Modelos de Regresión Robustos

Radiación trás el accidente de la Central Nuclear Fukushima

Flavia Felicioni[1] Clara Villalba[2]

[1]flaviafelicioni@gmail.com

[2]claraofvillalba@hotmail.com

Enfoque Estadístico del Aprendizaje - diciembre 2021

Contenido

1. Resumen
2. Motivación
3. Métodos de regresión robustos
4. Referencias

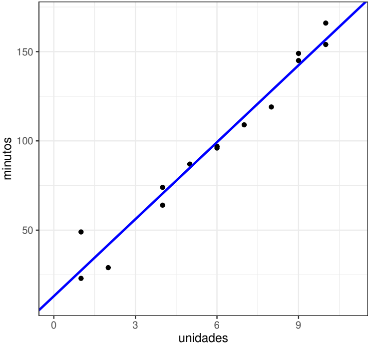
Resumen

- El modelo de regresión lineal simple sirve para vincular 2 variables aleatorias, la variable predictora con la variable dependiente o de respuesta.
- La regresión múltiple se extiende a más variables explicativas o predictoras
- Variables predictoras pueden ser numéricas y/o categóricas
- Modelos requieren satisfacer supuestos
- Modelos simples pero poco flexibles
- Sensibles ante la presencia de datos atípicos (outliers).
- Los métodos robustos surgen como alternativa de solución.

Motivación

Ejemplo introductorio: [1]

- Duración de una llamada al servicio técnico (a predecir)
- Cantidad de componentes electrónicos que deben ser reparados

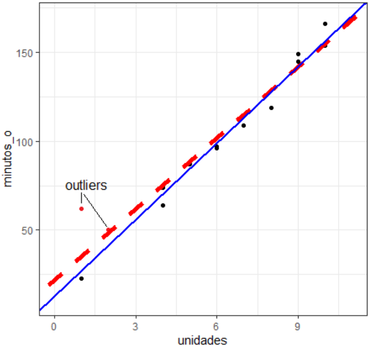


[1] datos originales tomados de [2]

Motivación

Ejemplo introductorio: [1]

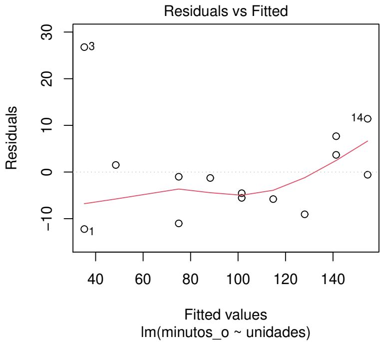
- Duración de una llamada al servicio técnico (a predecir)
- Cantidad de componentes electrónicos que deben ser reparados



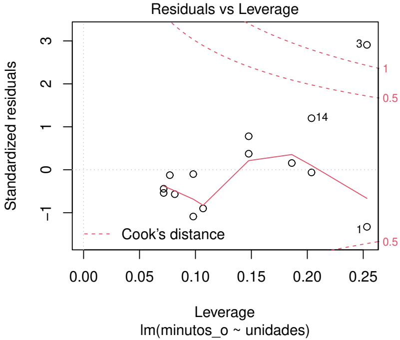
[1] datos originales tomados de [2]

Motivación - Residuos

Diagnóstico de modelo



Se observan outliers



D3 es observación es muy influyente

Regresión Lineal

Modelo lineal múltiple

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \tag{1}$$

y en forma matricial

$$Y = X\beta + \epsilon \tag{2}$$

donde $Y = [y_1, y_2 ...]^T$, β es el vector de parámetros (p*1) y

$$X = \begin{bmatrix} 1 & x_{11} & .. & x_{1,p-1} \\ 1 & x_{n1} & .. & x_{n,p-1} \end{bmatrix}$$

Ajuste de parámetros

OLS - Mínimos cuadrados a (1) es

$$g(\beta_0, .., \beta_j) = \textstyle\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_i)^2 \tag{3}$$

Ecuaciones normales

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}) = 0 \tag{4}$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}) x_{ij} = 0 \tag{5}$$

Ajuste de parámetros

OLS - Mínimos cuadrados a (1) es

$$g(\beta_0, .., \beta_j) = \textstyle\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_i)^2 \tag{3}$$

Ecuaciones normales

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}) = 0 \tag{4}$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}) x_{ij} = 0 \tag{5}$$

cuya solución

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{6}$$

Ajuste robusto

M-Estimadores de Regresión

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} g(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_i) \tag{7}$$

Ajuste robusto

M-Estimadores de Regresión

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} g(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_i) \tag{7}$$

y entonces

$$\sum_{i=1}^{n} W_i(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}) X_i = 0 \tag{8}$$

donde $\psi = g'$

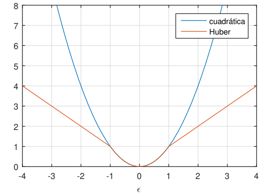
$$W_i = \frac{\psi(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})}{(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})} \tag{9}$$

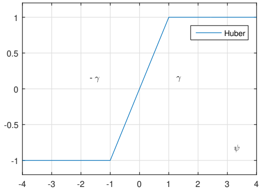
Estimadores comúnmente utilizados [3]

Tipo	Función objetivo $f_\phi(x)$	Función de influencia $f'_\phi(x)$	Función de peso $g_\phi(x)$
L_2	$\frac{x^2}{2}$	x	1
L_1	$\lvert x\rvert$	$\operatorname{sign}(x)$	$\frac{1}{\lvert x\rvert}$
L_p	$\frac{\lvert x\rvert^v}{v}$	$\operatorname{sign}(x)\lvert x\rvert^{v-1}$	$\lvert x\rvert^{v-2}$
Huber $(L_1 - L_2)$	$\begin{cases} \frac{x^2}{2} & \lvert x\rvert \le \gamma \\ \gamma\left(\lvert x\rvert - \frac{\gamma}{2}\right) & \lvert x\rvert \ge \gamma \end{cases}$	$\begin{cases} x & \lvert x\rvert \le \gamma \\ \gamma\operatorname{sign}(x) & \lvert x\rvert \ge \gamma \end{cases}$	$\begin{cases} 1 & \lvert x\rvert \le \gamma \\ \frac{\gamma}{\lvert x\rvert} & \lvert x\rvert \ge \gamma \end{cases}$
Cauchy	$\frac{c^2}{2}\log(1 + (x/c)^2)$	$\frac{x}{1 + (x/c)^2}$	$\frac{1}{1 + (x/c)^2}$
Gema y McClure	$\frac{x^2/2}{1 + x^2}$	$\frac{x}{(1 + x^2)^2}$	$\frac{1}{(1 + x^2)^2}$
Welsch	$\frac{c^2}{2}\left(1 - e^{-(x/c)^2}\right)$	$xe^{-(x/c)^2}$	$e^{-(x/c)^2}$
Tukey	$\begin{cases} \frac{c^2}{2}(1 - (1 - (x/c)^2)^3) & \lvert x\rvert \le c \\ \frac{c^2}{6} & \lvert x\rvert \ge c \end{cases}$	$\begin{cases} x(1 - (x/c)^2)^2 \\ 0 \end{cases}$	$\begin{cases} (1 - (x/c)^2)^2 \\ 0 \end{cases}$

Implementaciones en R

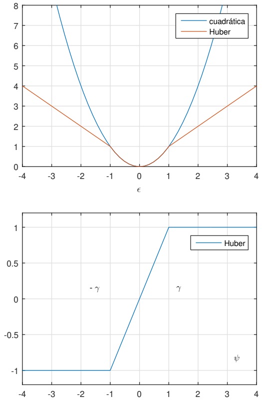
Función de Huber





Implementaciones en R

Función de Huber



- Paquete MASS - *rlm*: *iterated re-weighted least squares (IWLS)- psi functions are supplied for the Huber, Hampel and Tukey*
- Paquete robustbase- *lmrob*: *se propone estimador inicial, se calcula estimador de escala y finalmente se obtienen los estimadores de los parámetros, minimizando la función objetivo*

Referencias

M. E. Szretter Noste
Apunte de Regresión Lineal
Carrera de Especialización en Estadística para Ciencias de la Salud, Universidad de Buenos Aires, 2017

Samprit Chatterjee, Ali S. Hadi.
Regression Analysis by Example.
Fourth Edition. ISBN:9780471746966 —DOI:10.1002/0470055464. 2006. John Wiley Sons, Inc.

F. Parra Rodríguez
Estadística y Machine Learning con R: Ejercicios resueltos con R (Spanish Edition)
ISBN-10 : 6202252162- 2017