DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico

Trabajamos y nos divertimos

16 de julio de 2022 — Métodos Numéricos

Integrante	LU	Correo electrónico
Embon, Eitan	610/20	eembon1@gmail.com
Imperiale, Luca	436/15	luca.imperiale95@gmail.com

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. Resumen

En este trabajo, vamos a hacer el ejercicio de leer dos paper y entenderlos, vamos a tratar de replicar una sección de uno de ellos, tratando llegar a sus mismos resultados.

Vamos a estudiar y tratar de implementar una técnica de regresión lineal local conocida como *loess*. En la introducción hacemos un planteamiento teórico de del *loess* en su versión univariada basándonos en las explicaciones de [1], y en su versión multivariada [2]. Luego estos replicamos la sección cinco de [2] y detallamos cada uno de sus gráficos desde nuestra interpretación en la sección desarrollo. En la sección de experimentación, variamos los distintos parámetros que constituyen al método de *loess* y observamos en que resulta. Por último, en la conclusión tratamos de diferenciar la regresión local de la regresión global desde los conocimientos adquiridos en este trabajo.

2. Introducción

2.1. Modelo

En este trabajo vamos a aplicar y entender una técnica de regresión conocida como **loess** (*"locally estimated scatterplot smoothing"*).

Se puede hacer una analogía del *loess* con las *series de tiempo*, las cuales es hacer varias mediciones en diferentes momentos de tiempo, todas bajo las mismas circunstancias, y hallar una relación entre los parámetros dentro de nuestras mediciones. Queremos, en general, hacer un modelo de predicción de una variable a partir de una relación, en nuestro caso lineal, o polinomial, de otros parámetros. A la variable que se quiere predecir se la conoce como variable dependiente, y a las demás como variables independientes.

Este problema hace uso de lo que se conoce como regresión lineal, es decir, poner la variable dependiente en función del la o las variables independientes. Cuando el modelo es univariado, de una sola variable independiente se le cono regresión lineal simple, y cuando es multivariado, es decir con mas de una variable independiente, se lo conoce como regresión lineal múltiple o multivariada.

En general contaremos con $(x_i, y_i)i = 1...m$, m mediciones independientes, si es univariado, de la forma:

$$l(y_i) = \beta_0 + \beta_1 h(xi) + \epsilon_i, \forall i = 1...m \tag{1}$$

Siendo $h(x)$ y $l(x)$ funciones con misma dimensión de salida como de entrada, por ejemplo una estandarización, y ϵ el error de la aproximación que, importante, se necesita que sea de una distribución con varianza constante y una media centrada en el cero, muchas veces se asume que el error tiene distribución normal con varianza uno y media cero.

En el problema multivariado no vamos a tener, unas sola variable, si no varias, en general:

Aclaramos que damos por hecho que ya se aplicaron las funciones, por ejemplo de estandarización, a las variables.

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... + \beta_n x_i^{(n)} + \epsilon_i, \forall i = 1...m \tag{2}$$

El problema ahora se transforma en encontrar los coeficientes de cada ecuación, lo podemos resolver de la misma manera que cuadrados mínimos. Es decir, si llamamos $x_i = (1, x_i^1, x_i^1, ..., x_i^n)$ la fila i-ésima de una matriz X, a $B^t = (\beta_0, \beta_1, ..., \beta_n)$, y a $y^t = (y_1, ..., y_n)$ Podemos pasar a resolver cuadrados mínimos lineales del sistema.

$$XB = y \tag{3}$$

Esto así nos va a otorgar una solución, que va a ser una función lineal, que es la solución a la regresión lineal que teníamos en el principio.

Ahora, si llamamos $g(x)$ una función tal que aproxime a y, siendo $g(x_i) = y_i - \epsilon_i$, entonces tenemos que en el caso anterior $g(x)$ era un polinomio lineal. Esta función $g(x)$ en principio podría ser cualquier tipo de función, pero para nuestra conveniencia, elegiremos una función lineal o polinomial. Esto por las cualidades de las funciones lineales o polinomiales, que entran dentro de un grupo conocido en álgebra como transformaciones lineales.

En principio, en este trabajo $g(x)$ va a ser un polinomio en $\mathbf{R}$ de grado uno o dos. Para grado uno el problema quedaría como antes planteado en 4, para grado dos, quedaría, en el caso de una regresión múltiple, si tuviera tres variables independientes:

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \beta_3 {x_i^{(1)}}^2 + \beta_4 {x_i^{(2)}}^2 + \beta_5 x_i^{(1)} x_i^{(2)} + \epsilon_i, \forall i = 1...m \tag{4}$$

De esta forma pasamos a tener, de $n + 1$ ([1]) incógnitas en el caso lineal a tener $n + 1$, los casos de una variable mas

[1] Suponiendo que hay n variables independientes

el coeficiente independiente, $+n$, las variables al cuadrado, $+\binom{n}{2}$, la multiplicación de las variables cruzadas.

Llegado al caso, podríamos suponer que a mayor grado de $g(x)$, al tener mas incógnitas, vamos a poder adaptar o la función de regresión a interpolar mayor cantidad de puntos de la distribución real, y por ende aproximar mejor. La ecuación matricial nos quedaría de la misma forma que en 3, solo que la matriz X tendría igual cantidad de columnas que incógnitas a encontrar y, el vector B, que sería el que contenga las incógnitas $\beta_0...\beta_k$.

La idea general detrás de *loess* es aplicar una regresión lineal local, es decir, para cada medición se va a hacer una regresión local pesada poniendo como eje, o centro de gravedad, a la medición actual. Esto es, que multiplicaríamos a cada medición, por el peso que le queremos dar en nuestra regresión. El peso viene dado, en general, por la distancia de norma dos a la medición tomada como pivote actual. A ese peso se le aplicará una función $w_i(x_j)$, la función de peso para la j-ésima medición tomando como eje o pivote a la i-ésima medición, tal que, siendo $\rho(x_j)_i = |x_i - x_j|$, y $d(x_i) = \text{máx}_j\, |x_i - x_j|$:

$$w_i(x_j) = W(\frac{\rho(x_j)_i}{d(x_i)}) \tag{5}$$

$W(\mu)$ deberá cumplir las siguientes condiciones.

- $W(\mu) > 0$ para $|\mu| < 1$
- $W(-\mu) = W(\mu)$
- $W(\mu)$ no es una función creciente para $\mu \geq 0$
- $W(\mu) = 0$ para $\mu \geq 1$

Para este trabajo se usa para la W a la función tricúbica definida como:

$$W(\mu) = \begin{cases} (1 - |\mu|^3)^3, 0 \leq |\mu| < 1 \\ 0, 1 \leq |\mu| \end{cases} \tag{6}$$

Además de ser pesada la regresión, también es local, esto es, se define una variable $f = q/m$ siendo q el numero de mediciones o vecinos que se toman para cada regresión distinta. El parámetro m sería como definimos con anterioridad, la cantidad de mediciones distintas hechas. Así, la cantidad de participantes en cada regresión estará acotada por f que será una proporción de m.

Ahora de esta forma, la ecuación para $y_i, i = 1...m$ seria, para el pivote k-ésimo:

$$w_k(x_i)y_i = w_k(x_i) * g(x_i) + e_i^{(k)} \tag{7}$$

Observar que de esta forma, tendríamos m vectores soluciones $(\beta_0...\beta_n)$, para cada pivote tomado.

La ecuación matricial sería la misma que 3, solo que multiplicaríamos a ambos lados de la igualdad, la matriz W diagonal, que contiene en su iésmo elemento de la diagonal, al numero $w_k(x_i)$ en la regresión que tiene como eje al pivote k-ésimo.

$$W_kXB = W_ky \tag{8}$$

Aplicando ecuaciones normales, la solución $(\beta_0...\beta_n)$ k-ésima sería $B \in \mathbf{R^{n+1}}$ tal que

$$X^tW_kXB = X^tW_ky \tag{9}$$

3. Desarrollo

3.1. Replicamos sección 5 de [2]

3.1.1. Conociendo el dataset

Los datos para el desarrollo de esta sección son 111 mediciones de cuatro variables (ozono, radiación solar, temperatura y velocidad del viento). Se analiza la posible dependencia de la variable ozono en función de las demás, con lo que se va a poder predecir la cantidad de ozono en la atmósfera dependiendo de las otras 3 variables. En la figura (1) podemos ver graficadas las 111 mediciones de las respectivas variables.

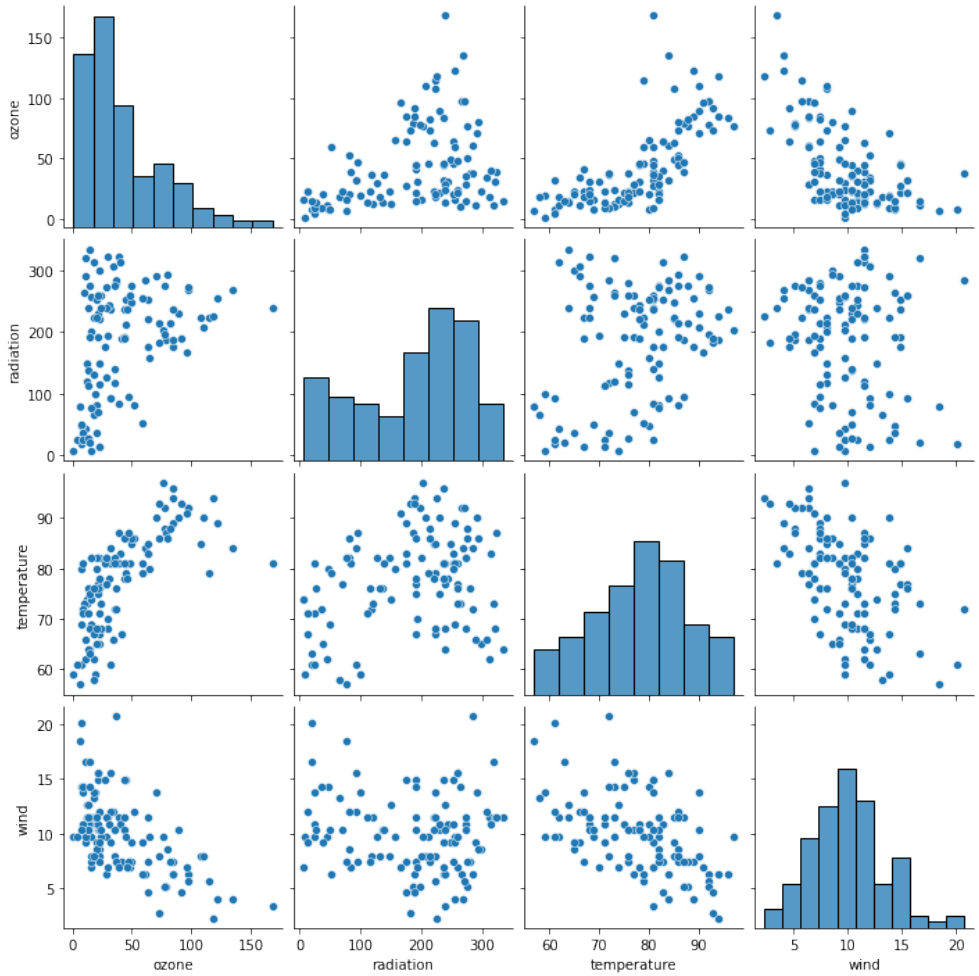


Figura 1: Scatterplot de las 111 mediciones

3.1.2. Asunciones criticas

Para aplicar el método de *loess*, es importante reconocer que se deben tener ciertas asunciones criticas. Una es que los errores ϵ_i son independiente y normalmente distribuidos con varianza constante. Otra es que la función entrenada o resultante del *loess* siga el patrón de los datos, lo que provee probablemente un estimador insesgado. Estas asunciones deben ser chequeadas.

Es importante observar que los errores sean con varianza constante, eso nos va a otorgar un predictor que no varié el error en su resultado en función del tamaño de los datos, o del valor del parámetro f. Para verificar esto, se puede hacer un *qqplot* del residuo en la regresión versus una distribución normal, para chequear la asunción de normalidad.

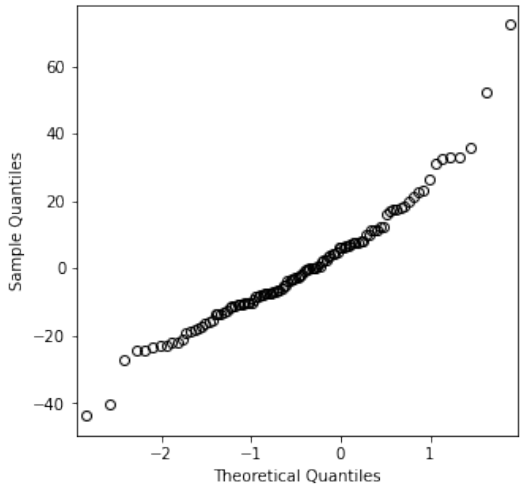


Figura 2: Esta imagen muestra los cuantiles del residuo $\hat{\epsilon}$ contra una distribución normal, se usaron parámetros de $f = 0{,}4$ y grado$(g(x)) = 1$

Luego para chequear una varianza constante se puede hacer un plot de $|\hat{\epsilon}|$ contra $\hat{y}$.

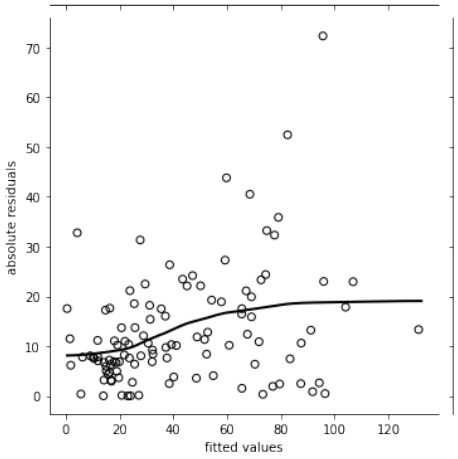


Figura 3: Esta imagen muestra a los errores $|\hat{\epsilon}|$ contra los valores entrenados $\hat{y}$, se usaron parámetros de $f = 0{,}4$ y grado$(g(x)) = 1$

Podemos observar que hay una pequeña alteración en la curva del error, con lo que podemos suponer que no es totalmente constante para diferentes resultados de los valores entrenados.

Se puede ahondar mas en la causa de esta distorsión, viendo como afectan los cambios en las variables independientes al error. Esto también permitirá entender el sesgo del estimador.

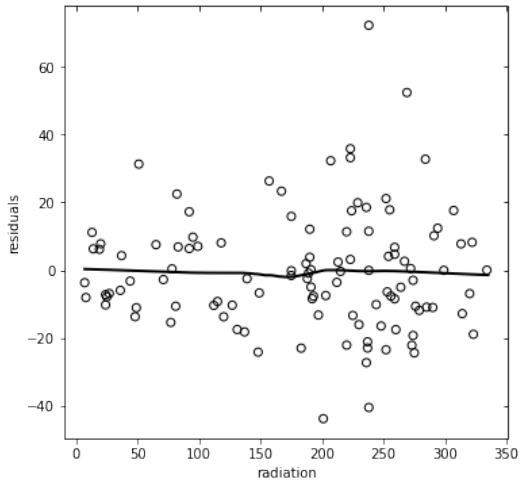


Figura 4: Esta imagen muestra a los errores $\hat{\epsilon}$ contra las mediciones de radiación solar, se usaron parámetros de $f = 0{,}4$ y grado$(g(x)) = 1$ para el scatterplot y utilizamos *loess* con $f = 2/3$ y grado$(g(x)) = 1$ para la regresión lineal.

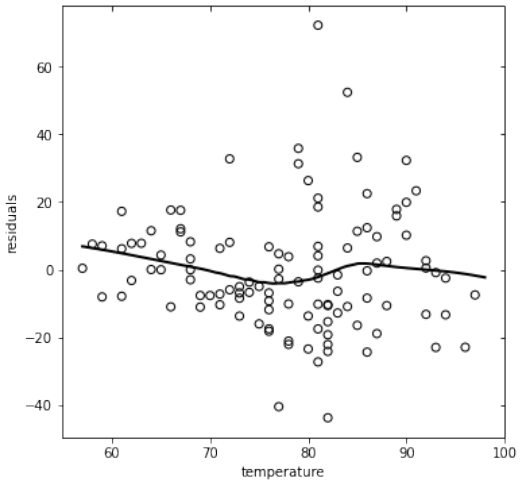


Figura 5: se usaron parámetros de $f = 0{,}4$ y grado$(g(x)) = 1$ para el scatterplot y utilizamos *loess* con $f = 2/3$ y grado$(g(x)) = 1$ para la regresión lineal.

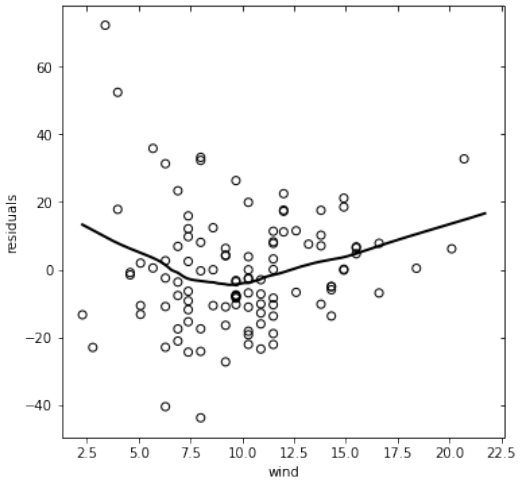
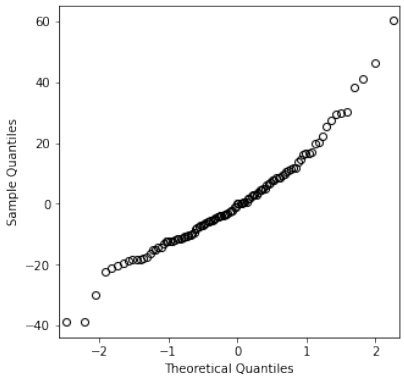


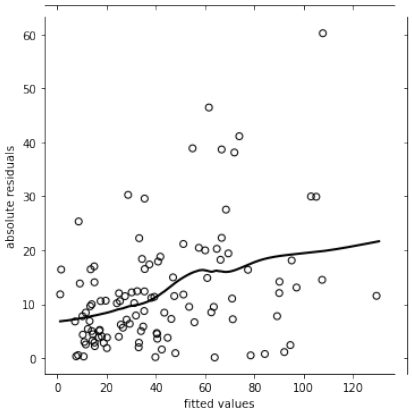
Figura 6: Esta imagen muestra a los errores $\hat{\epsilon}$ contra las mediciones de velocidad de viento, se usaron parámetros de $f = 0{,}4$ y grado$(g(x)) = 1$ para el scatterplot y utilizamos *loess* con $f = 2/3$ y grado$(g(x)) = 1$ para la regresión lineal.

A primera vista se puede notar que el error es constante a lo largo de la variable radiación pero que tiene ligeras distorsiones al variar las otras dos variables, con lo que podemos suponer que es, en parte, debido a estas dos desviaciones por lo que se produce una distorsión en la figura 3.

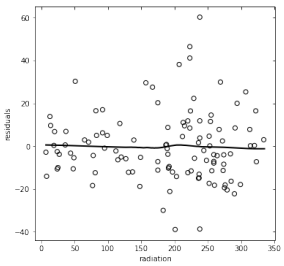
En [2] consideran que al decrecer el parámetro f se podría reducir aún mas la distorsión en la figura anteriores, pero esto supondría arriesgar la suavidad de la función regresora. Con lo cual decidieron aumentar el grado de g a cuadrático y utilizaron $f = ,8$ lo cual eliminó la distorsión, pero las inadecuaciones en la figura 2 permanecieron.



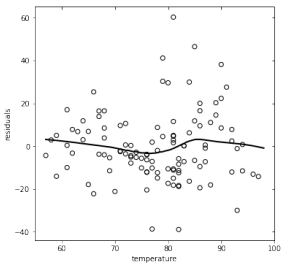
(a) Qqplot de los residuos.



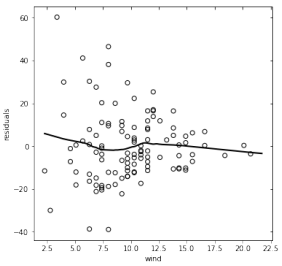
(b) Valores absolutos del residuo contra los valores entrenados



(c) El residuo contra la radiación



(d) El residuo contra la temperatura



(e) El residuo contra el viento

Figura 7: Se tomo devuelta el experimento con los valores de $f = 0{,}8$ y grado($g(x)$)= 2.
La curva en las figuras es un *loess* fit con $f = 2/3$ y grado($g(x)$)= 1

Efectivamente se puede observar que las distorsiones en en los plot del valor residual contra los datos de las variables independientes ha mejorad y ya no es tan abrupta, pero la inadecuación en la figura 7a se mantiene.

Aún así se decidió continuar con dichos parámetros, que encontraron que eran los que mejor ajustaban, y los usaron para aproximar al ozono en los gráficos 8, 9 y 10. Se realizó una aproximación cuadrática sobre la raíz cúbica del ozono con $f = ,8$. Creemos que se sacó la raíz cúbica para que las curvas de la regresión sean mas estrechas o puntiagudas, y por ende sea mas visual.

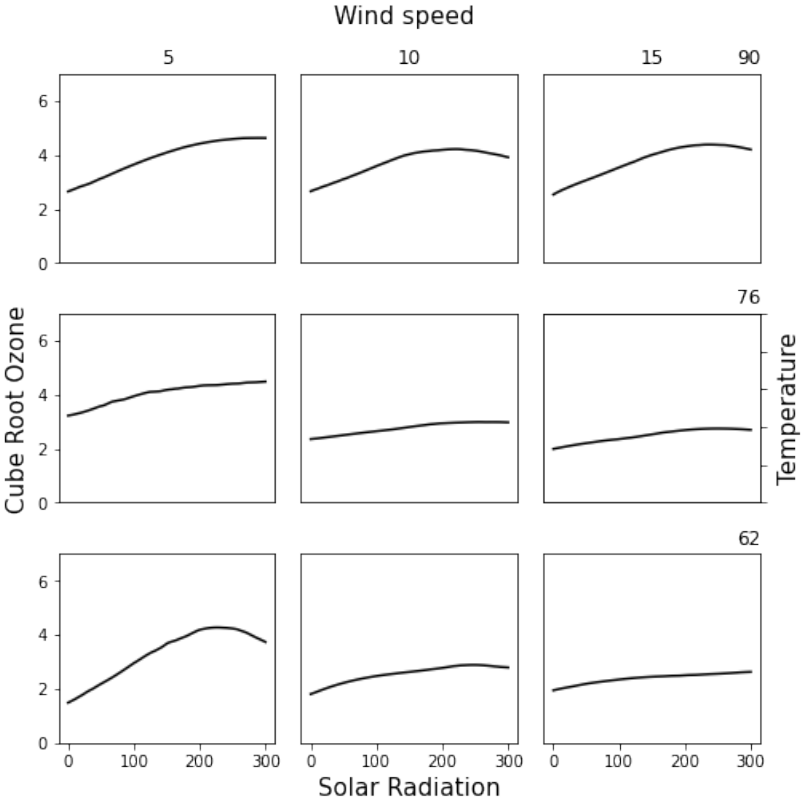


Figura 8: Ozono y los datos meteorológicos. Ozono fue transformada aplicándole la raíz cúbica y luego se realizó la curva usando los parámetros de $f = 0{,}8$ y grado$(g(x)) = 2$. Se grafica la raíz cúbica del Ozono en función de la radiación solar mientras se dajan en valores constante a los demás parámetros.

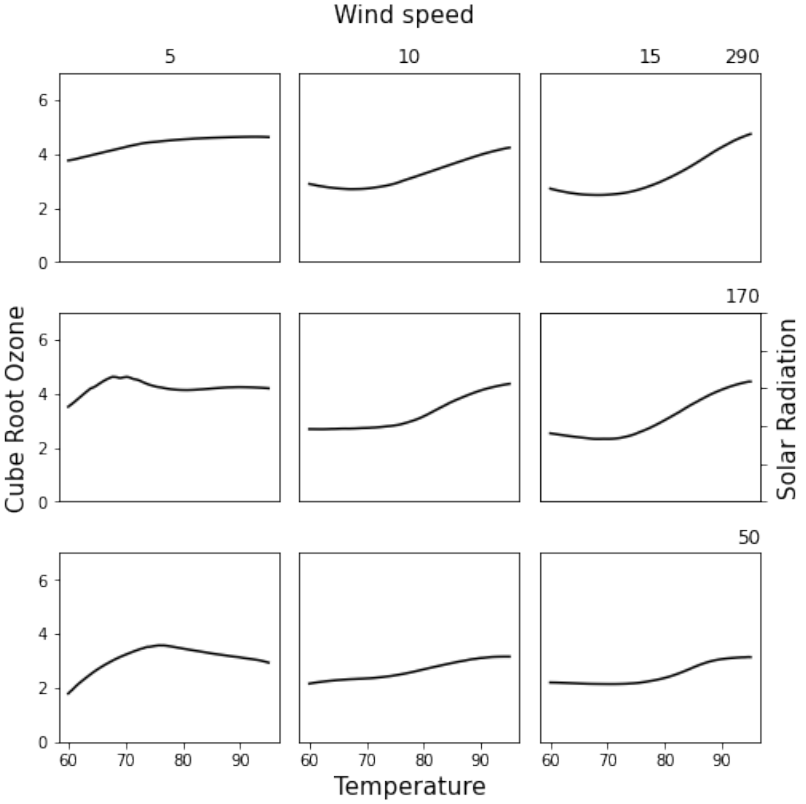


Figura 9: Ozono y los datos meteorológicos. Ozono fue transformada aplicándole la raíz cúbica y luego se realizó la curva usando los parámetros de $f = 0{,}8$ y grado$(g(x)) = 2$. Se grafica la raíz cúbica del Ozono en función de la temperatura mientras se dajan en valores constante a los demás parámetros.

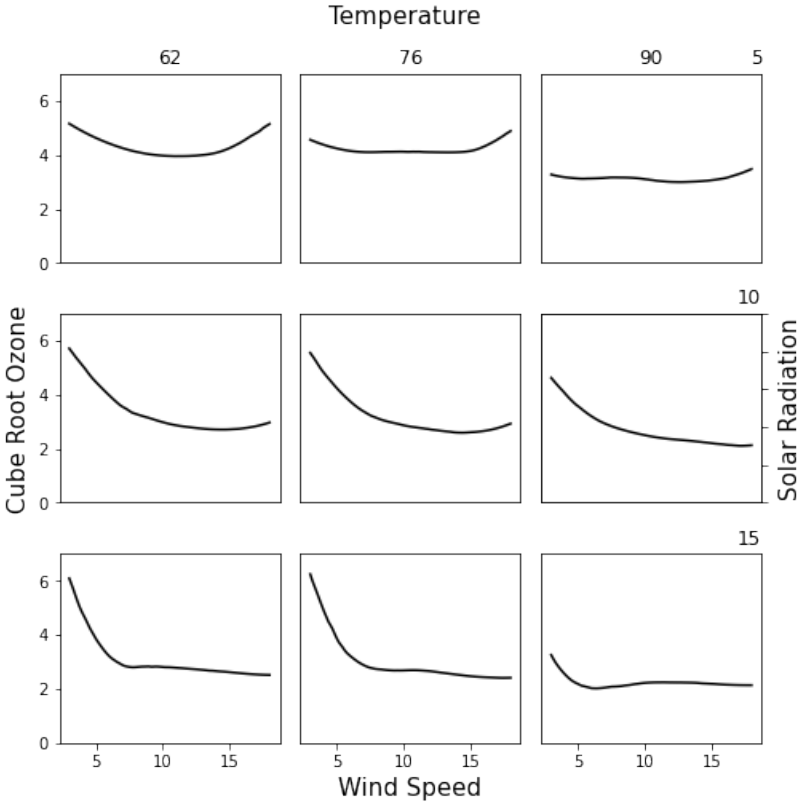


Figura 10: Ozono y los datos meteorológicos. Ozono fue transformada aplicándole la raíz cúbica y luego se realizó la curva usando los parámetros de $f = 0{,}8$ y grado$(g(x)) = 2$. Se grafica la raíz cúbica del Ozono en función del viento mientras se dajan en valores constante a los demás parámetros.

3.2. Interpretación de los gráficos

En esta sección vamos a tratar de explicar, a partir de nuestra interpretación, las razones y causas que llevaron a realizar los distintos gráficos que de la sección 5 de [2], las cuales replicamos en la sección 3.1.

Importante señalar que los gráficos explicados en 3.2.2, 3.2.3 y 3.2.4 son para observar asunciones críticas sobre los errores a la hora de estimar al ozono. Además también sirven de guía para encontrar o ajustar los parámetros del *loess* como f y g, de modo que el residuo sea lo mas chico posible y además cumplan las asunciones explicadas en las diferentes secciones arriba mencionadas. Luego en la sección 3.2.5 se aborda la cuestión de graficar la función regresora.

Una observación interesante para hacer es que antes de de realizar la regresión se requiere una estandarización de las variables independientes de de todas las mediciones. Hemos observado que las figuras de diagnóstico no tienen hecha esta estandarización, o por lo menos, si no estandarizamos podemos replicarlas correctamente, en cambio para la figura explicada en 3.2.5 si se necesita de esta estandarización. Pensamos que es posible que esta decisión se haya tomado en base a que en las figuras de diagnóstico tan solo se evalúa que se cumplan las hipótesis de normalidad, y la independencia del error con las variables independientes. En cambio en la la aproximación en si del ozono, donde se muestra la curva de regresión, si se estandariza.

3.2.1. Scatterplot de las 111 mediciones, figura 1

Antes de comenzar a experimentar y generar hipótesis se analiza el dataset y las variables de las mediciones, y sus relaciones entre si. Esto se hace para intuir el comportamiento que hay entre si, y además poder observar la distribución de cada una de las variables con los datos obtenidos.

En esta figura se presentan 16 scatterplots enfrentando todas las variables dos a dos, y viendo unas en función de otras y viceversa. En las figuras de la diagonal del medio se presentan los histogramas de las variables por separado, estos muestran la distribución de cada variable a partir de los datos obtenidos.

Observamos que la velocidad del viento, la temperatura y la radiación solar tienen una distribución parecida a la normal, en cambio el ozono tiene una distribución mas semejante a la distribución gamma.

3.2.2. qqplot del residuo contra una distribución normal, figura 2

Una asunción importante a la hora de realizar una estimación es que los errores sean normales, igualmente distribuidos e independientes. Se asume independencia porque es una propiedad que permite realizar ciertas operaciones que en algunos casos son preferibles. Por otro lado también es importante que los errores estén igualmente distribuidos, es decir que dependan de la misma distribución y con esto que se comporten de la misma manera. Por último se observa cuales son las distribuciones de los errores, en este caso se asume normalidad. Estas asunciones son imprescindibles para la predicción de errores, y que esta no sea caótica.

3.2.3. Valores absolutos del residuo contra los valores obtenidos de *loess*, figura 3

Otra asunción importante que se debe verificar es que la varianza de los errores sea constante, es decir que para todo $\hat{e_i}$ medido, tengan la misma varianza y por ende un error parecido.

Para verificar una varianza constante se grafica a $\hat{e_i}$ en función de $\hat{y_i}$, y queremos ver que la curva del error resultante es constante, es decir que para todo y_i los errores en sus aproximaciones son aproximadamente el mismo.

3.2.4. Residuo contra variables independientes, figuras 4, 5 y 6

El objetivo de estas figuras es verificar que los valores de la función regresora (*"the fitted function"*) sigan el patrón de las variables independientes, y por lo tanto que $\hat{y}$ resulte en un estimador cercano a insesgado. Esto se realiza graficando a el residuo o error que resulta de la estimación ($\hat{e} = y - \hat{y}$) contra las variables independientes. Se quiere ver que el error no dependa del valor de las variables independientes, es decir, que $\hat{e}(x) = c, c \in \mathbf{R}$, que el error sea constante para valores diferentes de las variables independientes. El objetivo va a ser encontrar esos parámetros f, g tales que el error sea lo mas constante posible en función de diferentes valores para las variables independientes.

3.2.5. Regresión del ozono en función de las demás variables independientes, figuras 8, 9 y 10

En las secciones 3.2.2, 3.2.3 y 3.2.4 la idea principal es encontrar o ajustar los parámetros de f y g tales que se cumplan ciertas asunciones criticas distintas, que hemos mencionado a lo largo de esta sección. Una vez encontrados, lo siguiente es observar el comportamiento de los valores predichos de la variable independiente (Ozono) para diferentes valores de las variables dependientes (*Wind speed, temperature, solar radiation*).

Estos gráficos tienen el inconveniente de estar en dos dimensiones, con lo qué para graficarlos se termina dejando valores fijos para dos variables independientes y valores libres para la variable restante. Así quedan tres gráficos mostrando la variabilidad del ozono en función de las variables independientes por separado.

Una observación interesante es que se grafica la aproximación a la raíz cúbica del ozono, esto en su escencia no es un cambio de experimento, ya que el resultado de esta aproximación termina siendo la raíz cúbica,(aproximadamente) , del valor aproximado normal ($\sqrt[3]{\hat{y}}$). Esta transformación sobre la curva resultante termina achatando el gráfico como se puede ver en la figura 11. Podría ser que hubiera una decisión de que el gráfico sea de esta forma para una mejor observación visual desde alguna perspectiva.

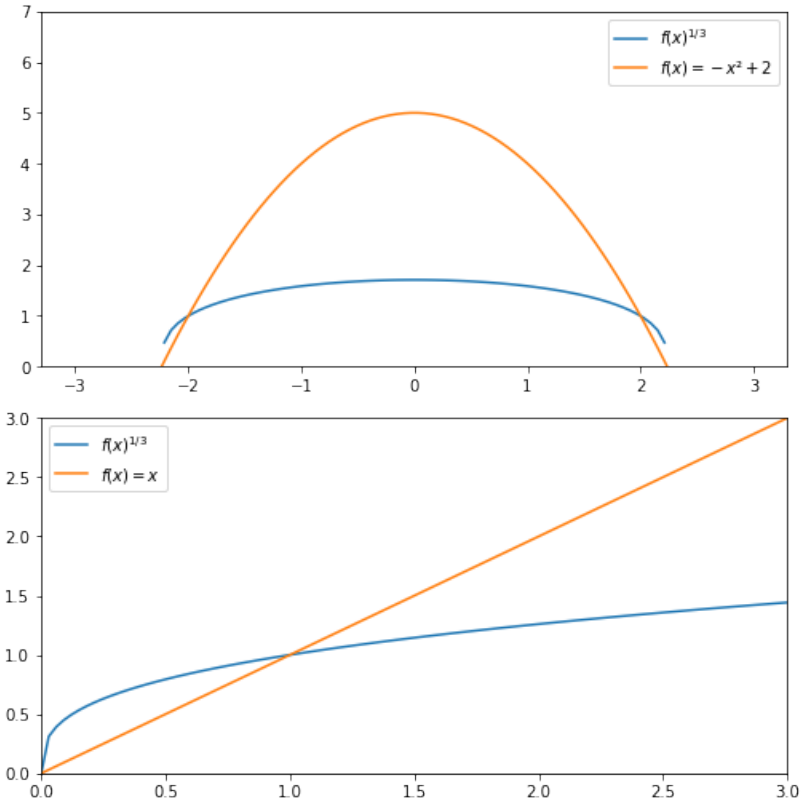


Figura 11: Raíz cúbica aplicada diferentes funciones

4. Experimentación

4.1. Reproducción de experimentos

En el archivo *experimentacion.ipynb* está toda la experimentación, así como los gráficos, implementados a lo largo de este trabajo.

Los experimentos fueron realizados en la siguentes condiciones:

- Ubuntu 20.04.4 LTS
- Intel Core i7-8550U CPU @1.80Ghz x 8
- 16 GB RAM
- GCC version 9.4.0 compilada con -03 flags

4.2. Data Sintética

Para la experimentación usamos datos creados sintéticamente por nosotros, con el objetivo de que la variación en los parámetros resulte mas expresiva en los gráficos.

Nuestros datos a experimentar constan de dos variables, una independiente y otra dependiente, lo que hicimos fue encontrar una función conocida y agregarle ruido para que el scatterplot no sea muy exacto, y usando *loess* deberíamos llegar a una representación de la función original.

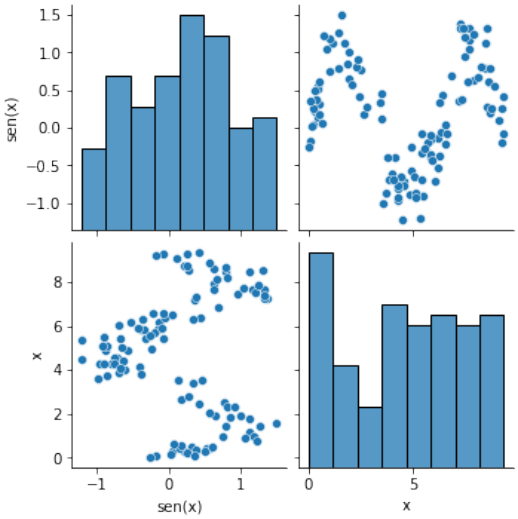


Figura 12: Scatterplot de la data sintética

Como se puede observar, la función que implementamos es el $sen(x)$.

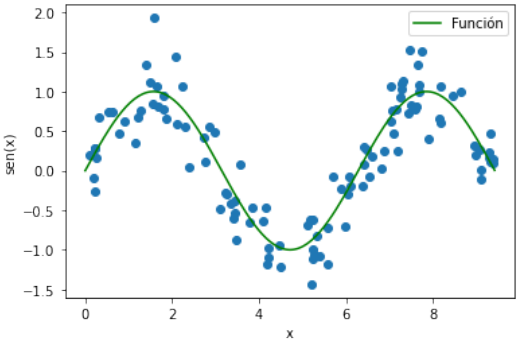


Figura 13: Scatterplot mas la función original de la data sintética

Cabe aclarar que el ruido que se le agrego a los puntos de la función original fueron de una distribución normal, con lo que cabe esperar que el error en la aproximación sea normal.

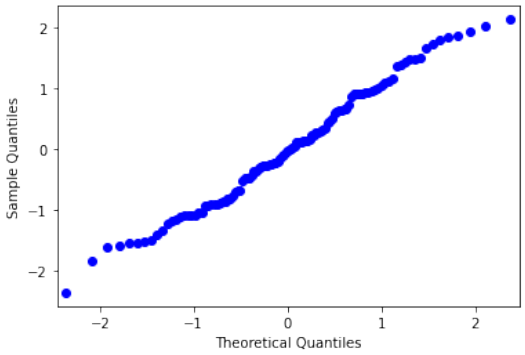


Figura 14: Qqplot del residuo contra una normal, se tomo $f = 0{,}5$, grado$(g) = 2$ y se usó la función tricúbica como peso

4.3. Experimentación con el grado g(x)

En general, tiene sentido que el polinomio que usemos para la regresión sea del mayor grado posible, ya que, a mayor grado, mayor es la facilidad de interpolar puntos mas variados, ya que una función con mas grados tiene mas facilidad para curvarse. En nuestro caso, experimentaremos solo con una función $g(x)$ de grado 1 y luego una de grado 2.

Una observación posible, es que el el agregar un grado a la función g nos aumenta la cantidad de columnas, y tener una matriz X con mas columnas que filas podría resultar innecesario y costoso de conseguir.

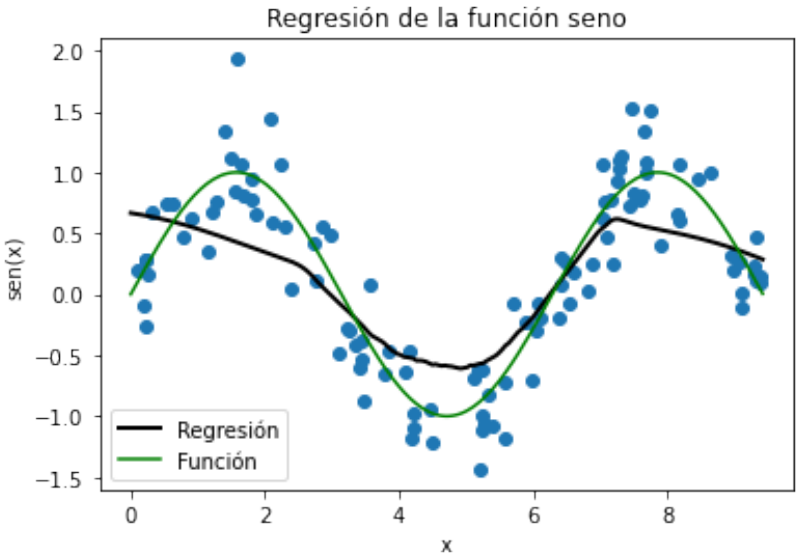


Figura 15: Regresión con $f = 0{,}5$, y grado$(g) = 1$

En el anterior gráfico podemos observar que para un $f = 0{,}5$ aceptable la función de regresión no se adecua totalmente a la función original.

Si miramos la figura 16 podemos encontrar una pequeña distorsión en la variación del error.

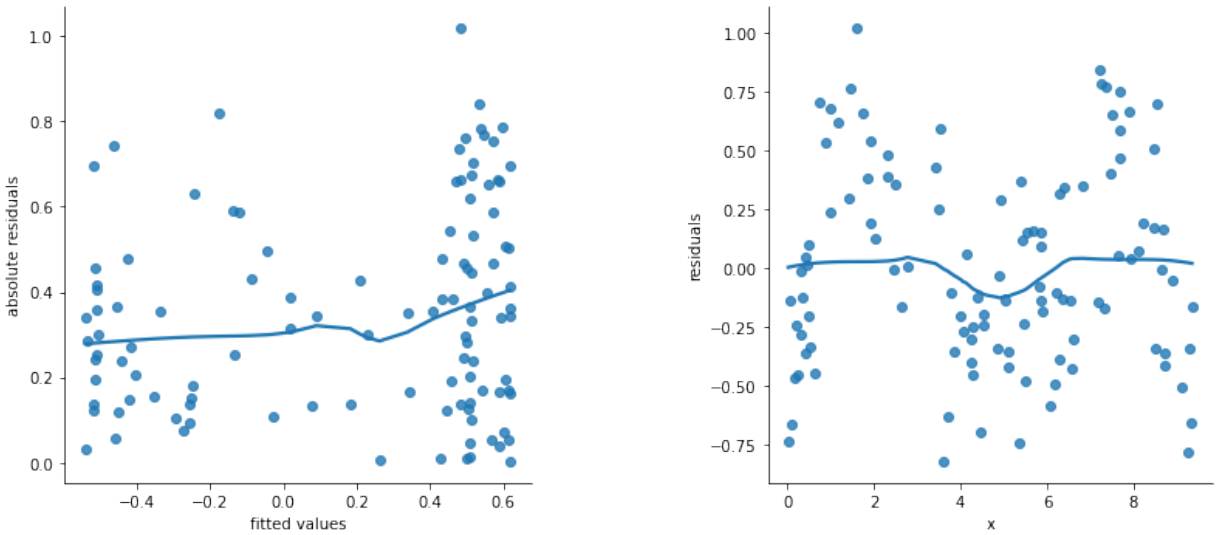


Figura 16: Variación del residuo

En cambio para grado dos, hay una mayor adaptabilidad de la función regresora, con lo que su error de aproximación tiende a ser mas constante e invariable.

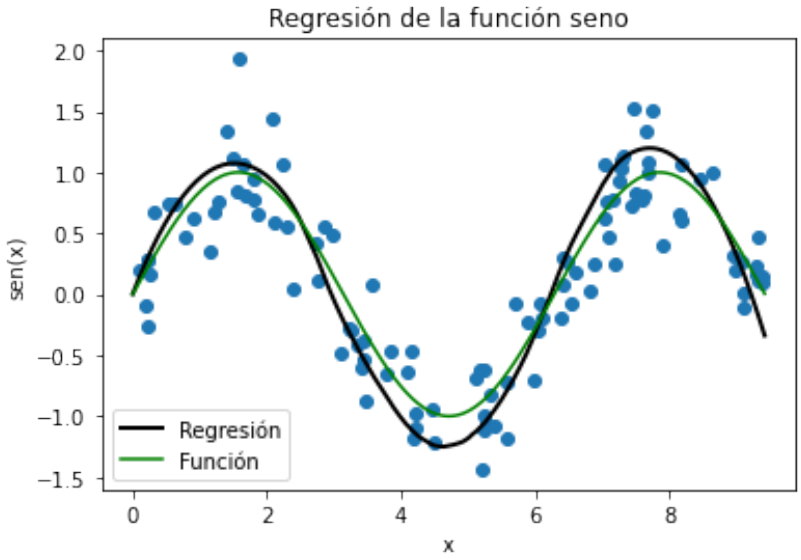


Figura 17: Regresión con $f = 0{,}5$, y grado$(g) = 2$

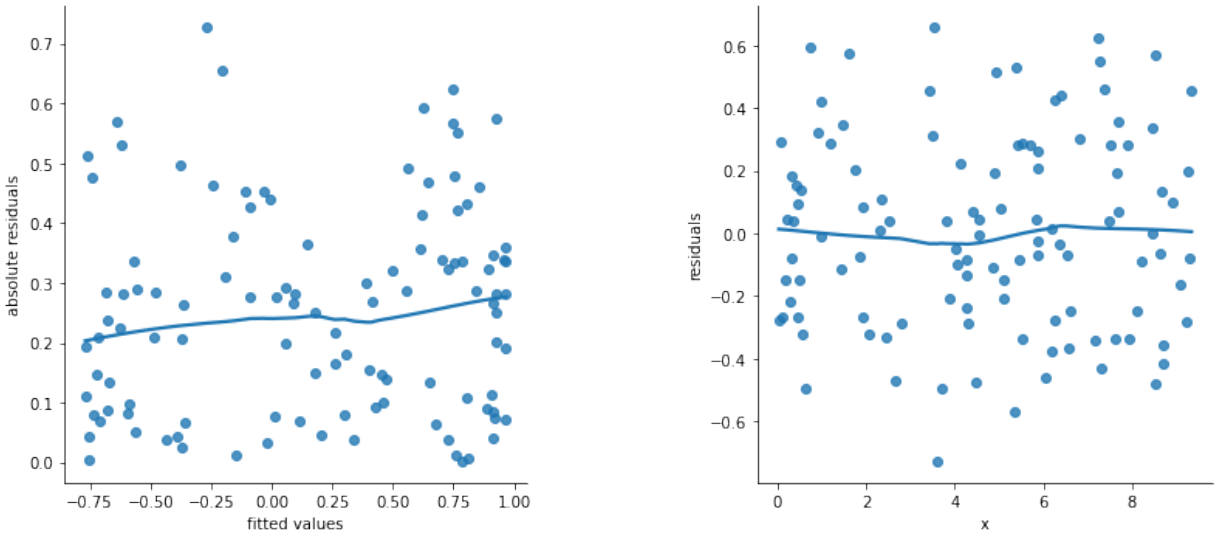


Figura 18: Variación del residuo para grado 2

4.4. Variamos la función de peso

A la hora de experimentar en el cambio de la función de peso hay que tener en cuenta ciertas propiedades, o restricciones en este caso, que debe cumplir, específicamente estas (2.1).

Analicemos la función tricúbica y cual sería su efecto en el peso de cuadrados mínimos. Recordemos la ecuación 6.

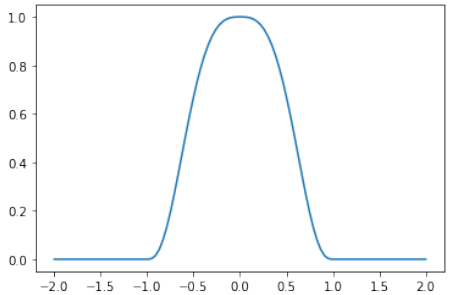


Figura 19: Función tricúbica

Esta función es simétrica, además si el dato de entrada está mas cerca del cero, devuelve casi uno, y si el dato es mas grande que uno, devuelve 0.

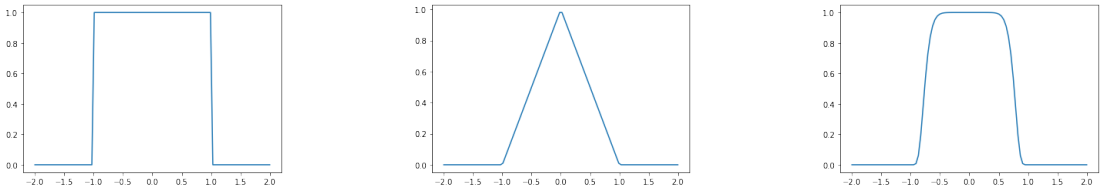
Como acá, en la ecuación (5) esta definido, el dato de entrada $0 \leq \mu \leq 1$ y μ esta definido en función de la distancia al pivote normalizado. Con lo que la función lo que hace es darle un mayor peso cuando la distancia es muy chica, y cuando es grande le da un menor peso.

Sea W_p la función de peso generalizada tal que,

$$W_p(\mu) = \begin{cases} (1 - |\mu|^p)^p, 0 \leq |\mu| < 1 \\ 0, 1 \leq |\mu| \end{cases} \tag{10}$$

Se cumple que $W_3 = W$, es decir que la función tricúbica es un caso particular de la función de peso generalizada.

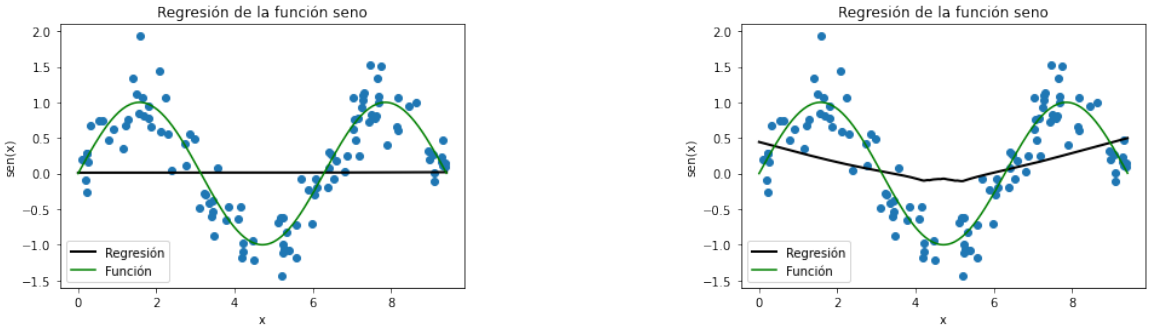
Veamos algunos ejemplos de funciones de esta familia:



(a) W_0 (b) W_1 (c) W_{10}

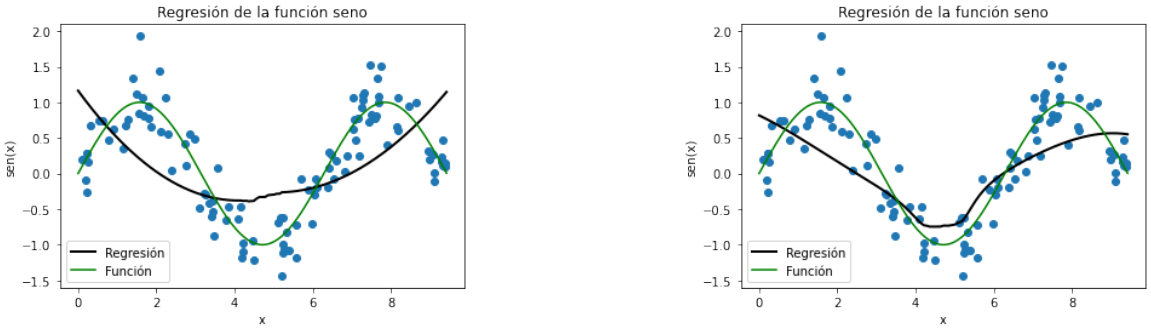
Figura 20: Ejemplos de la función W_p

En principio, vemos que, si $p \in [1; +\infty]$ vemos que a medida que p crece, el peso se distribuye mas igualitariamente, es decir, se les da un peso mayor a elementos dentro del vecindario que antes no se les daba. Tenemos dos casos extremos, que son el caso $p = 1$ y $p = 0$, en uno aumenta la relación peso-distancia cercana, y el otro la disminuye, o en realidad casi la desaparece, si no contamos a los elementos que están más lejos que el q-ésimo elemento, ya que se transforma en la función constante uno.



(a) W_0 (b) W_3

Figura 21: Regresión con grado$(g) = 1$ y $f = 0{,}9$



(a) W_0 (b) W_3

Figura 22: Regresión con grado$(g) = 2$ y $f = 0{,}9$

En las figuras 21 y 22 se aprecian algunas diferencias al variar la función de peso. Elegimos variar entre W_0 y W_3 porque creemos que alguna función que sea muy antagónica con la función tricúbica pero que aún así siga cumpliendo el rol de función de peso es W_0.

Al elegir $p = 0$ estamos casi eliminando el rasgo de regresión local, ya que le damos un mismo peso a todos los elementos en el vecindario menos al vecino a distancia q-ésima o mayor. Con lo que en nuestro caso, donde es muy poco probable que hayan dos mediciones en X iguales, podríamos suponer que a lo sumo una medición, es decir la q-ésima medición, es la única que no es tomada en cuenta para la regresión del pivote actual. Por esta razón también elegimos tomar un $f = 0{,}9$ suficientemente alto para que satisfagan algunas conclusiones.

Entre las conclusiones podemos suponer que, si tenemos una función de peso mas "laxa", que distribuya mas igualitariamente el peso, la regresión tiende a parecerse mucho a la función $g(x)$. En las dos figuras anteriores, podemos notar muy fuertemente este aspecto. En cambio, cuando el peso en la función se distribuye de una manera menos uniforme, la regresión final tiende a acomodarse mas a los puntos del dataset en detrimento de la suavidad de la función resultante, es decir pierde *smoothness*.

4.5. Variación de f o q

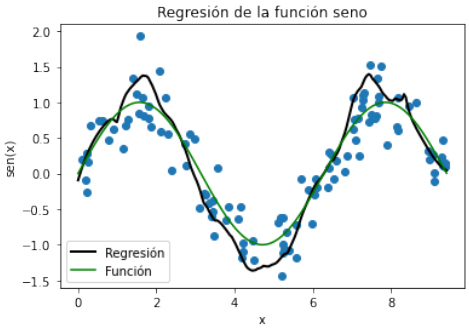
La variable f la definimos como la proporción de cantidad de vecinos sobre puntos totales. O sea,

$$f = \frac{q}{n} \tag{11}$$

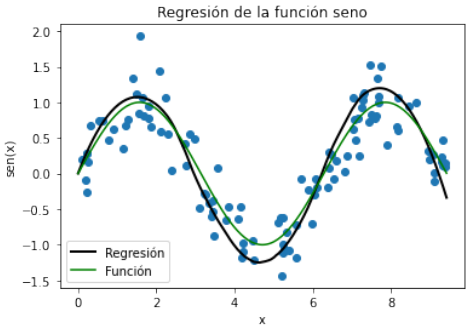
por lo tanto, solo vamos a experimentar variando f , y no q, ya que una depende de la otra.

Lo que nos proponemos ver es que f nos va a dar una idea de cuantos puntos de la muestra original toma en cuenta para estimar la función. Por lo que a valores bajos, la función generada se dejará llevar por cada punto individualmente, lo que va a provocar mucho ruido. En cambio, para valores altos, todos los puntos influirán mucho en la aproximación, por lo que el resultado final va a ser suave pero puede no parecerse a la función original. Los mejores resultados se encontrarán usando valores intermedios de f.

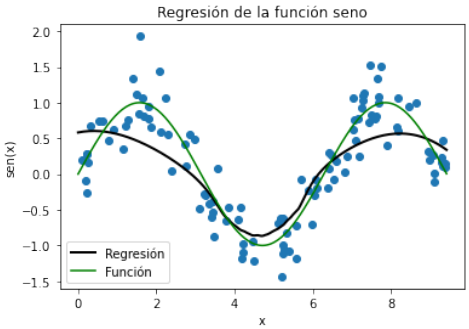
Para verlo dejamos fijo la función de peso y el grado, y vemos que pasa con distintos valores de f.



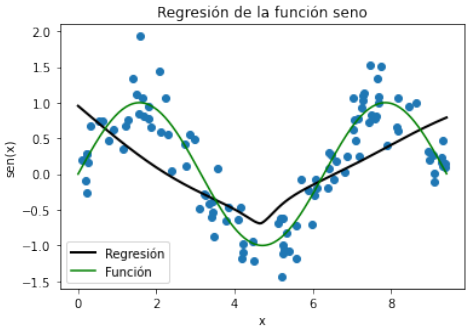
(a) $f = 0{,}2$



(b) $f = 0{,}5$



(a) $f = 0{,}8$



(b) $f = 1$

Figura 24: Regresión con grado y W fijos, para distintas f

Se puede comprobar lo dicho anteriormente en la figura 24, vemos como la función que mejor aproxima al seno original es la de $f = 0,5$

5. Conclusiones

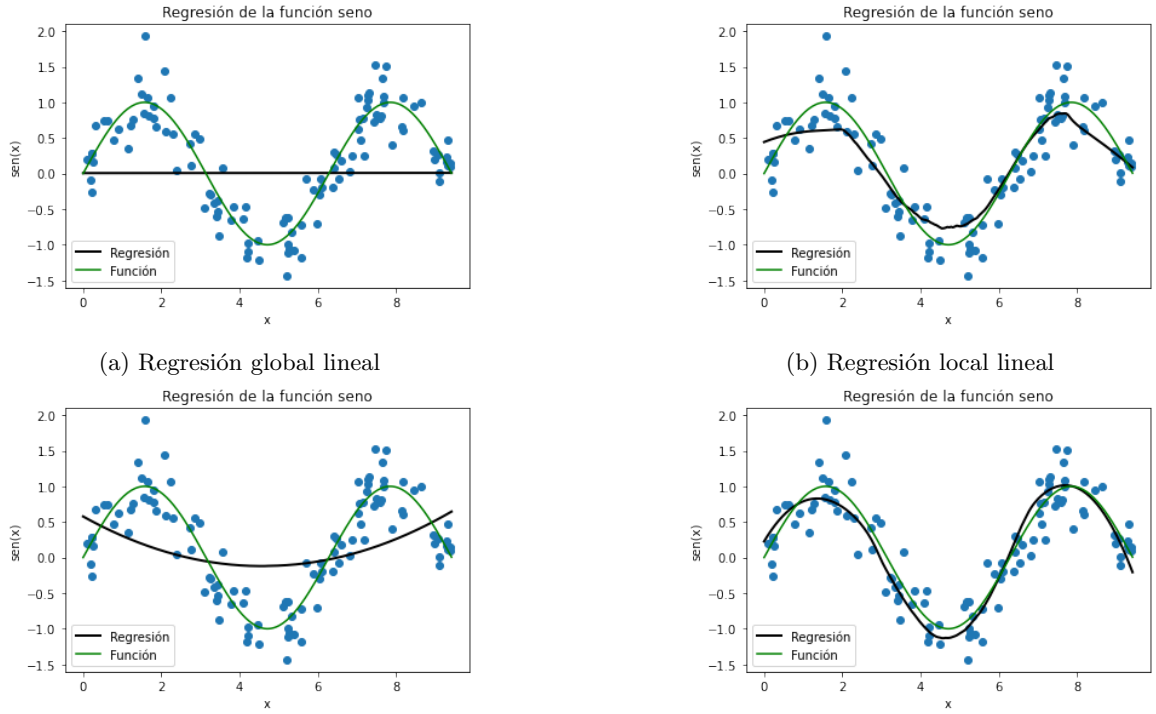
En este trabajo, nos hemos enfrentado a la lectura científica, a la replicación de experimentos, y al desarrollo propio de experimentos tratando de sacar conclusiones personales. Dentro de los contenidos adquiridos en este trabajo, fueron la regresión en todas sus formas (univariada o multivariada, local o global, lineal o cuadrática), pero principalmente hemos adquirido conocimiento de la regresión local pesada o *loess*.

A lo largo de este trabajo, hemos hecho un análisis extendido sobre lo que implica implementar *loess*, sus diferentes componentes y variables y formas de administrarlas para que den un mejor resultado.

Ahora que hemos concluido ¿vale la pena aplicar hoy en día una regresión local en vez de una regresión global?.

Al utilizar una regresión global, la regresión depende fuertemente de lo que nosotros llamaríamos $g(x)$, si esta función tiene un grado chico, entonces la regresión no aproximara los suficientes puntos para parecerse lo suficiente a la "función.ºriginal de las mediciones. Necesitamos, en algunos casos una función $g(x)$ con grado mas grande, que lineal o cuadrático. En cambio con una regresión local podemos adaptar lo suficientemente bien la regresión resultante a las mediciones con un grado de $g(x)$ relativamente mas bajo. Aún así, hoy en día, no es muy caro hacer una regresión con un grado de la función $g(x)$ alto, entonces estos problemas no ocurrirían, pero el costo en tiempo de implementar o programar una regresión lineal seguiría siendo mas costoso al igual que antes, por lo que habría que ver para que casos específicos realmente conviene usar regresión local.

En la figura siguiente se puede observar lo dicho anteriormente.



(a) Regresión global lineal

(b) Regresión local lineal

(c) Regresión global cuadrática

(d) Regresión local cuadrática

Figura 25: Regresión local vs global de la función $sen(x)$

Referencias

[1] *Robust Locally Weighted Regression and Smoothing Scatterplots*, William S. Cleveland. (1979).

[2] *Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting*, Cleveland, William S. and Devlin, Susan J. (1988).