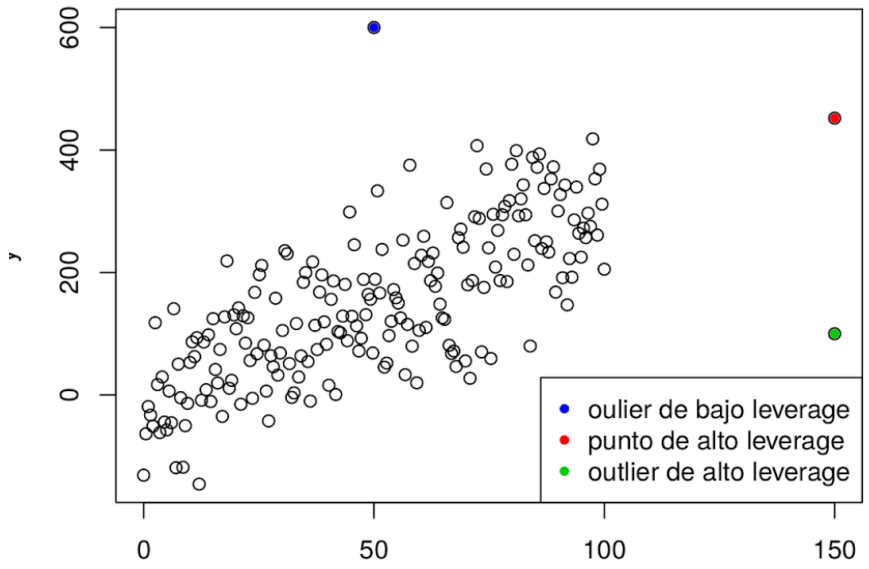
Robustez

Outliers: Datos que se desvían. Es un concepto subjetivo al problema. El mecanismo por el cual se generó ese dato fue distinto.
En algunas situaciones nos interesa detectar los outliers. En algunas situaciones no nos van a interesar, porque van a hacerle creer al modelo que tiene que prestarle atención a un evento muy improbable.

¿Cuándo un modelo es robusto? cuando por sí solo soporta de forma aceptable anomalías.

Outliers en regresiones: Son observaciones distantes del resto de los datos. Puede deberse a un error de medición, ruido, aleatoriedad, etc.

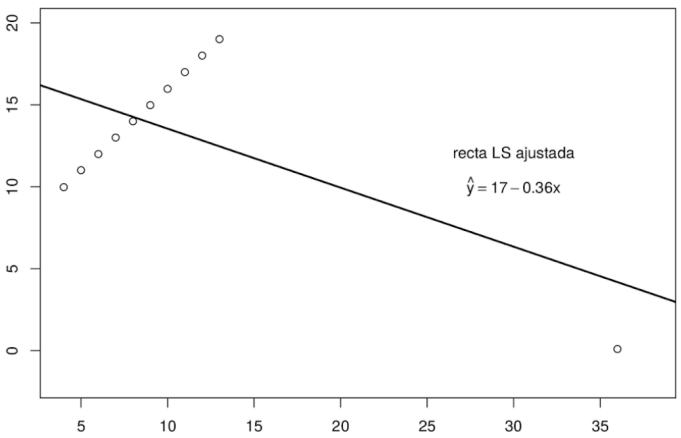


- Punto de alto leverage: no es un outlier. Se escapa de una de mis dimensiones, pero tiene sentido dentro del eje Y.
- Outlier de alto leverage: se aleja en el eje X de mis datos.
- Outlier de bajo leverage: se aleja en el eje Y de mis datos.

Problema: puede romper el ajuste. Un solo outlier puede hacer funcionar muy mal a mi regresion. Es parte relevante identificar estos puntos.

Si ajusto por cuadrados mínimos una regresion lineal estas buscando minimizar la funcion de perdida cambiando tus valores de b. entonces al ser al cuadrado algo que le erras por mucho te queda gigante

Por cuadrados mínimos obtenemos la siguiente recta:



Ocurre que el RSS (función de pérdida) se minimiza y para eso nos da esta recta, es lo que te da menos pérdida sumando todos los errores.

Minimizar la función de pérdida es equivalente a hallar los B que minimizan dicha función. Necesitamos cambiar la funcion de perdida:

> Estimador L1: podemos sacar el cuadrado y usar modulo,mejora un poco pero sigue siendo sensible a outliers y al tener módulo la derivada es muy complicada.
>
> Estimador LMS: usa la mediana pero es más dificil de usar ocmputacionalmente porque no es derivable

Ninguna de estas se usa en la práctica, se usa la siguiente:

M-estimadores de regresión: Es una generalización de las anteriores funciones de pérdida

Función de pérdida a minimizar

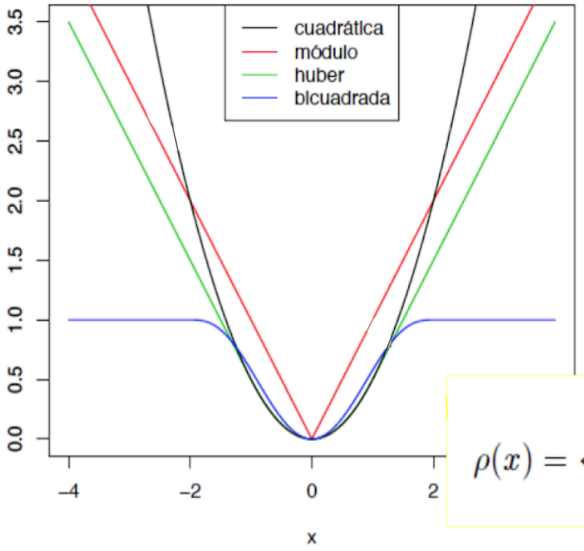
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}}{\hat{\sigma}}\right)$$

Estimador robusto de escala de errores

Agrega la función p y un estimador robusto que escala mi error. Lo que está en el estimador robusto está estandarizado.

¿Qué funciones pueden ser mi función p?

- **Cuadrática**
- **Módulo**: crece más lento que la cuadrática pero en el vertice no es derivable, no la queremos.
- **Huber**: función partida: recibe error y si es menor a un K devuelve una cuadratica pero si es mayor devuelve una lineal. Entonces crece más suave el error.
- **Bicuadrada**: a partir de cierto valor es una constante.



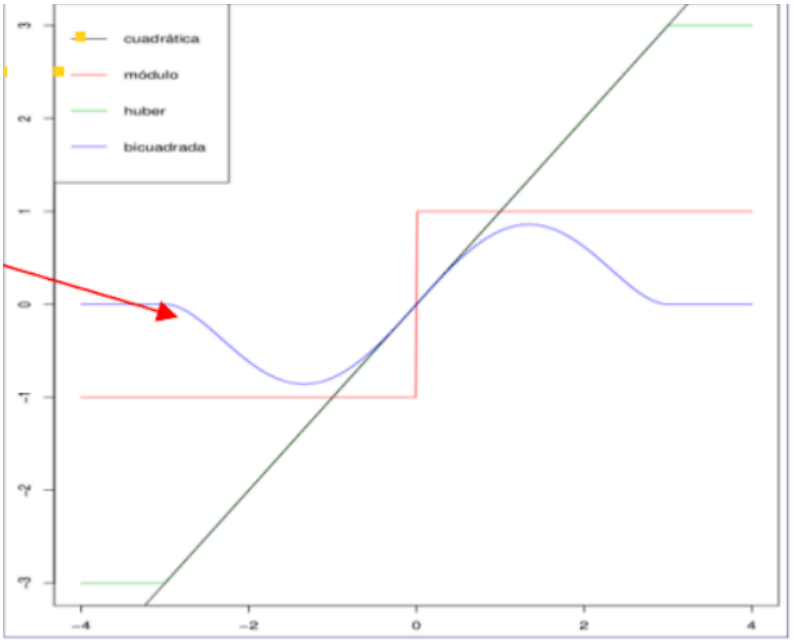
Función de Huber

$$\rho(x) = \begin{cases} x^2 & \text{si} \quad x \leq k \\ 2k\,|x| - k^2 & \text{si} \quad x > k \end{cases}$$

El K lo tengo que definir yo.

Una vez que determinamos la función de pérdida, buscamos los B que la minimizan. Para esto derivamos.

Las posibles **derivadas** que podemos obtener son las siguientes:

La 

derivada de la bicuadrada en los extremos vale cero.

Wi pesa la atipicidad(outlier):

$$W_i = \frac{\psi\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}{\hat{\sigma}}\right)}{\frac{y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}{\hat{\sigma}}}.$$

Es la derivada sobre la desviación estandarizada. Pesa la atipicidad porque si la derivada se hace cero usando la función bicuadrada no le va a dar mucha importancia porque es un punto lejano.

Si el wi me da cero, le voy a estar diciendo que ese punto no importa, que no le de importancia. La función de huber le va a dar importancia al punto pero hasta ahí.

La huber me restringe la magnitud de mis outliers, la bicuadrada no les da para nada importancia porque se hace cero.

Si la observación se aleja mucho de la media de nuestro datos (outlier) el numerador tomará valores extremos.

Dependiendo la función f p elegida Wi tomará valores muy grandes o muy chicos.

Lo más común en la práctica es usar huber.

Algoritmos:

RANSAC RANdom SAmple Consensus

Hay tres variables que voy a tener que settear yo, conociendo el problema en cuestión.

1. Seleccionar “n” valores random de los datos.
2. Ajustar la regresión con esos “n” valores.
3. Evaluar de los puntos restantes cuales son posibles “inliers”, puntos que se encuentran a menos de una distancia “d” de mi regresión.

4. Si #inliers > t => “posible buen modelo”
5. Realizar un nuevo ajuste con todos los inliers y evaluar el error del modelo en #inliers.
    a. Si este error < Mejor error => Mejor error = error y Mejor modelo = modelo actual
    b. Sino volver al paso 1

La idea de este algoritmo es que con suficientes iteraciones se van a tomar datos que no sean outliers.
No es un método determinista porque elijo datos al azar. Es un método greedy. No se sabe si llega a la mejor solución.
Estimación robusta incluso con mucha cantidad de outliers.
No es bueno este método con mainfold de distintas densidades. Puede no ser óptimo.

ESTIMADOR THEIL-SEN

1. Tomo de a pares de puntos del dataset y calculo la pendiente de la recta que pasa por esos dos puntos.
2. Ordeno las pendientes calculadas (sort) y me quedo con la mediana
3. Con la mediana calculada anteriormente evalúo todas las ordenadas posibles, bi = yi-mxi
4. Ordenar las ordenadas al origen calculadas
5. Quedarse con la mediana de las ordenadas al origen median (yi -mxi).

Ventajas: Insensitivo a outliers, no necesita hiper parámetros.
Desventaja: Dependiendo de la cantidad de puntos puede tardar O(n^2).

<u>Resumen</u>

- Modificar la función pérdida de forma tal de penalizar los residuos de las observaciones atípicas.
- Mediante enfoque algorítmicos hacer muchas estimaciones buscando aquellas que no sean influenciadas por los outliers.

Notas:
RANSAC es bueno para valores atípicos grandes en la dirección Y.

TheilSen es bueno para pequeños valores atípicos, tanto en la dirección X como en Y, pero tiene un punto de quiebre por encima del cual se desempeña peor que Ordinary least squares, igual que RANSAC.

Es posible que las puntuaciones de Huber Regressor no se comparen directamente con TheilSen y RANSAC porque no intenta filtrar completamente los valores atípicos, sino que reduce su efecto.