calcular los pesos de todas las muestras $(x^{(j)}, y^{(j)}), (i \neq j)$ con respecto a $(x^{(i)}, y^{(i)})$ utilizando $\omega_i(x^{(j)})$ y el tamaño del vecindario $f \cdot n$.
  - Armar la matriz de pesos.
  - Resolver el sistema de ecuaciones $(A^t \cdot W \cdot A) \cdot x = A^t \cdot y$
  - Utilizando la solución del sistema, calcular el valor ajustado $\hat{(y)}^i$ (el cálculo depende de si usamos regrsión cuadrática o lineal)
- Todos los valores ajustados nos permiten armar la superficie de regresión.

1.7 Regresión local pesada, sección 5.

Se busca replicar la sección 5 del paper [1]. A lo largo de 111 días, desde el primero de Mayo al treinta de Septiembre de 1973, se tomaron mediciones de cuatro variables aleatorias: ozono, temperatura, velocidad del viento y radiación solar.

El objetivo fue analizar la dependencia del ozono con respecto a las demás variables, para poder realizar predicciones con respecto a la concentración del ozono. Consecuentemente, la variable dependiente es la concentración de ozono en la atmósfera, mientras que las demás variables serán los predictores.

1.7.1 Matriz de scatterplots

La figura 1 muestra una matriz de Scatterplots de 4×4, donde se puede ver la disposición inicial de las muestras para cada par de variables.

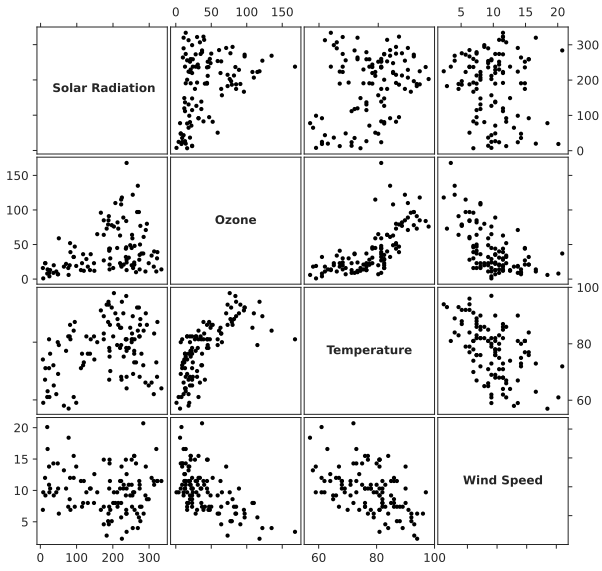

Figura 1: *Scatterplot de cada par de variables de entrada.*

Con estos datos, se realizó loes lineal con $f = 0.4$. Sin embargo, antes de analizar la dependencia del ozono, hay que recordar que para poder realizar la regresión lineal se requiere partir de una suposición: las variables estudiadas tienen que depender linealmente de los predictores, y las muestras deben presentar un grado de error con distribución normal (con varianza constante). Consecuentemente, hay que chequear que estas suposiciones se cumplan.

1.8 *qqplot*: Estudiando la distribución del error.

Para estudiar la distribución del error de muestreo, se grafica para cada muestra el valor $\epsilon^{(i)} = \hat{y^{(i)}} - y^{(i)}$ (residuos) y se hace el análisis en la figura 4. Este es un qqplot (quantile-quantile plot [3]) y herramienta de visualización que nos permite ver la disposición de los datos en función a la de una distribución teórica. Su objetivo es graficar los cuantiles (o percentiles) del conjunto de mediciones y graficarlos en función a los cuantiles de la distribución teórica.

Recordamos que los cuantiles son muestras que marcan que cierta proporción de datos se encuentran por debajo de la misma. Por ejemplo, el cuantil 0.25 indica que el 25% de datos tienen un valor menor al mismo (denotado $x_{(0.25)}$).

Luego, al estudiar el *qqplot* se debe analizar si se forma una "linea recta" entre los puntos. De ser así, se puede suponer que el conjunto de datos (más o menos) se distribuye de la misma forma que la distribución teórica. En caso de que no se forme una recta, no habrá ninguna relación lineal entre las muestras y la misma. En el caso en que forme una linea recta, y haya cierta pendiente

mayor o menor a los 45 grados, se puede decir que los datos tienen cierto sesgo.

En este caso, al graficar el error en relación a la distribución normal, se ve que los errores están sesgados a la derecha, indicando que el error no tiene distribución normal (ya que una variable aleatoria normal no tiene sesgo), y presentan asimetría estadística. Hay que tener en cuenta que un *qqplot* es una herramienta de visualización, y no es una prueba infalible que demuestre la distribución de los datos.

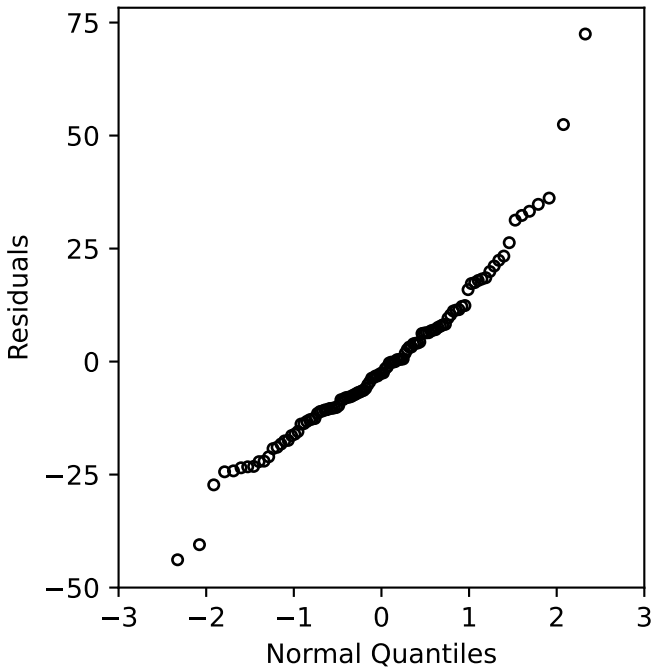

Figura 2: qqplot *entre los residuos absolutos del ozono y una distribución* $\mathcal{N}(0,1)$.

1.9 Scatterplot: Residuos vs Valores ajustados

Estudiando la distribución del error, observamos en la figura 3 el valor absoluto de los residuos contra los valores ajustados, utilizando un scatterplot.

Este tipo de gráfico *Residuals vs Fitted values* nos deja comparar la naturaleza del error en función a los valores ajustados: expone la linealidad, varianza y disposición de los errores.

Al realizar *loess* ($f = 2/3$) (tomamos como único predictor a los valores ajustados y como variable independiente a los errores) la función de regresión aumenta en función a los valores ajustados. Se puede decir que la varianza de los errores aumenta junto con los valores ajustados, ya que para los valores más bajos hay una alta concentración de puntos mientras que a partir del valor ajustado igual a 50 la varianza aumenta drásticamente (se ve como se "disipan" los puntos). Si el error tuviera distribución normal, la varianza debería ser constante.

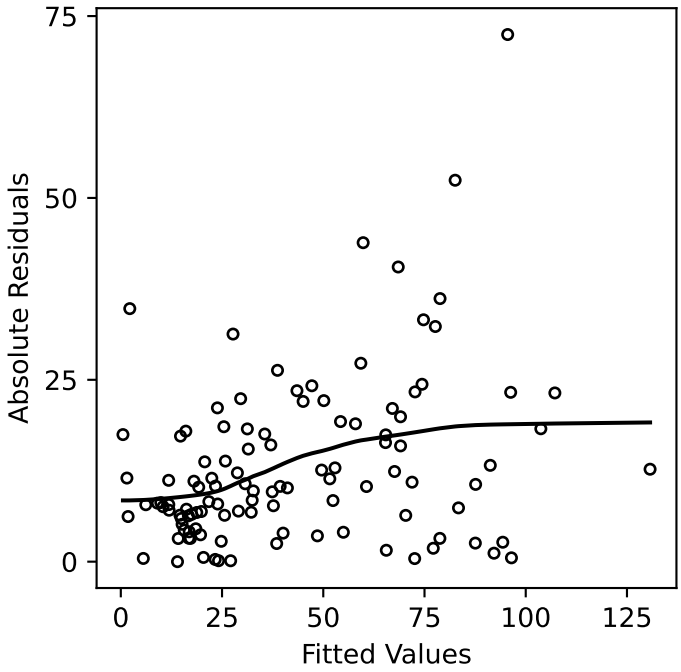

Figura 3: *Scatterplot de los residuos absolutos vs los valores ajustados del ozono.*

Finalmente, se concluye que el error no tiene distribución normal. Esto choca con las condiciones básicas para aplicar la metodología.

1.10 Scatterplots: Residuos vs Predictores

En la figura 4 se pueden observar tres scatterplots (*component-residual plot*), en cada uno se encuentra la disposición del residuo contra cada variable independiente. Un *component-residual plot* nos sirve para estudiar la dependencia del residuo en función a las variables independientes del modelo. En el primero, se observa que a medida que aumentaban las mediciones de la radiación solar, también aumentaba proporcionalmente la varianza de los residuos. Al aplicar *loess* univariado con $f = 2/3$ parece formarse una linea casi recta, indicando que la varianza aumentó de la misma forma tanto para los residuos positivos como negativos. Sin embargo, la recta también parece casi perfecta, asemejándose mucho a $y = 0$. Aun así, hay muchos puntos cerca del cero y tenemos valores bajos para el error. En la figura central se ve una distorsión más pronunciada cerca del centro de la figura. La regresión del *loess* está relativamente cerca del cero, pero parece indicar que los residuos no son tan buenos como en el caso anterior. En la última figura la distorsión es mucho mayor.

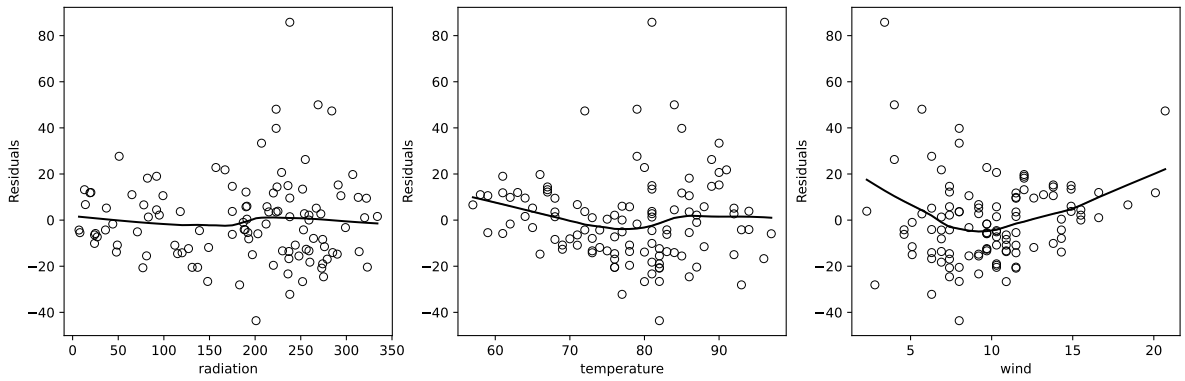

Figura 4: *Scatterplot de los residuos absolutos vs los valores ajustados del ozono.*

Finalmente, esto indica que al haber realizado *loess* multivariado con $f = 0.4$, la superficie de regresión no es buena.

1.11 La regresión adecuada

Para poder encontrar la mejor función de regresión, tal que siga el patrón de los datos, se discute en el paper volver a buscar la superficie de estimación utilizando *loess* con un valor menor para f, sin embargo 0.4 ya es un valor lo suficientemente chico. Hay que notar, que tener un f muy chico devolverá una curva buenisima, que probablemente (casi) interpole a todos los puntos. Sin embargo, sería completamente inservible al trabajar con datos desconocidos. Este fenomeno se llama *overfitting*. Consecuentemente, se decide efectuar la regresión ajustando de forma cuadrática en vez de lineal, y se indica que las distorsiones desaparecen. También, al considerar las raíces cúbicas de las mediciones de ozono, el error comenzó a comportarse de forma normal estándar. Finalmente, utilizando *loess* ajustado cuadráticamente y con $f = 0.8$, se consigue una superficie de regresión adecuada.

1.12 *Conditioning Plots*

Los *Conditioning Plots* [2] son gráficos de dos variables condicionadas al valor de las restantes. Por lo tanto, graficaremos a la variable target en función a un predictor, mientras fijamos los valores de las dos variables independientes restantes.

La figura 5 muestra una grilla de 3×3 de *conditioning plots*, donde se fija un par de las variables independientes y se estudia raíz cúbica del ozono. En este se gráfica el ozono en función a la temperatura fijando a la velocidad del viento en $[5, 10, 15]$ y fijando la radiación solar en los valores $[50, 170, 290]$. Luego, hay un gráfico en la grilla por cada combinación de valores fijos de los predictores.

Al estar trabajando con tres variables independientes, la superficie de regresión g será tal que $g : \mathbb{R}^3 \to \mathbb{R}$. g será una superficie de cuatro dimensiones, por lo tanto, si fijamos una variable se

tendrá una traza tridimensional. Luego, fijar dos variables nos dará como resultado a una función bidimensional.

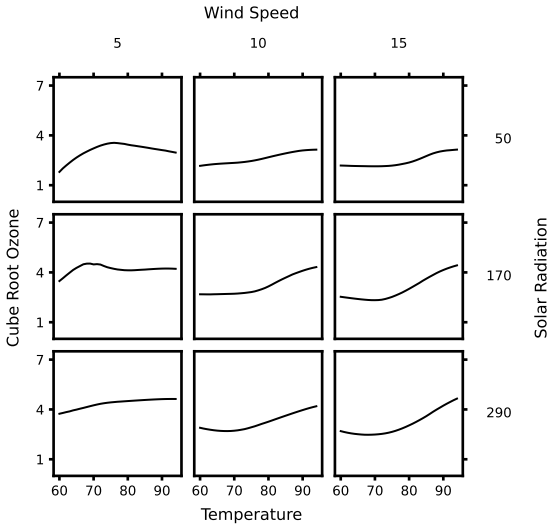

Figura 5: *Estudio de la raíz cúbica del ozono en función a la temperatura, fijando en tres valores distintos a la velocidad del viento y la radiación solar.*

Análogamente, la figura 7 trata de realizar el mismo estudio de la superficie de regresión, pero fijando la temperatura y velocidad del viento. Se estudia el comportamiento de la raíz cúbica del ozono en función a la radiación solar.

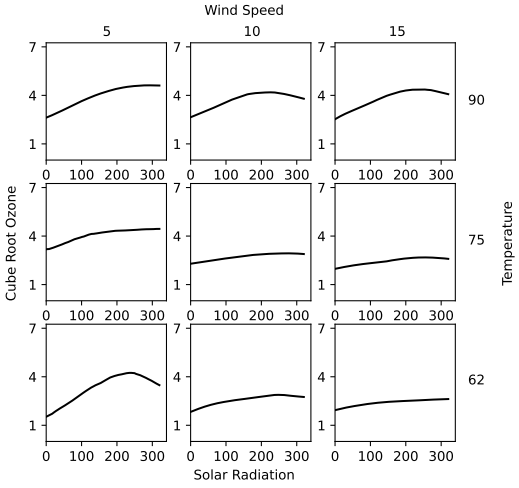

Figura 6: *Estudio de la raíz cúbica del ozono en función a la radiación solar, fijando en tres valores distintos a la temperatura y velocidad del viento.*

Por último, se grafica la única configuración restante: fijamos la temperatura y la radiación solar, graficamos la raíz cúbica del ozono en función a la velocidad del viento. Se ven los resultados en 7.

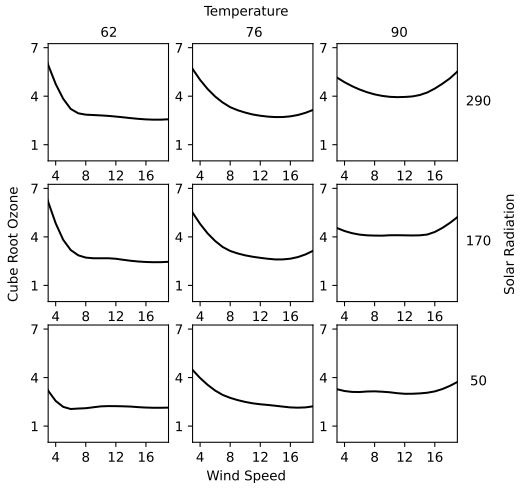

Figura 7: *Estudio de la raíz cúbica del ozono en función a la velocidad del viento, fijando en tres valores distintos a la temperatura y radiación solar.*

Para generar estos gráficos se tuvo que aplicar la regresión pesada sobre puntos que no se encontraban en el conjunto de datos base. Por lo tanto, graficar los "slices" de la superficie de regresión implicó generar predicciones a partir del conjunto de datos.

2 Experimentación

2.1 Ajustando en función de f

Objetivo

El objetivo de este experimento es ver como, para un conjunto de datos sintético, cambia la función de regresión. Se tomarán 100 muestras $(x^{(i)}, f(x^{(i)})$, y se realizara la regresión pesada con $f \in \{0.05,\ 0.25,\ 0.5,\ 0.99\}$. La función sintética es $f(x) = (\frac{1}{\sqrt{2}} \cdot x + \sqrt{5}) + \epsilon$, donde ϵ tiene una distribución normal. Para las 100 muestras sintéticas, los predictores se generaron de forma aleatoria utilizando el método de numpy, random.randint. Todos los x tienen un valor entre 150 y 250. Se llevó a cabo el experimento utilizando un ϵ con varianza igual a 10 y 50. Se cambian los valores de épsilon, para trabajar con distintas disposiciones de datos y ver como opera *loess* cuando se trabaja con regiones más densas y con datos más dispersos. Se ejecutará el experimento utilizando regresión lineal y regresión cuadrática.

Hipótesis

Se espera ver como, para los valores de f más bajos, se produce *overfitting*. Esto es darle mucho peso solo a las muestras más cercanas. Debería producirse una función de regresión que pase muy cerca de los puntos, solo en las partes más densas del scatterplot. Esto es porque, a los datos más lejanos no se les dará un peso mayor a cero. A medida que crece f la función de regresión debería volverse cada vez más suave.

Al utilizar regresión lineal, para ambas varianzas de ϵ $f = 0.05$ hará *overfitting*, para $f = 0.99$ se espera ver que con poca varianza sea casi una recta, mientras que con mucha varianza debería tener cierta curvatura. Esto es debido a que al tener regiones densas de datos las funciones de regresión deberían suavizarse más rápido (deberían comenzar a parecerse a una recta para valores de f más bajos) mientras que al tenerlos dispersos debería costar más conseguir una función suave.

Con la regresión cuadrática se espera ver que se produce overfitting también para los primeros valores de f, sin embargo también tenemos la intuición de que esta regresión será mucho más propensa a las curvaturas: por eso para $f = 0.99$ esperamos ver que se producirá una función de regresión muy suave, pero con más curvas que en el caso anterior.

Resultados

En la figura 8 se ven los resultados del experimento, la figura de la izquierda expone el caso en que $\epsilon = 10$, mientras que en la derecha $\epsilon = 50$. En el primer gráfico se puede ver como las funciones de regresión se suavizan rápidamente a medida que crece f. En la figura de la derecha se puede ver como cuesta mucho más conseguir una función suave. Esto se ve fácilmente con $f = 0.5$, en el primer caso ya casi parece una recta que presenta dos suaves curvaturas, mientras que en el segundo caso se notan mucho más las distorsiones. Con respecto a $f = 0.05$, como era de esperarse, se produce el *overfitting* en ambos casos. Para $f = 0.95$, en la primer figura es una recta, casi equivalente a haber hecho regresión lineal con cuadrados mínimos sobre todo el conjunto de datos, mientras que en el segundo caso se nota una curvatura más pronunciada.

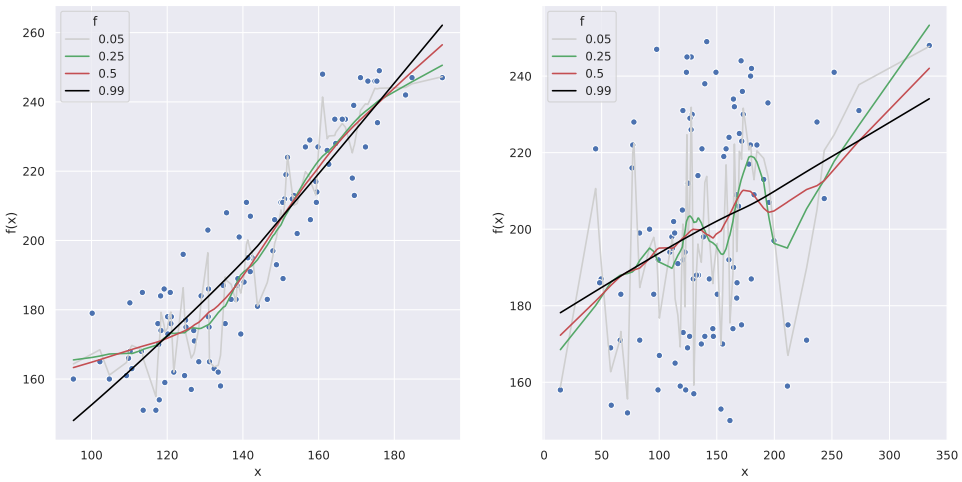

Figura 8: *Comparación de las funciones de regresión para distintos valores de f, ejecutando* loess *lineal sobre datos sintéticos. En la primer figura se tuvo un error de varianza 10, y en la segunda fue 50.*

Finalmente, en 9 tenemos la regresión cuadrática con poca varianza y con mucha varianza. Podemos observar que se produce el *overfitting* para los valores más bajos, y que efectivamente la regresión cuadrática es más propensa a formar curvas. Con valores mucho más elevados de f se producen de todas formas curvas pronunciadas.

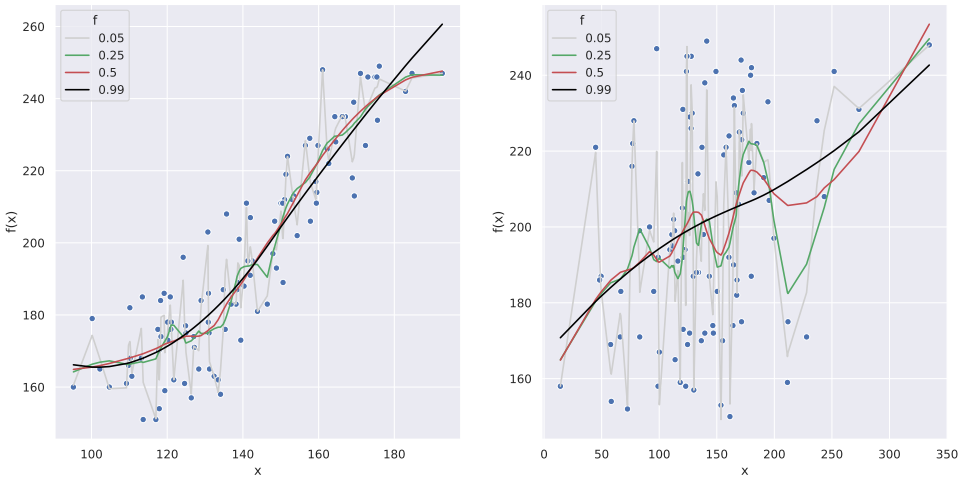

Figura 9: *Comparación de las funciones de regresión para distintos valores de f, ejecutando* loess *cuadrático sobre datos sintéticos. En la primer figura se tuvo un error de varianza 10, y en la segunda fue 50.*

2.2 Regresión lineal vs Cuadrática

Objetivo

La idea es estudiar cuando es mejor utilizar loess con regresiones locales lineales o cuadráticas. Se correrá el experimento para dos funciones sintéticas, la primera será una función lineal mientras que la segunda será un polinomio de grado igual a 3. A ambas funciones se le sumará un ϵ con distribución normal y varianza v. Luego, se realizará loess utilizando cuadrados mínimos lineales y cuadráticos para ambos conjuntos de datos con $f = 0.5$ y se compararán los errores realizados haciendo scatterplot de los valores absolutos de los residuos contra los valores ajustados (sobre estos también se correra *loess* para estimar los valores de los errores). Luego, se producirán 100 muestras sintéticas de la forma $(x^{(i)}, f/g(x^{(i)}))$. Los predictores serán generados de forma uniforme, con valores reales dentro del intervalo $[0, 6]$. Las funciones son:

$$f(x) = \frac{1}{\sqrt{2}} \cdot x + \sqrt{5} + \epsilon(v)$$

$$g(x) = (x-4)^4 + 4(x-4)^2 + x - 4 + \epsilon(v)$$

Primero, el experimento se correrá con una varianza v bajo, para estudiar las regresiones en los casos mas "obvios". Adicionalmente, se correrá el experimento con una varianza tal que no se note fácilmente de que función sintética se trata.

Hipótesis

Creemos que utilizar parábolas para aproximar los datos dará mejores resultados cuando las variables dependientes son valores de una función sintética con mayor grado. Esto se debe, a que si se tienen tres puntos en el plano (no en linea recta), siempre serán interpolables por una parábola. Consecuentemente, y con el mismo razonamiento, si las variables dependientes fueron generadas con una función lineal debería presentar mejores resultados utilizar cuadrados mínimos lineales, ya que el mejor polinomio interpolante será una recta.

En el caso donde hay poca varianza, calculamos que al utilizar a $f(x)$ para generar las variables dependientes, la regresión debería ser casi igual: ambas deberían parecer una recta. Sin embargo, la regresión lineal debería ser un poco mejor, ya que en el caso cuadrático es inevitable tener cierta curvatura. Al utilizar a $g(x)$, suponemos que la regresión cuadrática debería ser mejor, ya que es más obvio que se trata de un polinomio.

Con mucha varianza, tenemos la hipótesis de que la regresión lineal debería ser mejor cuando la función es lineal, y la cuadrática debería presentar menos residuos en el caso cuadrático.

Resultados

La figura 10 expone dos scatterplots, en el de la izquierda se muestra la disposición de los puntos al utilizar como función de las variables dependientes a $f(x)$. En la imagen de la derecha se encuentran otras 100 muestras pero con la función cúbica $g(x)$. En ambos se agrego "poco" ruido utilizando una varianza igual a uno. La función de color anaranjado es la función de regresión local lineal con

$f = 0.5$, mientras que la azul es la cuadrática. Se puede ver en el primer caso como se forma casi la misma recta para la regresión, sin embargo la función cuadrática presenta cierta curvatura en el medio. En el segundo caso es notable como la función cuadrática tiene un rendimiento mucho mejor con un f no tan bajo.

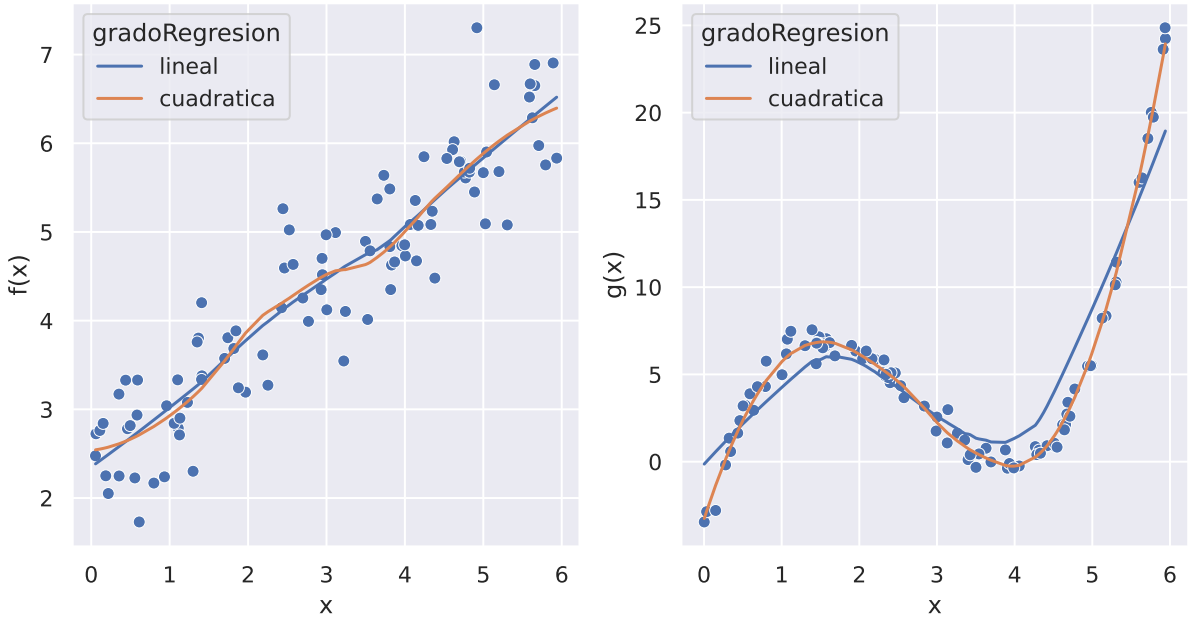

Figura 10: *Loess cuadrático y lineal para f(x)(izquierda) y g(x) (derecha). Los errores tenían varianza baja.*

En la figura 11 podemos ver cuatro scatterplots. En la primer fila se encuentran los residuos al haber hecho los datos con $f(x)$, y en la segunda se encuentras los de la función polinómica. En la primer columna estan los residuos del loess lineal, y en la segunda colúmna los residuos del loess cuadrático.

A todos se les graficó el loess univariado, con $f = 0.5$. Se puede ver como, al haber hecho producido los datos con la función lineal el error es casi idéntico. La diferencia al haber hecho los datos con $g(x)$ es mucho más notable: se expone como al haber hecho la regresión cuadrática los errores se encuentran por debajo del 1.5, mientras que además son más balanceados en ese intervalo, produciendo un loess más "pegado" a la recta $y = 0$.

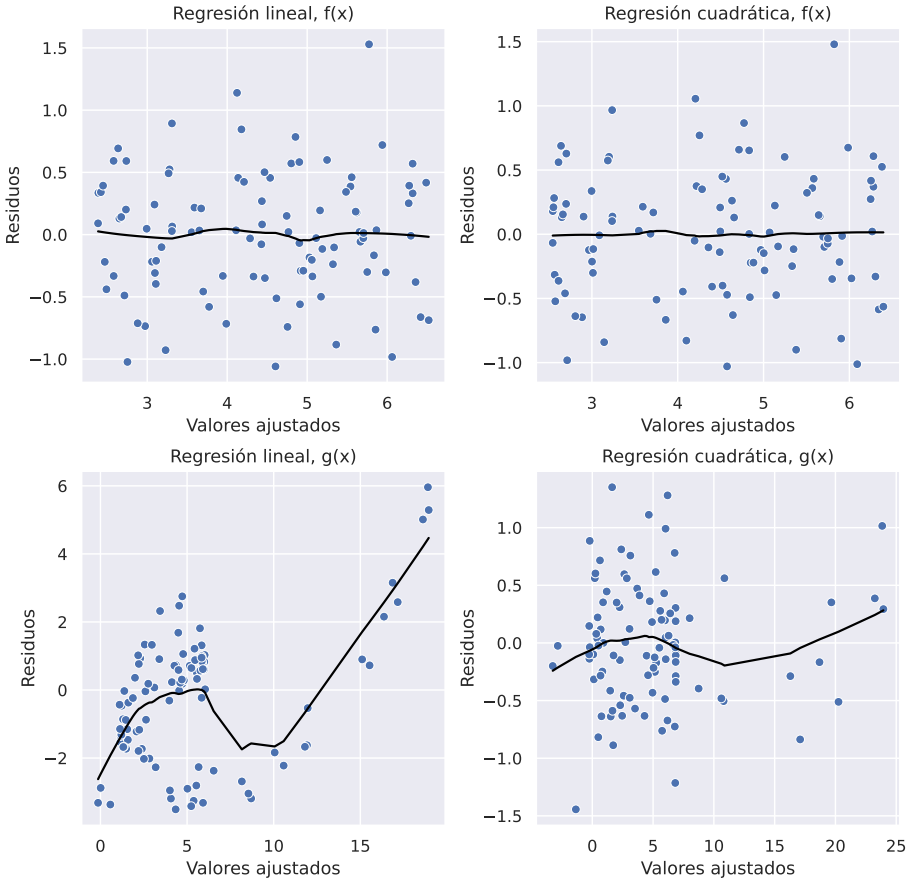

Figura 11: *Estudio de los residuos para cada función de regresión, para f(x) y g(x) con varianza baja.*

En la imagen 12 podemos ver al scatterplot con las mismas funciones, dispuestas de la misma forma de antes, pero con un error con varianza igual a 50. La varianza es tal que, sin conocer los ordenes de las funciones que generaron los datos, no nos deja asumir nada de la misma. En ambos se expone loess, con el mismo parámetro f visto anteriormente. En la primer figura la diferencia entre las regresiones sigue sin ser demasiado significativa (es más grande que antes, pero no es mucha), mientras que en el segundo caso tampoco lo es (hasta son bastante parecidas). Para ver que regresión es mejor, acudimos a los gráficos se los errores.

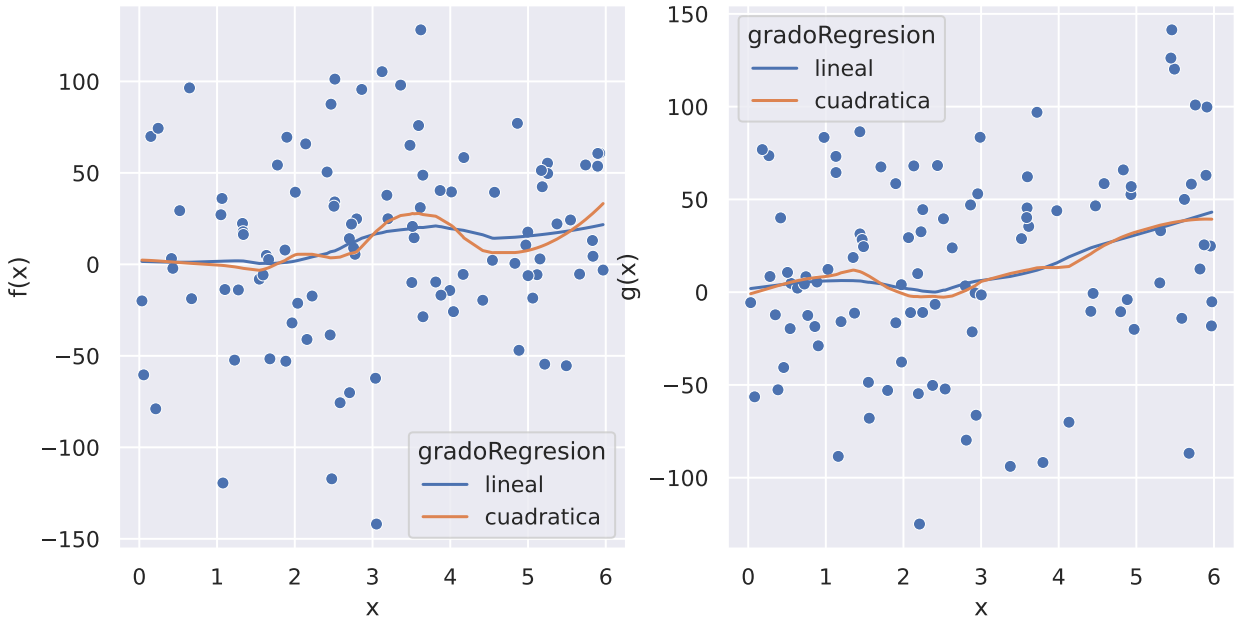

Figura 12: *Regresión lineal y cuadrática utilizando f(x) y g(x) con alta varianza.*

La figura 13 muestra los errores de las regresiones al utilizar ambas funciones para producir los datos, con un error con varianza igual a 50. A cada scatterplot se le graficó el *loess* $f = 0.5$. En la primer figura de todas se puede ver como el error de nuevo es bastante parecido, pero el loess de la regresión lineal se asemeja más a la recta $y = 0$ que el loess cuadrático. Con respecto al error al utilizar el polinomio, el *loess* expone que la regresión cuadrática es mejor. Sin embargo, la diferencia en errores nunca es demasiado significativa.

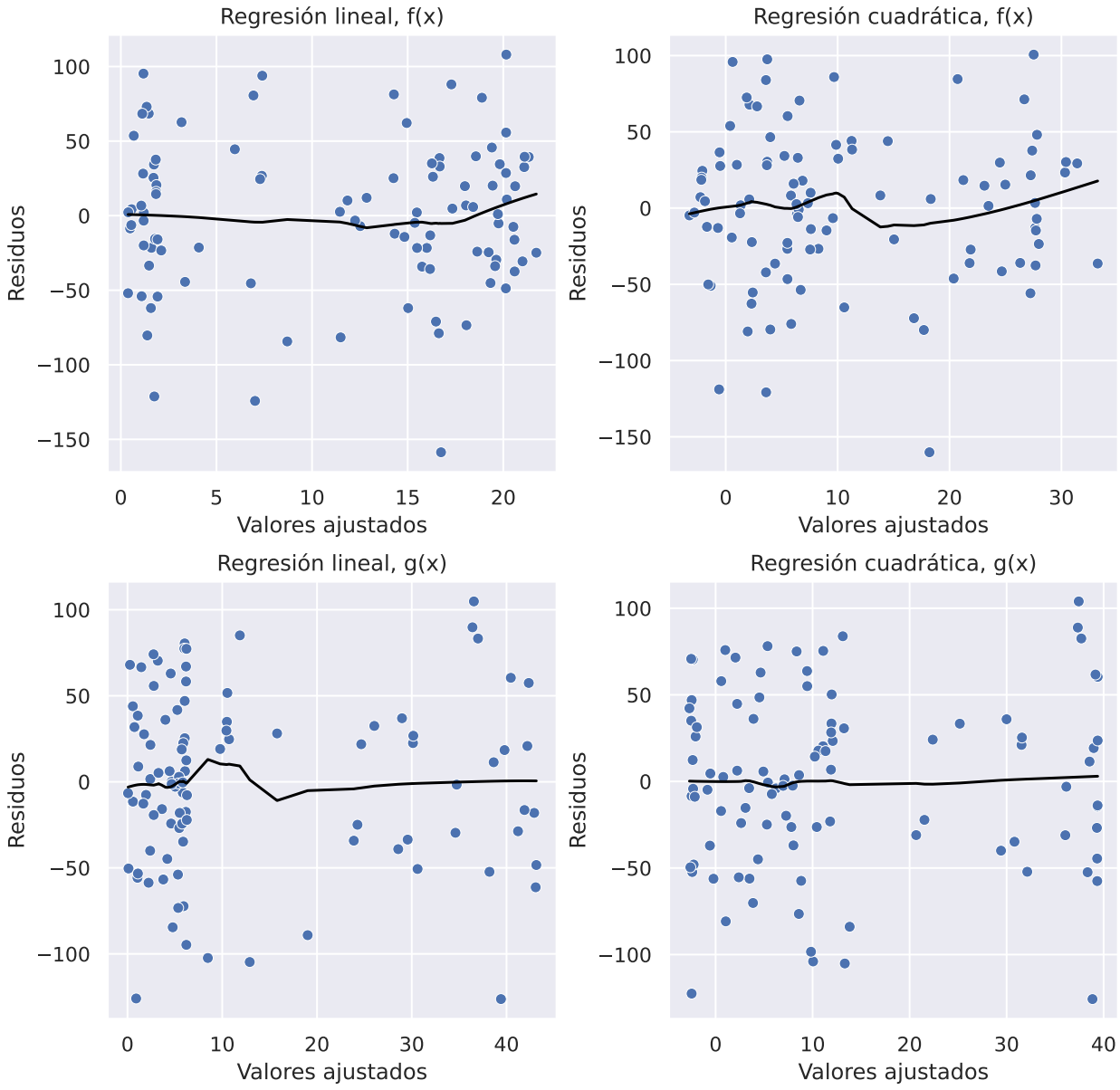

Figura 13: *Residues vs Fitted values, para la regresión cuadrática y lineal al utilizar como función para las muestras a f(x) y g(x).*

2.3 Funciones de peso W

Objetivo

Observar el comportamiento de la regresión utilizando las distintas funciones de peso enumeradas en la figura 14. Se analizará la función trigonométrica $\sin(x)$ y el polinomio $\frac{-\frac{1}{3}x^3+\frac{1}{2}x^2+x}{5}$ como dos tipos de clase de funciones distintas. Se le sumará un "ruido" ϵ normalmente distribuido tal que $\epsilon \sim \mathcal{N}(0,1)$. Luego se comparará visualmente, (aunque un análisis de residuos también podría ser apropiado).

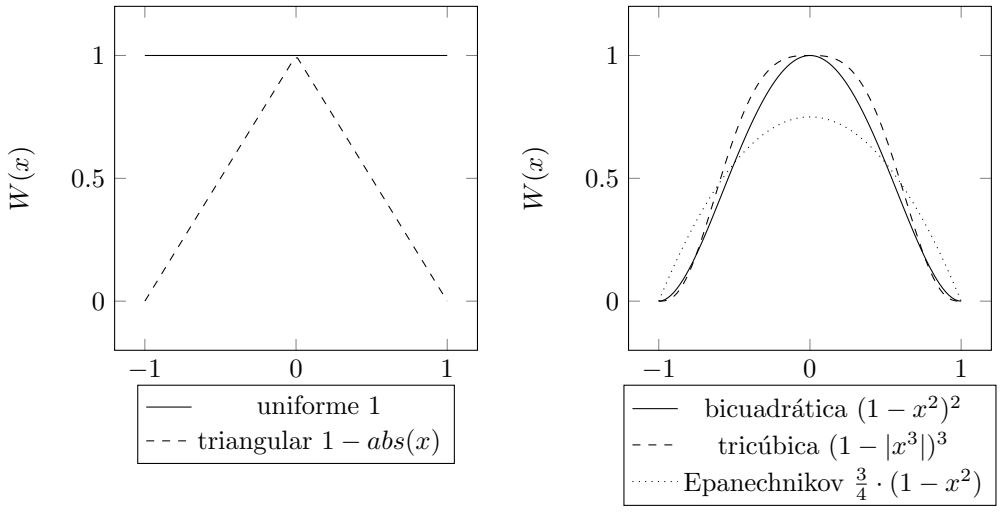

Figura 14: *Funciones de peso, a la izquierda uniforme y triangular, a la derecha bicuadrática tricúbica y Epanechnikov.*

Hipótesis

Las funciones uniforme y triangular devolverán regresiones poco "suaves" y más alejadas del valor real de la función. Entre las otras funciones, bicuadrática, tricúbica y Epanechnikov se espera un comportamiento parecido, con Epanechnikov siendo una función un poco más "suave". Notemos que la función uniforme no califica como una función de peso debido a que no es monotona decreciente, sin embargo la agregamos de todas formas para poder experimentar con una función no adecuada.

Resultados

La figura 15 confirma que la función uniforme tiene un mal ajuste, teniendo muchos saltos y no asemejándose a las funciones originales. Por otro lado la función triangular tuvo mucho mejor rendimiento del esperado obteniendo regresiones que se asemejan a las de las otras funciones más complejas. Entre el resto de las funciones no hay grandes diferencias, con las funciones bicuadráticas y tricúbicas siendo casi identicas. Como extensión se podría experimentar sobre más tipso de

funciones, quizas aquellas con cambio más bruscos arrojen resultados distintos.

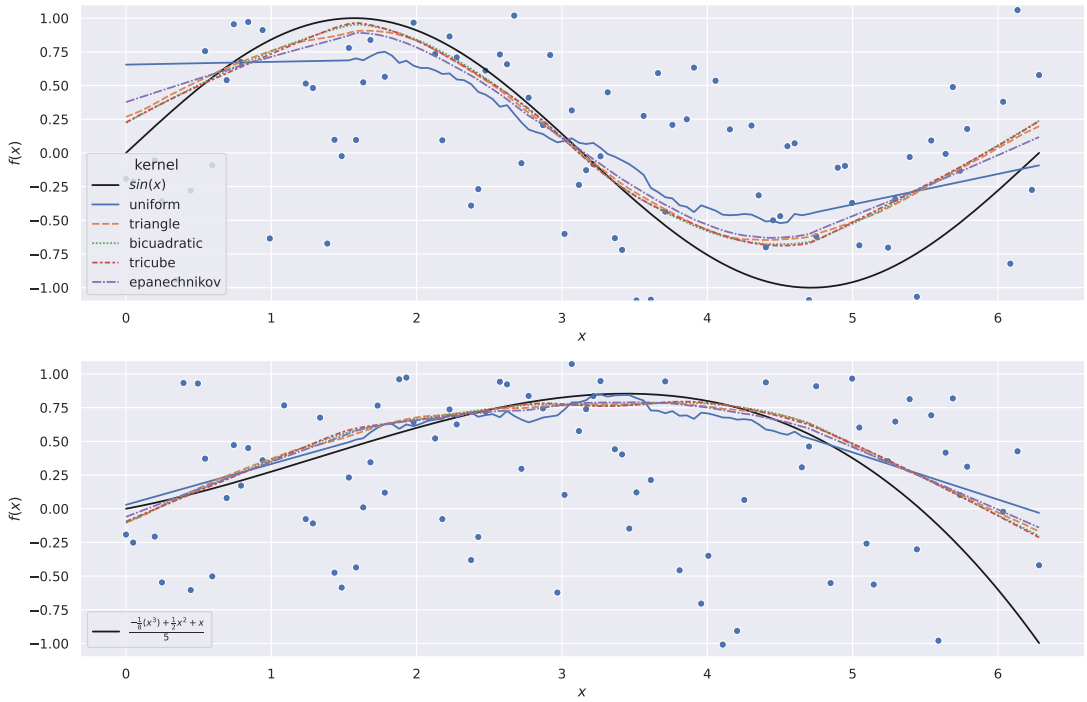

Figura 15: *Regresiones locales utilizando distintos kernels sobre una función trigonométrica (arriba) y un polinomio cúbico (abajo)*

2.4 Manipulando el ozono

Objetivo

Este experimento tiene como objetivo exponer como debería haberse comportado el error para poder llevar a cabo el *loess* $f = 0.4$ en el estudio de la dependencia del ozono en función a la temperatura, velocidad del viento y radiación solar (visto anteriormente en la sección 1.7). Para esto, se utilizó el conjunto de mediciones llevado a cabo por los autores del paper. Como ya se estudió, el error de las mediciones no tenían una distribución normal y como consecuencia no era sensato utilizar la regresión local lineal pesada con $f = 0.4$. En este experimento se reemplazaron los valores del ozono por la siguiente función $g : \mathbb{R}^3 \to \mathbb{R}$

$$g(r, t, v) = 2 \cdot r + t - v + \epsilon.$$

Donde ϵ es una variable aleatoria con distribución normal estándar. Luego, se realiza el *loess* con $f = 0.4$ y realizaremos el mismo estudio del error, mostrando el caso "óptimo".

Hipótesis

Al haber modificado los datos, si graficamos el qqplot se debería poder observar como los residuos en función a los percentiles de la función normal forman una línea "casi" recta. Como no debería haber ningún tipo de sesgo, la recta debería asemejarse a $y = x$. Con respecto al análisis de residuos en función de los valores ajustados (Residues vs Fitted values), los errores deberían ser tales que los residuos sean "balanceados". Esto implica, que al graficar la regresión con *loess* debería formarse una función casi igual a la recta $y = 0$. En los *Component-residual plots*, al igual que en el gráfico anterior, los errores deberían tener la misma disposición al graficarlos en función a los predictores. De la misma forma, el loess expone la semejanza a $y = 0$.

Resultados

En la figura 16 podemos observar como los residuos en función a los percentiles de la distribución normal forman una recta con pendiente uno, indiciando que no hay ningún tipo de sesgo y muestra que el error tiene una distribución normal.

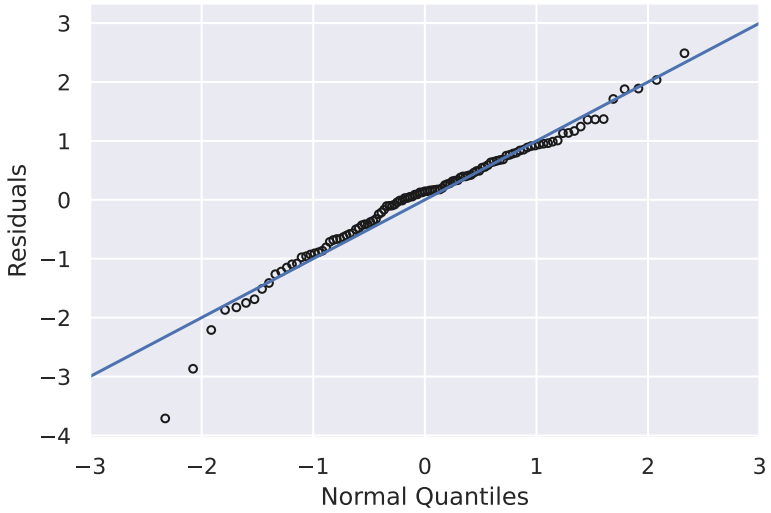

Figura 16: qqplot*, cuantiles de los residuos al haber hecho* loess *con* $f = .4$ *en función a los cuantiles de la distribución normal.*

En la figura 17 se ve que los residuos en función a los valores ajustados son "balanceados"(tenemos más o menos las mismas densidades de residuos negativos y positivos). Al graficar *loess* también exponemos que la regresión permite predecir que el residuo para un punto nuevo estará cerca del cero.

Figura 17: *Residues vs Fitted values*

Finalmente, en la figura 18 se puede apreciar que sucede lo mismo que en el gráfico anterior, al ver los residuos en función de las variables independientes se puede predecir que el residuo será cero.

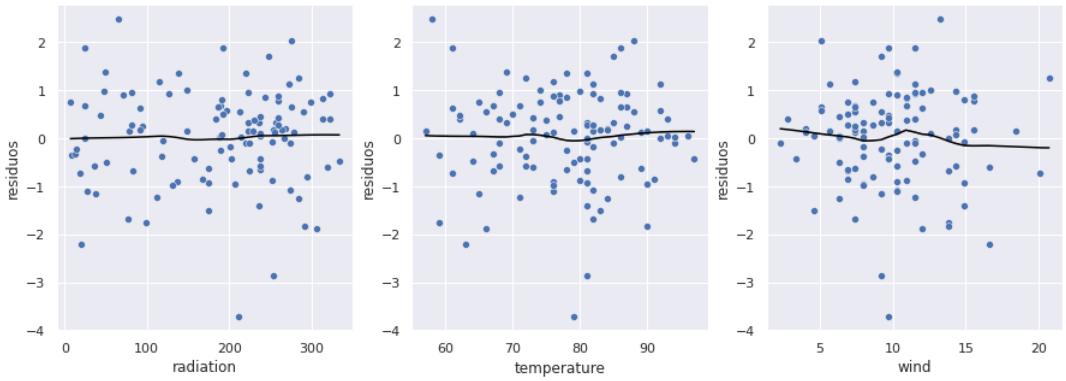

Figura 18: Component-residual plots

2.5 Experimentando con qqplots.

Objetivo

Se harán tres conjuntos de 1000 muestras sintéticas. Las muestras estarán formadas por tres variables aleatorias X, Y y Z y una variable dependiente que será una combinación lineal de las

anteriores más un ϵ, que agregue ruido. Luego, todas las muestras son de la forma

$$(x^{(i)}, y^{(i)}, z^{(i)}, f(x^{(i)}, y^{(i)}, z^{(i)}))$$

donde

$$f(x, y, z) = x + y - z + \epsilon.$$

El objetivo es experimentar con el ϵ, corriendo el experimento para tres distribuciones distintas, y viendo como es el *qqplot* en relación a la distribución normal. La idea es, ver como se verían los *qqplot* al chequear si este tiene distribución normal para ver si se puede aplicar *loess* lineal al conjunto de datos base. Se lo correrá para la distribución exponencial con $\lambda = 50$, y para la uniforme con valores en el intervalo [0, 100].

Hipótesis

Como se discutió antes, en el caso en que el ϵ tenga una distribución normal se observará que se forma una recta en el *qqplot* análoga a $y = x$. Sin embargo, la idea es descubrir como se verían los gráficos en el caso en que el error tiene distribución uniforme y exponencial, donde sabemos que no se asemejará a una recta el gráfico.

Resultado

Se puede ver en la figura 19 que, como era de esperar, los únicos datos que forman una recta son aquellos en los cuales el ϵ tiene distribución normal estándar.

Cuando el error se distribuye de forma exponencial, se puede observar mucha mayor densidad para los primeros cuantiles, mientras que a medida que se toman en cuenta los cuantiles positivos de la función normal baja la densidad de los errores. Sin embargo los residuos toman valores sumamente grandes. En el caso de la distribución uniforme, se puede ver como todos los cuantiles siempre se mantienen igual de densos.

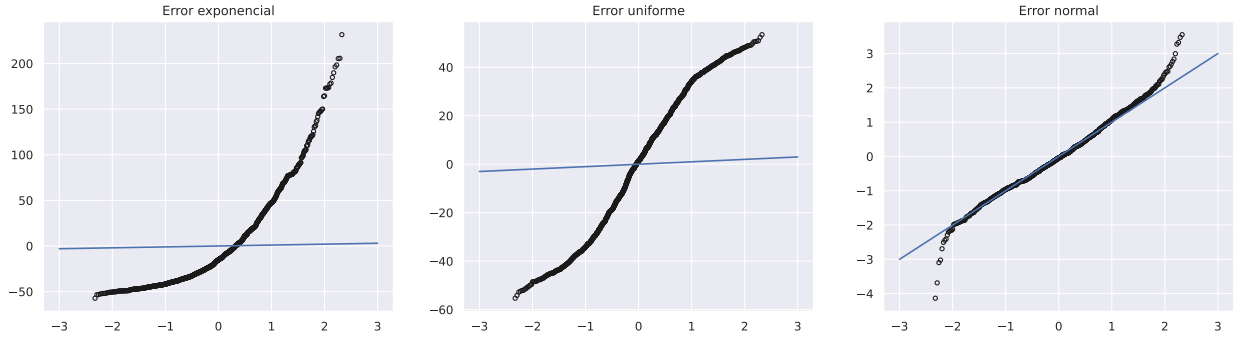

Figura 19: qqplots *para cuando el error tiene distribución exponencial, uniforme y normal.*

3 Conclusiones

El método de regresión *loess* nos permitió realizar regresiones lineales o cuadráticas locales, para vecindarios de distintos tamaños. Luego, se pudo experimentar con las variaciones de la regresión junto con el parámetro f (proporción de los vecindarios). Pudimos ver que, para valores muy bajos de f independientemente de la regresión que se utilice se producirá *overfitting* sobre los puntos del conjunto de datos. Adicionalmente, tener un f cerca de uno producirá una recta en el caso lineal y una función suave con más curvas en el caso cuadrático.

También se pudo comparar los errores al utilizar las regresiones en el caso en que la variable target fuera lineal o cuadrática. Se pudo ver que la diferencia no es demasiado significativa, pero fueron mejores la regresión lineal en el caso lineal y la cuadrática en el caso cuadrático.

También se pudo experimentar con las funciones de peso, vimos que la función uniforme tiene muy se ajusta muy mal a las funciones (de todas formas, la uniforme al no ser monótona decreciente no tiene que ser utilizada para *loess*), la función triangular fue mejor de lo esperado y todas las restantes fueron lo suficientemente buenas.

Adicionalmente se modificaron los datos provistos por los autores del paper, y vimos como debería verse el estudio del error cuando sabemos que tiene la distribución adecuada. Por último, pudimos experimentar con los *qqplots* para errores con distintas distribuciones.

Referencias

[1] William S. Cleveland and Susan J. Devlin. “Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting”. In: *Journal of the American Statistical Association* (1988).

[2] *Conditioning plots*. URL: `https://www.itl.nist.gov/div898/handbook/eda/section3/condplot.htm`.

[3] *Undersanding qqplots*. URL: `https://data.library.virginia.edu/understanding-q-q-plots/`.