Universidad de Guanajuato
CIMAT

Proyecto Final

Regresión Robusta

Autores:

Roberto Vásquez Martínez

Victor Daniel Alvarado Estrella

Profesor:

Dr. Enrique Villa Diharce

3 de junio de 2021

Índice

1. Introducción

Los estimadores por mínimos cuadrados son los estimadores insesgados más eficientes para los coeficientes de regresión cuando los errores están normalmente distribuidos. Sin embargo, no son muy eficientes cuando la distribución de los errores es de cola larga. Bajo estas circunstancias, tendremos outliers en los datos, es decir, observaciones cuyos errores ϵ_i son muy grandes en valor absoluto.

Cuando ajustamos un problema de regresión, minimizamos alguna medida de tendencia central del tamaño de los residuos. Mínimos cuadrados por ejemplo minimiza la media de los residuos al cuadrado (o equivalentemente, la suma de los residuos al cuadrado). Así, minimos cuadrados resuelve el problema de minimización

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

donde $e_i(\beta) = y_i - x_i^T \beta$.

La sensibilidad de mínimos cuadrados ante la presencia de outliers se debe a dos factores. Primero, si medimos el tamaño de los residuos usando los residuos al cuadrado, cualquier residuo con una magnitud grande tendrá un tamaño muy grande en comparación a los demás. Segundo, si usamos una medida de localización como la media que no es robusta, cualquier residuo grande tendrá un impacto muy grande en el criterio, resultando en puntos extremos teniendo una influencia desproporcionada en el ajuste.

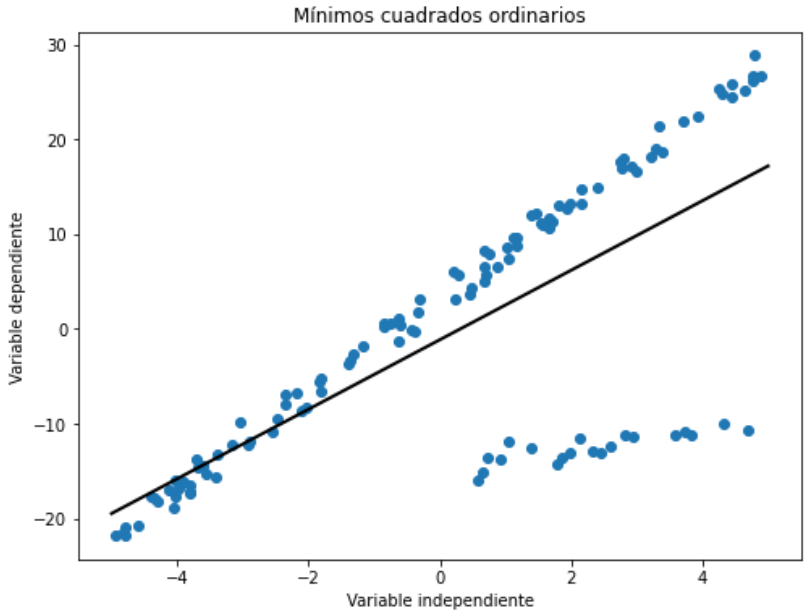

Figura 1: Ajuste por mínimos cuadrados cuando hay outliers.

Existen dos remedios populares para este problema. Primero, podemos medir el tamaño de los residuos de alguna otra manera, reemplazando el cuadrado e^2 por alguna otra función $\rho(e)$ que refleje el tamaño de los residuos de una forma menos extrema. 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 $|e_1| \geq |e_2|$. Segundo, podemos reemplazar la suma (o equivalentemente la media) por una medida de localización más robusta como la mediana o una media truncada.

2. M-Estimadores

Supongamos que las respuestas observadas 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 es solo el modelo estándar de regresión y σ es la desviación estándar de las respuestas.

La función de log verosimilitud correspondiente a esta función de densidad es

$$\begin{aligned} l(\beta,\sigma) &= -n\log\sigma + \sum_{i=1}^{n} \log f\left(\frac{y_i - x_i^T\beta}{\sigma}\right) \\ &= -\left\{ n\log\sigma + \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T\beta}{\sigma}\right)\right\}, \qquad \rho = -\log f. \end{aligned}$$

Por lo tanto, para estimar β y σ por máxima verosimilitud, debemos minimizar

$$n\log s + \sum_{i=1}^{n} \rho\left(\frac{e_i(b)}{s}\right)$$

en función de b y s.

2.1. Mínimas desviaciones absolutas

Sea $\rho(x) = |x|$. Los estimadores correspondientes son valores de s y b que minimizan

$$n\log s + \frac{1}{s}\sum_{i=1}^{n} |e_i(b)|.$$

Un valor de b que minimiza la expresión anterior es también un valor de b que minimiza

$$\sum_{i=1}^{n} |e_i(b)|$$

y se llama el estimador de mínimas desviaciones absolutas (en inglés *Least Absolute Deviations*).

Ahora bien, observemos que

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

Así pues, el estimador de mínimas desviaciones absolutas puede calcularse de manera numérica utilizando mínimos cuadrados reponderados iterativamente (en inglés *Iterative Reweighted Least Squares*).

Mínimas desviaciones absolutas (IRLS)

Inicializar los pesos como $w_i^{(0)} = 1$.

Para $t = 1, 2, \ldots, N$ hacer:

1. Calcular los coeficientes como $\hat{\beta}^{(t)} = (X^T W^{(t-1)} X)^{-1} X^T W^{(t-1)} y$ donde $W^{(t-1)} = \text{diag}\left(w_1^{(t-1)}, \ldots, w_n^{(t-1)}\right)$.
2. Actualizar los pesos como $w_i^{(t)} = \frac{1}{\max\left\{\delta, |y_i - x_i^T \hat{\beta}^{(t)}|\right\}}$ donde $\delta > 0$ es algún valor pequeño.

Devolver $\hat{\beta}^{(N)}$.

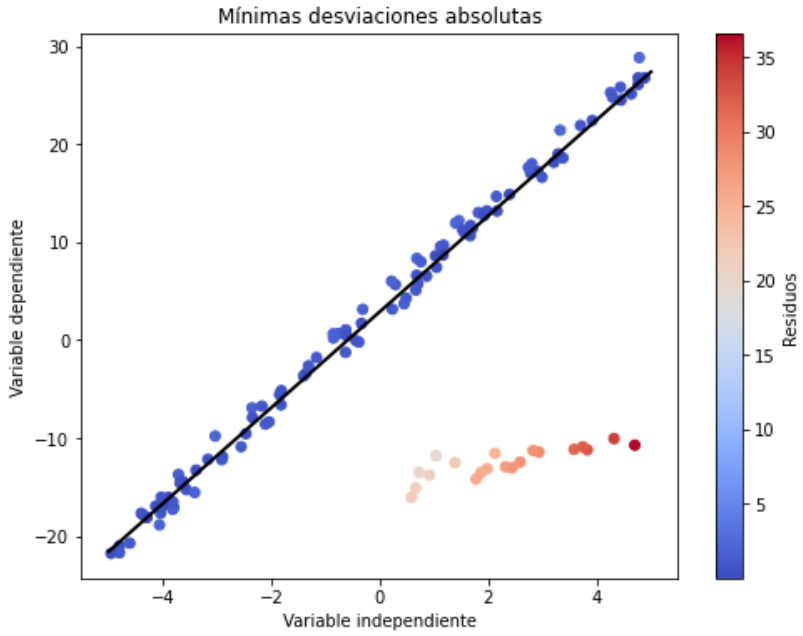

Figura 2: Ajuste por mínimas desviaciones absolutas. Los puntos están coloreados según el valor de los residuos en valor absoluto.

2.2. Regresión de Huber

Sea

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

de donde

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

Aquí, α es una constante que se debe escoger. El valor de α usualmente se escoge como 1.5, lo cual da un compromiso razonable entre mínimos cuadrados (que es la opción que da la mayor eficiencia en el modelo normal) y mínimas desviaciones absolutas, que dará mayor protección ante outliers.

Figura 3: Ajuste por regresión de Huber.

3. Medidas robustas de localización

En alternativa a los M-estimadores, podemos reemplazar la media por una medida de localización más robusta pero manteniendo los residuos al cuadrado como medida del tamaño. Esto lleva a la mínima mediana de cuadrados (en inglés *Least Median of Squares*), que minimiza

$$\text{mediana}_i \; e_i(b)^2.$$

Otra alternativa es utilizar la media truncada en lugar de la mediana, que resulta en mínimos cuadrados truncados (en inglés *Least Trimmed Squares*), que minimiza

$$\sum_{i=1}^{h} e_{(i)}(b)^2,$$

donde h se escoge para obtener un estimador robusto y $e_{(1)}(b)^2 \leq \ldots \leq e_{(n)}(b)^2$ son los residuos al cuadrado ordenados. El grado de truncamiento debe ser un tanto severo para hacer el estimador robusto. Una elección popular es $h = \lfloor n/2 \rfloor + 1$, lo cual equivale a truncar el 50 % de los residuos.

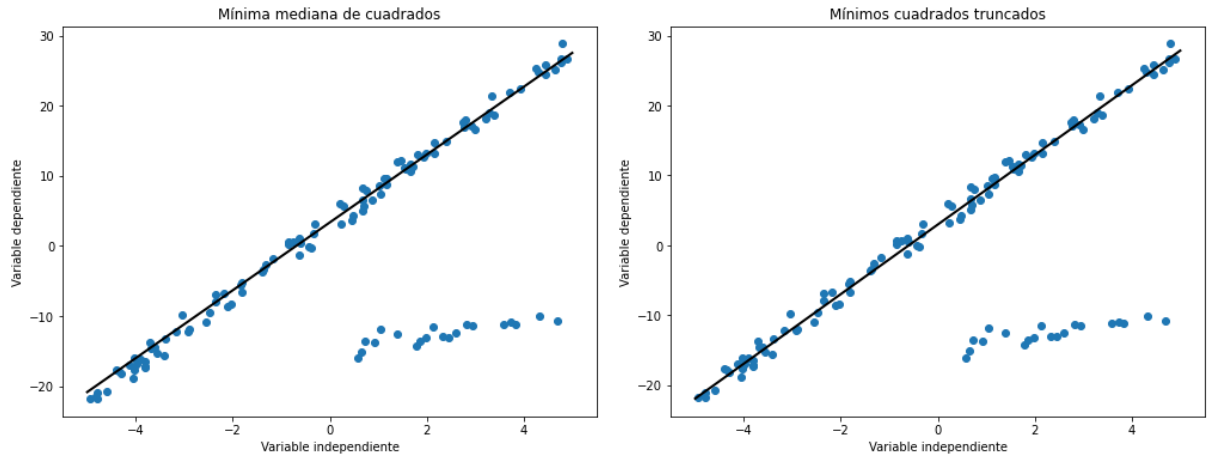

Figura 4: Ajuste por mínima mediana de cuadrados (izquierda) y mínimos cuadrados truncados (derecha).

4. RANSAC

El algoritmo *Random sample consensus (RANSAC)* es un método general de estimación de parámetros diseñado para lidiar con una gran proporción de outliers en los datos. RANSAC es una técnica de re-muestreo que genera soluciones candidatas utilizando el mínimo número de observaciones necesarias para estimar los parámetros del modelo.

RANSAC

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).

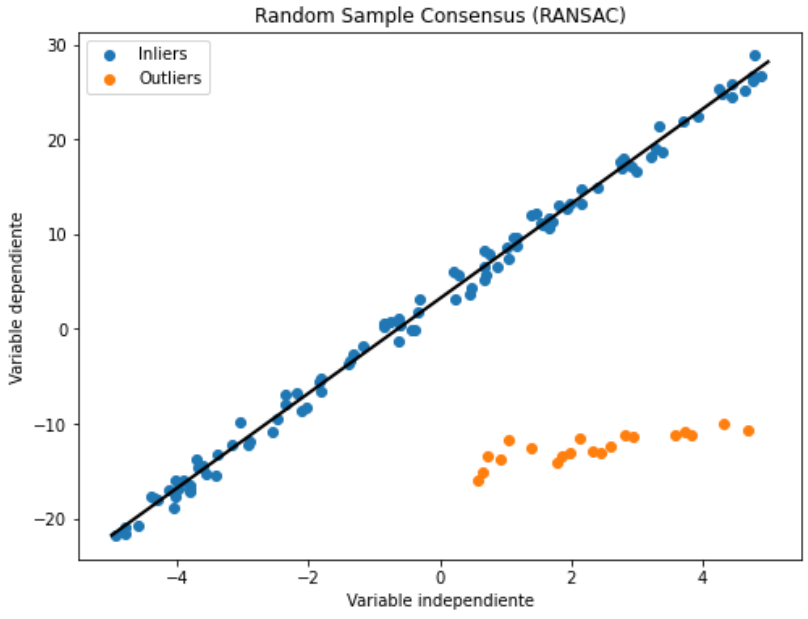

Figura 5: Ajuste por el algoritmo RANSAC.

El número de iteraciones N se escoge lo suficientemente grande para asegurar con probabilidad p (usualmente igual a 0.99) de que al menos uno de los conjuntos de

las muestras aleatorias no incluya un outlier. Sea u la probabilidad de que un punto seleccionado al azar sea un inlier. Sea m el mínimo número de puntos requeridos. Entonces

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

de donde

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

5. Medidas de Robustez

El propósito de esta sección es medir que tan robustas son las propuestas anteriores respecto al grado de contaminación o cantidad de outliers en la muestra de estudio. Discutiremos dos medidas de robustez que nos ayudarán a describir el comportamiento de los dos enfoque anteriores: el punto de ruptura (en inglés *breakdown* point) y la curva de influencia.

5.1. Punto de ruptura

En lo que sigue consideraremos a n como el tamaño de la muestra que estamos analizando.

Definición 1

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.

A partir de la definición no es difícil ver que el punto de ruptura de la media muestral es $1/n$ donde n es el número de observaciones mientras que el punto de ruptura de la mediana es cercano a $1/2$.

Observación 1

El mejor valor del punto de ruptura para un estimador es $1/2$, pues si más del $50\,\%$ de la muestra está contaminada sería imposible distinguir entre buenas y malas observaciones, además de que los outliers no son típicos en la muestra.

Como el estimador de mínimos cuadrados es combinación lineal de la respuestas, se sigue que un gran cambio arbitrario en la respuesta provoca un gran cambio en el estimador, por lo que el estimador de mínimos cuadrados tiene un punto de ruptura de $1/n$.

A pesar de que la mediana tenga un alto punto de ruptura y la mediana de las variables respuesta minimice la función objetivo del estimador de mínimas desviaciones absolutas en función del vector de parámetros se podría pensar que el estimador de mínimas desviaciones absolutas también tiene un alto de ruptura sin embargo no es así, se puede probar que en este caso también el punto de ruptura será $1/n$.

Se puede ver que los estimadores basados en medidas de localización robusta son ineficientes cuando hay normalidad en comparación con los M-estimadores, compensan esta ineficiencia brindado puntos de ruptura altos, cercanos a $1/2$.

5.2. Curva de Influencia

Comenzamos con la siguiente definición:

Definición 2

Supongamos tenemos una distribución F k-dimensional y θ un vector de parámetros que dependen de la F seleccionada, entonces podemos escribir

$$\theta = T(F),$$

decimos que T es un funcional estadístico.

El ejemplo más simple de un funcional estadístico es la media de una variable aleatoria, si $X : \Omega \to \mathbb{R}$ es una variable aleatoria fija entonces

$$T(F) = \int_{\Omega} x dF(x),$$

donde F es una función de distribución, la esperanza de X depende de la distribución F.

Supongamos que F es una distribución fija, podemos modelar un pequeño cambio en en un valor fijo x_0 considerando la mezcla

$$F_t = (1-t)F + t\delta_{x_0} \quad \text{con} \quad t \in [0,1],$$

y δ_{x_0} es la medida de Dirac en x_0.

A partir de la noción de un funcional estadístico y el modelo anterior podemos definir otra medida de robustez

Definición 3

La curva de influencia (IC) de un funcional estadístico T es la derivada con respecto a t de $T(F_t)$ evaluada en $t = 0$. Es una medida del cambio de T respecto a una pequeña cantidad de contaminación en x_0.

A partir de la curva de influencia se puede medir la robustez de los estimadores propuestos en secciones respecto a puntos palanca.

Observación 2

Las curvas de influencia de los M-Estimadores son no acotadas, luego no son robustas a puntos palanca, es decir, a outliers respecto a las variables explicativas.

Los M-estimadores resolvían el problema de tener outliers respecto a las variables respuesta no respecto a las variables explicativas que por la observación anterior podríamos esperar un pobre desempeño, además a pesar de que los estimadores basados en medidas de localización robusta tienen altos puntos de ruptura tienden a ser más inestables.

6. GM-estimadores

El propósito de esta sección es obtener un estimador con un punto de ruptura más alto en comparación con los M-estimadores y mucho más eficientes que los estimadores basados en medidas de localización robustas.

Si $\psi = \rho'$ en el contexto de M-estimadores, el GM-estimador es la solución a las ecuaciones normales formadas por

$$\sum_{i=1}^{n} \pi_i \psi \left(\frac{y_i - x_i^T \hat{\beta}}{s\pi_i} \right) x_i = 0.$$

Para valores apropiados de π_i el GM-estimador puede reducir el efecto de outliers con grandes puntos palanca. Otra vez para el problema de estimación se usa mínimos cuadrados reponderados iterativamente. Varios autories, sigieren tomar π_i como

$$\pi_i = \left[\frac{1 - h_{ii}}{h_{ii}} \right]^{1/2}$$

GM-Estimador (IRLS)

Inicializar los $\hat{\beta}^{(0)}$ como la estimación de algún método con alto punto de ruptura (ej. LTS).

Para $t = 1, 2, \ldots, N$ hacer:

1. Obtener un estimación robusta $s^{(t-1)}$ de la escala.
2. Calcular los pesos $w_i^{(t-1)} = \psi(e_i^{(t-1)}/s^{(t-1)}\pi_i)/(e_i^{(t-1)}/s^{(t-1)}\pi_i^2)$
3. Definir $W^{(t-1)} = \text{diag}\left(w_1^{(t-1)}, \ldots, w_n^{(t-1)}\right)$.
4. Obtener $\hat{\beta}^{(t)} = (X'WX)^{-1}X'Wy$.

Devolver $\hat{\beta}^{(N)}$.

Elección popular para el parámetro de escala robusto es

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

la constante 0.6745 se usa para hacer a s un estimador insesgado de σ.

Los pesos π mencionados anteriormente en general incluyen criterios típicos de análisis de residuos en mínimos cuadrados, el peso usado se basa en DFFITS. Combinaciones de pesos π y funciones ψ pueden producir excelentes estimadores.

A continuación mostramos el comportamiento de los GM-estimadores respecto a los M-estimadores cuando introducimos puntos palanca.

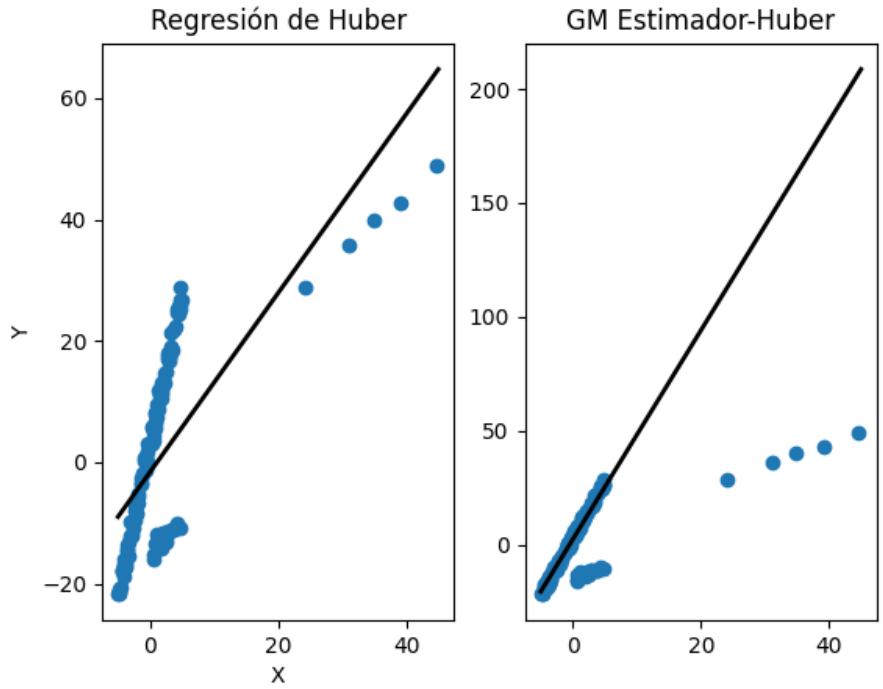

Figura 6: Regresión de Huber y GM-estimador correspondiente con puntos palanca.

Los GM-estimadores poseen la misma eficiencia y propiedades distribucionales de los M-estimadores, además se puede probar que el punto de ruptura de los GM-estimadores es mayor que el $1/n$ de los M-estimadores, pero no puede ser mayor a $1/p$ donde p es el número de regresores. La condición anterior anterior puede llevar al mismo problema que los M-estimadores cuando tenemos una gran cantidad de regresores.

7. Conclusiones

A lo largo de este reporte hemos presentado métodos que permiten lidiar con outliers en un problema de regresión. El propósito fue comparar estos distintos métodos sus bondades y deficiencias, además de mencionar algunos casos en las que estas deficiencias no pueden pasar desapercibidas.

En primer lugar, las ventajas que tienen los M-estimadores es su estabilidad y sus propiedades distribucionales, sin embargo al tener puntos de ruptura bajos estos métodos son sensibles a outliers en las variables explicativas, es decir, a puntos palanca. En contraste, los métodos basados en medidas de localización tienen puntos de ruptura altos pero cuando hay normalidad o pequeñas variaciones en puntos no extremos llegan a hacer ineficientes.

Con el motivo de encontrar un punto medio entre las ventajas y desventajas de los métodos anteriores se investigaron otras dos propuestas: El método RANSAC y los GM-estimadores.

La ventaja adicional de RANSAC respecto a los otros métodos descritos es su bondad de poder identificar outliers, de hecho este método funciona bien en el ejemplo de datos que presenta puntos palanca donde también se tiene un buen ajuste por parte del GM-estimador Huber y en el que falla la Regresión Huber estándar.

Por otro lado, el GM-estimador tiene como objetivo incrementar el punto de ruptura y conservar las bondades de los M-estimadores como lo son las propiedades distribucionales, estabilidad y facilidad de cómputo. Sin embargo, en muchos contextos la mejor que ofrece el GM-estimador respecto al punto de ruptura puede ser insuficiente cuando hay una gran cantidad de covariables en nuetro modelo.

Una propuesta que no se menciona en este reporte serían los MM-estimadores cuya motivación se centra en combinar métodos con altos puntos de ruptura y alta eficiencia.

Finalmente, el método seleccionado siempre debe considerar el contexto del problema que se intenta modelar, si se puede prescindir de algunas bondades para obtener otras más importantes en el problema a resolver, además de mencionar que los métodos presentados son ejemplos de muchos otros métodos que surgen como combinaciones y variaciones de estas ideas, cuyo objetivo central será obtener estimaciones razonables en muestras contaminadas por outliers.

Referencias

Derpanis, K. (2010). Overview of the RANSAC Algorithm. http://www.cse.yorku.ca/~kosta/CompVis_Notes/ransac.pdf

Seber, A., George & Lee. (2003). *Linear Regression Analysis* (2.ª ed.). Wiley & Sons. Inc.

Sidney, C. (s.f.). Iterative Reweighted Least Squares. https://cnx.org/exports/92b90377-2b34-49e4-b26f-7fe572db78a1@12.pdf/iterative-reweighted-least-squares-12.pdf

Simpson, J. (1995). New methods and comparative evaluations for robust and biased-robust regression estimation. https://apps.dtic.mil/dtic/tr/fulltext/u2/a298578.pdf