Regresión Robusta

Proyecto Final

Roberto Vásquez Martínez   Victor Daniel Alvarado Estrella

CIMAT
Universidad de Guanajuato

CIMAT

Profesor: Dr. Enrique Villa Diharce

3 de junio de 2021

Contenidos

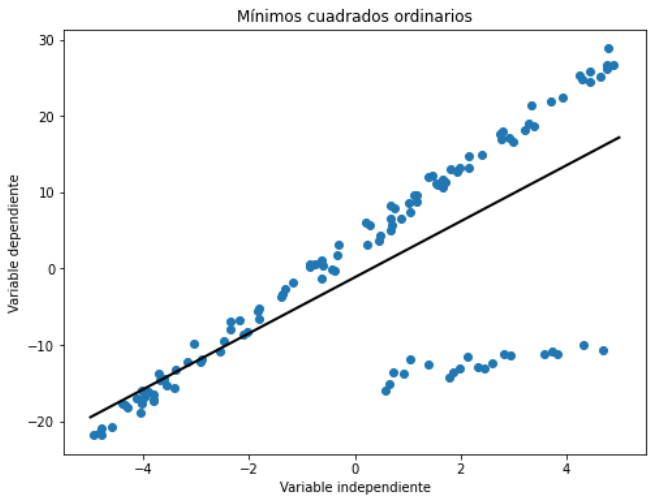
Introducción

Mínimos cuadrados resuelve el problema de minimización

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} e_i^2(\beta)$$

donde $e_i(\beta) = y_i - x_i^T \beta$.
Sin embargo, MC es sensible ante la presencia de *outliers*.

Introducción



Dos enfoques

- En lugar de tomar los residuos al cuadrado e^2, tomar alguna otra función de los residuos $\rho(e)$ que refleje la magnitud de los residuos de manera menos extrema.
- Reemplazar la suma (o equivalentemente la media) por una medida de localización más robusta como la mediana o media truncada.

Mínimas desviaciones absolutas

Mínimas desviaciones absolutas resuelve el problema de minimización

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left| y_i - x_i^T \beta \right| .$$

Observemos que

$$\sum_{i=1}^{n} \left| y_i - x_i^T \beta \right| = \sum_{i=1}^{n} \frac{1}{\left| y_i - x_i^T \beta \right|} \left(y_i - x_i^T \beta \right)^2 .$$

Mínimas desviaciones absolutas

El estimador de mínimas desviaciones absolutas puede calcularse entonces utilizando *mínimos cuadrados reponderados iterativamente*. Es un método iterativo en donde cada paso implica resolver un problema de mínimos cuadrados ponderados:

$$\begin{aligned}\hat{\beta}^{(t+1)} &= \arg\min_{\beta} \sum_{i=1}^{n} w_i^{(t)} \left(y_i - x_i^T \beta\right)^2 \\ &= \left(X^T W^{(t)} X\right)^{-1} X^T W^{(t)} y\end{aligned}$$

donde $W^{(t)} = \operatorname{diag}\left(w_1^{(t)}, \ldots, w_n^{(t)}\right)$.

Mínimas desviaciones absolutas

Los pesos se inicializan como

$$w_i^{(0)} = 1$$

y se actualizan después de cada iteración como

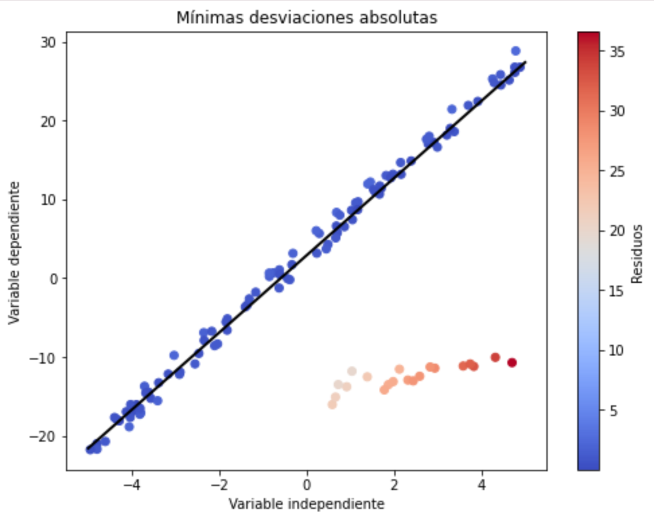
$$w_i^{(t)} = \frac{1}{\left| y_i - x_i^T \hat{\beta}^{(t)} \right|}.$$

Para evitar dividir entre cero, se debe hacer una regularización:

$$w_i^{(t)} = \frac{1}{\text{máx}\left\{ \delta, \left| y_i - x_i^T \hat{\beta}^{(t)} \right| \right\}}$$

donde δ es algún valor pequeño.

Mínimas desviaciones absolutas



M-estimadores

Supongamos que las respuestas Y_i son independientes y tienen función de densidad

$$f_i(y_i;\beta,\sigma) = \frac{1}{\sigma} f\left(\frac{y_i - x_i^T\beta}{\sigma}\right)$$

donde σ es un parámetro de escala. Por ejemplo, si f es la función de densidad normal estándar, entonces el modelo descrito corresponde al modelo de regresión usual.

La función de log verosimilitud es

$$l(\beta,\sigma) = -n\log\sigma - \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T\beta}{\sigma}\right)$$

donde $\rho = -\log f$. Debemos minimizar $-l(\beta,\sigma)$.

M-estimadores

La función ρ debe ser simétrica: $\rho(e) = \rho(-e)$; no negativa: $\rho(e) \geq 0$; y monótona: $\rho(|e_1|) \geq \rho(|e_2|)$ si $|e_1| \geq |e_2|$.
Ejemplo (Mínimos cuadrados): $\rho(x) = \frac{1}{2}x^2$ y

$$l(\beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 .$$

Ejemplo (Mínimas desviaciones absolutas): $\rho(x) = |x|$ y

$$l(\beta, \sigma) = -n \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} \left|y_i - x_i^T \beta\right| .$$

M-estimadores

Ejemplo (Huber): Sea

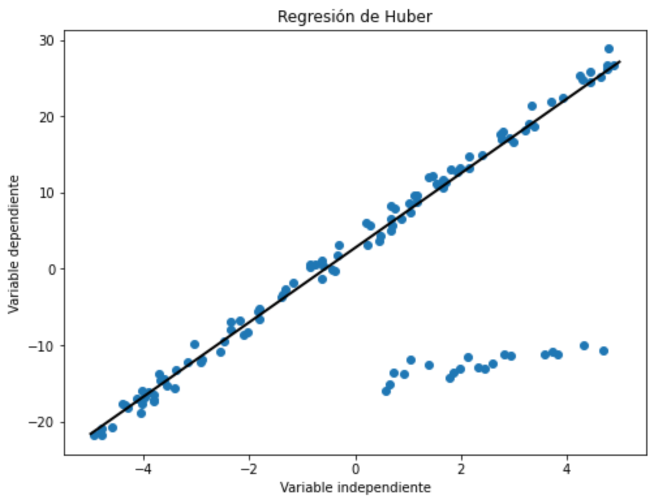
$$\rho'(x) = \begin{cases} -a, & \text{si } x < -a, \\ x, & \text{si } -a \le x \le a, \\ a, & \text{si } x > a, \end{cases}$$

de donde

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{si } |x| \le a, \\ a|x| - \frac{1}{2}a^2, & \text{si } |x| > a. \end{cases}$$

El valor de a usualmente se escoge como 1.5, lo cual da un compromiso razonable entre mínimos cuadrados y mínimas desviaciones absolutas.

M-estimadores



Observación

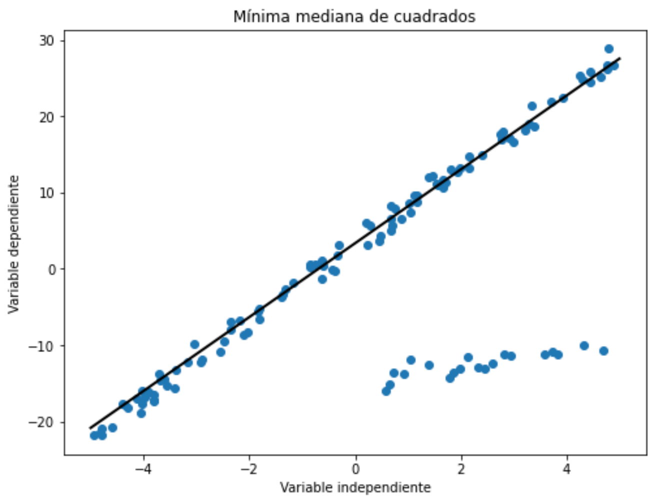
- Los M-estimadores son igual de de vulnerables a outliers que mínimos cuadrados en las variables explicativas
- Los M-estimadores son casi igual de eficientes que mínimos cuadrados cuando los errores son normales

Mínima mediana de cuadrados

Como alternativa a los M estimadores, podemos reemplazar la media por una medida de localización más robusta, conservando los residuos al cuadrado.
Por ejemplo, la *mínima mediana de cuadrados* minimiza

$$\hat{\beta} = \arg\min_{\beta} \text{mediana}_i \, e_i^2(\beta).$$

Medidas robustas de localización



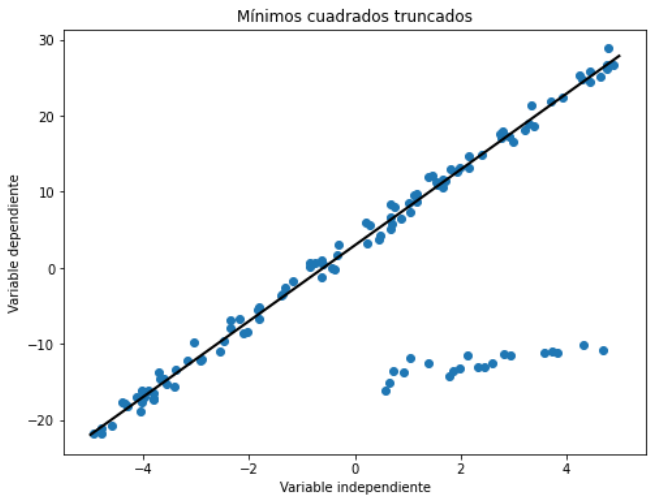
Mínimos cuadrados truncados

Otra alternativa es usar la media truncada en lugar de la mediana. *Mínimos cuadrados truncados* minimiza

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{h} e^2_{(i)}(\beta)$$

donde h se escoge para obtener un estimador robusto y $e^2_{(1)}(b) \leq \ldots \leq e^2_{(n)}(b)$ son los residuos al cuadrado ordenados. El grado de truncamiento debe ser severo para hacer la estimación robusta. Una elección popular es $h = \lfloor n/2 \rfloor + 1$.

Medidas robustas de localización



Observaciones

- Estos estimadores son muy robustos respecto a los errores y las variables explicativas.
- Pueden ser inestables, un cambio pequeño en puntos no extremos puede generar un cambio grande en el ajuste.
- Estos estimadores son ineficientes en comparación con mínimos cuadrados cuando los errores son normales.

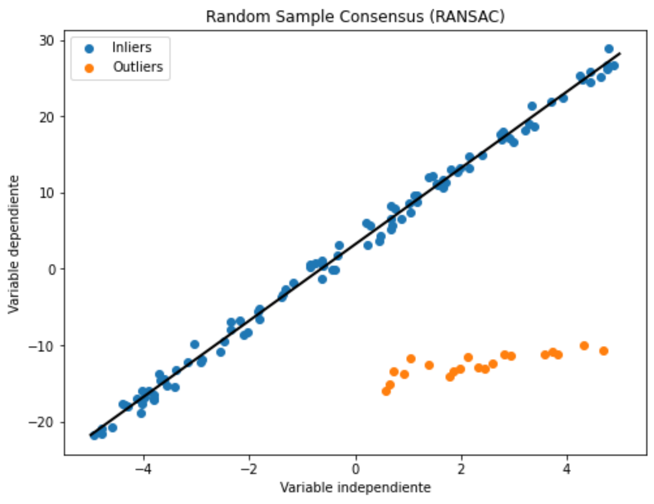
RANSAC

Random sample consensus (RANSAC) es un método iterativo para estimar los parámetros de un modelo sobre un conjunto de datos que contiene outliers.
También puede ser utilizado como un método de detección de outliers.

RANSAC (Algoritmo)

1. Seleccionar al azar el mínimo número de puntos requeridos para estimar los parámetros del modelo.
2. Estimar los parámetros del modelo.
3. Determinar cuántos puntos de todo el conjunto de puntos se ajustan con una tolerancia predefinida ϵ.
4. Si la proporción entre el número de inliers y el número total de puntos en el conjunto excede un umbral predefinido τ, re-estimar los parámetros del modelo usando todos los inliers identificados y terminar.
5. De lo contrario, repetir los pasos 1 - 4 (un número máximo de veces N).

RANSAC


Random Sample Consensus (RANSAC)
Inliers
Outliers
Variable dependiente
Variable independiente
30
20
10
0
−10
−20
−4
−2
0
2
4

RANSAC

El número de iteraciones N se escoge lo suficientemente grande para asegurar, con probabilidad p (usualmente 0.99), que al menos una de las muestras aleatorias no contenga outliers.
Sea u la probabilidad de que un punto escogido al azar sea un inlier y m el tamaño de cada muestra. Entonces

$$1 - p = (1 - u^m)^N$$

de donde

$$N = \frac{\log(1-p)}{\log\left(1-u^m\right)}.$$

- La media muestral se puede hacer arbitrariamente grande haciendo suficientemente grande un dato.
- Para hacer arbitrariamente grande a la mediana se necesita al menos el 50 % de los datos.

Punto de ruptura

Definición

El punto de ruptura de un estimador es la mínima fracción de datos que se pueden cambiar por un valor arbitrariamente grande y causar un cambio arbitrariamente grande en el estimador.

Observación

Mientras mayor punto de ruptura la estimación es menos sensible a outliers, el mejor valor posible del punto de ruptura es $1/2$, pues si más del 50 % de la muestra está contaminada, es imposible distinguir entre inliers y outliers, además los outliers no están típicamente en la muestra.

Para una muestra de tamaño n

- El punto de ruptura para la media muestral es de $\frac{1}{n}$
- La mediana tiene un punto de ruptura cercano al la $\frac{1}{2}$.

M-Estimadores

A pesar de que la mediana tiene un alto punto de ruptura y la mediana de $Y_1, \ldots, Y_n$ minimiza a

$$\sum_{i=1}^{n} |Y_i - \partial| \, ,$$

como función de ∂ se puede probar que el punto de ruptura de mínimas desviaciones absolutas tiene el mismo de ruptura que mínimos cuadrados. Los mismo es cierto para los M-estimadores.

Localización Robusta

Las medidas de localización son ineficientes en comparación con los *M*-estimadores. Compensan esa ineficiencia con altos puntos de ruptura.

Motivación

- Obtener un estimador en el que se incremente el punto de ruptura en comparación con los M-estimadores
- Conservar la eficiencia de los estimadores que no se tiene en los métodos de Localización Robusta.
- Obtener un estimador que sea más robusto respecto a las variables explicativas.

GM-estimadores

Si $\psi = \rho'$ en la notación de los M-estimadores. El GM-estimador es sólo de las ecuaciones normales formadas por

$$\sum_{i=1}^{n} \pi_i \psi\left(y_i - x_i'\hat{\beta}\right) x_i = 0,$$

donde π_i apropiados disminuyen la influencia de puntos palanca. Estos valores involucran diagnósticos típicos de outliers en OLS.
Varios autores sugieren

$$\pi_i = \left(\frac{1 - h_{ii}}{h_{ii}}\right)^{1/2}.$$

GM-estimadores

El GM-estimador se obtiene usando *mínimos cuadrados reponderados iterativamente* de la siguiente forma

$$\hat{\beta}_{GM} = (X'WX)^{-1}X'W \cdot y,$$

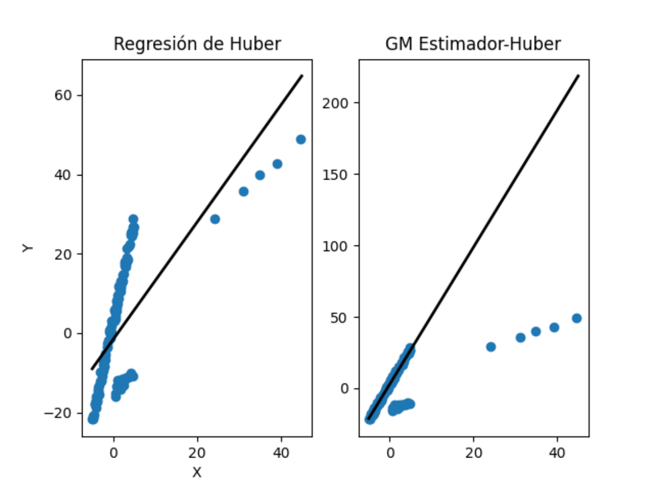
donde W es una matriz diagonal con pesos

$$w_i = \frac{\psi((y_i - x_i'\hat{\beta}_{GM})/\pi_i s)}{(y_i - x_i\hat{\beta}_{GM})/\pi_i^2 s},$$

y en cada iteración se utiliza una medida robusta de la escala s como

$$s = \text{mediana}\{\left|e_i(\hat{\beta}_{GM}) - \text{mediana}(e_i(\hat{\beta}_{GM}))\right|\}/0.6745.$$

GM-Estimador Huber



Observaciones

- El GM-estimador mantiene la eficiencia del M estimador y es más robusto a outliers respecto a las variables explicativas.
- Mantiene las propiedades distribucionales asintóticas del M-estimador.
- Mejora el punto de quiebre de $1/n$ a $1/p$ donde p es el número de variables.
- Se puede mejorar usando mejores π-pesos o modificando la función objetivo (Mallows, Schweppe, etc).

Conclusiones

- Los M-estimadores tienen alta eficiencia y propiedades distribucionales pero puntos de ruptura bajos.
- Los estimadores basados en medidas de localización robusta tienen puntos de ruptura altos pero en algunos contextos son inestables.
- El GM-estimador conserva la alta eficiencia y propiedades distribucionales del M-estimador aumentando el punto de ruptura. En algunos contextos esta mejora puede ser insuficiente.
- RANSAC tiene la bondad de detectar outliers.
- Existen estimadores que intentan tener alta eficiencia y puntos de ruptura alto, como el MM-estimador.

Referencias

George A. F. Seber & Alan J. Lee
Linear Regression Analysis.
Wiley, 2003

C. Sidney Burrus
Iterative Reweighted Least Squares.
Extraído de: https://cnx.org/exports/92b90377-2b34-49e4-b26f-7fe572db78a1@12.pdf/iterative-reweighted-least-squares-12.pdf.

Konstantinos G. Derpanis
Overview of the RANSAC Algorithm.
Extraído de:
http://www.cse.yorku.ca/ kosta/CompVis_Notes/ransac.pdf.
2010.

Referencias

James R. Simpson
New Methods and Comparative Evaluations for Robust and Biased-Robust Regression Estimation.
Extraído de: https://apps.dtic.mil/dtic/tr/fulltext/u2/a298578.pdf.
1995.