PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Regresión Robusta

Profesor: Alex de la Cruz H.

MODELOS LINEALES 1

EST631

Contenido

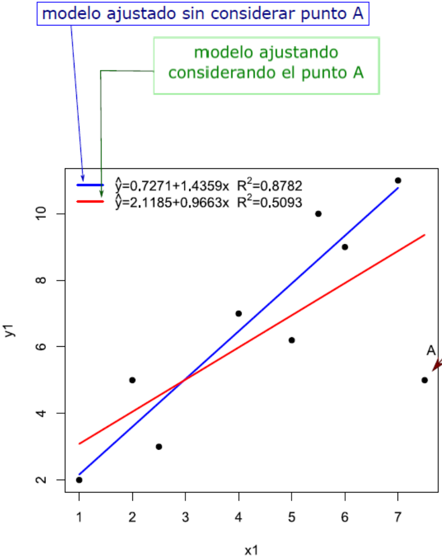
Introducción

- En el siguiente contexto $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, el MCO es la técnica más utilizada, bajo el supuesto de que no existen valores atípicos influyentes.

- Sin embargo, en la práctica, estos supuestos pueden no cumplirse: la presencia de observaciones extremas o datos anómalos puede distorsionar significativamente los estimadores MCO, haciéndolos poco representativos y menos eficientes.

- En la práctica, la distribución de la variable de respuesta puede no ser normal y presentar valores atípicos que afectan el modelo de regresión.
- Cuando la distribución tiene colas largas, se generan observaciones extremas que influyen en exceso sobre los estimadores por mínimos cuadrados, sesgando la ecuación de regresión.

- La regresión robusta surge como una alternativa para enfrentar estas limitaciones, ofreciendo estimadores menos sensibles a valores atípicos y estructuras de error no normales.

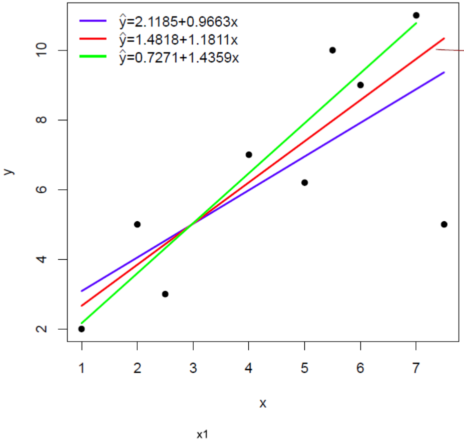


El punto A esta al extremo derecho del espacio de X,pero su valor de respuesta queda cerca del promedio de las otras observaciones.
Es claro que este punto tuvo un efecto radical sobre el modelo de regresión y sobre el R^2

- Si se elimina la observación A se obtiene una recta que pasa muy bien por el resto de los datos. Sin embargo, esta práctica no es recomendable. A veces, los datos se pueden eliminar (o modificar) con base en el conocimiento de la materia.

- En casos más complicados, donde intervienen muchos regresores y la muestra es mayor, no será tan simple identificar que el modelo de regresión se ha distorsionado por observaciones como A.

- Un procedimiento de regresión robusta es aquel que amortigua el efecto de las observaciones que serían muy influyentes si se usaran los mínimos cuadrados.
- Además de la insensibilidad a los valores atípicos, un procedimiento de estimación robusta debería producir, en esencia, los mismos resultados que los mínimos cuadrados cuando la distribución básica es normal y cuando no hay valores atípicos.
- Otro objetivo deseable de la regresión robusta es que los procedimientos de estimación sean relativamente fáciles de llevar a cabo.



Los valores de los coeficientes son intermedios y el punto A no jala mucho la recta hacia abajo

Regresión Robusta

DEFINICIÓN 1 (REGRESIÓN ROBUSTA).

La *regresión robusta* es una técnica de estimación que busca limitar la influencia de valores atípicos en un modelo de regresión, produciendo estimadores más estables cuando los datos no cumplen los supuestos ideales de los mínimos cuadrados.

Los métodos de regresión robusta son técnicas que en potencia se pueden usar cuando hay valores atípicos. Una clasificación frecuente de los valores atípicos es la siguiente:

- Valor atípico de regresión
- Valor atípico residual
- Valor atípico en el espacio X
- Valor atípico en el espacio Y
- Valores atípicos en los espacios X y Y

Valor atípico de regresión

Es un punto que se desvía de la regresión lineal que se determina con las $n-1$ observaciones restantes.

Valor atípico residual

Es un punto que tiene un residual estandarizado o studentizado grande, cuando se usa en la muestra de observaciones con que se ajusta un modelo.

Valor atípico en el espacio X

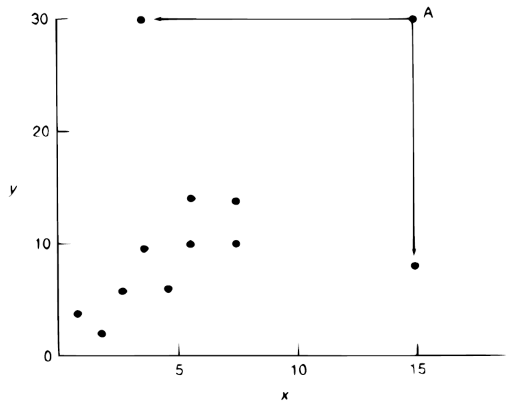
Es una observación remota en una o más coordenadas x.

Valor atípico en el espacio Y

Es una observación con coordenada y inusual. El efecto que tiene la observación sobre el modelo de regresión depende de su coordenada x y de la disposición general de las demás observaciones en la muestra.

Valores atípicos en los espacios X y Y

Es una observación que es atípica en sus coordenadas x e y al mismo tiempo. El efecto de esos puntos depende por completo de la disposición de las demás observaciones de la muestra.


30
20
10
0
y
5
10
15
x
A

- La observación identificada con A en la figura ilustra un caso en donde en la muestra hay un valor atípico en el espacio x y en y. Este punto no tiene de hecho impacto sobre la línea de regresión; tan sólo es un punto de gran influencia.
- Algunas personas podrían pensar que ese punto es de gran influencia "buena", porque está aproximadamente sobre la recta que pasa por las demás observaciones, y tiene el efecto de aumentar el valor de R^2 y bajar el error estándar de la pendiente; por otra parte, podría no ser aconsejable pasar por alto la presencia de esas observaciones aparentemente atípicas en la muestra.

- Nótese que si se cambia la coordenada x de este punto, para que quede cerca del recorrido de los demás valores de x, como se ve en la figura, la observación sería un valor atípico de espacio y, atípico residual y atípico de regresión. Si sólo se cambia la coordenada y para hacerla más o menos igual al promedio de las demás respuestas, como se ve en la figura, se obtendría un valor atípico de espacio x, atípico residual y atípico de regresión.

Considérese el modelo de regresión lineal simple

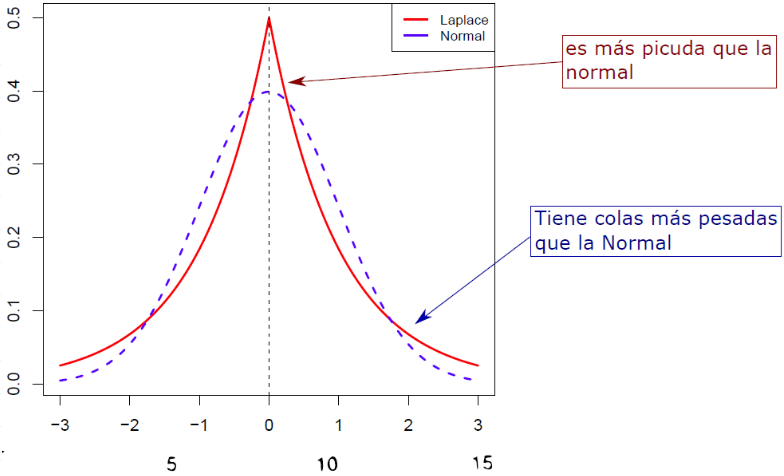
$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i, \quad i = 1, 2, \ldots, n$$

los errores son v.a. independientes con distribución doble exponencial (Laplace)

$$f(\epsilon_i) = \frac{1}{2\sigma} e^{-|\epsilon_i|/\sigma}, \quad -\infty < \epsilon_i < \infty.$$

La función de verosimilitud para β_0 y β_1, es:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{2\sigma} e^{-|\epsilon_i|/\sigma} = \frac{1}{(2\sigma)^n} \exp\left(-\sum_{i=1}^{n} |\epsilon_i|/\sigma \right).$$


Laplace
Normal
es más picuda que la normal
Tiene colas más pesadas que la Normal
0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
3
5
10
15

Para la función: $L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{2\sigma} e^{-|\epsilon_i|/\sigma} = \frac{1}{(2\sigma)^n} \exp\left(-\sum_{i=1}^{n} |\epsilon_i|/\sigma\right)$

- Maximizar consiste en minimizar $\sum_{i=1}^{n} |\epsilon_i|$.
- Recuerde: EMV aplicado al modelo con errores de normal, conduce al MCO.
- La hipótesis de una distribución de error con colas más pesadas implica que el MCO ya no es una técnica óptima.
- Nótese que el criterio del error absoluto ponderaría los valores atípicos con mucho menos severidad que los MCO.

DEFINICIÓN 2 (ESTIMADORES M).

Los estimadores de $\boldsymbol{\beta}$ se obtiene *minimizando* una *función* particular en $\boldsymbol{\beta}$:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(\epsilon_i) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\beta}) \tag{1}$$

donde ρ es una función de la *contribución* de cada residuo a la función objetivo. Un ρ razonable debe cumplir las siguientes propiedades:

- Siempre no negativo: $\rho(\epsilon) \geq 0$.
- Igual a cero cuando su argumento es cero: $\rho(0) = 0$.
- Simétrico: $\rho(\epsilon) = \rho(-\epsilon)$, aunque en algunos problemas la simetría es indeseable.
- Monótono en $|\epsilon_i|$: $\rho(\epsilon_i) \geq \rho(\epsilon_i')$ para $|\epsilon_i| > |\epsilon_i'|$.

El valor mínimo de la ecuación 1 se puede encontrar derivando con respecto al argumento $\boldsymbol{\beta}$, e igualando las derivadas parciales resultantes a:

$$0 = \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$$

$$0 = \sum_{i=1}^{n} \psi(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})\, \boldsymbol{x}_i^\top$$

donde $\psi = \rho'$.

Para reducirlo a un problema de mínimos cuadrados ponderados, definimos la función de peso $w_i = w(\epsilon_i) = \psi(\epsilon_i)/\epsilon_i$. Así, las ecuaciones de estimación pueden escribirse como

$$\sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\,\mathbf{x}_i^\top = 0.$$

Resolver estas ecuaciones de estimación es equivalente a un problema de mínimos cuadrados ponderados, minimizando $\sum w_i^2 \epsilon_i^2$. Los pesos, sin embargo, dependen de los residuos, los residuos dependen de los coeficientes estimados, y los coeficientes estimados dependen de los pesos. Por lo tanto, se requiere una solución iterativa llamada *mínimos cuadrados reponderados iterativamente* (IRLS).

Para reducirlo a un problema de mínimos cuadrados ponderados, definimos la función de peso $w_i = w(\epsilon_i) = \psi(\epsilon_i)/\epsilon_i$. Así, las ecuaciones de estimación pueden escribirse como

$$\sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\,\mathbf{x}_i^\top = 0.$$

Resolver estas ecuaciones de estimación es equivalente a un problema de mínimos cuadrados ponderados, minimizando $\sum w_i^2 \epsilon_i^2$. Los pesos, sin embargo, dependen de los residuos, los residuos dependen de los coeficientes estimados, y los coeficientes estimados dependen de los pesos. Por lo tanto, se requiere una solución iterativa llamada *mínimos cuadrados reponderados iterativamente* (IRLS).

Algoritmo 1: Algoritmo del IRLS

1. Seleccione los valores iniciales para $\boldsymbol{\beta}^{(0)}$, como las estimaciones de mínimos cuadrados.
2. En cada iteración t, calcule los residuos $\epsilon_i^{(t-1)}$ y los pesos asociados $w_i^{(t-1)} = w[\epsilon_i^{(t-1)}]$ de la iteración anterior.
3. Resolver para los nuevos estimadores obtenidos por mínimos cuadrados ponderados:

$$\boldsymbol{\beta}^{(t)} = \left[\boldsymbol{X}^\top \boldsymbol{X}^{(t-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{X}^{(t-1)} \boldsymbol{y}$$

donde X es la matriz del modelo, con $x_i^\top$ como su i-ésima fila, y $W^{(t-1)} = \text{diag}\{w_i^{(t-1)}\}$ es la matriz de ponderación actual.

Los pasos 2 y 3 se repiten hasta que los coeficientes estimados convergen.

La **matriz de covarianza asintótica** de $\boldsymbol{\beta}$ es

$$V(\boldsymbol{\beta}) = \frac{E(\psi^2)}{[E(\psi')]^2}(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Usando $\sum[\psi(\epsilon_i)]^2$ para estimar $E(\psi^2)$ y $\left[\sum \psi'(\epsilon_i)/n\right]^2$ para estimar $[E(\psi')]^2$ se obtiene la matriz de covarianza asintótica estimada, $\widehat{V}(\boldsymbol{\beta})$ (que no es confiable para muestras pequeñas).

Tabla 1: Funciones objetivo con su respectivo peso y método

Método	**Función Objetivo**	**Función Peso**
Mínimos Cuadrados Ordinarios	$\rho_{MCO}(e) = e^2$	$w_{MCO}(e) = 1$
Huber	$\rho_H(e) = \begin{cases} \frac{1}{2}e^2 & \lvert e\rvert \leq k \\ k\lvert e\rvert - \frac{1}{2}k^2 & \lvert e\rvert > k \end{cases}$	$w_H(e) = \begin{cases} 1 & \lvert e\rvert \leq k \\ \frac{k}{\lvert e\rvert} & \lvert e\rvert > k \end{cases}$
Bisquare	$\rho_B(e) = \begin{cases} \frac{k^2}{6}\left[1 - \left(1 - (e/k)^2\right)^3\right] & \lvert e\rvert < k \\ \frac{k^2}{6} & \lvert e\rvert \geq k \end{cases}$	$w_B(e) = \begin{cases} \left[1 - (e/k)^2\right]^2 & \lvert e\rvert < k \\ 0 & \lvert e\rvert \geq k \end{cases}$

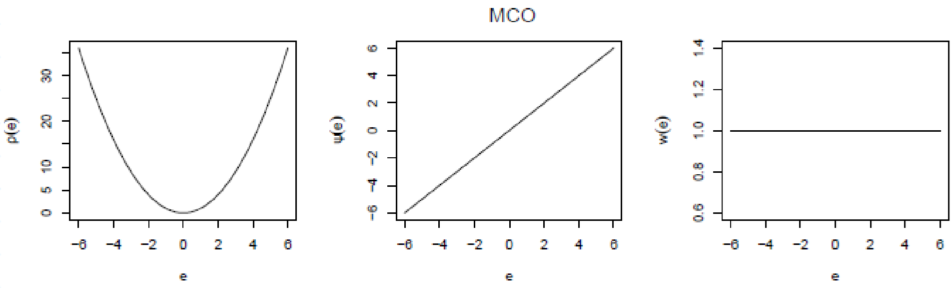
- El valor k para los estimadores de Huber y bisquare se conoce como *constante de sintonización*; valores más pequeños de k producen más resistencia a los valores atípicos, pero a expensas de una menor eficiencia cuando los errores se distribuyen normalmente.

- La constante de sintonización generalmente se selecciona para proporcionar una eficiencia razonablemente alta en el caso normal; en particular, $k = 1.345\,\sigma$ para Huber y $k = 4.685\,\sigma$ para bisquare (donde σ es la desviación estándar de los errores) produce una eficiencia del 95 % cuando los errores son normales, y aún ofrece protección contra valores atípicos.

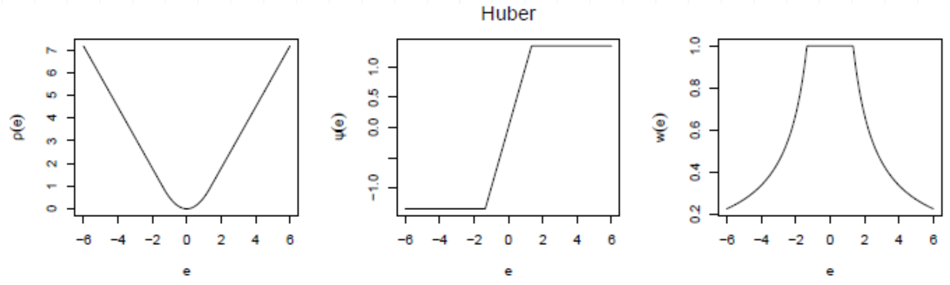
En una aplicación, necesitamos una estimación de la desviación estándar de los errores para usar estos resultados. Por lo general, se emplea una medida robusta de dispersión con preferencia a la desviación estándar de los residuos. Por ejemplo, un enfoque común es tomar

$$\hat{\sigma} = \frac{MAR}{0.6745},$$

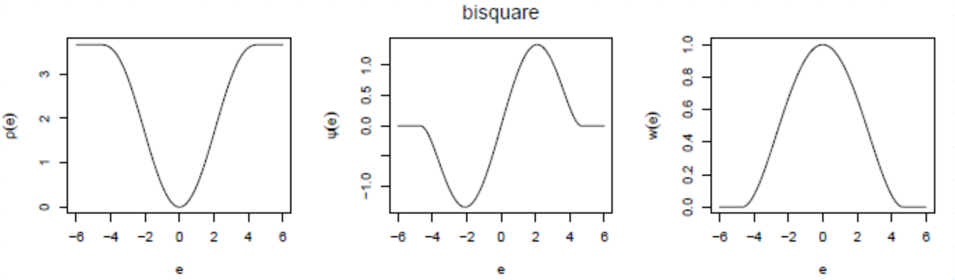
donde MAR es la mediana residual absoluta.



Mínimos cuadrados asigna el mismo peso a cada observación.



- Las funciones objetivo de mínimos cuadrados y Huber aumentan sin límite a medida que el residual ϵ se aleja de 0, pero la función objetivo de mínimos cuadrados aumenta de manera más rápida.
- Los pesos para el estimador de Huber disminuyen cuando $|\epsilon| > k$.



- Niveles de la función objetivo bisquare eventualmente se nivelan (para $|\epsilon| > k$).
- Los pesos para el bisquare disminuyen conforme ϵ se aleja de 0, y son 0 para $|\epsilon| > k$.
- La función ψ del estimador bisquare se vuelve 0 para residuos muy grandes.

Aplicación de regresión Robusta

Aplicación

Para ilustrar la aplicación, usaremos la base de datos stackloss que contiene observaciones recolectadas en una planta química de oxidación de amoníaco, donde se midió la pérdida de masa del gas (stack.loss) bajo diferentes condiciones operativas.

- *stack.loss*: Pérdida de masa de gas en la torre de absorción (variable respuesta).
- *Air.Flow*: Flujo de aire operando en el sistema.
- *Water.Temp*: Temperatura del agua en el proceso.
- *Acid.Conc.*: Concentración de ácido utilizada.

Desarrollo en clase

Ejercicio

Implemente el modelo de regresión robusta y justifique por qué es adecuado para la base de datos `Prestige` que reúne información sobre diversas ocupaciones en Canadá a mediados del siglo XX, disponible en el paquete `carData`:

- *prestige*: Puntuación de prestigio ocupacional (variable respuesta).
- *income*: Ingreso promedio de la ocupación (en dólares canadienses).
- *education*: Años promedio de educación de los trabajadores en la ocupación.
- *women*: Porcentaje de mujeres en la ocupación.
- *census*: Código del censo.
- *type*: Tipo de ocupación (profesional, manual, etc.).