UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

UNA INTRODUCCIÓN A LA REGRESIÓN ROBUSTA

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

LUIS ÁNGEL MARTÍNEZ ZARAZÚA

DIRECTORA DE TESIS:
MAT. MARGARITA ELVIRA CHÁVEZ CANO
2015

Ciudad Universitaria, D. F.

Nombre del Alumno: Luis Ángel Martínez Zarazúa

Carrera: Actuaría

Número de cuenta: 303090476

Correo Electrónico: mtzarazua@gmail.com

Grado y nombre del propietario: Dra. Ruth Selene Fuentes García

Grado y nombre del propietario: M. en C. Inocencio Rafael Madrid Ríos

Grado y nombre del propietario: Mat. Margarita Elvira Chávez Cano

Grado y nombre del propietario: Act. Francisco Sánchez Villareal

Grado y nombre del propietario: Act. Edna Gabriela López Estrada

Título del trabajo: Una introducción a la Regresión Robusta
Número de páginas: 113
Año: 2015

A mis padres que me brindaron la oportunidad de tener una educación universitaria

A mi hermano y Mauricio por siempre estar presentes

A mi familia por el apoyo incondicional

A la maestra Margarita por su paciencia e invaluable ayuda

A mis amigos por hacer de esto la mejor experiencia

A los que no vemos pero que sé que están

A Diana y Ernesto por el tiempo sacrificado y contribuir para poder concluir este trabajo

A Dios ...

Luis

Dicen que soy héroe,
yo débil, tímido, casi cobarde,
si siendo como soy
pude hacer lo que hice,
imagínense lo que pueden hacer
todos ustedes juntos.

Mahatma Gandhi

Índice general

Introducción

A menudo existen fenómenos en los que el comportamiento de una variable se puede explicar a través de otras variables, el análisis de regresión es una importante herramienta estadística utilizada para el análisis de dichos fenómenos. Existen diferentes técnicas de análisis de regresión sin embargo, desde hace más de 200 años, la regresión lineal por mínimos cuadrados es el método de regresión más popular ya que proporciona estimadores con expresiones explícitas y que son óptimos al suponer una serie de hipótesis sobre los datos, la más importante es suponer que los datos observados siguen una distribución normal.

Sin embargo, en la realidad no siempre ocurre que todos los datos sigan el mismo patrón, existen algunas observaciones que se comportan de manera diferente, ya sea por comportamiento natural del fenómeno o un simple error al momento de registrar el resultado del fenómeno; a estas observaciones se les conoce como outliers o datos atípicos y pueden influir de manera determinante en la estimación del modelo. Uno podría esperar que si se tiene un cumplimiento aproximado de los supuestos del modelo de regresión lineal no debería de causar grandes variaciones en el resultado de la estimación. Sin embargo, en el primer capítulo de este trabajo se muestra de manera ilustrativa lo sensible que son estos estimadores a la presencia de observaciones atípicas.

A partir de la década de los años 60's se han desarrollado diferentes técnicas de regresión que no se vean afectadas y sean estables en la presencia de outliers. Los métodos de regresión robusta tienen como objetivo proporcionar estimadores que no sufran un cambio significativo cuando no se cumple alguna de los supuestos del modelo de regresión lineal simple (robustez), es decir, que los estimadores robustos sean muy semejastes a los estimadores por mínimos cuadrados al no considerar los outliers; y además, en caso de que se cumplan todos los supuestos del modelo de regresión lineal, los estimadores robustos sean muy aproximados a los estimadores por mínimos cuadrados (eficiencia).

En el capítulo 2 se presentarán las diferentes medidas robustas de localización y escala que se han desarrollado, aplicados a la versión más simple del modelo de regresión lineal con el objetivo de mostrar de manera gráfica las ventajas de cada una de las

diferentes medidas. En el capítulo 3 se presenta la generalización de diferentes métodos de estimación en el modelo de regresión lineal múltiple. Los estimadores que logran tener un balance entre la robustez y la eficiencia son los M-estimadores, presentados por Huber en 1973. Lamentablemente, los M-estimadores pierden sus propiedades de robustez cuando las variables regresoras son aleatorias y se tiene presencia de datos atípicos (puntos de palanca).

En el capítulo 4 se presentan algunos de los estimadores robustos, presentados a partir de las década de los 80's, que buscan resolver los problemas presentados por los M-estimadores en fenómenos con variables regresoras aleatorias: GM-estimadores, S-estimadores y MM-estimadores. Los MM-estimadores, presentados por Yohai en 1987, son los estimadores que logran tener, simultáneamente, los atributos deaseados para un estimador robusto: alta resistencia a los outliers y ser altamente eficientes.

En el capítulo final, se presenta un ejempo estudiado por diferentes autores. Se trata de un conjunto de datos generados de manera artificial con el objetivo de mostrar las propiedades de los MM-estimadores y se lleva a cabo la comparación del desempeño de los diferentes métodos presentados.

Con el fin de ilustrar las ventajas y debilidades de cada uno de los diferentes estimadores presentados, a lo largo del trabajo se presentan diferentes ejemplos con la ayuda del paquete estadístico R.

Capítulo 1

Regresión Lineal

1.1. Regresión Lineal Simple

A menudo nos encontramos con fenómenos en los que el comportamiento de una variable y se puede explicar a través de una variable x

$$y = f(x)$$

Si consideramos que la relación $f(x)$ es una línea recta entonces, la explicación de y a través de x se puede representar de la siguiente manera

$$y = \beta_0 + \beta_1 x \tag{1.1}$$

donde el parámetro β_0 representa la ordenada al origen y β_1 la pendiente de la recta.

En la mayoría de las ocasiones esta relación no describe de manera exacta al conjunto de datos dados debido a que se han omitido muchas variables de poca importancia. Para considerar lo anterior, se agrega un término de error aleatorio, ϵ, que refleja todos los factores diferentes a la variable regresora x y que influyen sobre la variable de respuesta y pero que ninguno de ellos es relevante individualmente.

Considerando lo anterior, el modelo de regresión lineal simple se define de la siguiente manera

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1.2}$$

donde a β_0 y β_1 se les llama coeficientes de regresión y son parámetros desconocidos que deben ser estimados.

Para comprender el modelo de regresión lineal simple, supongamos que se puede fija el valor de la variable regresora x y sólo se observa el valor de la variable de respuesta y. Si x es fija, el término de error ϵ determina las propiedades de la variable y. Supongamos que la media y la varianza de ϵ son 0 y σ^2, respectivamente, entonces la respuesta media en cualquier valor de la variable regresora x es

$$E(y|x) = \mu_{y|x} = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x$$

como se puede observar, es la misma relación que establece el modelo de regresión lineal simple. Por otro lado, la varianza de y para cualquier valor de x es

$$\text{Var}(y|x) = \sigma^2_{y|x} = \text{Var}(\beta_0 + \beta_1 x + \epsilon) = \sigma^2$$

Por lo tanto, el verdadero modelo de regresión es una línea recta de valores promedio $\mu_{y|x} = \beta_0 + \beta_1 x$, es decir, la altura de la línea de regresión en cualquier valor de x es el valor esperado de y para dicho valor de x. Se puede interpretar que la pendiente β_1 es el cambio de la media de y para cada cambio unitario de x. Además, la variabilidad de y para un valor en particular de x queda determinada por la varianza del término de error del modelo, σ^2. Esto significa que hay una distribución de valores de y para cada x, y que la varianza de esta distribución es igual en cada x, para después hacer predicción.

El objetivo principal del análisis de regresión es estimar el valor de β_0 y β_1 a partir de la información con la que se dispone, de tal manera que la recta obtenida se ajuste lo mejor posible a dicha información.

1.2. Hipótesis del modelo

Consideremos un conjunto de n datos $y_1, y_2, \ldots, y_n$ de una variable de respuesta y, los cuales están relacionados a un conjunto de valores $x_1, x_2, \ldots, x_n$ de una variable regresora x. Cada valor y_i representa la medición de un experimento que se replica bajo las mismas condiciones n veces, es decir, la variable y es una variable aleatoria. Supondremos que la relación entre ambas variables satisface el modelo de regresión lineal simple

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1.3}$$

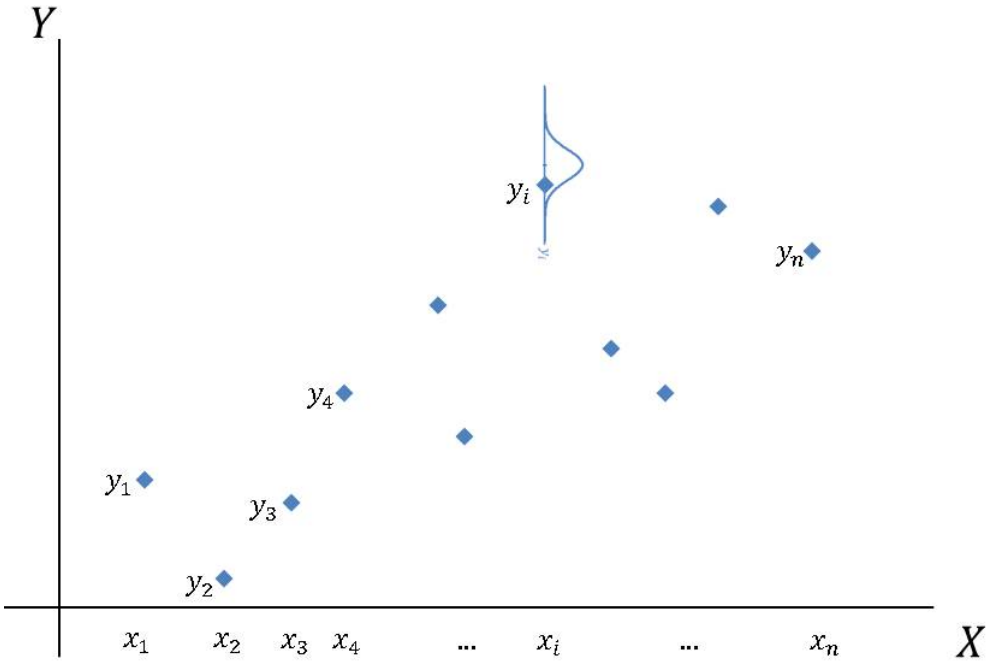


FIGURA 1.1: Cómo se generan las observaciones en el modelo de regresión lineal simple

donde ϵ_i representa un término de error aleatorio. Es conveniente considerar que la variable regresora x está controlada por el investigador, es decir, no es una variable aleatoria. Por lo tanto, la distribución de la variable de respuesta y está determinada sólo por la distribución del error aleatorio ϵ.

Dado que las observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ son mediciones de un mismo experimento realizado bajo las mismas condiciones, se tienen supuestos sobre el término de error aleatorio y en los cuales descansa el modelo de regresión lineal simple:

a) La esperanza de los errores aleatorios es cero

$$E(\epsilon_i) = 0 \qquad i = 1, 2, \ldots, n$$

Con este supuesto se establece que el efecto individual de las variables incluidas en el término de error tienden a compensarse.

b) Los errores tienen la misma varianza

$$\text{Var}(\epsilon_i) = \sigma^2 \qquad i = 1, 2, \ldots, n$$

Esta hipótesis establece que todos los elementos aleatorios tienen la misma varianza y que ésta es constante.

c) Los errores no están correlacionados

$$E(\epsilon_i \epsilon_j) = 0 \qquad i \neq j$$

d) Los errores son identicamente distribuidos con distribución Normal

$$\epsilon_i \sim N(0, \sigma^2) \qquad i = 1, 2, \ldots, n$$

Esta es la hipótesis más común en el modelo de regresión lineal simple. Bajo la hipótesis de normalidad la covarianza cero lleva a considerar a los errores como variables independientes.

1.3. Estimación de los coeficientes de regresión

Existen varios métodos para estimar el valor de los coeficientes del modelo de regresión lineal simple, entre estos, el método de mínimos cuadrados es el más conocido y utilizado. Este método fue desarrollado por Legendre en 1805 y tuvo un éxito inmediato debido a que era el único con el que se podía realizar fácilmente el cálculo de los estimadores de los coeficientes del modelo de regresión lineal simple antes de que existieran las computadoras.

La estimación de los parámetros β_0 y β_1 con el método por mínimos cuadrados se realiza de tal manera que denotando a la recta estimada:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1.4}$$

la suma de los cuadrados de las diferencias entre las observaciones y_i y la recta es la mínima posible.

A la diferencia entre el valor observado y_i y el valor ajustado por el modelo de regresión lineal simple $\hat{y}_i$ se define como residual.

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \qquad i = 1, 2, \ldots, n$$

Por lo tanto, si $\hat{\beta}_0$ y $\hat{\beta}_1$ corresponden a los estimadores por mínimos cuadrados de los parámetros β_0 y β_1, respectivamente, cumplen el criterio:

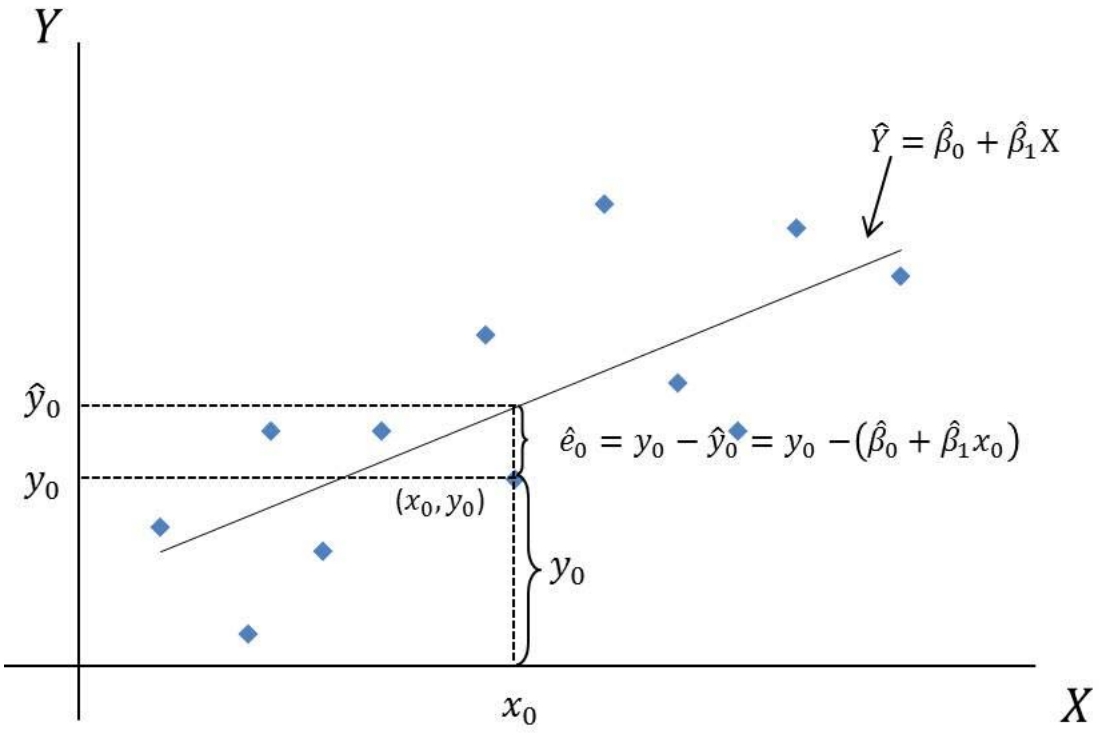


FIGURA 1.2: Recta estimada por Mínimos Cuadrados

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{1.5}$$

Dado que la ecuación (1.5) es de grado par y positiva, la expresión para cada uno de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ se puede obtener, equivalentemente, derivando la función $S(\beta_0, \beta_1)$ respecto a cada parámetro β_0 y β_1 e igualar a cero para obtener el punto donde se logra el mínimo absoluto.

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

y

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Al simplificar las igualdades anteriores, se obtiene el siguiente sistema de ecuaciones llamado "ecuaciones normales"

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\ \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i &= 0 \end{aligned} \tag{1.6}$$

La solución del sistema de ecuaciones normales es:

$$\begin{aligned} \hat{\beta}_0 &= \frac{\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i x_i - \frac{1}{n}(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned} \tag{1.7}$$

1.3.0.1. Propiedades de los estimadores por mínimos cuadrados

Los estimadores por mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen algunas propiedades importantes; antes de enunciarlas, es importante hacer algunas simplificaciones que facilitarán la comprensión de estas propiedades.

Primero, reescribiremos al estimador $\hat{\beta}_1$ como una combinación lineal de la variable y.

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) y_i = \sum_{i=1}^{n} c_i y_i \end{aligned} \tag{1.8}$$

para $i = 1, 2, \ldots, n$.

Si sustituimos el valor de y_i de acuerdo al modelo de regresión lineal simple se tiene la siguiente igualdad

$$\begin{aligned} \hat{\beta}_1 &= \sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n} c_i(\beta_o + \beta_1 x_i + \epsilon_i) \\ &= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i \epsilon_i \end{aligned} \tag{1.9}$$

Por otro lado, dado que la variable regresora x no es una variable aleatoria, se cumplen las siguientes igualdades

$$\begin{aligned}\sum_{i=1}^{n} c_i x_i &= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) x_i \\ &= \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2)} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 1\end{aligned} \tag{1.10}$$

note que

$$\begin{aligned}\sum_{i=1}^{n} c_i &= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \\ &= \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 0\end{aligned} \tag{1.11}$$

Sustituyendo las igualdades de las ecuaciones (1.10) y (1.11) en la ecuación (1.9) se obtiene la siguiente expresión

$$\hat{\beta}_1 = \beta_0 * 0 + \beta_1 * 1 + \sum_{i=1}^{n} c_i \epsilon_i = \beta_1 + \sum_{i=1}^{n} c_i \epsilon_i \tag{1.12}$$

Por lo tanto, el estimador $\hat{\beta}_1$ se puede representar como el verdadero valor de la pendiente de la recta de regresión (desconocido) más una combinación lineal del error aleatorio.

1.3.0.2. Esperanza

Considerando la ecuación (1.12), la esperanza del estimador $\hat{\beta}_1$ de obtiene de la siguiente manera

$$E\left[\hat{\beta}_1\right] = E\left[\beta_1 + \sum_{i=1}^{n} c_i \epsilon_i\right] = \beta_1 + \sum_{i=1}^{n} E\left[c_i \epsilon_i\right] \tag{1.13}$$

debido a que los c_i's no son variables aleatorias,se obtiene que $\hat{\beta}_1$ es un estimador insesgado de β_1, es decir,

$$E\left[\hat{\beta}_1\right] = \beta_1 + \sum_{i=1}^{n} c_i E\left[\epsilon_i\right] = \beta_1 + \sum_{i=1}^{n} c_i * 0 = \beta_1 \tag{1.14}$$

De igual manera, $\hat{\beta}_0$ es un estimador insesgado de β_0. Para demostrarlo calculamos la esperanza de $\hat{\beta}_0$

$$\begin{aligned} E\left[\hat{\beta}_0\right] &= E\left[\bar{y} + \hat{\beta}_1\bar{x}\right] = E\left[\frac{\sum_{i=1}^{n} y_i}{n}\right] - E\left[\hat{\beta}_1\right]\bar{x} \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[\beta_0 + \beta_1 x_i + \epsilon_i\right] - \beta_1\bar{x} \\ &= \frac{1}{n}\beta_0 + \beta_1\frac{\sum_{i=1}^{n} x_i}{n} - \beta_1\bar{x} = \beta_0 \end{aligned} \tag{1.15}$$

1.3.0.3. Varianza

Consideremos nuevamente la ecuación (1.12)

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} c_i\epsilon_i$$

La varianza de $\hat{\beta}_1$ es:

$$\begin{aligned} c\text{Var}\left(\hat{\beta}_1\right) &= \text{Var}\left(\sum_{i=1}^{n} c_i\epsilon_i\right) = \sum_{i=1}^{n}\text{Var}\left(c_i\epsilon_i\right) \\ &= \sum_{i=1}^{n} c_i^2\text{Var}\left(\epsilon_i\right) = \sigma^2\sum_{i=1}^{n} c_i^2 \end{aligned} \tag{1.16}$$

La igualdad anterior se cumple considerando el supuesto de varianza constante, σ^2, e independencia del término de error aleatorio, ϵ_i para $i = 1, \ldots, n$.

Por otro lado, la varianza del estimador $\hat{\beta}_0$ es

$$\text{Var}\left(\hat{\beta}_0\right) = \text{Var}\left(\bar{y} - \hat{\beta}_1\bar{x}\right) = \text{Var}(\bar{y}) + \bar{x}^2\text{Var}\left(\hat{\beta}_1\right) - 2\bar{x}\text{Cov}\left(\bar{y}, \hat{\beta}_1\right) \tag{1.17}$$

Consideremos las siguientes igualdades:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i + \epsilon_i)}{n} = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon} \tag{1.18}$$

entonces

$$\begin{aligned} \text{Var}(\bar{y}) &= \text{Var}(\beta_0 + \beta_1 \bar{x} + \bar{\epsilon}) \\ &= \text{Var}(\beta_0) + \text{Var}(\beta_1 \bar{x}) + \text{Var}(\bar{\epsilon}) = \frac{\sigma^2}{n} \end{aligned} \tag{1.19}$$

Por otro lado

$$\begin{aligned} \text{Cov}(\bar{y}, \epsilon_i) &= \text{Cov}(\beta_0 + \beta_1 \bar{x} + \bar{\epsilon}, \epsilon_i) \\ &= \text{Cov}(\beta_0, \epsilon_i) + \bar{x}\text{Cov}(\beta_1, \epsilon_i) + \text{Cov}(\bar{\epsilon}, \epsilon_i) \\ &= 0 + 0 + \frac{1}{n} \sum_{i=1}^{n} \text{Cov}(\epsilon_j, \epsilon_i) = \frac{\sigma^2}{n} \end{aligned} \tag{1.20}$$

y

$$\begin{aligned} \text{Cov}\left(\bar{y}, \hat{\beta}_1\right) &= \text{Cov}\left(\bar{y}, \hat{\beta}_1 + \sum_{i=1}^{n} c_i \epsilon_i\right) \\ &= \text{Cov}(\bar{y}, \beta_1) + \sum_{i=1}^{n} c_i \text{Cov}(\bar{y}, \epsilon_i) \\ &= 0 + \frac{\sigma^2}{n} \sum_{i=1}^{n} c_i = 0 \end{aligned} \tag{1.21}$$

Sustituyendo las igualdades (1.17), (1.19) y (1.20) en la ecuación (1.12) se tiene:

$$\text{Var}\left(\hat{\beta}_0\right) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{n S_x^2} + 2\bar{x} * 0 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2}\right) \tag{1.22}$$

1.3.0.4. Estimación de σ^2

El estimador de σ^2 se obtiene apartir de la suma de cuadrados de residuales $\sum_{i=1}^{n} e_i^2$ por lo que debemos de considerar las siguientes igualdades

$$e_i = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x}) \tag{1.23}$$

y

$$(y_i - \bar{y}) = \beta_1(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon}) \tag{1.24}$$

A partir de las ecuaciones (1.23) y (1.24), la suma de cuadrados de residuales se puede reescribir de la siguiente manera

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(\hat{\beta}_1 - \beta_1)^2(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2 - 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(x_i - \bar{x})^2(\epsilon_i - \bar{\epsilon}) \tag{1.25}$$

La esperanza de la ecuación (1.25) es la siguiente

$$\begin{aligned} E\left[\sum_{i=1}^{n} e_i^2\right] &= \sum_{i=1}^{n}(\hat{\beta}_1 - \beta_1)^2 E\left[x_i - \bar{x}\right]^2 + E\left[\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2\right] \\ &\quad -2E\left[\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(x_i - \bar{x})^2(\epsilon_i - \bar{\epsilon})\right] \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + (n-1)\sigma^2 - 2\sigma^2 = (n-2)\sigma^2 \end{aligned} \tag{1.26}$$

Por lo tanto, un estimador insesgado para σ^2 está dado por

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{1.27}$$

Debido a que el valor de $\hat{\sigma}^2$ depende de la suma de cuadrados de residuales, cualquier violación de los supuestos sobre el error aleatorio del modelo pueden dañar gravemente la utilidad de $\hat{\sigma}^2$ como estimador de σ^2 y producir un modelo inestable, en el sentido de que una muestra distinta de observaciones podría conducir a un modelo totalmente diferente y obtener conclusiones opuestas.

1.3.1. Propiedades de los residuales en el modelo de regresión lineal simple

A partir de la aplicación del método por mínimos cuadrados para la estimación de los coeficientes de regresión surgen las siguientes propiedades:

1. La suma de los residuales es igual a cero:

$$\sum_{i=1}^{n} e_i = 0 \tag{1.28}$$

2. La recta de regresión por mínimos cuadrados siempre pasa por el punto $(\bar{x}, \bar{y})$

3. La suma de los residuales ponderados por el valor correspondiente de la variable regresora es igual a cero.

$$\sum_{i=1}^{n} x_i e_i = 0$$

4. La suma de los residuales ponderados por el valor ajustado correspondiente es igual a cero.

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

5. La suma de los valores observados y_i es igual a la suma de los valores ajustados $\hat{y}_i$:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

1.3.2. Eficiencia de los estimadores

El siguiente teorema establece que los estimadores por mínimos cuadrados son los de mínima varianza dentro de los estimadores lineales e insesgados.

Teorema 1.3.1. Teorema de Gauss-Markov.- Bajo los supuestos del modelo de regresión lineal, los estimadores por mínimos cuadrados de los coeficientes de regresión son los estimadores insesgados con mínima varianza del conjunto de estimadores que se pueden representar como combinación lineal de las observaciones.

Demostración:

Primero demostraremos que el teorema se cumple en el caso del estimador $\hat{\beta}_1$. Para ellos supondremos que existe un estimador $\hat{\beta}_1'$ diferente a $\hat{\beta}_1$ que también cumple con

la condición de que es una combinación lineal de las y_i's y que, además, es de varianza minima. Anteriormente se reexpresó al estimador $\hat{\beta}_1$ como una combinación lineal de las observaciones y_i:

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$$

Por lo tanto suponemos que el estimador $\hat{\beta}_1'$ se puede expresar de la siguiente forma:

$$\hat{\beta}_1' = \sum_{i=1}^{n} \alpha_i y_i \tag{1.29}$$

donde los valores de las α_i's son constantes y, dado que $\hat{\beta}_1'$ es diferente al estimador por mínimos cuadrados, por lo menos alguna o algunas de las constantes α_i son diferentes a la constante c_i para una misma i.

Por otro lado, dado que $\hat{\beta}_1'$ es insesgado se debe cumplir que

$$E\left(\hat{\beta}_1'\right) = \beta_1 \tag{1.30}$$

de donde:

$$\begin{aligned}
E\left(\bar{\beta}_1\right) &= E\left(\sum_{i=1}^{n} c_i y_i\right) = E\left(\sum_{i=1}^{n} c_i \left(\beta_0 + \beta_1 x_i + \epsilon_i\right)\right) \\
&= E\left(\beta_0 \textstyle\sum_{i=1}^{n} c_i\right) + E\left(\beta_1 \sum_{i=1}^{n} c_i x_i\right) + E\left(\sum_{i=1}^{n} c_i \epsilon_i\right) \\
&= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i E(\epsilon_i) \qquad (1.31) \\
&= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i * 0 \\
&= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i
\end{aligned}$$

Por lo que el estimador $\hat{\beta}_1'$ será insesgado si se cumplen las siguientes igualdades

$$\sum_{i=1}^{n} \alpha_i = 0$$

$$\sum_{i=1}^{n} \alpha_i x_i = 1$$

Por otro lado, la varianza del estimador $\hat{\beta}_1'$ es

$$\text{Var}\left(\hat{\beta}_1'\right) = \text{Var}\left(\sum_{i=1}^{n} \alpha_i y_i\right) = \sum_{i=1}^{n} \alpha_i^2 \text{Var}\left(y_i\right) = \sum_{i=1}^{n} \alpha_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} \alpha_i^2 \tag{1.32}$$

donde σ^2 es un parámetro desconocido. Por lo tanto, el problema puede plantearse como minimizar $\sum_{i=1}^{n} \alpha_i^2$ sujeto a las condiciones de insesgamiento, es decir,

$$\text{mín} \sum_{i=1}^{n} c_i^2 \text{ sujeto a } \sum_{i=1}^{n} \alpha_i = 0 \text{ y } \sum_{i=1}^{n} \alpha_i x_i - 1 = 0 \tag{1.33}$$

Aplicaremos el método de multiplicadores de Lagrange:

Sea

$$\phi = \sum_{i=1}^{n} \alpha_i^2 - 2\lambda \left(\sum_{i=1}^{n} \alpha_i\right) - 2\gamma \left(\sum_{i=1}^{n} \alpha_i x_i - 1\right) \tag{1.34}$$

entonces

$$\begin{aligned}
\frac{\partial \phi}{\partial \lambda} &= -2 \sum_{i=1}^{n} (\alpha_i) &&= 0 \Rightarrow \sum_{i=1}^{n} \alpha_i = 0 \\
\frac{\partial \phi}{\partial \gamma} &= -2 \left(\sum_{i=1}^{n} \alpha_i x_i - 1\right) &&= 0 \Rightarrow \sum_{i=1}^{n} \alpha_i x_i = 1 \\
\frac{\partial \phi}{\partial \alpha_i} &= -2\alpha_i - 2\lambda - 2\gamma x_i &&= 0 \Rightarrow \alpha_i - \lambda - \gamma x_i = 0
\end{aligned} \tag{1.35}$$

Haciendo la suma sobre $i = 1, \ldots, n$ para $\dfrac{\partial \phi}{\partial \alpha_i}$ se tiene

$$\sum_{i=1}^{n} \alpha_i - n\lambda - \gamma \sum_{i=1}^{n} x_i = 0 \tag{1.36}$$

Sustituyendo la condición $\sum_{i=1}^{n} \alpha_i = 0$ en la ecuación (1.36) se tiene lo siguiente:

$$\lambda = -\gamma \bar{x} \tag{1.37}$$

Consideremos nuevamente la ecuación $\frac{\partial \phi}{\partial \alpha_i}$ en (1.35), multiplicando por x_i y sumando sobre i se tiene:

$$\begin{aligned} \alpha_i x_i - \lambda x_i - \gamma x_i^2 &= 0 \\ \sum_{i=1}^{n} \alpha_i x_i - \lambda \sum_{i=1}^{n} x_i - \gamma \sum_{i=1}^{n} x_i^2 &= 0 \end{aligned} \tag{1.38}$$

Considerando que $\sum_{i=1}^{n} \alpha_i x_i = 1$ y sustituyendo (1.37) en (1.38):

$$1 + \gamma \bar{x} \sum_{i=1}^{n} x_i - \gamma \sum_{i=1}^{n} x_i^2 = 0 \tag{1.39}$$

Despejando γ tenemos

$$\begin{aligned} \gamma \left(\bar{x} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2 \right) &= -1 \\ \Rightarrow \gamma = \frac{1}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} &= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})} \end{aligned} \tag{1.40}$$

Por lo tanto, de la tercera ecuación de (1.35):

$$\begin{aligned} \alpha_i &= \lambda + \gamma x_i \\ &= -\gamma \bar{x} + \gamma x_i = \gamma (x_i - \bar{x}) \\ &= \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned} \tag{1.41}$$

Dado que $\hat{\beta}_1' = \sum_{i=1}^{n} \alpha_i y_i$ se tiene que

$$\hat{\beta}_1' = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} y_i = \hat{\beta}_1 \tag{1.42}$$

que era lo que se quería demostrar. De manera análoga se contruye el estimador para $\hat{\beta}_0$.

A partir del resultado anterior, se dice que los estimadores por mínimos cuadrados son los estimadores lineales insesgados óptimos , donde "óptimos" significa que son de varianza mínima.

Pero los estimadores por mínimos cuadrados tiene una propiedad aún más importante, si el término aleatorio tiene una distribución Normal, el estimador es eficiente, es decir, tiene la menor varianza posible dentro de la clase de los estimadores insesgados, sean lineales o no. Lo anterior se valida al probar que los estimadores por mínimos cuadrados coinciden con los estimadores máximo verosímil cuando los errores del modelo tienen una distribución Normal, $\epsilon_i \sim N(0, \sigma^2)$ para $i = 1, \ldots, n$.

Consideremos el modelo de regresión lineal simple donde los errores siguen una distribución $\epsilon_i \sim N(0, \sigma^2)$. La función de verosimilitud se define de la siguiente manera:

$$L(y_i) = \prod_{i=1}^{n} \frac{1}{\sigma^2\sqrt{2\pi}} e^{\frac{-e_i^2}{2\sigma^2}} \tag{1.43}$$

Aplicando logaritmo a la ecuación (1.43) tenemos

$$\begin{aligned} \ln L(y_1, \ldots, y_n) &= -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \sum_{i=1}^{n}\frac{e_i^2}{\sigma^2} \\ &= -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \sum_{i=1}^{n}\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \end{aligned} \tag{1.44}$$

Para encontrar el estimador máximo verosímil de los parámetros (β_0, β_1 y σ^2) se deriva la ecuación (1.44) respecto a cada parámetro y se iguala a cero.

En el caso de β_0

$$\left.\frac{\partial \ln L(y_i)}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \tag{1.45}$$

despejando $\hat{\beta}_0$ de la igualdad anterior tenemos

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1.46}$$

Realizando el mismo procedimiento, el estimador máximo verosímil del coeficiente β_1 es:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.47}$$

Por otro lado, a partir de la tercera ecuación se obtiene el estimador máximo verosimil del parámetro σ^2, la variaza del error:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 \end{aligned} \tag{1.48}$$

Se puede notar que tanto el estimador $\hat{\beta}_0$ como $\hat{\beta}_1$ coinciden con los estimadores por mínimos cuadrados sin embargo, el estimador $\hat{\sigma}^2$ tiene una pequeña variación. A continuación veremos que el estimador máximo verosímil del parámetro σ^2 no es insesgado. Para probar esto debemos calcular la esperanza del estimador máximo verosimil del parámetro σ^2 y validar que es diferente del parámetro.

$$E\left(\hat{\sigma}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) \tag{1.49}$$

donde

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= \beta_0 + \beta_1 x_i + \epsilon_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \\ &= \left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) x_i + \epsilon_i \end{aligned} \tag{1.50}$$

por lo que, elevando al cuadrado y sumando sobre i

$$\begin{aligned} \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left(-\left(\hat{\beta}_0 - \beta_0\right) - \left(\hat{\beta}_1 - \beta_1\right) x_i + \epsilon_i\right)^2 \\ &= \sum_{i=1}^{n} \left(\left(\hat{\beta}_0 - \beta_0\right) + \left(\hat{\beta}_1 - \beta_1\right) x_i - \epsilon_i\right)^2 \end{aligned} \tag{1.51}$$

sustituyendo (1.51) en (1.49) y desarrollando el cuadrado tenemos lo siguiente:

$$\begin{aligned} E\left(\sum_{i=1}^{n} e_i^2\right) &= E\left(\sum_{i=1}^{n} \left(\hat{\beta}_0 - \beta_0\right)^2 + \left(\hat{\beta}_1 - \beta_1\right)^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \epsilon_i^2 \right. \\ &\quad \left. +2\left(\hat{\beta}_0 - \beta_0\right)\left(\hat{\beta}_1 - \beta_1\right) \sum_{i=1}^{n} x_i - 2\left(\hat{\beta}_0 - \beta_0\right) \sum_{i=1}^{n} \epsilon_i - 2\left(\hat{\beta}_1 - \beta_1\right) \sum_{i=1}^{n} x_i \epsilon_i \right) \\ &= nE\left(\hat{\beta}_0 - \beta_0\right)^2 + \sum_{i=1}^{n} x_i^2 E\left(\hat{\beta}_1 - \beta_1\right)^2 + \sum_{i=1}^{n} E\left(\epsilon_i^2\right) \\ &\quad +2\sum_{i=1}^{n} x_i \left(\left(\hat{\beta}_0 - \beta_0\right)\left(\hat{\beta}_1 - \beta_1\right)\right) - 2E\left(\left(\hat{\beta}_0 - \beta_0\right) \sum_{i=1}^{n} \epsilon_i\right) \\ &\quad -2E\left(\left(\hat{\beta}_1 - \beta_1\right) \sum_{i=1}^{n} x_i \epsilon_i\right) \\ &= n\mathrm{Var}\left(\hat{\beta}_0\right) + \sum_{i=1}^{n} x_i^2 \mathrm{Var}\left(\hat{\beta}_1\right) + n\sigma^2 + 2\sum_{i=1}^{n} x_i \mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_0\right) \\ &\quad -2E\left(\left(\hat{\beta}_0 - \beta_0\right) \sum_{i=1}^{n} \epsilon_i\right) - 2E\left(\left(\hat{\beta}_1 - \beta_1\right) \sum_{i=1}^{n} x_i \epsilon_i\right) \end{aligned} \tag{1.52}$$

sustituyendo el valor de $\mathrm{Var}\left(\hat{\beta}_0\right)$, $\mathrm{Var}\left(\hat{\beta}_1\right)$ y $\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right)$ obtenemos lo siguiente:

$$
\begin{aligned}
E\left(\sum_{i=1}^{n} e_i^2\right) &= n\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\sigma^2 + \sum_{i=1}^{n} x_i^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + n\sigma^2 \\
&\quad -2\sum_{i=1}^{n} x_i \frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} - 2E\left(\left(\hat{\beta}_0-\beta_0\right)\sum_{i=1}^{n}\epsilon_i\right) - 2E\left(\left(\hat{\beta}_1-\beta_1\right)\sum_{i=1}^{n} x_i\epsilon_i\right)
\end{aligned}
\tag{1.53}
$$

como vimos en la ecuación (1.12), dado que $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, podemos reexpresar al estimador $\hat{\beta}_1$ como el parámetro β_1 más una combinación lineal del error aleatorio

$$
\hat{\beta}_1 = \beta_1 + \left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})\,\epsilon_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)
$$

los mismo se puede realizar para el estimador $\hat{\beta}_0$

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\
&= \beta_0\beta_1\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i - \hat{\beta}_1\bar{x} \\
&= \beta_0 + \sum_{i=1}^{n} x_i^2\left(\frac{1}{n} - \frac{\bar{x}\,(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\epsilon_i
\end{aligned}
\tag{1.54}
$$

Considerando las igualdades anteriores se tiene lo sigueinte:

$$
E\left(\left(\hat{\beta}_0-\beta_0\right)\sum_{i=1}^{n}\epsilon_i\right) = E\left(\left(\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}\,(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\epsilon_i\right)\left(\sum_{i=1}^{n}\epsilon_i\right)\right) \tag{1.55}
$$

Desarrollando la ecuación (1.55) y aplicando la esperanza se tiene

$$E\left(\left(\hat{\beta}_0-\beta_0\right)\sum_{i=1}^{n}\epsilon_i\right)=\sigma^2\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{x}\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right)=\sigma^2 \tag{1.56}$$

De manera similiar tenemos

$$\begin{aligned}E\left(\left(\hat{\beta}_1-\beta_1\right)\sum_{i=1}^{n}x_i\epsilon_i\right) &= E\left(\left(\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\epsilon_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right)\left(\sum_{i=1}^{n}x_i\epsilon_i\right)\right)\\ &= \frac{1}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\left(\sigma^2\sum_{i=1}^{n}x_i\left(x_i-\bar{x}\right)\right)\\ &= \sigma^2\end{aligned} \tag{1.57}$$

Por lo tanto, sustituyendo (1.56) y (1.57) en (1.53) se tiene:

$$\begin{aligned}E\left(\sum_{i=1}^{n}e_i^2\right) &= \sigma^2+\frac{n\bar{x}^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sigma^2+\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sigma^2+n\sigma^2-2n\frac{n\bar{x}^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sigma^2\\ &\quad -2\sigma^2-2\sigma^2\\ &= \left(n-3\right)\sigma^2+\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sigma^2-\frac{n\bar{x}^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sigma^2\\ &= \frac{\sum_{i=1}^{n}x_i^2-n\hat{x}^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sigma^2+\left(n-3\right)\sigma^2\\ &= \left(n-2\right)\sigma^2\end{aligned} \tag{1.58}$$

Por lo tanto $E\left(\hat{\sigma}^2\right) \neq \sigma^2$, lo que significa que el estimador de σ^2 obtenido por máxima verosimilitud no es insesgado. Sin embargo, se puede construir un estimadore insesgado para σ^2 a partir de la ecuación anterior de tal manera que si

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{1.59}$$

entonces $E\left(\hat{\sigma}^2\right) = \sigma^2$

Generalmente, los estimadores de máxima verosimilitud tienen mejores propiedades estadísticas que los estimadores por mínimos cuadrados ya que además de cumplir con los supuestos del modelo de regresión lineal sobre la media y la varianza de los errores, deben cumplir el supuesto de normalidad. Los estimadores para β_0 y β_1 por máxima verosimilitud son insesgados y tienen varianza mínima comparados con todos los demás estimadores insesgados. También son estimadores consistentes (consistencia es una propiedad que indica que cuando se tiene una muestra grande los estimadores difieren del valor verdadero del parámetro en una cantidad muy pequeña, cuando n se hace grande). Además, son un conjunto de estadísticas suficientes (esto significa que contienen toda la "información" de la muestra original).

1.4. Medidas de bondad de ajuste. Coeficiente de determinación

Una vez que se ha realizado el ajuste por mínimos cuadrados, es necesario medir el grado de ajuste entre el modelo y los datos. En el caso en el que se hayan estimado varios modelos alternativos podrían utilizarse medidas de bondad de ajuste para determinar cuál es el modelo más adecuado.

El coeficiente de determinación, R^2, es la medida de bondad de ajuste más conocida. Esta medida se basa en una descomposición de la variabilidad total de la variable de respuesta y, como se ve a continuación.

Consideremos lo que se llama la igualdad fundamental del análisis de varianza (o la partición de $\sum_{i=1}^{n} (y_i - \bar{y})^2$), para lo cual consideraremos la siguiente igualdad

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \tag{1.60}$$

La expresión $y_i - \hat{y}_i$ no indica la distancia vertical entre y_i y $\bar{y}$. Elevando al cuadrado ambos lados de la igualdad y sumando sobre i se tiene

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 \tag{1.61}$$

desarrollando el lado derecho de la igualdad se tiene lo siguiente

$$\sum_{i=1}^{n} ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \tag{1.62}$$

Consideremos las siguientes simplificaciones:

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\ \Longrightarrow \hat{y}_i - \bar{y} &= \hat{\beta}_1 (x_i - \bar{x}) \end{aligned} \tag{1.63}$$

Por otro lado

$$\begin{aligned} y_i - \hat{y}_i &= (y_i - \bar{y}) + (\bar{y} - \hat{y}_i) \\ &= (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \end{aligned} \tag{1.64}$$

Por lo tanto, considerando las ecuaciones (1.63) y (1.64) tenemos la siguiente igualdad

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) \Big((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \Big) \tag{1.65}$$

Desarrollando la expresión anterior y sustituyendo $\hat{\beta}_1$ se tiene:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \hat{\beta}_1\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) - \hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\
&\quad - \frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= 0
\end{aligned} \tag{1.66}$$

Por lo tanto la ecuación (1.61) se reduce a la siguiente expresión:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i) + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}e_i^2 \tag{1.67}$$

Esta igualdad nos indica que la variabilidad total de las observaciones se puede descomponer en dos partes: la cantidad de variabilidad en las observaciones y_i explicada por la línea de regresión, más la variación residual que queda sin explicar por la línea de regresión.

Notese lo siguiente:

$$\begin{aligned}
(\hat{y}_i-\bar{y}) &= \hat{\beta}_1(x_i-\bar{x}) \\
\Rightarrow \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 &= \hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
\Rightarrow \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} &= \hat{\beta}_1^2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \\
&= \frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}
\end{aligned} \tag{1.68}$$

El coeficiente de determinación se define como:

$$\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})\right)^2\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}=R_{xy}^2 \tag{1.69}$$

En el caso del modelo lineal simple, el coeficiente de correlación al cuadrado, denotado por R^2, es igual a la proporción de la variación total en torno a la media $\bar{y}$ explicada por la regresión; idealmente se espera que la suma de cuadrados debida a la regresión sea mucho mayor que la suma de cuadrados del error, es decir, que R^2 tenga un valor muy cercano a uno.

Ahora bien,

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\\ \Rightarrow \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} &= \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}+\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\end{aligned} \tag{1.70}$$

Por lo tanto

$$R_{xy}^2=1-\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{1.71}$$

Entonces se cumple que $0 \leq R_{xy}^2 \leq 1$

- Cuando el valor de R_{xy}^2 es cercano a 0, el modelo explica una baja proporción de la variabilidad de la variable y
- Cuando el valor de R_{xy}^2 es cercano a 1, el modelo explica la mayor parte de la variabilidad de y

1.5. Modelo de localización

La versión más sencilla del modelo de regresión lineal simple es aquella donde se pretende estimar el centro μ de una distribución a partir de las observaciones y_i, es decir, se considera que la variable bajo estudio fluctúa alrededor de un cierto valor.

$$y_i = \mu + \epsilon_i \tag{1.72}$$

donde ϵ_i representa las fluctuaciones aleatorias, las cuales no son correlacionadas y tienen la misma dispersión (σ^2) alrededor de su media ($E(\epsilon_i) = 0$). Este modelo es conocido como el modelo de localización y prentende estimar el valor de μ.

Si para la estimación del parámetro μ nos restringimos a la consideración de estimadores lineales e insesgados, de acuerdo con el teorema de Gauss Markov, el estimador por mínimos cuadrados es el estimador lineal insesgado óptimo, donde "óptimo" significa que son los estimadores con mínima varianza.

Los estimadores lineales son de la forma

$$\sum_{i=1}^{n} a_i y_i \tag{1.73}$$

Por otro lado, la condición de insesgamiento da como resultado lo siguiente:

$$\begin{aligned} E\left(\sum_{i=1}^{n} a_i y_i\right) &= \sum_{i=1}^{n} a_i E\left(y_i\right) \\ &= \sum_{i=1}^{n} a_i \left(\mu + E\left(\epsilon_i\right)\right) \\ &= \sum_{i=1}^{n} a_i \mu \\ \Leftrightarrow \qquad &= \sum_{i=1}^{n} a_i = 1 \end{aligned} \tag{1.74}$$

La varianza del estimador está dada por:

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} a_i y_i\right) &= \sum_{i=1}^{n} a_i^2 \text{Var}\left(y_i\right) + \sum_{i \neq j} a_i a_j \text{Cov}\left(y_i, y_j\right) \\
&= \sum_{i=1}^{n} a_i^2 \text{Var}\left(\mu + \epsilon_i\right) + \sum_{i \neq j} a_i a_j \text{Cov}\left(\mu + \epsilon_i, \mu + \epsilon_j\right) \\
&= \sum_{i=1}^{n} a_i^2 \sigma^2 + \sum_{i \neq j} a_i a_j E\left(\epsilon_i \epsilon_j\right) \\
&= \sum_{i=1}^{n} a_i \sigma^2
\end{aligned} \tag{1.75}$$

Por lo tanto, la expresión anterior será mínima cuando lo sea $\sum_{i=1}^{n} a_i^2$, sujeta a la restricción $\sum_{i=1}^{n} a_i = 1$. Entonces se debe minimizar la siguiente expresión:

$$\text{mín} \sum_{i=1}^{n} a_i^2 + \lambda \left(\sum_{i=1}^{n} \left(a_i - 1\right)\right) = Q \tag{1.76}$$

Para minimizar, se obtiene la derivada respecto a a_i y λ y se iguala a cero

$$\begin{aligned}
\frac{\partial Q}{\partial a_i} 2a_i + \lambda = 0 \qquad i = 1, \ldots, n \\
\frac{\partial Q}{\partial \lambda} \sum_{i=1}^{n} a_i - 1 = 0
\end{aligned} \tag{1.77}$$

sumando las i derivadas de Q con respecto a a_i se tiene

$$2\sum_{i=1}^{n} a_i + n\lambda = 0 \Rightarrow \lambda = -\frac{2}{n} \text{ , pues } \sum_{i=1}^{n} a_i = 1 \tag{1.78}$$

sustituyendo (1.78) en (1.77) se tiene:

$$\frac{\partial Q}{\partial a_i} = 2a_i + \lambda = 2a_i + \frac{2}{n} = 0 \tag{1.79}$$

entonces

$$a_i = \frac{1}{n} \tag{1.80}$$

además $\frac{\partial^2 Q}{\partial^2 a_i} 2 > 0$. Por lo tanto, considerando (1.80) se minimiza la expresión de la varianza del estimador

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y} \tag{1.81}$$

Por otro lado, el estimador por mínimos cuadrados es:

$$\begin{aligned} Z &= \sum_{i=1}^{n} \epsilon_i \\ &= \sum_{i=1}^{n} (y_i - \mu)^2 \end{aligned} \tag{1.82}$$

el estimador por mínimos cuadrados $\hat{\mu}$ de μ es aquel que cumple la siguiente condición

$$\begin{aligned} \frac{\partial Z}{\partial \mu} &= -2 \sum_{i=1}^{n} y_i - \hat{\mu} = 0 \\ &\Rightarrow \sum_{i=1}^{n} y_i - n\hat{\mu} = 0 \\ &\Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y} \end{aligned} \tag{1.83}$$

Con lo que se valida el teorema de Gauss Markov para el modelo de localización.

1.6. Modelo de regresión múltiple

Ahora se presentará la generalización del modelo de regresión lineal para más de una variable regresora. Consideremos un conjunto de n observaciones que dependen de p variables regresoras, la observación se indicará a través del subíndice i y las variables por j. El modelo de regresión lineal múltiple se puede expresar de la siguiente manera

$$y_i = \sum_{j=1}^{p} x_{ij} \beta_j + \epsilon_i \tag{1.84}$$

donde $x_{i1}, \ldots, x_{ip}$ representan las variables regresoras, y_i la variable de respuesta, $\beta_1, \ldots, \beta_p$ los parámetros desconocidos que deben ser estimados y ϵ_i el error aleatorio.

Si se expresa a los vectores p-dimensionales $(x_{i1}, \ldots, x_{ip})$ y $(\beta_1, \ldots, \beta_p)$ como $\boldsymbol{x_i}$ y $\boldsymbol{\beta}$, respectivamente, se puede reescribir el modelo de regresión lineal múltiple de la siguiente manera

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.85}$$

Los valores ajustados $\hat{\boldsymbol{y}}$ y los residuales $\boldsymbol{e}$ a partir del estimador $\hat{\boldsymbol{\beta}}$ se definen como

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \text{ y } \boldsymbol{e} = \hat{\boldsymbol{y}} - \boldsymbol{y}$$

Como se revisó anteriormente, el estimador por mínimos cuadrados de $\boldsymbol{\beta}$ se obtiene de:

$$\min_{\hat{\beta}} \sum_{i=1}^{n} e_i^2 \tag{1.86}$$

a partir de la ecuación anterior se derivan las ecuaciones normales

$$\boldsymbol{X'X}\hat{\boldsymbol{\beta}} = \boldsymbol{X'y} \tag{1.87}$$

De donde los estimadores por mínimos cuadrados están dados por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \tag{1.88}$$

Si se cumplen los supuestos del modelo de regresión lineal, $E(\epsilon_i) = 0$ y $\text{Var}(\epsilon_i) = \sigma^2$ para $i = 1, \ldots, n$, y que la matriz $\boldsymbol{X}$ es fija y de rango completo, es decir, que no existe colinealidad; los estimadores por mínimos cuadrados son insesgados.

La varianza de $\hat{\boldsymbol{\beta}}$ es:

$$\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2(\boldsymbol{X'X})^{-1} \tag{1.89}$$

donde la varianza del error se estima a partir de la varianza de los residuales

$$\sigma^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 \tag{1.90}$$

El vector de valores ajustados de la variable de respuesta $\hat{\boldsymbol{Y}}$ se pueden obtener a partir de

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{1.91}$$

1.7. Los outliers en el modelo de regresión

En las secciones anteriores se mostraron las propiedades que tiene los estimadores por mínimos cuadrados cuando se cumplen los supuestos del modelo de regresión lineal, pero ¿qué ocurre cuando alguno o algunos de estos supuestos no se cumplen?. En esta sección se mostrará de manera ilustrativa la afectación que sufren los estimadores por mínimos cuadrados cuando existe la presencia de datos atípicos en las observaciones, es decir, cuando no todas las observaciones cumplen los supuestos del modelo de regresión lineal o simplemente son producto de un error de medición.

Consideremos el modelo de regresión lineal simple

$$y_{=}\beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n \tag{1.92}$$

para tener la facilidad de representar las observación (x_i, y_i) en una gráfica de dispersión y visualizar la estructura de los datos. Para el modelo de regresión lineal múltiple con un número de variables $p > 4$ esto no es posible.

En la Figura (1.3) inciso a) se muestra la gráfica de dispersión de 5 observaciones que se encuentran perfectamente alineadas. Por lo tanto, al realizar la estimación de los estimadores por mínimos cuadrados se obtiene un ajuste perfecto de la recta de regresión $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Ahora, supongamos que una de las observaciones tiene un error de medición y esto origina que obtenga un valor fuera del patrón en y_4. En la Figura (1.3) inciso b) se puede observar que el punto (x_4, y_4) ya no forma parte del escenario ideal (indicado por el círculo). A este punto se le conoce como outlier vertical, como se puede notar este tipo de outliers tiene en este caso una fuerte influencia sobre la recta estimada pues origina que la nueva recta ajustada a los datos sea significativamente diferente a la obtenida en la Figura (1.3) inciso a). Este tipo de outliers han recibido mucha atención ya que al tener una diferencia importante en el valor de la variable de respuesta se origina un incremento, en términos absolutos, en el valor del residual y, dado que los estimadores por mínimos cuadrados minimizan la suma de cuadrados de residuales, estos se ven afectados fácilmente. En el caso de regresión lineal múltiple, cuando se tiene un gran número de variables explicativas y no es fácil visualizar los datos, es posible identificar este tipo de outliers a partir de los residuales o a través de sus gráficas, en la siguiente sección se presentarán algunos de los métodos de diagnóstico de outliers más conocidos.

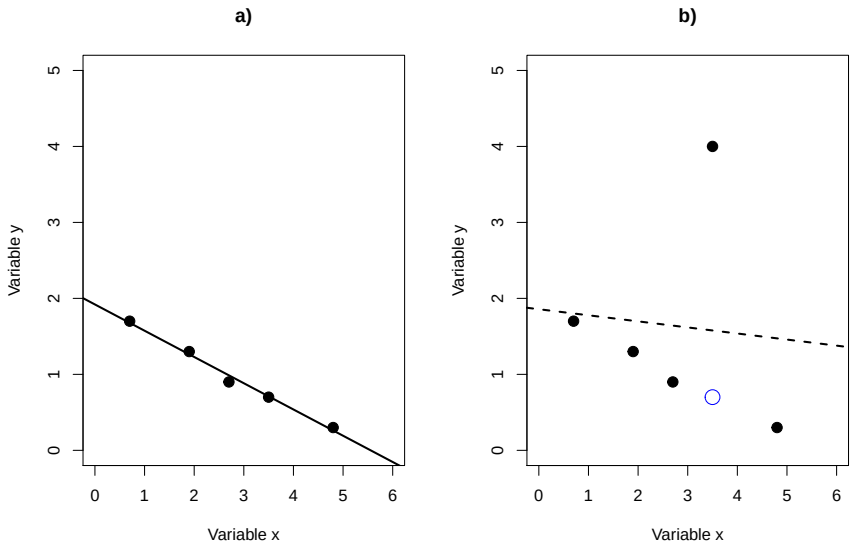


FIGURA 1.3: Outlier vertical

Por otro lado, usualmente se considera que el valor de $x_{i1}, \ldots, x_{ip}$ es un número fijo sin embargo, esto sólo ocurre cuando el experimento es diseñado. En realidad, en la mayoría de los casos se cuenta con un listado de variables entre las cuales se encuentra las variables regresoras y la variable de respuesta que se determinan en función del objetivo de la investigación. Esto significa que no sólo se puede tener errores en la variable de respuesta, de hecho, es más probable tener datos atípicos en las variables regresoras $x_{i1}, \ldots, x_{ip}$ pues generalmente $p > 1$.

En la Figura (1.4-a) se tienen nuevamente cinco puntos, que se ajustan perfectamente a la recta estimada. Ahora consideraremos que el error de medición ocurre en el valor de x_1, a este tipo de datos atípicos se los conoce como outliers horizontales y también tiene una fuerte influencia sobre los estimadores por mínimos cuadrados ya que jalan la recta estimada hacia ellos. La razón de que la observación (x_1, y_1) atrae a la recta de regresión se debe a que el valor del residual, e_1, se incrementa ya que x_1 se encuentra muy lejos de la posición original y ocasiona que se incremente el valor de $\sum_{i=1}^{5} e_i^2$ respecto a esa recta. Así, la recta original no puede ser elegida como la recta de mínimos cuadrados pues no minimiza la suma de cuadrados de residuales. En la Figura (1.4-b) se observa que la recta de mínimos cuadrados se inclina hacia la observación (x_1, y_1) con la finalidad de reducir el valor de e_1^2 sin importar que el resto de los residuales $e_2^2, \ldots, e_5^2$ incrementen.

A diferencia de los outliers verticales, es posible que algunos de los outliers horizontales no necesariamente sean outliers de regresión. Una observación (x_k, y_k) se considera outlier horizontal en función sólo del valor de x_k pero no significa necesariamente que tendrá un gran impacto sobre los estimadores por mínimos cuadrados pues el valor de y_k puede

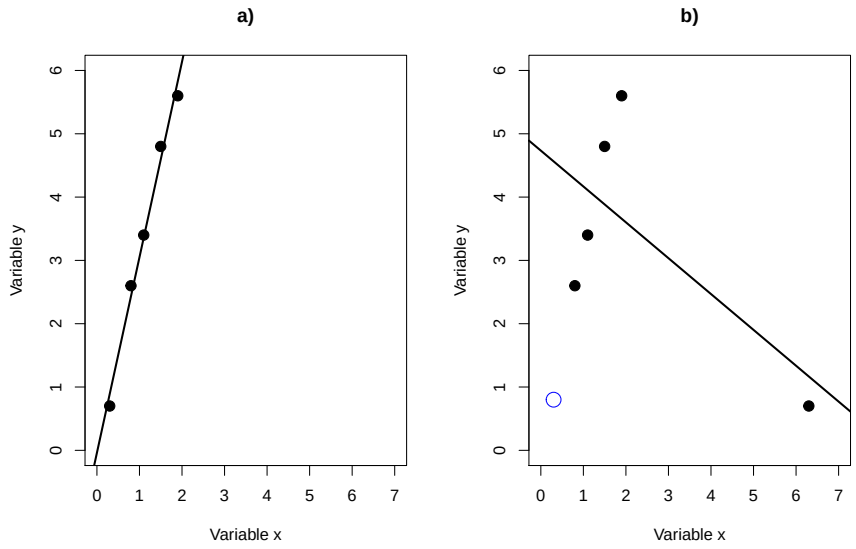


FIGURA 1.4: Outlier horizontal

hacer que se mantenga la tendencia del resto de las observaciones y así beneficiar el análisis de regresión al reducir las regiones de confianza.

La manera más comúnmente utilizada para hacer la identificación de los outliers es ajustar una recta de mínimos cuadrados con todas las observaciones, identificar las observaciones que tienes los residuales más grandes, eliminarlas o corregirlas y volver a realizar la estimación de mínimos cuadrados con las observaciones restantes. Aplicar este método puede tener algunos problemas. Como se mencionó anteriormente, existe la posibilidad de que alguna de las observaciones con un residual grande sea ocasionado por un fenómeno que no esté considerado en la investigación o si dentro de las observaciones se encuentran outliers horizontales que atraen a la recta de regresión, como en el Figura (1.4-b). Al identificar las observaciones con los residuales más grandes se pueden eliminar observaciones que sí cumplen el comportamiento del fenómeno estudiado y quedarse con la observación errónea.

1.7.1. Métodos clásicos para la detección de datos atípicos

Existen técnicas conocidas como "Diagnóstico de regresión" las cuales identifican a aquellas observaciones que no se comportan como la mayoría de los datos y tienen una gran influencia sobre los estimadores por mínimos cuadrados.

En general, estas técnicas se basan en realizar un ajuste por mínimos cuadrados y a partir de los residuales obtenidos, identificar a las observaciones con mayor influencia para corregirlas o excluirlas. Dado que los estimadores por mínimos cuadrados minimizan el

valor de la suma de cuadrados de residuales, una observación con una gran influencia y que no se ajuste a la mayoría de los datos puede jalar la recta de regresión ocasionando que observaciones que sí describen el fenómeno que estamos estudiando tengan un residual grande.

A continuación se presentarán algunas herramientas que ayudan a identificar a las observaciones con mayor influencia sobre la recta de regresión con el fin de que se dedique especial atención a estas observaciones.

1.7.1.1. Análisis de residuales

Consideremos nuevamente la definición de los residuales

$$e_i = y_i - \hat{y}_i \qquad i = 1, \ldots, n \tag{1.93}$$

donde y_i es una observación y $\hat{y}_i$ su valor estimado correspondiente. Como se puede considerar que los residuales son la desviación entre los datos y el ajuste, también es una medida de la variabilidad de la variable de respuesta que no explica el modelo de regresión. Por otro lado, también se puede interpretar a los residuales como los valores observados del término de error del modelo, por lo que toda violación a los supuestos del término de error debe reflejarse en los residuales.

En ocasiones es de utilidad trabajar con residuales estandarizados, los cuales son útiles para detectar observaciones atípicas.

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma}}} \qquad i = 1, \ldots, n \tag{1.94}$$

Los residuales estandarizados tienen media cero y varianza unitaria, por lo que, cuando se tiene un residual estandarizado grande ($|d_i| > 2.5$)[1] podría indicar de que se trata de un dato atípico.

1.7.1.2. Gráficas de residuales

El análisis gráfico de los residuales es un otra forma de validar la no violación de los supuestos del modelo de regresión lineal e identificar posibles observaciones atípicas.

Las gráficas básicas de los residuales son las siguientes:

[1] Ver Andersen Robert, "Modern Methods for Robust Regression", pág. 69

- Gráfica de probabilidad normal: a partir de esta gráfica se puede identificar cuando no se cumple el supuesto de normalidad, es decir, que los errores provengan de una distribución de colas más o menos pesadas que la distribución normal.
- Gráfica de residuales contra de los valores ajustado: con esta gráfica podemos análizar si no se viola el supuesto sobre la varianza constante de los errores.
- Gráfica de residuales contra cualquier variable regresora

1.7.1.3. Matriz sombrero

Una herramienta de diagnóstico comúnmente utilizada, son los elementos de la diagonal de la matriz de proyección por mínimos cuadrados o matriz sombrero.

Consideremos el modelo de regresión lineal múltiple

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.95}$$

donde $\boldsymbol{y}$ y $\boldsymbol{\epsilon}$ son vectores de $n \times 1$ que representan la variable de respuesta y el vector de errores, respectivamente; $\boldsymbol{\beta}$ el vector de parámetros no conocidos de $p \times 1$ y la matriz $\boldsymbol{X}$ de dimensión $n \times p$ los valores de las variables regresoras.

La matriz sombrero $\boldsymbol{H}$ se define como

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'} \tag{1.96}$$

bajo el supuesto de que $\boldsymbol{X}$ es de rango completo.

A la matriz $\boldsymbol{H}$ se le conoce como matriz sombrero porque pone un "sombrero" al vector $\boldsymbol{y}$, es decir, transforma el vector de observaciones $\boldsymbol{y}$ en su valor estimado utilizando los estimadores por mínimos cuadrados $\hat{\boldsymbol{y}}$.

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y} \tag{1.97}$$

A partir de la ecuación anterior se puede interpretar al elemento h_{ij} de $\boldsymbol{H}$ como el efecto que ejerce la observación y_i sobre $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$. En otras palabras, los elementos de la diagonal de $\boldsymbol{H}$ miden la influencia que tiene la observación y_i sobre el valor estimado $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$.

Es importante notar que

$$\operatorname{tr}(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ij} = p \tag{1.98}$$

Lo anterior se cumple debido a

$$\text{tr}(\boldsymbol{H}) = \text{tr}\left(\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}\right) = \text{tr}\left(\boldsymbol{X'X}(\boldsymbol{X'X})^{-1}\right) = \text{tr}\left(\boldsymbol{I}_p\right) = p^2$$

Además, la matriz $\boldsymbol{H}$ tiene las propiedades de ser idempotente $(\boldsymbol{HH} = \boldsymbol{H})$ y simétrica $(\boldsymbol{H'} = \boldsymbol{H})$ por lo que se cumplen la siguiente condición

$$h_{ii} = \sum_{i=1}^{n} h_{ij}h_{ji} = \sum_{i=1}^{n} h_{ij}h_{ij} \sum_{i=1}^{n} h_{ij}^2 \tag{1.99}$$

Por lo tanto

$$\begin{aligned} h_{ii} &= h_{ii}^2 + \sum_{i=1}^{n} h_{ij}^2 \sum_{i=1}^{n} h_{ij}^2 \\ \Rightarrow 1 &= h_{ii} + \frac{\sum_{j \neq i} h_{ij}^2}{h_{ii}} \geq h_{ii} \end{aligned} \tag{1.100}$$

Considerando las ecuaciones (1.99) y (1.100), la influencia que tiene una observación sobre su valor estimado se define en el rango $0 \leq h_{ii} \leq 1$. Por otro lado, a partir de la ecuación (1.100) podemos decir que en promedio cada h_{ii} influye p/n. La mayoría de los autores consideran que se debe poner especial atención en aquellas observaciones en las que $h_{ii} \geq 2p/n$.

A pesar de que la matriz sombrero es una de las herramientas más utilizadas para la detección de los outliers tiene dos graves problemas: puede ser afectada por la acción combinada de varios valores atípicos y sólo detecta outliers horizontales, no considera la posición de la observación en función de la varaible y, por lo que no identifica cuáles observaciones son buenas o malas.

1.7.1.4. Distancia de Cook

Otra medida muy popular para medir la influencia que tiene una observación sobre los estimadores por mínimos cuadrados es la "Distancia de Cook", la cual se define como

$$D_i = \frac{e_i}{s^2} \frac{h_{ii}}{p(1-h_{ii})^2} \tag{1.101}$$

[2] Si $\boldsymbol{A}$ y $\boldsymbol{B}$ son matrices cuadradas de n x n, entonces $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$

donde e_i es el i-ésimo residual, y s^2 el estimador insesgado del parámetro de escala calculado en la sección anterior. El primer término de la ecuación anterior representa el residual estandarizado de la observación i, por lo tanto mide la discrepancia de la observación. La segunda parte de la ecuación contiene el apalancamiento de la observación, se llama apalancamiento porque el dato atípico mueve la recta de regresión. Un valor grande en D_i sugiere que la observación i tiene un apalancamiento, presenta un residual estandarizado grande o ambos. Una gráfica que ayuda a complementar el análisis de los datos para localizar las observaciones atípicas es la "grafica de influencia", dicha gráfica compara el valor de D_i versus $h_{ii}/(1 - h_{ii})$.

En la práctica, a las observaciones que cumplan cualquiera de las siguientes condiciones se les debe de poner especial atención:

- $|e_i| > 2$
- $h_{ii} > 2p/n$
- $D_i > 8/(n - 2p)$

Capítulo 2

Regresión Robusta

Como se revisó en el capítulo anterior, el enfoque de regresión lineal basado en los estimadores por mínimos cuadrados es óptimo cuando se cumple que los errores son independientes e idénticamente distribuidos con distribución normal con media $\mu = 0$ y varianza constante σ^2. Sin embargo, estos estimadores son muy sensibles en presencia de observaciones atípicas, cuando la distribución normal en los errores es solamente aproximada, su sesgo y varianza pueden aumentar de manera considerable y ocasionar que los estimadores tomen valores erróneos.

Para solucionar este problema existen dos alternativas:

1. Detectar las observaciones atípicas previamente, utilizando alguna técnica de diagnóstico, modificarlas (revisarlas o suprimirlas) y luego efectuar el cálculo de los estimadores por mínimos cuadrados.
2. Utilizar métodos robustos, esto es, encontrar estimadores que:
    - Se vean poco afectados si una porción de las observaciones no provienen de una distribución normal.
    - Se comporten de forma parecida a los estimadores por mínimos cuadrados en caso de que no haya presencia de datos atípicos.

Considerando estas dos alternativas es necesario hacer la distinción entre los métodos de regresión robusta y los métodos de regresión resistentes.

Los métodos de regresión robusta son aquellos que utilizan la información de todas las observaciones pero le asignan un peso menor a las observaciones que son altamente inusuales. Estos métodos proporcionan estimadores que son poco afectados por los datos

atípicos pero, en caso de que los errores sigan una distribución normal, son estimadores relativamente eficientes respecto a los estimadores por mínimos cuadrados.

Por otro lado, los métodos de regresión resistente tienen como principal objetivo prevenir a aquellas observaciones inusuales que causan una drástica influencia en el valor de los estimadores de los coeficientes de regresión. Este tipo de métodos no sólo disminuyen el peso de las observaciones extremas sino que a menudo establecen criterios para la eliminación total de estas observaciones en el análisis. Como se mencionó anteriormente, el utilizar métodos de regresión resistentes puede ocasionar problemas ya que al considerar a los residuales como criterio para identificar las observaciones atípicas, pueden ser eliminadas observaciones "buenas" y trabajar con las observaciones atípicas si es que estas tienen una fuerte influencia sobre los estimadores por mínimos cuadrados, como en la figura (1.4-b).

En este trabajo se presentarán métodos que buscan ponderar la influencia de las observaciones inusuales o valores extremos, métodos de regresión robusta.

2.1. Robusticidad y Robustez

Un estimador es robusto debe de satisfacer las siguientes dos condiciones:

1. Si los datos u observaciones sufren un pequeño cambio, esto no debe causar un cambio sustancial en el estimador.
2. Si los datos u observaciones cumplen los supuestos del modelo de regresión lineal, el estimador se comporta de manera similar a los estimadores por mínimos cuadrados.

La primera condición establece la resistencia que debe tener el estimador a las observaciones inusuales, es decir, el estimador debe de proporcionar una buen modelo de estimación que refleje el comportamiento de la mayoría de los datos. Esta primera condición se puede considerar como la validación de la robusticidad. La segunda condición establece que el incumplimiento de la distribución hipotética del estimador tiene poco impacto en la precisión de éste. A esta condición se le considera como la eficiencia de la robustez.

Los métodos de regresión robusta se ocupan tanto de la validación de la robusticidad como de la eficiencia de la robustez. En cambio, los métodos de regresión resistente usualmente tienen poco interés en la eficiencia de la robustez.

En este trabajo se presentan métodos de regresión robusta que buscan ponderar la influencia de las observaciones inusuales.

2.2. Antecedentes importantes

Para poder evaluar la robustez de un estimador es necesario tomar en cuenta distintos conceptos: sesgo, consistencia, eficiencia, punto de ruptura (Breakdown Point) y función de influencia. A continuación se presentará la definición de cada uno de estos conceptos.

2.2.1. Sesgo

Consideremos una muestra aleatoria X de tamaño n. Sea $\hat{\theta}$ un estimador del parámetro θ con distribución de probabilidad P, el estimador $\hat{\theta}$ es insesgado si

$$E\left(\hat{\theta}\right) = \theta$$

es decir, el valor esperado de un estimador insesgado es el valor del parámetro desconocido. Se define el sesgo del estimador $\hat{\theta}$ como:

$$\text{Sesgo} = E\left[\hat{\theta} - \theta\right]$$

2.2.2. Error cuadrático medio

El error cuadrático medio (ECM) de un estimador $\hat{\theta}$ del parámetro θ se define como

$$E\left[(\hat{\theta} - \theta)^2\right]$$

El error cuadrático medio proporciona, en promedio, el error que se tiene al estimar el parámetro θ por medio de $\hat{\theta}$.

2.2.3. Consistencia

El insesgamiento es importante para determinar al mejor estimador, pero también es necesario considerar la consistencia del estimador. Un estimador $\hat{\theta}$ es consistente si este converge en probabilidad a θ cuando el tamaño de la muestra tiende a infinito. También se puede definir la consistencia de un estimador en términos del error cuadrático medio. Se dice que $\hat{\theta}$ es consistente si

$$\lim_{n\to\infty} \text{ECM}(\hat{\theta}) = 0$$

2.2.4. Punto de ruptura (Breakdown point)

El punto de ruptura (BDP, por sus siglas en inglés) es la medida global de resistencia de un estimador. Se define como la fracción o porcentaje más grande de datos discrepantes que un estimador puede tolerar sin producir resultados arbitrarios. Este concepto fue introducido por Hampel en 1971 con una definición asintótica para muestras infinitas y redefinido por Donoho y Huber en 1983 para muestras finitas.

Consideremos una muestra aleatoria $\boldsymbol{x}$ y sea $\hat{\theta}$ un estimador construido a partir de dicha muestra. Si realizamos la sustitución de m observaciones con valores arbitrarios $\boldsymbol{x}^*$, el máximo efecto que sufrirá el estimador $\hat{\theta}^*$ por la sustitución de las m observaciones se medirá de la siguiente manera:

$$\text{efecto}\left(m; \hat{\theta}, \boldsymbol{x}^*\right) = \sup_{\boldsymbol{x}} \|\hat{\theta}^* - \hat{\theta}\|$$

donde el supremo es sobre todo posible $\boldsymbol{x}^*$. Si el efecto $\left(m; \hat{\theta}, \boldsymbol{x}^*\right)$ es infinito, los m datos discrepantes tienen un gran impacto en el estimador $\hat{\theta}$. A partir de lo anterior, el punto de ruptura de un estimador $\hat{\theta}$ para una muestra finita se define como:

$$\text{BDP}\left(\hat{\theta}, \boldsymbol{x}\right) = \min\left\{\frac{m}{n} : \text{efecto}\left(m; \hat{\theta}, \boldsymbol{x}^*\right) \text{ es infinito}\right\}$$

donde n es el número de observaciones de la muestra y m es el número de observaciones discrepantes. Es claro que el punto de ruptura más alto que un estimador puede tener es del $50\,\%$.

2.2.5. Función de Influencia

El concepto de función de influencia fue dado originalmente por Hampel en 1974, esta función mide el impacto que tiene una sola observación atípica en la estimación de los coeficientes de regresión.

Supongamos que tenemos $(n-1)$ observaciones y agregamos una observación arbitraria x^*, llamemos $\hat{\theta}$ al estimador que se obtiene a partir de las primeras $(n-1)$ observaciones y $\hat{\theta}^*$ al obtenido cuando se agrega la nueva observación. La función de influencia (FI) para el estimador $\hat{\theta}$ se define como:

$$\text{FI}\left(\hat{\theta}, \boldsymbol{x}^*\right) = \lim_{n\to\infty} \frac{\hat{\theta} - \hat{\theta}^*}{1/n}$$

En palabras sencillas, la función influencia indica el cambio que sufre un estimador causado por la observación atípica x^*, estandarizado por la proporción de la contaminación.

Una función de influencia acotada es un atributo deseable para un estimador robusto, ya que la influencia de una observación en particular sólo puede ser tan alta como la cota. Por el contrario, una función de influencia no acotada permite que la influencia de las observaciones atípicas crezca independientemente de qué tan inusuales son dichas observaciones.

2.2.6. Eficiencia Relativa

Otro aspecto importante para poder entender la estimación robusta es la eficiencia. La eficiencia de un estimador está determinada por la relación que hay entre la varianza mínima posible y la varianza del estimador. Cuando se cumple que tiene la menor varianza posible, es decir cuando la relación es igual a 1, tenemos un estimador eficiente. Un estimador es asintóticamente eficiente si es eficiente cuando el tamaño de la muestra tiende a infinito. En general, un estimador es considerado eficiente cuando la relación entre las varianzas es relativamente cercana a 1.

Para la mayoría de los diferentes tipos de estimadores existe un estimador que tiene la menor varianza bajo algunas hipótesis, por lo que se puede utilizar este estimador para comparar la eficiencia del resto de estimadores.

Supongamos que tenemos dos estimadores, $\hat{\theta}_1$ y $\hat{\theta}_2$, del parámetro θ donde $\hat{\theta}_1$ tiene la máxima eficiencia y un error cuadrático medio pequeño. La eficiencia relativa de $\hat{\theta}_2$ está definida como

$$\text{eficiencia}\left(\hat{\theta}_1, \hat{\theta}_2\right) = \frac{E\left[\left(\hat{\theta}_2 - \theta\right)^2\right]}{E\left[\left(\hat{\theta}_1 - \theta\right)^2\right]}$$

En el modelo de regresión lineal, los estimadores mínimos por cuadrados son los estimadores insesgados que tienen la mayor eficiencia bajo el supuesto de normalidad. Esto significa que, bajo las hipótesis de que los errores tienen una varianza constante y siguen una distribución normal, ningún estimador robusto es más eficiente que los estimadores por mínimos cuadrados, sin embargo, existen estimadores que están muy cercanos a la eficiencia y que además, en caso de que no se cumpla el supuesto de normalidad, tienen una alta resistencia a las observaciones atípicas.

Capítulo 3

Localización y escala

Aunque existen diferentes métodos de regresión, en general todos tienen como objetivo predecir el valor de una variable dependiente a partir de una o varias variables predictoras ó regresoras tomando en cuenta una medida de localización y escala de la variable de respuesta.

Consideremos el modelo de localización

$$y_i = \mu + \epsilon_i \tag{3.1}$$

donde μ es el valor cierto de un parámetro desconocido y ϵ_i el término de error aleatorio que cumple las hipótesis distribucionales del modelo de regresión lineal simple. Dado que las observaciones $y_1, \ldots, y_n$ son una función de los errores, también son independientes con función de distribución común

$$F(y) = F_0(y - \mu) \tag{3.2}$$

Si F_0 es la función de distribución normal con media $\mu = 0$ y varianza σ^2 constante, es decir, no hay presencia de datos atípicos; los estimadores por mínimos cuadrado son los que tienen la menor varianza ya que además son los estimadores máximo verosímiles.

En la práctica, la normalidad exacta difícilmente se satisface, por lo que se dice que F_0 es aproximadamente una normal. Para formalizar la idea de normalidad aproximada, imaginaremos que una proporción $1-\epsilon$ de las observaciones son generadas por el modelo normal mientras que una porción ϵ es generada por un mecanismo no conocido, donde $0 < \epsilon < 1$. Esto se puede representar formalmente de la siguiente manera

$$F_0 = (1 - \epsilon)G + \epsilon H \tag{3.3}$$

donde G es una función de distribución $N(\mu, \sigma^2)$ y H es cualquier otra función de distribución. La función de distribución F_0 es llamada función de distribución normal contaminada. En el caso en el que las funciones de distribuciónes G y H son normales pero con diferentes parámetros, a la función de distribución F_0 se le conoce como normal mezclada.

La función de densidad de F_0 está dada por

$$f_0 = (1 - \epsilon)g + \epsilon h \tag{3.4}$$

donde g y h son las funciones de densidad de G y H respectivamente.

3.1. Medidas de Localización

Una medida de localización es una cantidad que caracteriza una posición en una distribución. Usualmente las medidas de localización más utilizadas son las de tendencia central; los estimadores por mínimos cuadrados, por ejemplo, utilizan a la media muestral de las variables independientes como medida de localización. Otras medidas de localización que se pueden considerar son los cuantiles o la mediana.

Definición: *Sea X una variable aleatoria con función de distribución F_0. Un estimador $T(X)$ es una medida de localización de F_0 si para cualesquiera constantes a y b se cumplen las siguientes condiciones:*

1. $T(X + a) = T(X) + a$
2. $T(-X) = -T(X)$
3. *Si* $X \geq 0$ *entonces* $T(X) \geq 0$
4. $T(bX) = bT(X)$

La primera condición establece que una medida de localización debe de ser equivariante (se utilizará el término "equivariante" para referirnos a estadísticas que se transforman adecuadamente, y el término "invariante" para aquellas que permanecen sin cambios) con respecto a la posición, es decir, si un valor es sumado a todos los valores de la variable

X entonces la medida de localización se incrementará en la misma proporción. Las condiciones 2 y 4 establecen que si todos los valores de la variable X son multiplicados por un cierto valor, entonces la medida de localización reflejará dicha alteración, es decir, la medida tiene una escala equivariante. Por otro lado, la condición 3 establece que la medida de localización debe de tener el mismo rango que la variable X.

A continuación se presentan algunas de las medidas de localización más utilizadas.

3.1.1. La Media

La media es la medida de localización más conocida ya que es la medida utilizada en el modelo de regresión lineal. Si consideramos el modelo de localización y además suponemos que la distribución de la que provienen las observaciones es normal, entonces la media simple $\bar{y}$ es el estimador de μ con la mayor eficiencia. Por lo tanto, las observaciones se pueden estimar de la siguiente manera

$$\hat{y}_i = \bar{y} \qquad i = 1, \ldots, n \tag{3.5}$$

Aunque la media es la medida de tendencia central más utilizada, no es una medida de localización robusta ya que en presencia de observaciones atípicas se ve seriamente afectada.

En el modelo de regresión lineal simple, la media minimiza a la función de mínimos cuadrados (función objetivo)

$$\text{mín} \sum_{i=1)}^{n} (y_i - \hat{\mu})^2 \tag{3.6}$$

La función de influencia de un estimador se obtiene al derivar la función objetivo respecto a la variable de respuesta y. En el caso de los estimadores por mínimos cuadrados su función de influencia se define como

$$FI_{\bar{x}}(Y) = 2y \tag{3.7}$$

Como se mencionó anteriormente, una función de influencia no acotada es un mal atributo para un estimador ya que en presencia de valores atípicos la perturbación que puede sufrir el estimador sería tan grande como se quisiera. Como se muestra en la ecuación (3.7) la función de influencia para la media no es acotada, por lo que no puede ser un estimador robusto.

3.1.2. La Media Ajustada

Un enfoque más robusto de una medida de localización es la media ajustada, en la cual se descartan algunos de los valores más pequeños y más grandes con la intención de reducir el impacto de los valores extremos.

Sea $y_1, \ldots, y_n$ una muestra aleatoria ordenada, se define el porcentaje de datos que se desea excluir como $\alpha \in \left[0, \frac{1}{2}\right)$. Se obtiene la media a partir de los datos sin considerar los m valores más pequeños y los m valores más grandes, con $m = [\alpha(n-1)]$, donde $[\cdot]$ representa la parte entera de $\alpha(n-1)$. Por lo tanto, se define la media ajustada como

$$\bar{y}_\alpha = \frac{1}{(n-2m)} \sum_{i=m+1}^{n-m} y_i \tag{3.8}$$

De esta manera se lleva a cabo una elección objetiva de las observaciones y no se excluyen observaciones arbitrariamente, lo que hace que el resultado sea una función de todas las observaciones, incluso de aquellas que no están consideradas en la suma. Los casos límite, $\alpha = 0$ y $\alpha = 1/2$, corresponden a la media simple y a la mediana, respectivamente.

El porcentaje o monto a cortar es el que determina el punto de ruptura de la media ajustada, es decir, BDP=α. Leger y Romano en 1990, sugirieron realizar el cálculo de la media ajustada para un α igual a 0, 0.1 y 0.2; y elegir aquella con la que obtenemos el menor error estándar.

La función de influencia de la media ajustada, al contrario de la media simple, sí es acotada y, al igual que el BDP, se determina en función del porcentaje a cortar.

La eficiencia de la media ajustada depende de la distribución de las observaciones, si la distribución tiene valores extremos, excluir las observaciones más grandes y las más pequeñas puede aumentar la eficiencia dado que la variabilidad de las observaciones se reduce sin embargo, si la distribución es normal y muchas observaciones son excluidas, entonces la precisión de la media ajustada disminuye.

3.1.3. La Mediana

La mediana es el valor de la variable y que se encuentra a la mitad de los datos cuando estos se ordenan de manera ascendente.

Para obtener la mediana, primero se deben ordenar los datos de manera ascendente

$$y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$$

Entonces, la mediana se define como

$$M = \begin{cases} M = y_{(n+2)/2} & \text{cuando } n \text{ es par} \\ M = 0.5 * y_{n/2} + 0.5 * y_{n/2+1} & \text{cuando } n \text{ es impar} \end{cases} \tag{3.9}$$

La mediana es el valor $\hat{\mu}$ que minimiza a la función objetivo de valor absoluto

$$\text{mín} \sum_{i=1}^{n} |y_i - \hat{\mu}| \tag{3.10}$$

La función de influencia para la mediana es

$$FI_M(y) = \begin{cases} 1 & \text{cuando } y > 0 \\ 0 & \text{cuando } y = 0 \\ -1 & \text{cuando } y < 0 \end{cases} \tag{3.11}$$

A diferencia de la media, la función de influencia para la mediana sí es acotada, de hecho, es uno de los estimadores más resistentes a valores atípicos ya que alcanza el máximo punto de ruptura posible, BDP=0.5. Sin embargo, la mediana tiene la desventaja de ser poco eficiente cuando se cumple el supuesto de normalidad, lo que ocasiona que no sea un buen estimador robusto.

3.2. Medidas de dispersión

Definición: *Sea X una variable aleatoria. Una medida de dispersión se define como cualquier función $\phi(X)$ no negativa, que satisface las siguientes condiciones para cualesquiera constantes $a > 0$ y b:*

1. $\phi(aX) = a\phi(X)$
2. $\phi(X + b) = \phi(X) + b$
3. $\phi(-X) = \phi(X)$

La primera condición establece que una medida de dispersión debe de ser equivariante con respecto a la escala, es decir, si todos los valores de X son alterados por un valor en particular, entonces la medida será alterada en la misma proporción. La segunda

condición establece que la medida de dispersión es invariante respecto a la localización, es decir, si un valor es sumado a todos los valores de X esto no debe afectar al estimador. Por último, el inciso c) establece que una medida de dispersión también debe de ser invariante respecto al cambio de signo.

A continuación se presentarán algunos de los estimadores de dispersión más utilizados y se analizará cuánto se afectan en presencia de observaciones atípicas.

3.2.1. Desviación Estándar

La medida de escala más utilizada es la desviación estándar S, que se define como

$$S_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}} \tag{3.12}$$

Al igual que ocurre con la media, en el caso de que las observaciones y_i provengan de una distribución normal, la desviación estándar es la medida de escala que tiene la mayor eficiencia, es decir, es el estimador máximo verosímil. Sin embargo, en caso de que la distribución de las observaciones sea de colas pesadas y se tenga la presencia de observaciones atípicas, la desviación estándar se ve seriamente afectada ya que posee una función de influencia no acotada y un punto de ruptura igual a 0.

3.2.2. La desviación media con respecto a la media o a la mediana

La desviación media con respecto a la media, comúnmente conocida como desviación media, se obtiene a partir del valor absoluto de la diferencia de la variable y y su media, ponderada por el tamaño de la muestra,

$$DM = \frac{\sum_{i=1}^{n} |y_i - \bar{y}|}{n} \tag{3.13}$$

Por otro lado, la desviación media con respecto a la mediana, como es de suponerse, se obtiene de la diferencia de la variable y respecto a su mediana

$$DM = \frac{\sum_{i=1}^{n} |y_i - M_y|}{n} \tag{3.14}$$

donde M_y representa la mediana de la variable y.

Aunque ambas desviaciones son relativamente más eficientes que la desviación estándar cuando existen observaciones atípicas, ambas tienen también el problema de contar con

una función de influencia no acotada y un punto de ruptura igual a 0, lo que hace que actualmente sean consideradas obsoletas. Sin embargo, estas medidas de escala jugaron un papel muy importante en las primeras técnicas de regresión robusta y en el desarrollo de medidas de escala más robustas.

3.2.3. La mediana del valor absoluto de la desviación con respecto a la mediana

La mediana del valor absoluto de la desviación con respecto a la mediana (MAD), se define como

$$\text{MAD} = \text{mediana}|y_i - M_y| \tag{3.15}$$

Aunque está basada en la desviación respecto a la mediana, la MAD es mucho más resistente a las observaciones atípicas que la medidas de despersión antes presentadas. La MAD tiene las propiedades deseadas para un estimador robusto, tiene una función de influencia acotada y logra el punto de ruptura más alto posible, BPD=0.5; lo que hace que sea una de las medidas de escala más utilizadas en los métodos de regresión robusta.

3.3. M-estimadores

En esta sección se presenta una familia de estimadores que generalizan la idea de la máxima verosimilitud para las medidas de localización y de escala con el fin de poder obtener estimadores más robustos, que tienen a la media y a la mediana como casos límite. Los M-estimadores fueron propuestos por Huber en 1964 y fueron los primeros estimadores en considerar un balance entre la eficiencia y la resistencia.

3.3.1. Generalización de la máxima verosimilitud

Consideremos el modelo de localización nuevamente

$$y_i = \mu + \epsilon_i \qquad i = 1, \ldots, n$$

donde los ϵ_i son variables aleatorias i.i.d. con función de distribución F_0 y μ un parámetro de localización desconocido. Se supone que $f_0 = F_0'$ es la función de densidad

correspondiente. La función de densidad conjunta de las observaciones y_i (función de verosimilitud) es

$$L(\mathbf{y};\mu) = \prod_{i=1}^{n} f_0(y_i - \mu) \tag{3.16}$$

El estimador máximo verosímil de μ es el valor $\hat{\mu}$ que maximiza a $L(\mathbf{y};\mu)$, o equivalentemente, minimiza la función $\rho(\mathbf{y};\mu)$

$$-\log(L(\mathbf{y};\mu)) = \min \sum_{i=1}^{n} \rho(y_i - \hat{\mu}) \tag{3.17}$$

donde $\rho = -\log(f_0)$.

En el caso en el que f_0 es la distribución normal estándar

$$f_0(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \tag{3.18}$$

aplicando la función logaritmo, obtenemos la función objetivo de los estimadores por mínimos cuadrados, $\rho(y) = y^2$.

Por otro lado, si f_0 es la distribución doble exponencial

$$f_0(y) = \frac{1}{2} e^{-|y|} \tag{3.19}$$

al aplicar el logaritmo obtenemos la función objetivo $\rho(y) = -|y|$ que corresponde a la mediana.

En general, un M-estimador se define como aquel estimador que minimiza una *ρ-función* o función de los residuales (función objetivo).

Definición: *Sin pérdida de generalidad, definiremos una ρ-función o función objetivo como una función que satisface las siguientes condiciones:*

a) $\rho(y)$ *es una función no decreciente de* $|y|$

b) $\rho(0) = 0$

c) $\rho(y)$ *es creciente para* $y > 0$*, tal que* $\rho(y) < \rho(\infty)$

d) Si $\rho(y)$ *es una función acotada entonces supondremos que* $\rho(\infty) = 1$

3.3.2. M-estimadores de Localización

Dada una función objetivo, $\rho(y)$, que mide la distancia de la variable Y a un estimador de la posición μ. Un M-estimador de localización se define como el valor $\hat{\mu}$ tal que:

$$\min \sum_{i=1}^{n} \rho\left(y_i - \hat{\mu}\right) \tag{3.20}$$

Es importante señalar que estos estimadores no necesariamente deben ser máximo verosímiles para alguna distribución específica.

Equivalentemente, si la *ρ-función* es diferenciable, un M-estimador de localización se define, como el valor $\hat{\mu}$ que es solución de la siguiente ecuación

$$\sum_{i=1}^{n} \psi\left(y_i - \hat{\mu}\right) = 0 \tag{3.21}$$

con $\psi(y) = \rho'(y)$.

La ecuación (3.21) siempre tiene solución si la función $\psi(y)$ es no decreciente. Por otro lado, si la función $\psi(y)$ es estrictamente creciente y continua, la solución es única.

En general, los M-estimadores pueden tener diferentes formas y propiedades dependiendo de la elección de la función $\rho(y)$ o, equivalentemente, de $\psi(y)$. La elección de la función objetivo se lleva a cabo considerando que el estimador resultante debe reducir el peso de las observaciones atípicas. Si $\psi(y)$ es no acotada entonces el punto de ruptura del estimador será asintóticamente 0 cuando el tamaño de la muestra tienda a infino. En cambio, si $\psi(y)$ es impar y acotada entonces el punto de ruptura será asintóticamente 0.5.

A continuación se mostrarán los M-estimadores de Huber y Biponderados, los cuales jugaron un papel muy destacado en el desarrollo de la teoría de regresión robusta.

3.3.2.1. Estimadores de Huber

Un tipo de funciones $\rho(y)$ y $\psi(y)$ con propiedades importantes son las introducidas por Peter Huber en 1964. La familia de funciones de Huber se definen de la siguiente manera

$$\rho_k(y) = \begin{cases} y^2 & \text{si } |y| \leq k \\ 2k|y| - k^2 & \text{si } |y| > k \end{cases} \tag{3.22}$$

Como se puede observar, la función $\rho_k(y)$ es cuadrática en la región central y aumenta linealmente para valores de $|y|$ mayores a k. Los M-estimadores de Huber correspondientes a los casos límite, cuando $k \to \infty$ y $k \to 0$, son la media y la mediana respectivamente.

Al calcular la derivada de la función (3.22) obtenemos la función de influencia $\psi_k(y)$ de los M-estimadores de Huber

$$\psi_k(y) = \begin{cases} k & \text{si } y > k \\ y & \text{si } y \leq |k| \\ -k & \text{si } y < k \end{cases} \tag{3.23}$$

La función de pesos para los M-estimadores de Huber, que indica el peso que se le otorga a cada una de las observaciones se define como:

$$W_k(y) = \begin{cases} 1 & \text{si } |y| \leq k \\ \dfrac{k}{|y|} & \text{si } |y| > k \end{cases} \tag{3.24}$$

En el centro, la distribución la función de pesos de Huber se comporta de la misma manera que la media, es decir, otorga el mismo peso a todas las observaciones. En los extremos se comporta parecido a la mediana, otorga pesos más pequeños conforme la observación es más atípica.

El valor de k para los estimadores de Huber es elegido en función de la eficiencia que se quiere lograr respecto a la distribución normal. Huber demostró que con k=1.345 se obtiene buena resistencia a los valores extremos y una eficiencia aproximada de 95 %.

3.3.2.2. Estimadores biponderados (Estimadores de Tukey)

Si se quiere una función objetivo que sea aún más resistente a los datos atípicos, es común que se elijan funciones de la familia biponderada, llamadas también re-descendientes. Este tipo de funciones fueron desarrolladas por John W. Tukey en 1974. La función $\rho(y)$ de la familia de funciones biponderadas se define como

$$\rho(y) = \begin{cases} 1 - \left[1 - \left(\dfrac{y}{k}\right)^2\right]^3 & \text{si } |y| \leq k \\ 1 & \text{si } |y| > k \end{cases} \tag{3.25}$$

Al calcular la derivada de la función anterior obtenemos la función $\psi(y)$, $\rho'(y) = (6\psi(y))/k^2$, donde

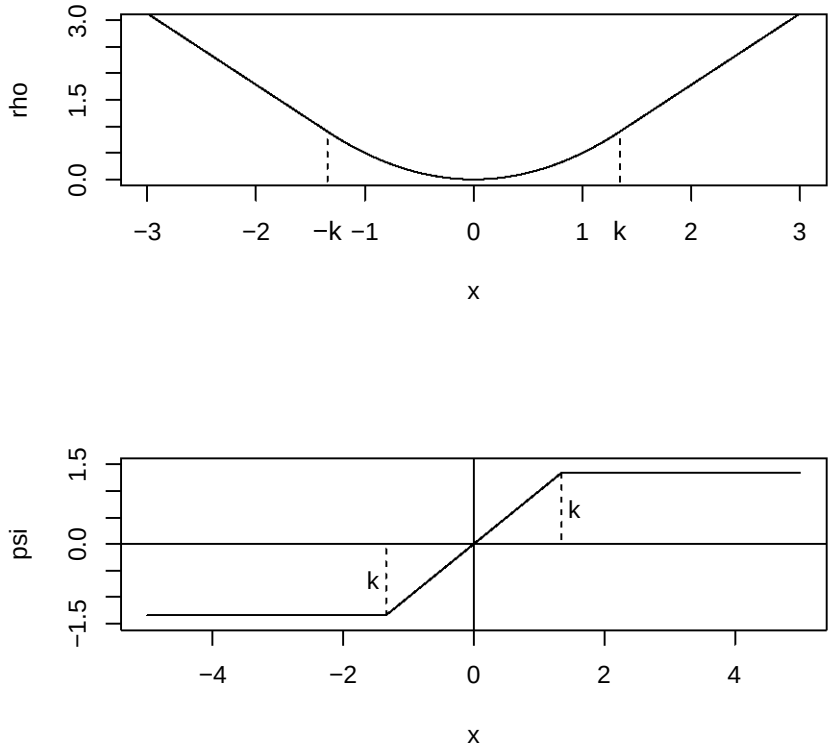


FIGURA 3.1: Funciones de Huber $\rho(y)$ y $\psi(y)$

$$\psi(y) = \begin{cases} y\left[1 - \left(\frac{y}{k}\right)^2\right]^2 & \text{si } |y| \leq k \\ 0 & \text{si } |y| > k \end{cases} \tag{3.26}$$

Es importante notar que la función de influencia para los estimadores biponderados tiende rápidamente a cero.

La función de pesos para esta familia de funciones se define como

$$W(y) = \begin{cases} \left[1 - \left(\frac{y}{k}\right)^2\right]^2 & \text{si } |y| \leq k \\ 0 & \text{si } |y| > k \end{cases} \tag{3.27}$$

Las funciones de pesos de los M-estimadores de Huber y de Tukey se comportan similarmente. Para valores de $|y| > k$ el estimador biponderado otorga un peso igual a cero y sólo el centro recibe un peso de uno. En cambio, el estimador de Huber no otorga peso de cero a ninguna observación y una proporción más significativa de observaciones reciben el peso de 1.

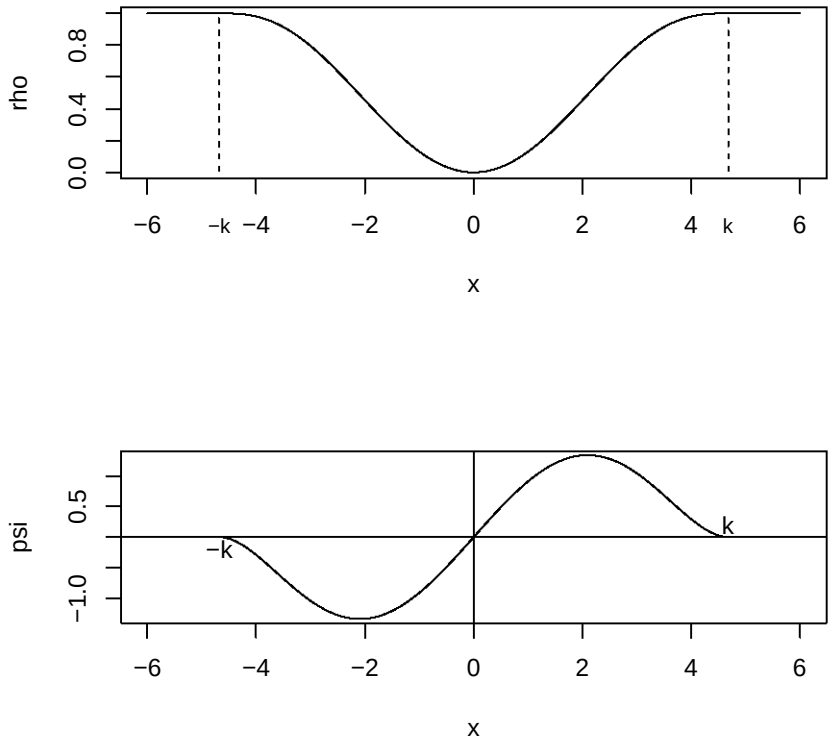


FIGURA 3.2: Funciones redescendientes $\rho(y)$ y $\psi(y)$

Al igual que los M-estimadores de Huber, el valor de la constante k es elegido en función de la eficiencia que se quiere lograr respecto a los estimadores por mínimos cuadrados. Con k=4.685, los M-Estimadores biponderados tiene una eficiencia de 95 % en caso de que los datos provengan de una distribución normal.

3.3.3. M-estimadores de dispersión

Consideremos que las observaciones y_i satisfacen el modelo multiplicativo, definido como:

$$y_i = \sigma \epsilon_i \tag{3.28}$$

donde las ϵ_i son variables aleatorias i.i.d. con función de distribución F_0 y σ es un parámetro desconocido mayor a 0. Si f_0 es la función de densidad de F_0, entonces las y_i tienen la siguiente función de densidad

$$\frac{1}{\sigma} f_0 \left(\frac{y_i}{\sigma} \right) \tag{3.29}$$

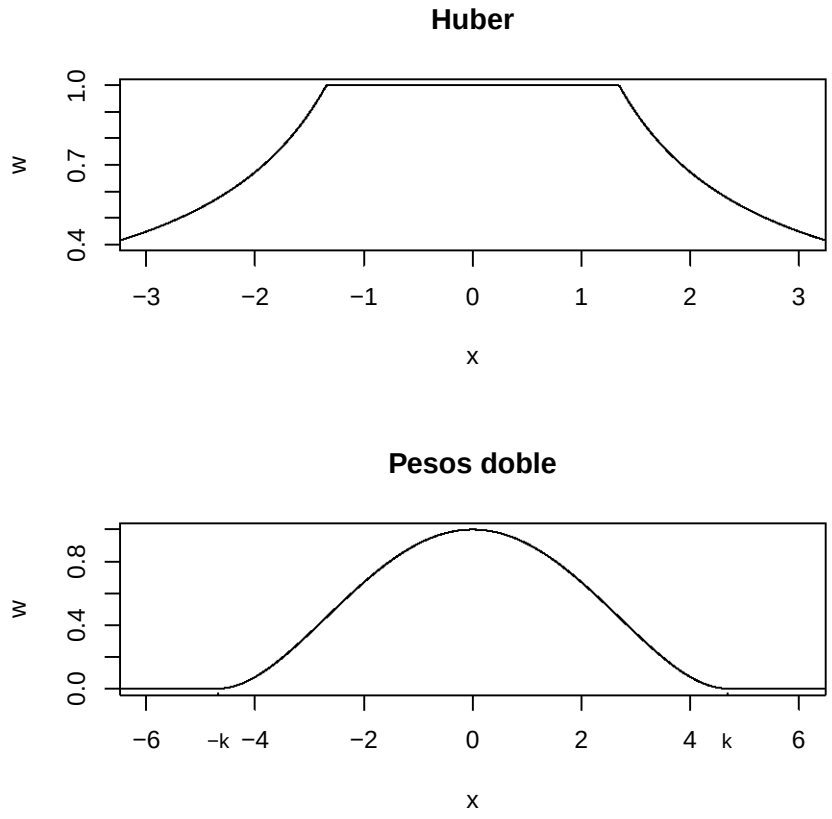


FIGURA 3.3: Funciones de peso de los M-estimadores de Huber y Tukey

El estimador máximo verosímil de σ, para el modelo multiplicativo, es el valor de $\hat{\sigma}$ que maximiza la siguiente expresión

$$\max \left[\frac{1}{\hat{\sigma}^n} \prod_{i=1}^{n} f_0 \left(\frac{y_i}{\hat{\sigma}} \right) \right] \tag{3.30}$$

o bien, aplicando el logaritmo y derivando (3.30) con respecto al parámetro σ, $\hat{\sigma}$ satisface la siguiente igualdad

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{y_i}{\hat{\sigma}} \right) = 1 \tag{3.31}$$

donde $\rho(t) = t\psi(t)$ con $\psi = -f_0'/f_0$

Si F_0 es la distribución $N(0,1)$, entonces $\rho(t) = t^2$ y, por lo tanto, la función de verosimilitud es

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\sigma}\right)^2 = 1 \tag{3.32}$$

El valor $\hat{\sigma}$ que satisface la ecuación (3.32) es

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=n}^{n} y_i^2}{n}} \tag{3.33}$$

Considerando la construcción de los estimadores máximo verosímiles para el parámetro de dispersión, los M-estimadores de dispersión se definen como el valor de $\hat{\sigma}$ que satisface la siguiente ecuación

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i}{\hat{\sigma}}\right) = \delta \tag{3.34}$$

donde ρ es una función objetivo que debe de cumplir las siguientes condiciones:

i. $\rho(t) = \rho(-t)$

ii. ρ es no decreciente

iii. ρ es continua excepto en un número finito de puntos

Para que la ecuación (3.34) tenga solución se debe cumplir que $0 < \delta < \rho(\infty)$, generalmente δ es igual a $E_{f_0}[\rho(x)]$. Cuando la función ρ es acotada, se supone que $\rho(\infty) = 1$.

Los M-estimadores de dispersión son equivariantes respecto a la escala ya que cumplen la condición

$$\hat{\sigma}(cy_1, \ldots, y_n) = c\hat{\sigma}(y_1, \ldots, y_n)$$

para cualquier $c > 0$. Además, si ρ es una función par, para cualquier constante c se cumple que

$$\hat{\sigma}(cy) = |c|\hat{\sigma}(y)$$

El grado de robustez o punto de ruptura, en el caso de los estimadores de escala, busca que el estimador de escala se mantenga alejado del cero ("implosión") o del infinito ("explosión"). Al igual que en los estimadores de localización, las funciones de la familia

de Huber o de Turkey son usualmente elegidas como función objetivo para obtener los M-estimadores de escala.

3.3.3.1. M-estimador de Escala Biponderado

El M-estimador de escala biponderado (biweight midvariance) es un estimador eficiente y altamente resistente a datos atípicos, alcanzando un punto de ruptura de 0.5. Este se obtiene utilizando como función objetivo la función de Tukey o Biponderada y las siguientes condiciones:

i. $k = 1$

ii. $\rho(y) = \min\left(1 - (1 - y^2)^3, 1\right)$

iii. $\delta = 0.5$

3.3.3.2. Estimadores invariantes bajo posición y escala

Los estimadores de dispersión juegan un papel muy importante en el desarrollo de los estimadores de localización. Como vimos en la sección anterior, los estimadores de localización se obtienen a partir de la minimización de la función objetivo lo que ocasiona que sean invariantes respecto a la posición pero no respecto a la escala.

En el siguiente ejemplo se muestra la afectación que sufre un estimador de localización cuando no es invariante respecto a la escala. El ejercicio consiste en obtener el estimador $\hat{\beta}$ a partir de un conjunto de observaciones, después modificar la variable de respuesta triplicando la distancia entre las observaciones y la recta ajustada; y por último, generar nuevamente el estimador $\hat{\beta}$ considerando las nuevas observaciones; si el estimador $\hat{\beta}$ es invariante respecto a la escala no debería sufrir ningún cambio.

Consideremos las siguientes observaciones

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
x_i	1	2	3	4	5	6	7	8
y_i	1.6	1.5	2	2	2.5	2.2	3	0

Utilizando el método iterativo de mínimos cuadrados ponderados (IRLS por sus siglas en inglés) obtenemos el M-estimador de Huber $\hat{\beta} = \begin{pmatrix} 1.8193 \\ 0.0257 \end{pmatrix}$[1].

[1] Estimación: anexo I.I

A partir del estimador $\hat{\beta}$ se obtienen los residuales e_i y se construye la nueva variable de respuesta y^* como $y^* = y_i + 3e_i$:

i	1	2	3	4	5	6	7	8
$\hat{y}_i$	1.845	1.8707	1.8964	1.9221	1.9478	1.9735	1.9993	2.025
e_i	-0.245	-0.3707	0.1036	0.0779	0.5522	0.2265	1.0007	-2.025
y_i^*	0.865	0.3878	2.3107	2.2336	4.1565	2.8794	6.0022	-6.0749

Considerando la nueva variable y^*, se calcula nuevamente el M-estimador de Huber $\hat{\beta} = \begin{pmatrix} 0.2779 \\ 0.5316 \end{pmatrix}$.

Como se puede observar, el estimador $\hat{\beta}$ es diferente a $\hat{\beta}^*$ lo cual nos muestra que los M-estimadores no son invariantes respecto a la escala. En las siguientes secciones se desarrollarán estimadores de localización robustos que sean invariantes respecto a la posición y escala.

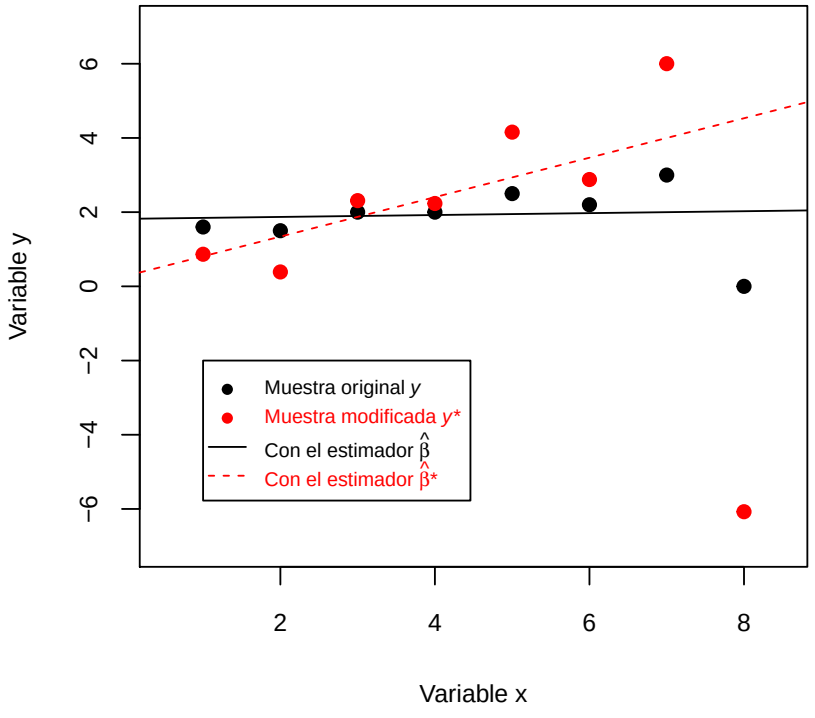


FIGURA 3.4: Los M-estimadores no son invariantes respecto a la escala

3.4. M-estimadores de localización con parámetro de dispersión desconocido

El problema que tienen los M-estimadores de localización de no ser equivariantes respecto a la escala se resuelve estandarizando el valor de los residuales utilizando de un estimador de su escala. Considerando lo anterior, el M-estimador de localización equivariante respecto a la escala se define como el valor $\hat{\mu}$ que minimiza la siguiente expresión:

$$\min \sum_{i=1}^{n} \rho \left(\frac{y_i - \hat{\mu}}{\sigma} \right), \tag{3.35}$$

donde σ es un parámetro de dispersión. En general σ no es conocido y debe ser estimado. Por lo tanto, para construir un M-estimador de localización equivariante respecto a la escala se debe de estimar el valor de σ de manera previa o simultáneamente.

3.4.0.3. Cuando se estima σ previamente

Para obtener el M-estimador de localización con escala equivariante, se debe hallar el valor $\hat{\mu}$ tal que minimice la siguiente expresión

$$\min \sum_{i=1}^{n} \rho \left(\frac{y_i - \hat{\mu}}{\hat{\sigma}} \right), \tag{3.36}$$

donde $\hat{\sigma}$ es un estimador de dispersión previamente calculado. Dado que $\hat{\sigma}$ no depende del parámetro μ, el valor de $\hat{\mu}$ se puede obtener, equivalentemente, a partir de la siguiente ecuación

$$\sum_{i=1}^{n} \psi \left(\frac{y_i - \hat{\mu}}{\hat{\sigma}} \right) = 0 \tag{3.37}$$

donde $\psi(y) = \rho'(y)$. Es claro que $\hat{\sigma}$ debe de ser un estimador robusto para que las observaciones atípicas no afecten las propiedades robustas del M-estimador de localización. Generalmente el estimador de escala utilizado es una variación de la medida de dispersión MAD con el fin de lograr un estimador eficiente en caso de que las observaciones provengan de una distribución normal. El estimador $MADN$ normaliza el estimador MAD

$$MADN = \frac{MAD}{0.675} \tag{3.38}$$

El factor 6.745 se obtiene considerando la siguiente equivalencia en caso de que se cumpla el supuesto de normalidad

$$MAD = \Phi^{-1}(0.75) * \sigma \approx 0.675 * \sigma$$

3.4.0.4. Cuando se estiman μ y σ simultáneamente

Para la estimación simultánea de los parámetros μ y σ es necesario considerar un modelo de localización-dispersión con dos parámetros desconocidos

$$y_i = \mu + \sigma\epsilon_i \tag{3.39}$$

donde ϵ_i tiene una función de densidad f_0. Por lo tanto, la variable y tiene la siguiente función de densidad

$$f(y) = \frac{1}{\sigma} f_0 \left(\frac{y - \mu}{\sigma} \right) \tag{3.40}$$

Considerando el proceso para obtener los estimadores máximos verosímiles se concluye que los estimadores $\hat{\mu}$ y $\hat{\sigma}$ satisfacen el siguiente sistema de ecuaciones

$$\begin{aligned} \sum_{i=1}^{n} \psi \left(\frac{y_i - \hat{\mu}}{\hat{\sigma}} \right) &= 0 \\ \frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{y_i - \hat{\mu}}{\hat{\sigma}} \right) &= \delta \end{aligned} \tag{3.41}$$

donde $\psi(y) = -\rho_0'$, $\rho_{\text{escala}}(y) = y\psi(y)$ y $\delta = 1$. En general, los estimadores $\hat{\mu}$ y $\hat{\sigma}$ que cumplen el sistema de ecuaciones anterior no necesitan ser estimadores máximo verosímiles para alguna distribución específica.

3.5. Interpretación de los M-estimadores como estimadores por mínimos cuadrados iterados

3.5.0.5. Cálculo del estimador de localización con parámetro de dispersión previamente calculado

Consideremos la definición de un M-estimador equivariante respecto a la escala utilizando el estimador MADN como el estimador del parámetro σ

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) = 0 \tag{3.42}$$

La ecuación anterior se puede reescribir de la siguiente manera

$$\sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) W\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) = 0 \tag{3.43}$$

donde:

$$W(u) = \begin{cases} \dfrac{\psi(u)}{u} & \text{si } y \neq 0 \\ \\ \psi'(0) & \text{si } y = 0 \end{cases},$$

entonces:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} W\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) y_i}{\sum_{i=1}^{n} W\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right)} \tag{3.44}$$

Esta expresión muestra un promedio ponderado de las y_i's por los pesos $W(u)$, es decir, el estimador de mínimos cuadrados ponderado con pesos $W(u)$.

A partir de las ecuaciones anteriores surgió un método iterativo para el cálculo de $\hat{\mu}$, el método iterativo de mínimos cuadrados ponderados (IRLS por sus siglas en inglés).

Primero, se considera un estimador inicial $\hat{\mu}_0$, generalmente es la media simple, y se calculan los pesos W_0 considerando los estimadores $\hat{\mu}_0$ y MADN

$$W_0 = W\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) \tag{3.45}$$

Posteriomente se calcula el nuevo estimador $\hat{\mu}_1$ utilizando W_0

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} W_0 \left(\frac{y_i - \hat{\mu}_0}{\hat{\sigma}} \right) y_i}{\sum_{i=1}^{n} W_0} \tag{3.46}$$

Con el valor de $\hat{\mu}_1$ se calculan los pesos W_1:

$$W_1 = W \left(\frac{y_i - \hat{\mu}_1}{\hat{\sigma}} \right), \tag{3.47}$$

y se obtiene $\hat{\mu}_2$:

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n} W_1 \left(\frac{y_i - \hat{\mu}_1}{\hat{\sigma}} \right)}{\sum_{i=1}^{n} W_1} \tag{3.48}$$

Iterando este procedimiento se construye una sucesión de estimadores $\hat{\mu}_i$ que converge a la solución de $\hat{\mu}$.

En resumen, el algoritmo para obtener el estimador $\hat{\mu}$ considerando una tolerancia ϵ es

1. Calcular $\hat{\sigma} = MADN(x)$ y $\hat{\mu}_0 = \bar{x}$.
2. Calcular los pesos y los estimadores $\hat{\mu}_{k+1}$ para $k = 0, 1, 2, \ldots$
3. Detener la iteración cuando $|\hat{\mu}_{k+1} - \hat{\mu}_k| < \epsilon\hat{\sigma}$.

3.5.0.6. Cálculo del estimador de localización y dispersión simultáneamente

Para la estimación simultánea de μ y σ. Primero se calculan los estimadores iniciales de μ y σ, $\hat{\mu}_0$ y $\hat{\sigma}_0$ respectivamente. Luego se obtiene W_0^1 tal que

$$W_0^1 = W^1 \left(\frac{y_i - \hat{\mu}_0}{\hat{\sigma}_0} \right), \tag{3.49}$$

donde:

$$W^1(u) = \begin{cases} \dfrac{\psi(u)}{u} & \text{si } y \neq 0 \\ \\ \psi'(0) & \text{si } y = 0 \end{cases}$$

Posteriormente se calcula $\hat{\mu}_1$:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} W_0^1 \left(\frac{y_i - \hat{\mu}_0}{\hat{\sigma}_0} \right)}{\sum_{i=1}^{n} W_0^1} \tag{3.50}$$

El proceso iterativo continúa, ahora calculando el valor del siguiente estimador de σ, $\hat{\sigma}_1$, utilizando la siguiente ecuación

$$\hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{n} W_0^2 \left(\frac{y_i - \hat{\mu}_1}{\hat{\sigma}_0} \right)^2}{n\delta}} \tag{3.51}$$

con $W_0^2 = W^2 \left(\frac{y_i - \hat{\mu}_1}{\hat{\sigma}_0} \right)$, y:

$$W^2(u) = \begin{cases} \dfrac{\rho''(u)}{u^2} & \text{si } y \neq 0 \\ \\ \rho''(0) & \text{si } y = 0 \end{cases}$$

Posteriormente se calcula la función de pesos W_1^1, $\hat{\mu}_2$, W_1^2 y $\hat{\sigma}_2$ hasta que la diferencia de dos iteraciones sucesivas sea menor a ϵ, para una ϵ dada.

3.6. Comparación de diferentes medidas de localización y escala

Ejemplo: Datos simulados

En este ejemplo se compara el comportamiento de las diferentes medidas de localización y escala presentadas anteriormente. La comparación se realiza bajo dos escenarios: en el primero, tenemos 29 observaciones "bien comportadas" que cumplen los supuestos del modelo de localización, es decir, se obtiene a partir de una función de distribución $N(0, 1)$. En el segundo escenario se consideran los mismos datos pero se añade una observación con valor de 80 para que sea considerada como outlier.

En la siguiente gráfica de dispersión se muestran los datos simulados y el estimador por mínimos cuadrados del parámetro de localización en cada escenario.

Como se explicó anteriormente, era de esperarse que con la presencia de un sólo dato atípico, el estimador de mínimos cuadrados se distorsiona ya que su punto de ruptura es igual a cero. Pero, ¿cómo se comportan los otros estimadores con propiedades robustas?

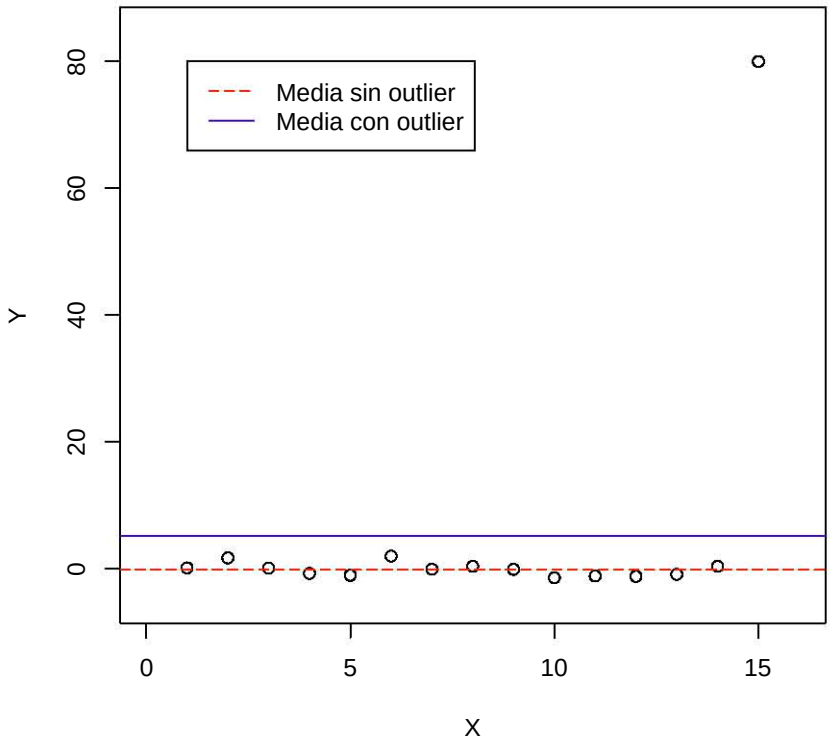


FIGURA 3.5: Datos simulados

En el siguiente cuadro se muestra el valor que toman las diferentes medidas de localización y escala para cada escenario:

Estimador	$\hat{\theta}^*$ **Con outlier**	$\hat{\theta}$ **Sin outlier**
Medidas de localización		
Media	5.43509	0.10903
Media Ajustada (*t=0.1*)	0.22894	0.15570
Mediana	0.28876	0.18546
M-estimador (Huber, *k=1.345*)	0.22894	0.12548
M-estimador (Tukey)	0.12823	0.12822
Medidas de Escala		
Desviación Estándar	20.6414	0.77866
Desviación media respecto a la media	9.94199	0.63319
Desviación media respecto a la mediana	5.90506	0.63319
Mediana de la desviación absoluta (MAD)	0.96238	0.96008
MAD normalizada (MADN)	1.42680	1.42339
M-estimador de escala redescendiente	0.63598	0.59374

En el caso de las medidas de localización, la media ajustada tiene un buen comportamiento pues sólo se tiene la presencia de un solo outlier, sin embargo esto cambiaría si la muestra tiene más de tres datos atípicos. Por otro lado, los M-estimadores son muy estables ya que no sufren un cambio importante por la presencia del dato atípico y son mucho más eficientes que los otros estimadores si no se cuenta con la presencia de la observación con valor de 80.

Respecto a las medidas de dispersión, podemos ver que el estimador MAD es el estimador menos afectado por la presencia del outlier.

Capítulo 4

Estimación robusta para el modelo de regresión lineal

En la sección anterior se desarrollaron diferentes estimadores de localización y escala para la versión más simple del modelo de regresión lineal, el modelo de localización y escala. En este capítulo, considerando que de los diferentes estimadores desarrollados anteriormente, los M-estimadores son los únicos que combinan las propiedades de robustez y eficiencia, se presenta la generalización de los M-estimadores para el modelo de regresión lineal múltiple y se presentan las diferentes soluciones que se han propuesto en respuesta a las limitaciones que presentan los M-estimadores cuando la variable predictoras es también una variable aleatoria y no cumple los supuestos del modelo de regresión lineal.

4.1. M-estimadores

Como se presentó anteriormente, los M-estimadores minimizan una función de los residuales y sus propiedades robustas se determinan a partir de la elección de la función objetivo.

Consideremos el modelo de regresión lineal múltiple

$$y_i = \boldsymbol{x_i'}\boldsymbol{\beta} + \epsilon_i \tag{4.1}$$

donde ϵ_i es el término de error aleatorio, $\boldsymbol{\beta}$ es el vector de parámetros no conocidos y $\boldsymbol{x_i'}$ es el vector de variables independientes. Los residuales e_i se definen como

$$e_i = y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}} \tag{4.2}$$

Supongamos que los e_i siguen una distribución conocida F_0 con densidad f_0. Por lo tanto, la función de verosimilitud para estimar al vector de parámetros $\boldsymbol{\beta}$ es:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_0\left(y_i - \boldsymbol{x_i'}\boldsymbol{\beta}\right) \tag{4.3}$$

El estimador máximo verosímil de $\boldsymbol{\beta}$ se obtiene maximizando la función de verosimilitud anterior o, equivalentemente, minimizando el menos logaritmo de la verosimilitud (4.3)

$$\ln L(\boldsymbol{\beta}) = \min \sum_{i=1}^{n} \rho_0\left(y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}\right) \tag{4.4}$$

con $\rho_0(y) = -\ln f_0(y)$.

Derivando la ecuación (4.4) respecto a $\hat{\boldsymbol{\beta}}$, se obtiene el sistema de ecuaciones normales

$$\sum_{i=1}^{n} \psi_0\left(y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}\right)\boldsymbol{x_i} = \mathbf{0} \tag{4.5}$$

donde $\psi_0(y) = \rho_0'(y) = -f_0'(y)/f_0(y)$. Cuando f_0 es la función de densidad normal, lo anterior se traduce en minimizar la suma de cuadrados de residuales y obtener los estimadores por mínimos cuadrados.

Como se revisó en el capítulo anterior, los M-estimadores de los coeficientes de regresión minimizan la suma de una función de los residuales que crece menos rápido que la función cuadrática o, mejor aún, de una función acotada. Los M-estimadores de los coeficientes de regresión se definen como el vector $\hat{\boldsymbol{\beta}}$ que minimiza la siguiente expresión:

$$\min \sum_{i=1}^{n} \rho\left(y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}\right) \tag{4.6}$$

o, equivalentemente, resuelve la siguiente ecuación:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\boldsymbol{x_i} = \mathbf{0} \tag{4.7}$$

donde $\psi_0(y) = \rho_0'(y)$

En la ecuación (4.7) se puede notar que los residuos están controlados por la función $\psi(y)$, de tal manera que a partir de la elección adecuada de $\psi(y)$, como las funciones de

Huber o Redescendiente, se puede limitar la influencia de una observación cuando tenga un residual relativamente grande.

Para obtener el valor de los M-estimadores es necesario emplear el procedimiento iterativo de mínimos cuadrados ponderados (IRLS) que se presentó en la sección (3.5). Sin embargo, es importante recordar que la solución $\hat{\boldsymbol{\beta}}$ no es invariante respecto a la escala, por lo que se deben de estandarizar los residuales utilizando un estimador robusto de escala $\hat{\sigma}$, el cual puede obtenerse previamente o de manera simultánea en el cálculo de $\hat{\boldsymbol{\beta}}$. Para el caso de la estimación previa del parámetro de escala, la mediana del valor absoluto de las desviaciones (MAD) es el estimador más utilizado.

Los M-estimadores de regresión tienen las propiedades deseadas para un estimador robusto ya que cuentan con una función de influencia acotada y pueden alcanzar un punto de ruptura de 0.5, además, bajo los supuestos en el teorema de Gauss-Markov, alcanzan una eficiencia del 95 % respecto a los estimadores por mínimos cuadrados.

La extensión de los M-estimadores al modelo lineal fue realizada por Relles (1968), Huber (1972), Yohai (1974), Yohai y Maronna (1979), Klein y Yohai (1979).

4.2. Comparación de los estimadores robustos en el modelo de regresión lineal

Ejemplo: Datos "Phones"[1]

En este ejemplo se utilizará el conjunto de datos "Phones" que se encuentran en la paquetería "MASS" de R, los cuales son una extracción del Estudio Estadístico de Bélgica realizado por el Ministerio de Economía Belga. Están conformados por 24 observaciones que registran el número de llamadas telefónicas internacionales (cifras en millones de llamadas) realizadas durante el periodo 1950-1973 en Bélgica. Los datos se muestran en la siguiente tabla y se puede apreciar que hay una tendencia creciente cada año. Sin embargo, el número de llamadas registradas para los años 1964-1969 parecen tener un error y, probablemente, los años 1963 y 1970 también pueden estar afectados. Actualmente se sabe que la discrepancia de estos datos se debe a que en estos años se utilizó otro sistema de registro, el cual proporcionaba el número total de minutos de las llamadas realizadas cada año en lugar de sólo el número de llamadas.

[1] Aunque formalmente estos datos no pueden ser utilizados para realizar un ajuste por medio del modelo de regresión lineal, se utilizan sólo para ilustrar el comportamiento de los estimadores antes presentados.

Obs	Año	Llamadas	Obs	Año	Llamadas
1	50	4.4	**13**	62	16.1
2	51	4.7	**14**	63	21.2
3	52	4.7	**15**	64	119.0
4	53	5.9	**16**	65	124.0
5	54	6.6	**17**	66	142.0
6	55	7.3	**18**	67	159.0
7	56	8.1	**19**	68	182.0
8	57	8.8	**20**	69	212.0
9	58	10.6	**21**	70	43.0
10	59	12.0	**22**	71	24.0
11	60	13.5	**23**	72	27.0
12	61	14.9	**24**	73	29.0

La figura (4.1) muestra el comportamiento de los datos, se puede apreciar el comportamiento discrepante de las observaciones de los años 1963-1970.

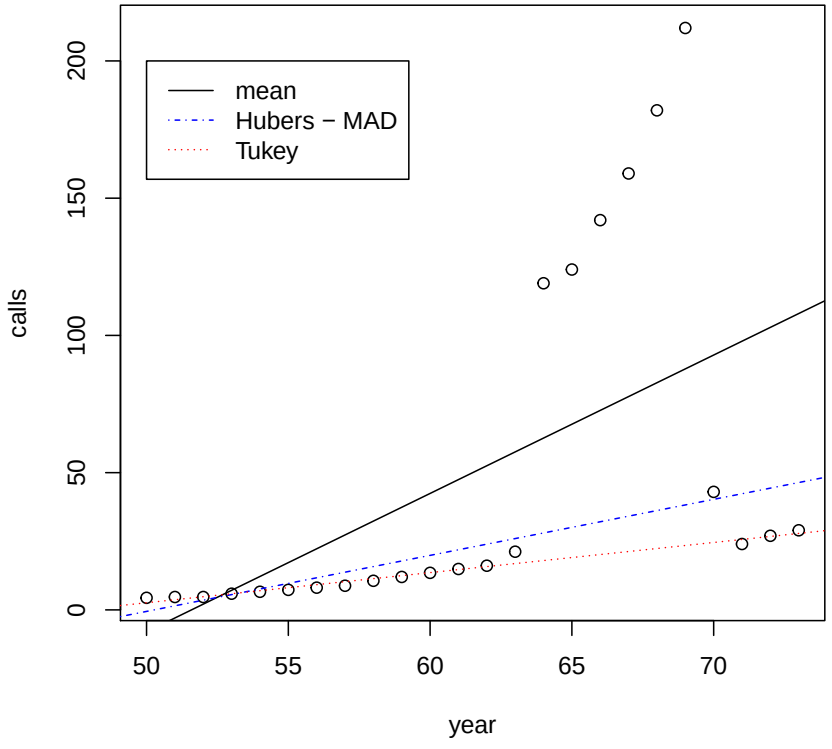


FIGURA 4.1: Diferentes ajustes a los datos “Phones”

Al realizar el análisis por mínimos cuadrados, se obtienen los siguientes resultados:

```
Call:
lm(formula = calls ~ year)

Residuals:
   Min     1Q Median     3Q    Max
-78.97 -33.52 -12.04  23.38 124.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -260.059    102.607  -2.535   0.0189 *
year           5.041      1.658   3.041   0.0060 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.22 on 22 degrees of freedom
Multiple R-squared:  0.2959,        Adjusted R-squared:  0.2639
F-statistic: 9.247 on 1 and 22 DF,  p-value: 0.005998
```

Como era de esperarse, el ajuste por mínimos cuadrados se ve considerablemente afectado por los outliers debido a que las observaciones de los años 60's jalan la recta de regresión.

En la figura (4.2) se muestran cuatro diferentes gráficas basadas en el ajuste por mínimos cuadrados: gráfica de los residuales vs los valores ajustados, el QQ-plot de los residuales estandarizados, la gráfica de la distancia de Cook y la relación de la distancia de Cook vs $h_{ii}/(1-h_{ii})$.

En la figura (4.2) se puede notar cuáles son las observaciones que más influyen en el ajuste por mínimos cuadrados. Las últimas dos observaciones, por ejemplo, tienen un punto de apalancamiento mayor a $2p/n$ pero no tienen un residual grande. Por otro lado, la observación número 20 tiene también un punto de apalancamiento elevado, su residual es grande y es la observación que tiene mayor influencia sobre la recta de regresión, es uno de los datos que influyen en la recta para que salga de la trayectoria de la mayoría de los datos.

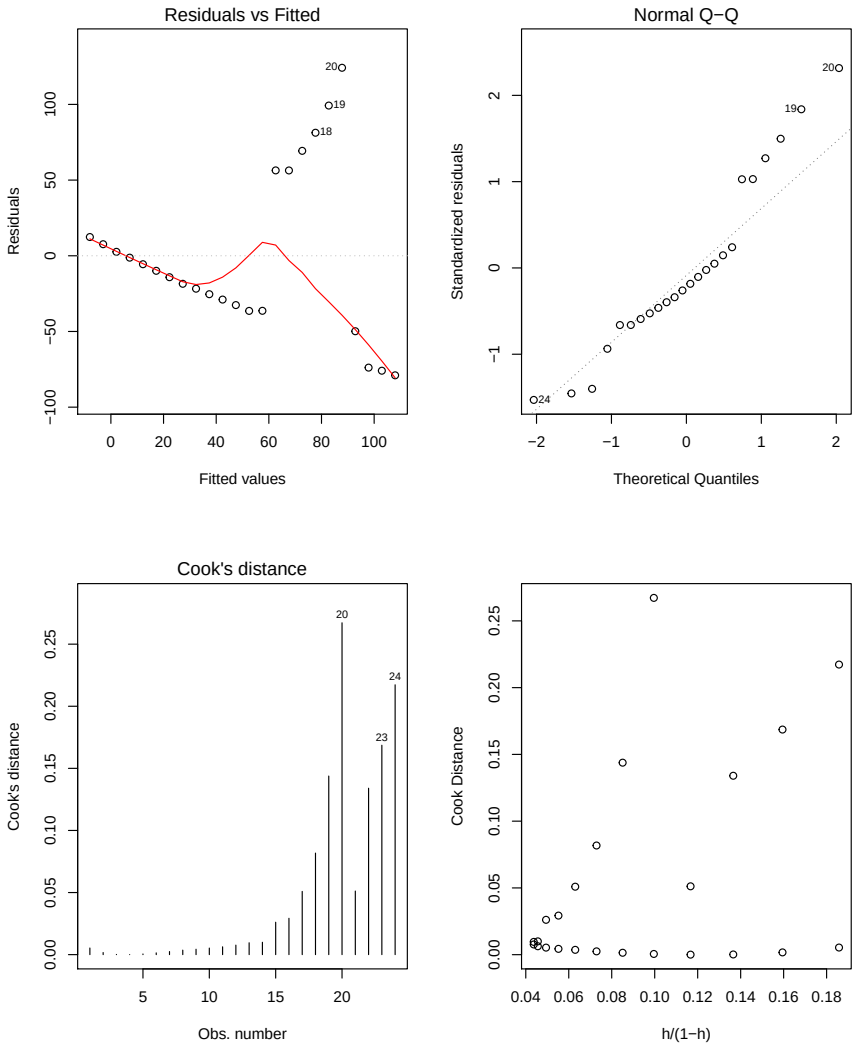


FIGURA 4.2: Análisis de residuales, ajuste por mínimos cuadrados

Considerando las observaciones de la década de los 60's, realizaremos el ajuste utilizando estimadores con propiedades robustas. El ajuste del M-estimador de Huber (con MAD como parámetro de escala) genera los siguientes resultados:

```
Call: rlm(formula = calls ~ year, maxit = 50)
Residuals:
    Min      1Q  Median      3Q     Max
-18.314  -5.953  -1.681  26.460 173.769

Coefficients:
            Value     Std. Error t value
(Intercept) -102.6222   26.6082    -3.8568
year           2.0414    0.4299     4.7480

Residual standard error: 9.032 on 22 degrees of freedom
```

Para el análisis del comportamiento de los residuales y los valores ajustados basados en el M-estimador de Huber consideramos la figura (4.3) que presenta las siguientes gráficas: gráfica de los valores ajustados vs valores observados, los residuales vs los valores ajustados, el QQ-plot de los residuales estandarizados y el peso que le asigna a cada una de las observaciones:

En la figura (4.3) se puede notar que el M-estimador de Huber asigna un peso menor a 1 a las observaciones a partir de la número 15, con excepción de la observación 21. Sin embargo, en la gráfica de los datos se observa que las últimas tres observaciones siguen la trayectoria de la mayoría de los datos. Dado que el M-estimador de Huber tampoco reconoce a la observación número 21 como atípica, la recta de regresión es atraída por la observación perdiendo la trayectoria correcta.

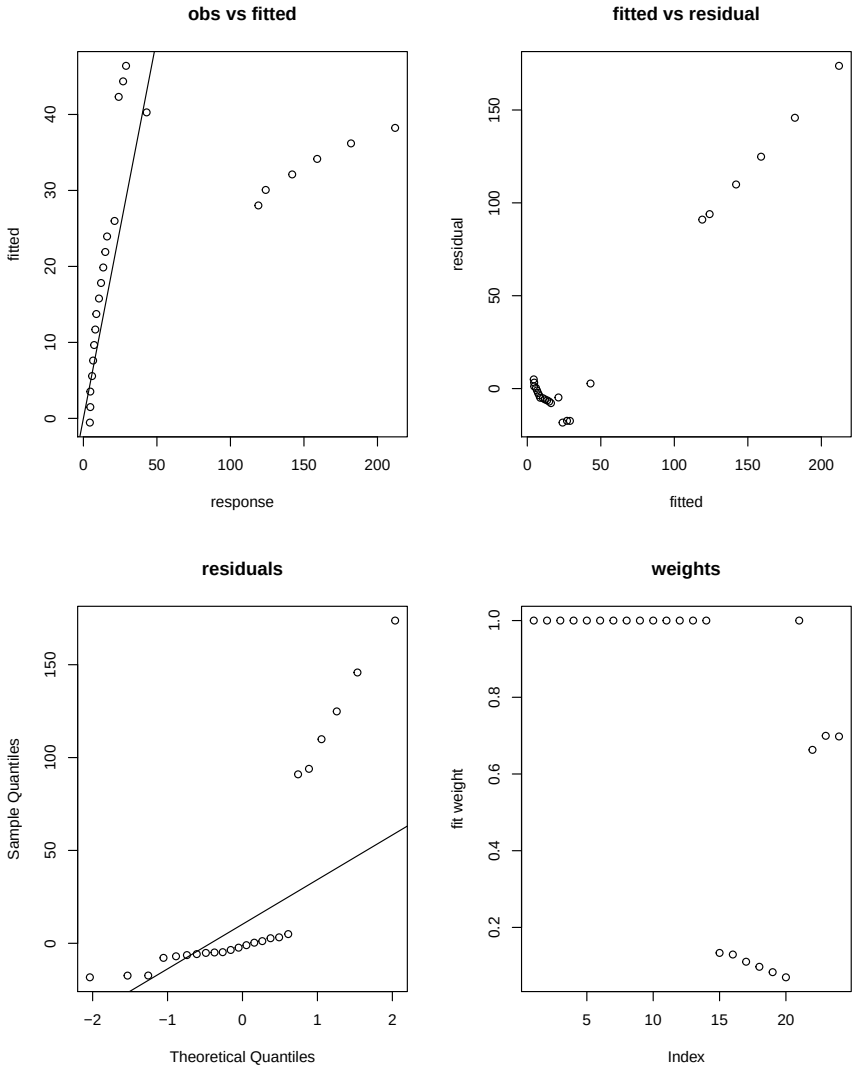


FIGURA 4.3: Análisis de residuales, ajuste M-estimador de Huber

Considerando los problemas de los ajustes realizados anteriormente, hacemos el mismo ejercicio considerando el M-estimador de Tukey con el estimador de escala MAD. El resultado del ajuste es el siguiente:

```
Call: rlm(formula = calls ~ year, psi = "psi.bisquare")
Residuals:
     Min       1Q   Median       3Q      Max
 -1.6585  -0.4143   0.2837  39.0866 188.5376

Coefficients:
            Value    Std. Error t value
(Intercept) -52.3025   2.7530   -18.9985
year          1.0980   0.0445    24.6846

Residual standard error: 1.654 on 22 degrees of freedom
```

La figura (4.4) muestra las gráficas de diagnóstico considerando el ajuste del M-estimador de Tukey el cual suprime la influencia de todas las observaciones extremas y logra un ajuste perfecto de los datos correctos como se mostró en la figura (4.1).

En la siguiente tabla se muestra el peso que le asignan los tres diferentes estimadores a cada una de las observaciones.

Obs	**OLS**	**Huber**	**Tukey**	**Obs**	**OLS**	**Huber**	**Tukey**
1	1.00000	1.00000	0.89479	**13**	1.00000	1.00000	0.99649
2	1.00000	1.00000	0.96671	**14**	1.00000	1.00000	0.47391
3	1.00000	1.00000	0.99971	**15**	1.00000	0.13354	0.00000
4	1.00000	1.00000	1.00000	**16**	1.00000	0.12933	0.00000
5	1.00000	1.00000	0.99494	**17**	1.00000	0.11055	0.00000
6	1.00000	1.00000	0.97942	**18**	1.00000	0.09730	0.00000
7	1.00000	1.00000	0.96109	**19**	1.00000	0.08332	0.00000
8	1.00000	1.00000	0.92798	**20**	1.00000	0.06991	0.00000
9	1.00000	1.00000	0.97971	**21**	1.00000	1.00000	0.00000
10	1.00000	1.00000	0.99232	**22**	1.00000	0.66294	0.91055
11	1.00000	1.00000	0.99979	**23**	1.00000	0.69951	0.99803
12	1.00000	1.00000	0.99835	**24**	1.00000	0.69783	0.95680

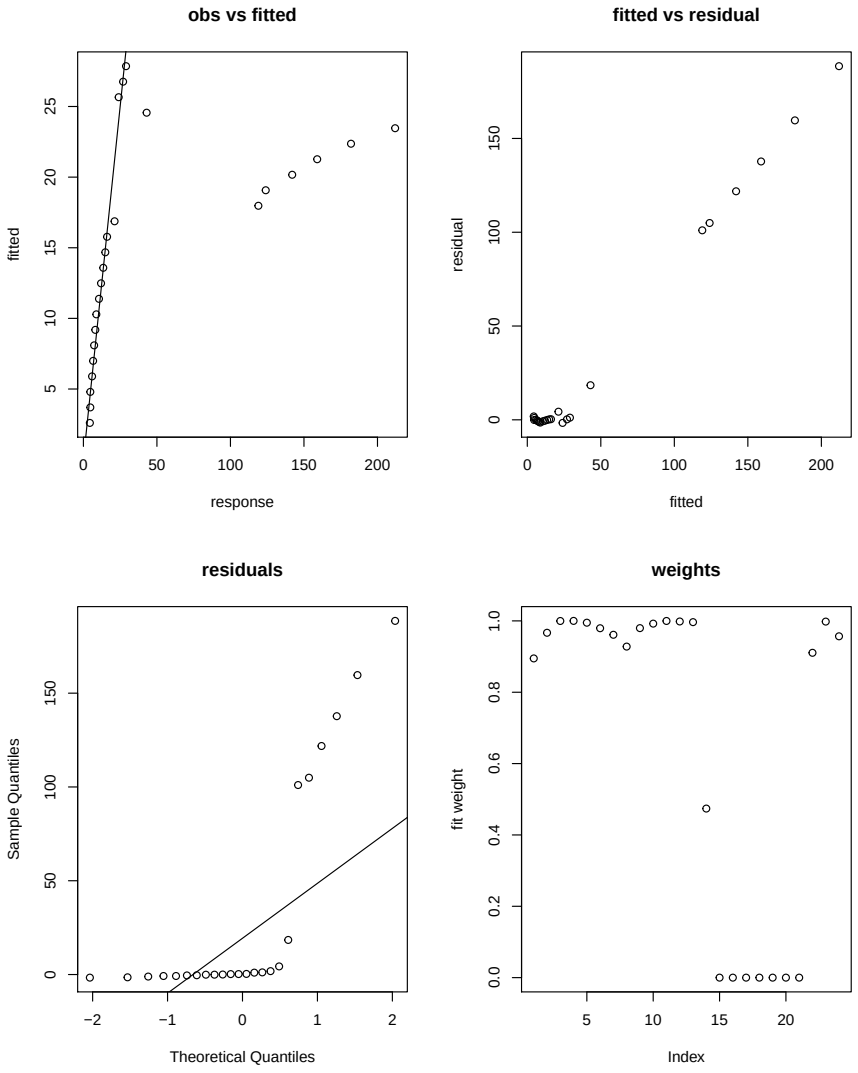


FIGURA 4.4: Análisis de residuales, ajuste M-estimador de Tukey

4.3. Modelo de Regresión Lineal con variables predictoras aleatorias

Hasta el momento el enfoque ha sido el desarrollo de estimadores robustos para situaciones en las que sólo hay datos atípicos en la variable de respuesta y se ha supuesto que las variables predictoras son constantes y no contiene outliers; en este caso los M-estimadores tienen las propiedades deseadas para un estimador robusto. Sin embargo, cuando las variables predictoras $\boldsymbol{X}$ es vector aleatorio, la existencia de puntos de palanca puede distorsionar el valor de los M-estimadores y, en ocasiones, lograr que se desempeñen peor que los estimadores por mínimos cuadrados. Lo anterior se debe a que para valores de $\boldsymbol{X}$ la función de influencia de los M-estimadores no es acotada y, por lo tanto, su punto de ruptura disminuye a 0.

En el siguiente ejemplo se muestra cómo los M-estimadores fallan cuando las matriz $\boldsymbol{X}$ es aleatorio y tiene puntos palanca.

Ejemplo: Datos "Stars CYG"

En la siguiente tabla se muestran los datos del diagrama de Hertzsprung-Russell de las estrellas pertenecientes al cluster CYG OB1; el cual pertenece a la constelación Cygnus y está formado por 47 estrellas. La primera variable contiene el logaritmo de la temperatura efectiva en la superficie de la estrella y la segunda el logaritmo de la intensidad de la luz de la estrella.

Obs	**Log. Temp**	**Log. Luz**	**Obs**	**Log. Temp**	**Log. Luz**
1	4.37	5.23	**25**	4.38	5.02
2	4.56	5.74	**26**	4.42	4.66
3	4.26	4.93	**27**	4.29	4.66
4	4.56	5.74	**28**	4.38	4.9
5	4.3	5.19	**29**	4.22	4.39
6	4.46	5.46	**30**	3.48	6.05
7	3.84	4.65	**31**	4.38	4.42
8	4.57	5.27	**32**	4.56	5.1
9	4.26	5.57	**33**	4.45	5.22
10	4.37	5.12	**34**	3.49	6.29
11	3.49	5.73	**35**	4.23	4.34
12	4.43	5.45	**36**	4.62	5.62
13	4.48	5.42	**37**	4.53	5.1
14	4.01	4.05	**38**	4.45	5.22
15	4.29	4.26	**39**	4.53	5.18
16	4.42	4.58	**40**	4.43	5.57
17	4.23	3.94	**41**	4.38	4.62
18	4.42	4.18	**42**	4.45	5.06
19	4.23	4.18	**43**	4.5	5.34
20	3.49	5.89	**44**	4.45	5.34
21	4.29	4.38	**45**	4.55	5.54
22	4.29	4.22	**46**	4.45	4.98
23	4.42	4.42	**47**	4.42	4.5
24	4.49	4.85			

El diagrama de Hertzsprung-Russell es un diagrama de dispersión donde el logaritmo de la temperatura corresponde al eje x. En la siguiente gráfica aparecen dos conjuntos de observaciones, la mayoría de los datos siguen una tendencia en la esquina inferior derecha y 4 puntos se encuentran en la esquina superior izquierda, alejados del resto de los puntos. En astronomía, se dice que las 43 observaciones siguen la secuencia principal

mientras que las otras cuatro estrellas se consideran como "gigantes" (observaciones 11, 20, 30 y 34).

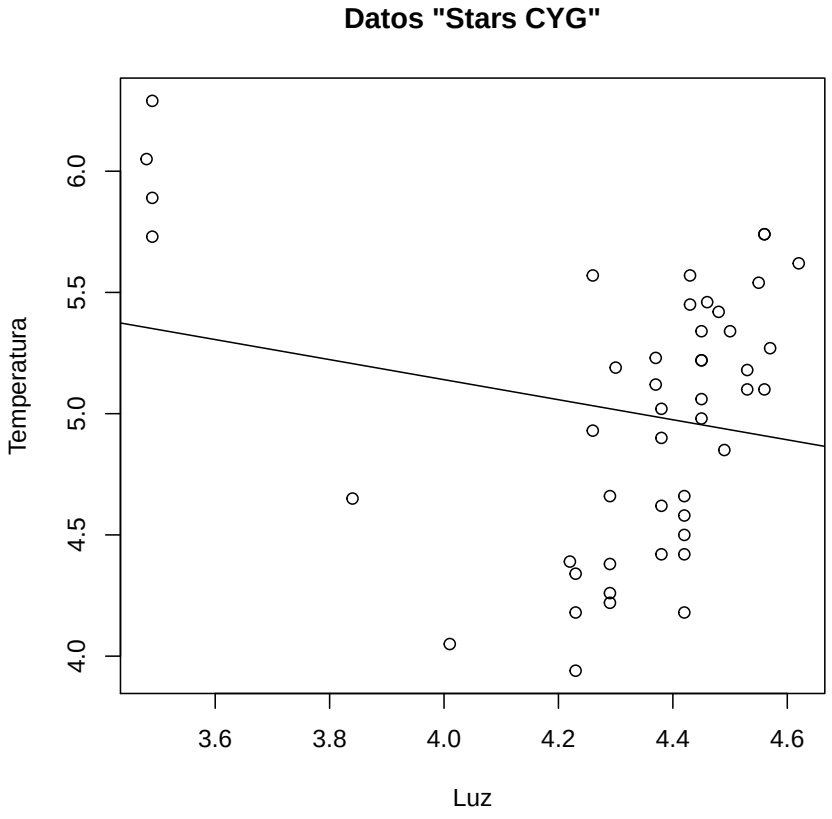


FIGURA 4.5: Diagrama de Hertzsprung - Russell, cluster CYG OB1

En la figura (4.5) se puede observar cómo los 4 puntos de la esquina superior izquierda atraen a la recta de regresión por mínimos cuadrados pero, ¿ocurrirá lo mismo con los M-estimadores?

Realizaremos el ajuste a partir de los M-estimadores de Huber y Tukey, el análisis de diagnóstico de residuales para ver el tratamiento que le dan los estimadores a cada observación y, finalmente, graficaremos las rectas de regresión para ver como se ajustan a los datos.

Ajuste del M-estimador de Huber

```
Call: rlm(formula = log.light ~ log.Te, maxit = 50)
Residuals:
    Min      1Q  Median      3Q     Max
-1.1132 -0.5118  0.1268  0.4382  0.9197

Coefficients:
            Value   Std. Error t value
(Intercept)  6.8659  1.2641     5.4315
log.Te      -0.4285  0.2926    -1.4643

Residual standard error: 0.7026 on 45 degrees of freedom
```

Análisis de residuales del M-estimador de Huber

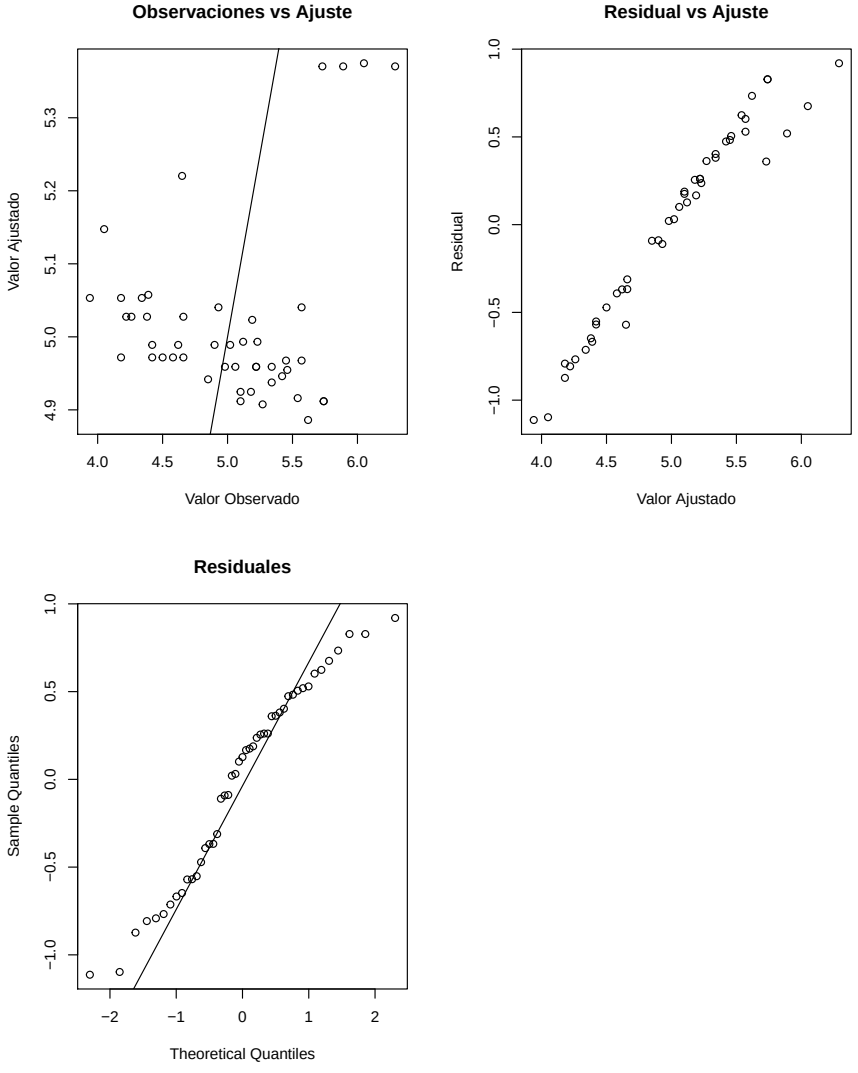


FIGURA 4.6: Análisis de residuales, ajuste M-estimador de Huber

Resultados obtenidos realizando el ajuste con el M-estimador de Tukey:

```
Call: rlm(formula = log.light ~ log.Te, psi = "psi.bisquare")
Residuals:
    Min      1Q  Median      3Q     Max
-1.1154 -0.5160  0.1231  0.4332  0.9253

Coefficients:
            Value   Std. Error t value
(Intercept)  6.8235  1.3200     5.1692
log.Te      -0.4180  0.3056    -1.3677

Residual standard error: 0.7058 on 45 degrees of freedom
```

Análisis de residuales del M-estimador de Tukey

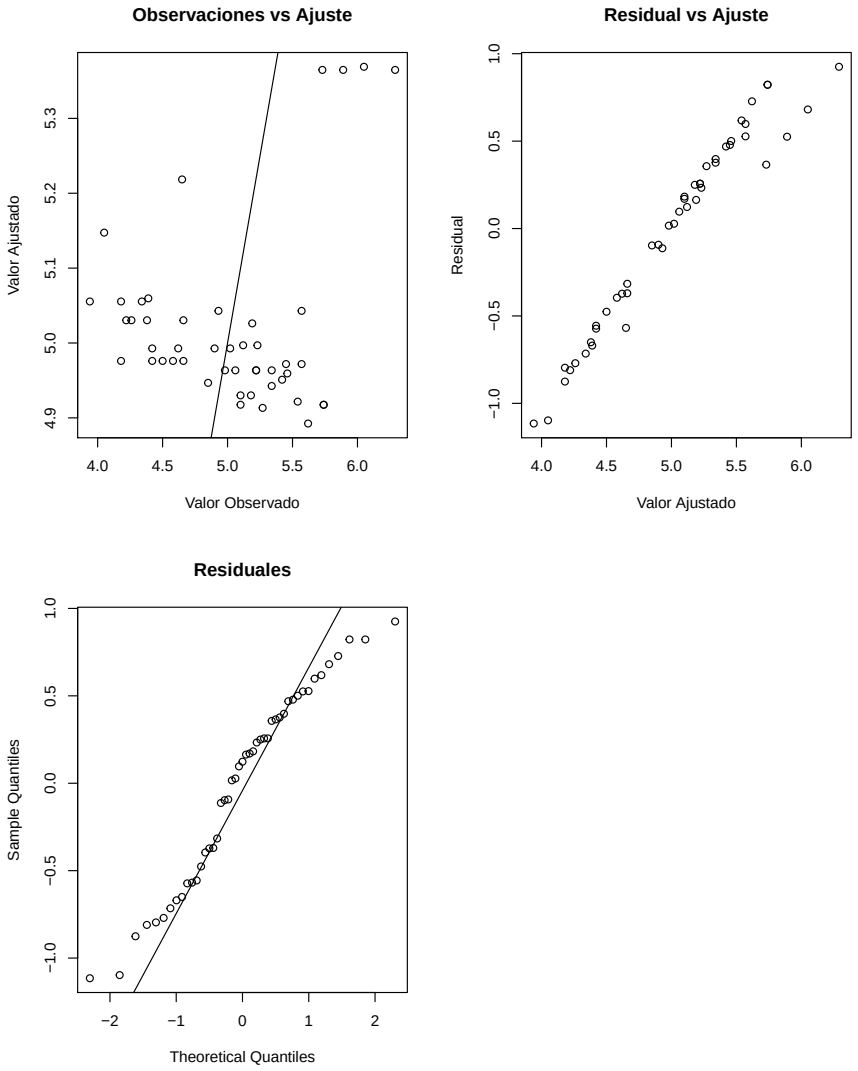


FIGURA 4.7: Análisis de residuales, ajuste M-estimador de Tukey

Finalmente, en la figura (4.8) se compara el peso que le asigna cada estimador a las diferentes observaciones.

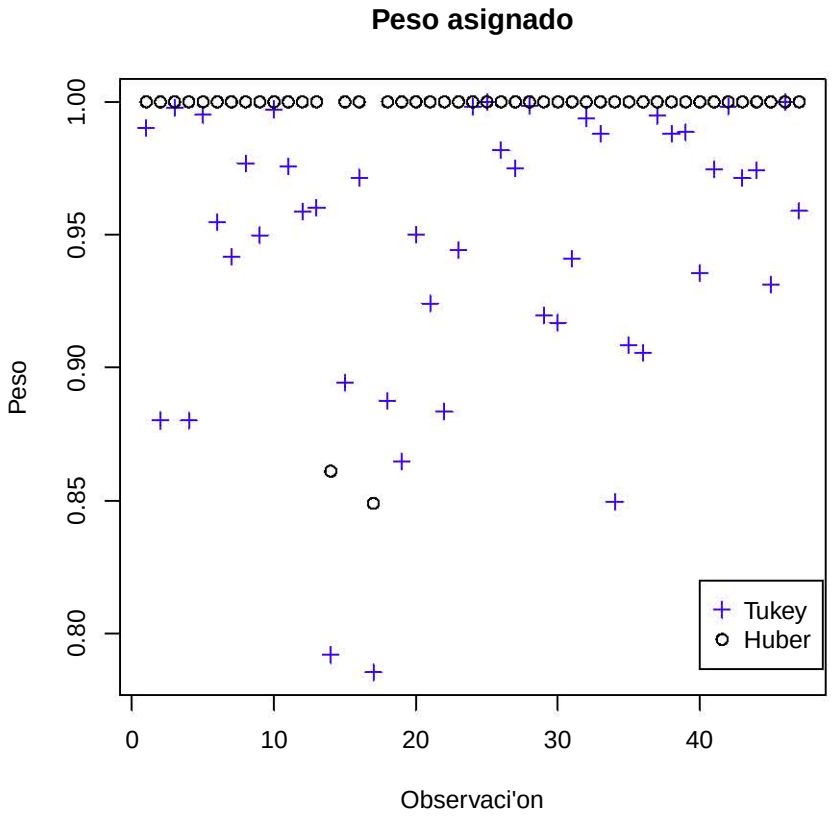


FIGURA 4.8: Comparativo del peso asignado a cada observación, M-estimador de Huber y Tukey

El peso que los M-estimadores asignaron a las observaciones que se encuentran claramente fuera de la tendencia de la mayoría de los datos son los siguientes:

Obs	**Huber**	**Tukey**
11	1.00000	0.97573
20	1.00000	0.95015
30	1.00000	0.91693
34	1.00000	0.84950

Por otro lado, ambos estimadores asignan el menor peso a las observaciones 14 y 17

Obs	**Huber**	**Tukey**
14	0.86104	0.79183
17	0.84888	0.78534

A partir del análisis anterior, se puede concluir que el M-estimador de Huber y el de Tukey se ven afectado por las observaciones 11, 20, 30 y 40, al igual que el estimador por mínimos cuadrado, ya que no identifican a estas observaciones como datos atípicos. El M-estimador de Huber les asigna peso de uno a estas observaciones y M-estimador de Tukey les asigna un peso muy cercano a uno.

En la figura (4.9) se encuentran las líneas estimadas de regresión generadas a partir de los ajustes realizados anteriormente.

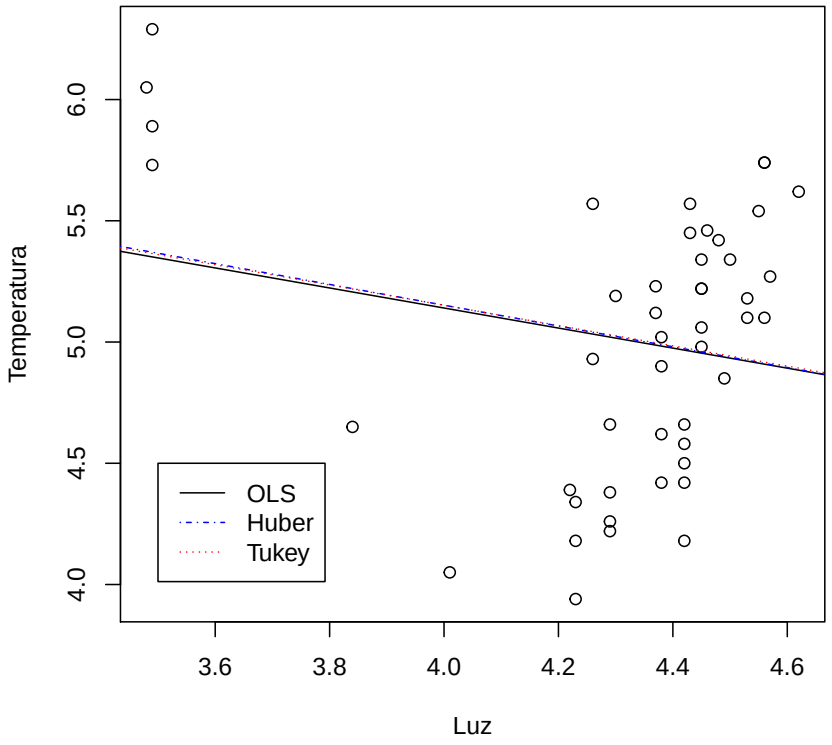


FIGURA 4.9: Diferentes ajustes para los datos "Stars CYG"

Como se mencionó anteriormente, los M-estimadores no reconocen a las observaciones de la esquina superior izquierda como datos atípicos ya que en función del valor en dirección "y" estas observaciones no son atípicas, lo cual no ocurre al considerar su valor en "x".

4.4. GM-estimadores

Como respuesta a los problemas que presentan los M-estimadores en presencia de puntos palanca, se desarrollaron los M-estimadores Generalizados (GM-estimadores). La idea de estos estimadores es ponderar el peso de los outliers y los puntos palanca de tal manera

que se acote la influencia de los datos atípicos (en cualquier dirección) sobre la recta estimada.

Considerando nuevamente el modelo de regresión múltiple, los GM-estimadores son aquellos que satisfacen la siguiente ecuación

$$\sum_{i=1}^{n} w_i\left(\boldsymbol{x}_i\right) \psi\left(\frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}}{v_i(\boldsymbol{x}_i)\hat{\sigma}}\right) \boldsymbol{x}_i = \mathbf{0} \tag{4.8}$$

donde $\psi(y)$ regularmente es una función de la familia de Huber o Biponderada, y los pesos $w_i(\boldsymbol{x_i})$ y $v_i(\boldsymbol{x_i})$ dependen de un ajuste inicial por medio de mínimos cuadrados y se reajustan utilizando el proceso iterativo de mínimos cuadrados ponderados (IRLS).

El primer GM-estimador fue propuesto por Mallows en 1975. Este estimador sólo considera la función de pesos $w_i(\boldsymbol{x_i})$, por lo que $v_i(\boldsymbol{x_i}) = 1$. La función $w_i(\boldsymbol{x_i})$ se define a partir de los valores de la diagonal de la matriz sombrero, h_{ii}. Dado que el valor de h_{ii} se encuentra entre 0 y 1, al definir $w_i(\boldsymbol{x_i}) = \sqrt{1 - h_{ii}}$, se asegura que a las observaciones con un alto punto de palanca se les otorga un menor peso que aquellas con un apalancamiento cercano a 0. El problema de esta primera propuesta es que castiga por igual a las observaciones con un alto punto de palanca sin considerar si estas son observaciones "buenas" o "malas", originando que en algunas ocasiones se afecte la eficiencia del estimador.

Una segunda propuesta, realizada por Schweppe en 1977, además de considerar $w_i(\boldsymbol{x_i}) = \sqrt{1 - h_{ii}}$, ajusta estos pesos de acuerdo al tamaño de los residuales e_i considerando $v_i(\boldsymbol{x_i}) = w_i(\boldsymbol{x_i})$. Aunque este estimador tiene un mayor punto de ruptura que los M-estimadores, puede verse distorsionado cuando se presentan errores asimétricos, por lo que en la mayoría de los casos resultan obsoletos.

Aunque los GM-estimadores resuelven el problema presentado por los M-estimadores al ponderar el efecto de los punto palanca; Maronna, Butos y Yohai demostraron en 1979 que el punto de ruptura para los GM-estimadores nunca es mayor a $1/(p+1)$, donde p es el número de parámetros estimados, incluso experimentos numéricos comprobaron que nunca alcanza un punto de ruptura igual a $1/(p+1)$. Lo anterior significa que el punto de ruptura decrece en función del incremento de los parámetros, esto representa un gran problema ya que cuando se incrementa el número de parámetros es, precisamente, cuando hay más oportunidad de la existencia de puntos palanca y la detección de los outliers de regresión se vuelve más complicada.

Varios estimadores han sido propuestos siguiendo la idea de los GM-estimadores: Theil (1950), Bronw y Mood (1951), Sen (1968), Jaeckel (1972) y Andrews (1974) sin embargo,

todos tienen un punto de ruptura menor al 30 % para el caso de la regresión simple ($p = 2$).

4.5. S-estimadores

Considerando el bajo punto de ruptura que pueden alcanzar los GM-estimadores cuando se incrementa el número de parámetros a estimar, se introdujo una nueva clase de estimadores con el objetivo de crear estimadores de regresión que tengan un punto de ruptura alto y que, además, compartan la flexibilidad y las buenas propiedades asintóticas de los M-estimadores. El enfoque de esta nueva clase de estimadores considera la escala de los residuales como parte central del problema y, a partir de ésta, se deriva el estimador $\hat{\boldsymbol{\beta}}$. Rousseeuw y Yohai introdujeron los S-estimadores en 1984, los cuales toman su nombre debido a que se obtienen a partir de la ecuación de un M-estimador de escala, este tipo de estimadores tienen la gran ventaja de que se pueden calcular directamente de los datos sin necesidad de realizar el cálculo de algún estimador inicial y llevar acabo un proceso iterativo.

Los M-estimadores de escala se obtienen a partir de la siguiente ecuación

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) = \delta \tag{4.9}$$

donde ρ es una *ρ-función* acotada, $\delta = E_\Phi[\rho(r)]$ y Φ representa la distribución normal estándar. Para cada vector $\boldsymbol{\beta}$ se puede calcular la dispersión de los residuales $\hat{\sigma}(e_1(\hat{\boldsymbol{\beta}}), \ldots, e_n(\hat{\boldsymbol{\beta}}))$. Los S-estimadores se definen como el valor $\hat{\boldsymbol{\beta}}$ tal que

$$\min \hat{\sigma}(e_1(\hat{\boldsymbol{\beta}}), \ldots, e_n(\hat{\boldsymbol{\beta}}))$$

Los S-estimadores pueden llegar a tener un punto de ruptura de 0.5 si la función objetivo cumple las siguientes tres condiciones:

1. $\rho(x)$ es simétrica, continuamente diferenciable y $\rho(0) = 0$
2. Existe una constante k tal que se cumple que $\rho(x)$ es estrictamente creciente en el intervalo $[0, k]$ y constante en $[k, \infty)$
3. $\delta/\rho(k) = 1/2$

La función objetivo comúnmente elegida es la función redescendiente.

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2k^2} + \dfrac{x^6}{2k^4} & |x| \leq k \\ \\ \dfrac{k^2}{6} & |x| > k \end{cases}$$

Rousseeuw y Leroy demostraron que en general el punto de ruptura de los S-estimadores, cuando la función objetivo cumple las tres condiciones señaladas anteriormente, es:

$$\frac{\frac{n}{2} - p + 2}{n}$$

por lo que tiene un valor asintótico de 0.5.

Por otro lado, la eficiencia asintótica de esta clase de estimadores depende de la función objetivo elegida. Desafortunadamente, dado que la constante "k" se elige para asegurar que se alcance un alto punto de ruptura, no es posible asegurar simultáneamente un alto punto de ruptura y una alta eficiencia asintótica, por lo que si se quiere asegurar un punto de ruptura elevado tendremos un alto costo en la eficiencia asintótica y viceversa. Para el caso de la función biponderada, con $k = 1.547$ la eficiencia es sólo del 28.7 % cuando se cumple el supuesto de normalidad. En la siguiente tabla se muestra la eficiencia asintótica de los S-estimadores para diferentes valores de punto de ruptura cuando se utiliza la función biponderada.

BDP	**Eficiencia**	k	δ
% 50	% 28.7	1.547	0.1995
% 45	% 37.0	1.756	0.2312
% 40	% 46.2	1.988	0.2634
% 35	% 56.0	2.251	0.2957
% 30	% 66.1	2.56	0.3278
% 25	% 75.9	2.973	0.3593
% 20	% 84.7	3.42	0.3899
% 15	% 91.7	4.096	0.4194
% 10	% 96.6	5.182	0.4475

Debido a la dependencia negativa que se tiene entre la eficiencia asintótica y en el punto de ruptura, los S-estimadores no pueden ser considerados como una buena elección para un estimador robusto de regresión. Sin embargo, dado que los S-estimadores pueden alcanzar la cota máxima que se puede esperar para el punto de ruptura, son un excelente estimador inicial para obtener estimadores de regresión robustos y eficientes.

Ejemplo: Datos "Stars CYG"

Utilizaremos nuevamente los datos del diagrama de Hertzsprung-Russell del Cluster de Estrellas CYG OB1 (datos "Stars CYG") para ilustrar cómo los GM-estimadores y los S-estimadores resuelven el problema de los M-estimadores respecto a la presencia de puntos palanca.

Recordemos que los datos se componen de 47 observaciones donde la variable explicativa contiene el logaritmo de la temperatura en la superficie de la estrella y la variable respuesta el logaritmo de la intensidad de la luz.

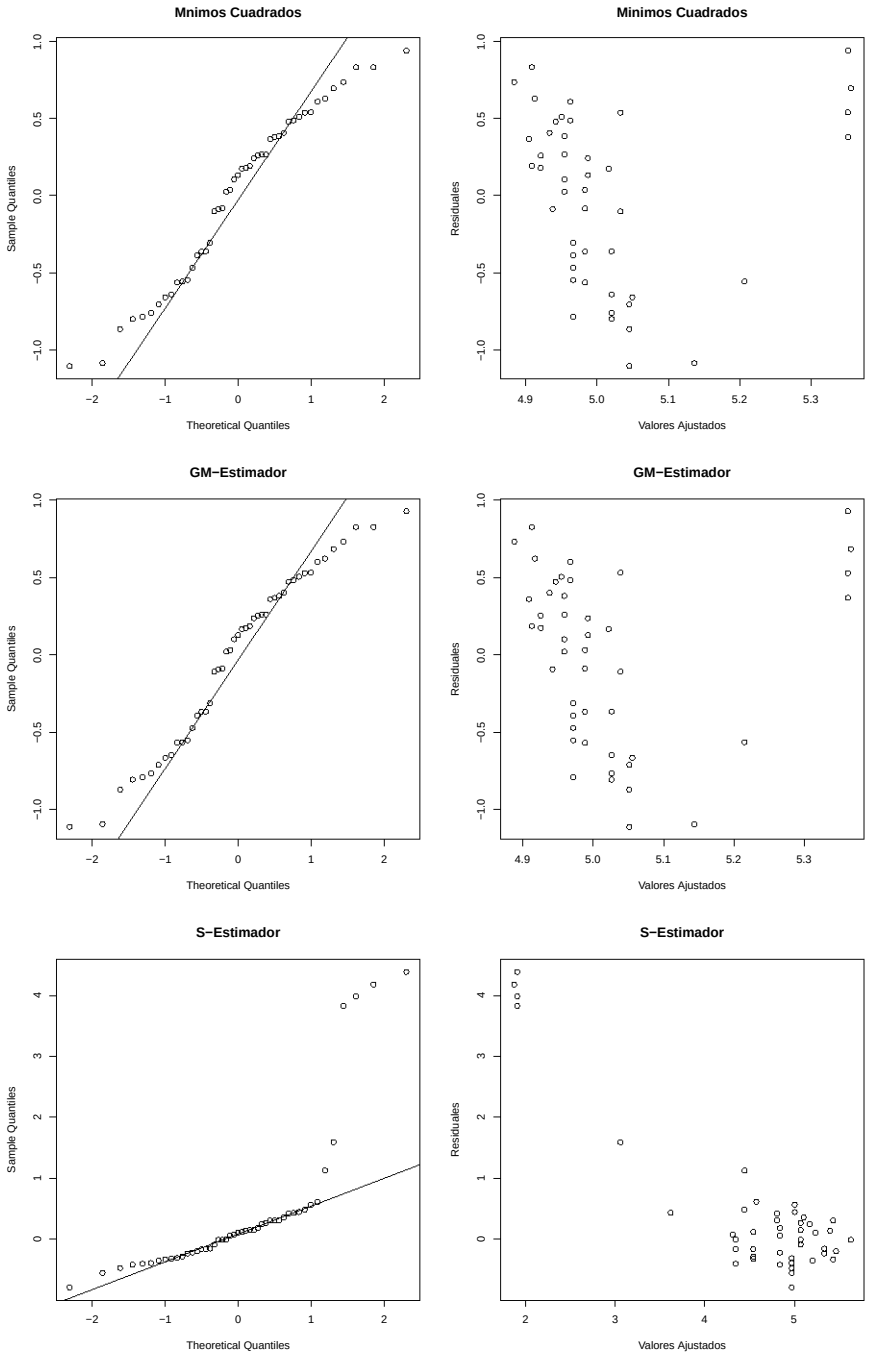
Los resultados al llevar a cabo el ajuste de los datos a través del GM-estimador son los siguientes:

```
Call:
wle.lm(formula = log.light ~ log.Te, raf = "SCHI2")

Root  1

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-1.0762 -0.5108  0.1273  0.4312  0.9276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.8240     1.2463   5.475 2.03e-06 ***
log.Te       -0.4191     0.2885  -1.452    0.154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5633 on 43.58231 degrees of freedom
Multiple R-Squared: 0.04617,        Adjusted R-squared: 0.02428
F-statistic:  2.11 on 1 and 43.58231 degrees of freedom,        p-value: 0.1535
```

Por otro lado, el ajuste a partir del S-estimador con la función biponderada ($k = 1.547$ y $\delta = 0.1995$) presenta los siguientes resultados:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.7915 -0.2270  0.1045  0.4035  0.3900  4.3870

(Intercept)      log.Te
  -9.543077    3.282051
```

Ahora, ¿cómo se comportan los residuales y los valores ajustados considerando el ajuste por mínimos cuadrados (como referencia), el ajuste del GM-estimador y el ajuste del S-estimador realizados anteriormente?



A partir del análisis anterior, podemos ver que tanto el GM-estimador como el S-estimador se comportan diferentes respecto al estimador por mínimos cuadrados, ajustándose de mejor manera a la mayoría de los datos. En la figura (4.10) se muestran la rectas de regresión utilizando los diferentes estimadores

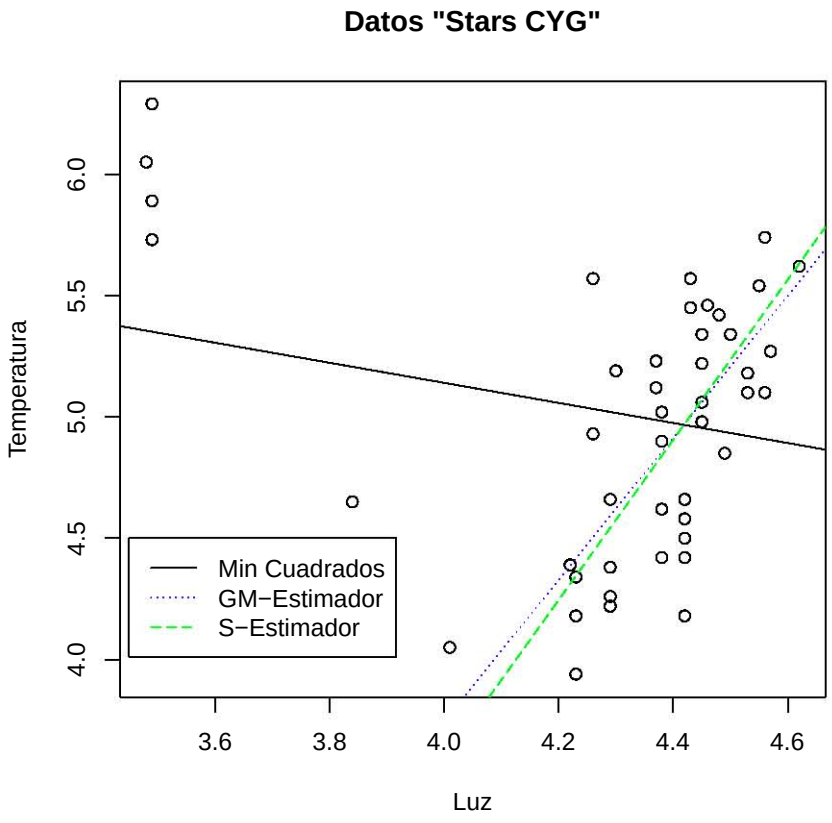


FIGURA 4.10: Diferentes ajustes a los datos "Stars CYG"

4.6. Combinando resistencia y eficiencia, MM-estimadores

En las secciones anteriores se presentaron diferentes clases de estimadores que desafortunadamente sólo cumplen de manera satisfactoria alguna de las propiedades necesarias para un estimador robusto: cuentan con un alto grado de resistencia a los outliers pero una baja eficiencia o viceversa. En esta sección se presentará una familia de estimadores de regresión, llamados MM-estimadores, que surgió a partir de los estimadores desarrollados anteriormente y que satisfacen, simultáneamente, las propiedades de un alto punto de ruptura (0.5) y una alta eficiencia bajo errores provenientes de una distribución normal (95 %). El término "MM" en el nombre de esta clase de estimadores se refiere al hecho de que es necesario más de un procedimiento para la obtención de un M-estimador para poder obtener el estimador final. Esta clase de estimadores fueron propuestos por Yohai en 1987 y son la técnica de regresión robusta más utilizada actualmente.

Los MM-estimadores se obtienen a través de un proceso de tres etapas. En la primera etapa se debe calcular un estimador que tenga un alto punto de ruptura y que no necesariamente debe ser eficiente. En la segunda etapa se obtiene los residuales considerando

el estimador inicial y se calcula un M-estimador de escala que tenga un punto de ruptura de 0.5. En la etapa final, se obtiene el M-estimador para el vector de parámetros $\boldsymbol{\beta}$ utilizando una *ρ-función* acotada y el estimador de escala obtenido en la segunda etapa para que el M-estimador sea invariante respecto a la escala.

4.6.1. Cálculo de los MM-estimadores

Etapa 1.- Calcular un estimador de regresión $\hat{\boldsymbol{\beta}}_s$ de $\boldsymbol{\beta}$ con un punto de ruptura alto, preferentemente 0.5. El estimador no necesita ser eficiente.

Etapa 2 .- A partir del estimador $\hat{\boldsymbol{\beta}}_s$ obtenemos los residuales y calculamos un M-estimador de escala $S_n = \hat{\sigma}\left(e_1(\hat{\boldsymbol{\beta}}), \ldots, e_n(\hat{\boldsymbol{\beta}})\right)$ con un punto de ruptura de 0.5, es decir, usando una función objetivo que denotaremos como ρ_0 y una constante $\delta = 0.5$ tal que

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i(\beta)}{\hat{\sigma}}\right) = 0.5$$

Etapa 3 .- Sea ρ_1 otra función objetivo tal que $\rho_0 \geq \rho_1$. Definimos el MM-estimador como cualquier solución de:

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{e_i(\hat{\boldsymbol{\beta}})}{s_n}\right) = 0$$

que cumple:

$$\sum_{i=1}^{n}\rho_1\left(\frac{e_i(\hat{\boldsymbol{\beta}})}{s_n}\right) \leq \sum_{i=1}^{n}\rho_1\left(\frac{e_i(\hat{\boldsymbol{\beta}}_s)}{s_n}\right)$$

Generalmente el estimador elegido como inicial es un S-estimador; es indispensable que el estimador inicial sea robusto pues con ello aseguramos que el estimador final $\hat{\boldsymbol{\beta}}$ herede la misma propiedad. En la segunda etapa la función objetivo utilizada es la función biponderada con $k_0 = 1.548$ para asegurar un punto de ruptura de 0.5. Finalmente, en la última etapa también se utiliza la función biponderada con una constante $k_1 = 4.685$ para lograr una eficiencia normal del 95 %.

Una propiedad importante con la que cuentan los MM-estimadores, además de alto punto de ruptura y elevada eficiencia, es el llamado "ajuste exacto". Un estimador tiene la propiedad de ajuste exacto si para cualquier muestra de n observaciones, donde al menos $n - \frac{n}{2} + 1$ de las observaciones cumplen exactamente que $y_i = \boldsymbol{x_i'}\boldsymbol{\beta}$, entonces el estimador obtenido es exactamente igual a $\boldsymbol{\beta}$ independientemente de las otras observaciones. Los MM-estimadores heredan esta propiedad del estimador inicial utilizado en la primera

etapa. Uno de los estimadores que tienen esta propiedad son los S-estimadores por lo que son elegidos para ser el estimador inicial.

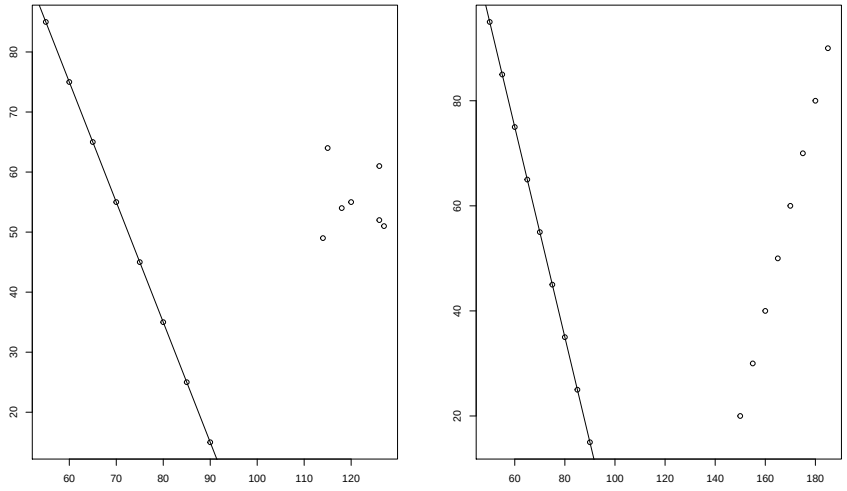


FIGURA 4.11: Propiedad ajuste perfecto

Ejemplo: Datos Simulados

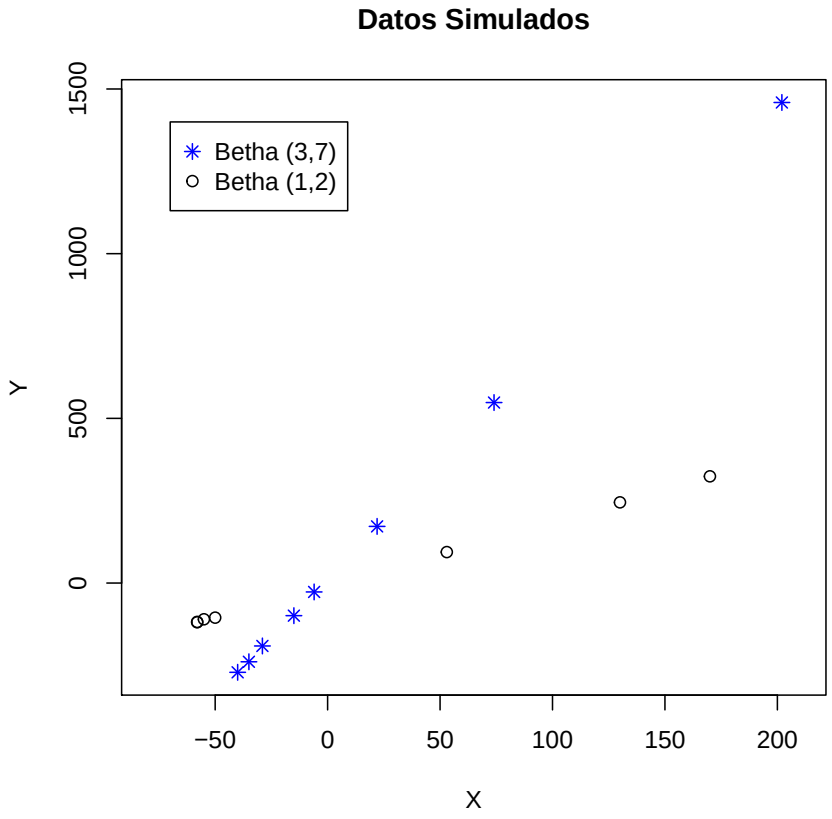
En el siguiente ejemplo se muestra la propiedad de ajuste perfecto que tienen los MM-estimadores cuando se utiliza como estimador inicial a un S-estimador. Se realizará el ajuste de los datos utilizando EMC, M-estimador Biponderado, S-estimador y MM-estimador.

Consideremos las siguientes 15 observaciones de las cuales 8 fueron generadas considerando $\boldsymbol{\beta} = \begin{pmatrix} 3 \\ 7 \end{pmatrix}$ y las 7 restantes con $\boldsymbol{\beta} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

X1	6	2	15	10	5	2	4	3	0	-2	-3	-5	-16	-12	-15
X2	22	-15	202	74	-6	-35	-29	-40	-55	-58	-58	-50	170	53	130
Y	172	-99	1459	548	-27	-239	-191	-271	-110	-118	-119	-105	324	94	245

En la siguiente gráfica se muestra el comportamiento de los datos, identificando las observaciones que provienen de cada uno de los modelos.

Los resultados de $\hat{\boldsymbol{\beta}}$ utilizando los diferentes métodos son los siguientes:



Tipo de Estimador	**Estimador $\hat{\beta}$**
Mínimos Cuadrados	$\begin{pmatrix} 21.07 \\ 4.50 \end{pmatrix}$
M-estimador	$\begin{pmatrix} 20.83 \\ 4.53 \end{pmatrix}$
S-estimador	$\begin{pmatrix} 3.00 \\ 7.00 \end{pmatrix}$
MM-estimador	$\begin{pmatrix} 3.00 \\ 7.00 \end{pmatrix}$

Como se puede observar, tanto el S-estimador como el MM-estimador estiman de manera adecuada el valor de los factores de los cuales se genera la mayoría de las observaciones (aunque esta mayoría se defina con sólo una observación de diferencia), esta propiedad reafirma el elevado punto de ruptura con el que cuentan estos estimadores. Por otro lado, se puede observar como el M-estimador biponderado se comporta de manera muy similar al estimador por mínimos cuadrados debido a la presencia de puntos palanca en los datos.

4.7. Consecuencias de una alta eficiencia

A pesar de que una de las propiedades deseadas para los estimadores robustos es una alta eficiencia respecto a los estimadores por mínimos cuadrados cuando las observaciones cumplen el supuesto de normalidad (al ser estos los estimadores máximo verosímiles), esto genera algunas consecuencias negativas ya que el estimador puede ser engañado por observaciones que sí provienen de una distribución normal e ignorar la presencia de datos atípicos.

Ejemplo: Datos simulados

Para mostrar el problema que pueden tener los estimadores con una alta eficiencia, realizaremos un ejemplo utilizando 20 datos simulados. Generaremos los 20 valores de la variable de respuesta considerando una distribución normal. Por otro lado, se generan números consecutivos del 1 al 20 y 8 de estos números serán sustituidos aleatoriamente por valores generados a partir de una distribución Cauchy. Los datos generados son los siguientes:

Obs	**x**	**y**	**Obs**	**x**	**y**
1	70.06	21.01	**11**	40.61	67.96
2	-109.51	21.27	**12**	12.00	61.61
3	3.00	24.09	**13**	-216.52	82.97
4	4.00	35.53	**14**	14.00	82.26
5	5.00	41.07	**15**	15.00	80.47
6	6.00	24.13	**16**	16.00	79.64
7	0.58	40.27	**17**	17.00	105.45
8	8.00	55.22	**18**	18.00	97.67
9	-2.16	42.97	**19**	19.00	100.44
10	-12.32	59.3	**20**	97.12	101.33

En la figura (4.12) se muestra el ajuste realizado a partir de los siguientes estimadores:

- Estimador mínimo cuadrado
- MM-estimador utilizando el S-estimador en la primera etapa y la función de Tukey en la segunda y tercera etapa con $k_0 = 1.548$ y $k_1 = 4.685$ respectivamente.
- MM-estimador utilizando el S-estimador en la primera etapa y la función de Tukey en la segunda y tercera etapa con $k_0 = 1.548$ y $k_1 = 3.44$ respectivamente.

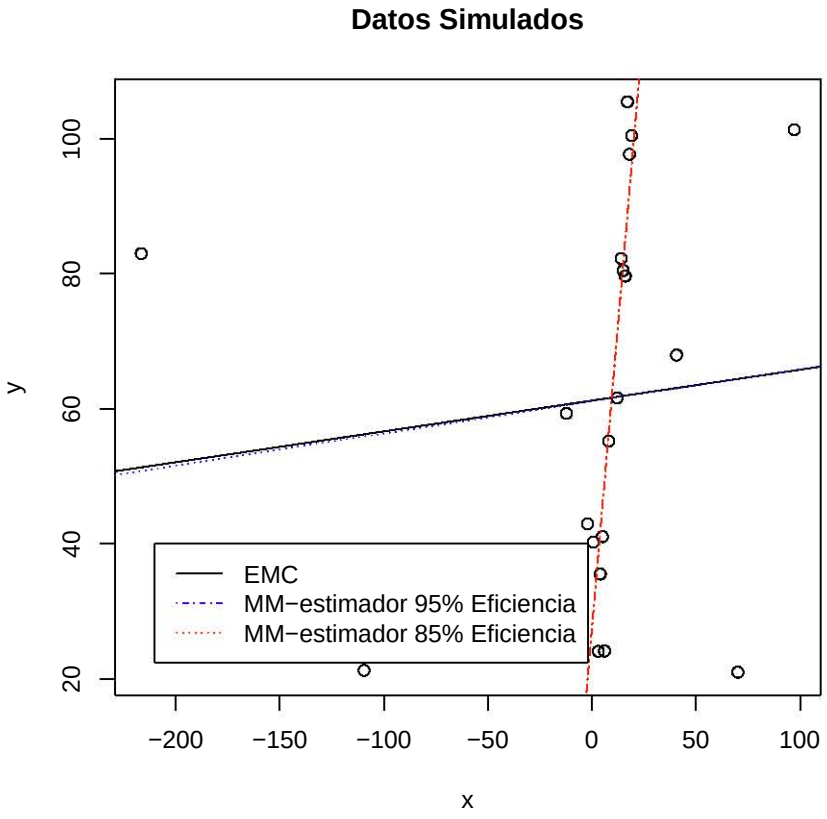


FIGURA 4.12: Diferentes ajustes a datos simulados

En la figura (4.12) se puede observar como el MM-estimador con una eficiencia de 95 % se comporta de manera similar que el estimador de mínimos cuadrados y el MM-estimador con eficiencia de 85 % tiene un mejor comportamiento ya que sí describe el comportamiento de la mayoría de los datos.

4.8. Intervalos de confianza robustos

Dado que las observaciones atípicas afectan a la media simple $\bar{y}$ y la desviación estándar S, los intervalos de confianza para $\mu = E(y)$ basados en la teoría normal pueden ser poco confiables.

Consideremos el intervalo de confianza para μ, el cual se justifica por el supuesto de que las observaciones son independientes e identicamente distribuidas con distribución normal, este intervalo se basa en la cantidad pivotal

$$T = \frac{\bar{y} - \mu}{s/\sqrt{n}} \tag{4.10}$$

a partir de esta se obtiene el intervalo para el parámetro μ con un nivel de confianza de $1-\alpha$:

$$\bar{y} \pm t_{n-i,1-\alpha/2}\frac{s}{\sqrt{n}} \tag{4.11}$$

donde $t_{(n-i,1-\alpha)}$ es el cuantil $1-\alpha$ de la distribución t-Students con $n-i$ grados de libertad.

El caso más simple es cuando se tiene una distribución de los datos simétrica respecto $\mu = E(y)$. Entonces $E(y)$ y el intervalo está centrado. En el caso de una distribución con colas pesadas la distribución puede originar que se incremente s y por lo tanto la longitud del intervalo también será afectado. Si los datos provienen de una distribución $(1-\epsilon)N(\mu,\sigma^2)+\epsilon H$ el intervalo $(1-\alpha)X100\,\%$ de confianza fallará en la robustez de la confianza y la longitud del intervalo.

Basándonos en la construcción de los intervalos de confianza basados en la t de Student para un estimador podemos obtener intervalos de confianza robustos de tal manera que no sean afectados por las observaciones atípicas. En general, los intervalos de confianza robustos los obtendremos sustituyendo la media y la desviación estándar por estimadores robustos de localización y de escala.

Consideremos el M-estimador $\hat{\mu}$ obtenido a partir de la función (3.4.0.3), si el elemento aleatorio de que provienen las observaciones es una función simétrica entonces para una n grande la distribución de $\hat{\mu}$ es aproximadamente $N(\mu, n/v)$ donde v se define como

$$v = \sigma^2\frac{E\left(\psi(y-\mu/\sigma)^2\right)}{E\left(\psi(y-\mu/\sigma)\right)^2} \tag{4.12}$$

Dado que v es desconocida un estimador $\hat{v}$ se puede obtener remplazando las esperanzas por el promedio simple, por lo que el parámetro puede estimarse de la siguiente manera

$$\hat{v} = \hat{\sigma}^2\frac{\text{average}\left[\psi(y-\hat{\mu}/\hat{\sigma})^2\right]}{\text{average}\left[\psi(y-\hat{\mu}/\hat{\sigma})\right]^2} \tag{4.13}$$

A partir de este estimador podemos definir una versión robusta de la cantidad pivotal para los M-estimadores de la siguiente manera

$$T = \frac{\hat{\mu}-\mu}{\sqrt{\hat{v}/n}} \tag{4.14}$$

de esta manera, su distribución es aproximadamente $N(0,1)$ cuando n es grande. Por lo tanto, una aproximación de un intervalo robusto se puede calcular como

$$\hat{\mu} \pm Z_{1-\alpha/2}\sqrt{\hat{v}/n}$$

donde $Z_{1-\alpha/2}$ denota el cuantil $1-\alpha/2$ de la distribución $N(0,1)$

4.9. Ejemplo MM-estimadores

A continuación se evaluará el comportamiento de los principales estimadores presentados a lo largo de este capítulo con la intención de mostrar las ventajas que presentan los MM-estimadores sobre los estimadores obtenidos previamente.

Se utilizará un conjunto de datos generados para este propósito por Hawkins, Bradu, y Kass en 1984 ("Artificial"), de tal manera que se conozca a priori cuáles son los datos atípicos y poder evaluar el desempeño de los diferentes estimadores.

Los datos "Artificial" se encuentran en la paquetería "robustbase" y consisten de 75 observaciones con tres variables explicativas. Las primeras 14 observaciones son datos atípicos, creados en dos grupos: las observaciones 1-10 son observaciones con un punto de palanca alto y que se encuentran fuera del patrón seguido por la mayoría de las observaciones, las observaciones 11-14 también tienen un punto de palanca alto pero sí son "buenas" observaciones. Con fines ilustrativos, agregaremos 5 observaciones de tal manera que no tengan un apalancamiento alto pero que sean datos atípicos. Las observaciones se muestran en la siguiente tabla.

Obs	x1	x2	x3	y	Obs	x1	x2	x3	y
1	2.6	0	1.9	23.8	**41**	0.1	2.2	2.7	-1
2	3.3	0.6	1.1	24.1	**42**	0.1	3	2.6	-0.6
3	0.5	1.7	3.3	22	**43**	1.5	1.2	0.2	0.9
4	2	1.6	0.8	26.3	**44**	2.1	0	1.2	-0.7
5	0.1	3.4	0.3	25.4	**45**	0.5	2	1.2	-0.5
6	10.1	19.6	28.3	9.7	**46**	3.4	1.6	2.9	-0.1
7	9.5	20.5	28.9	10.1	**47**	0.3	1	2.7	-0.7
8	10.7	20.2	31	10.3	**48**	0.1	3.3	0.9	0.6
9	9.9	21.5	31.7	9.5	**49**	1.8	0.5	3.2	-0.7
10	10.3	21.1	31.1	10	**50**	1.9	0.1	0.6	-0.5
11	10.8	20.4	29.2	10	**51**	1.8	0.5	3	-0.4
12	10.5	20.9	29.1	10.8	**52**	3	0.1	0.8	-0.9
13	9.9	19.6	28.8	10.3	**53**	3.1	1.6	3	0.1
14	9.7	20.7	31	9.6	**54**	3.1	2.5	1.9	0.9
15	9.3	19.7	30.3	9.9	**55**	2.1	2.8	2.9	-0.4
16	11	24	35	-0.2	**56**	2.3	1.5	0.4	0.7
17	12	23	37	-0.4	**57**	3.3	0.6	1.2	-0.5
18	12	26	34	0.7	**58**	0.3	0.4	3.3	0.7
19	11	34	34	0.1	**59**	1.1	3	0.3	0.7
20	3.4	2.9	2.1	-0.4	**60**	0.5	2.4	0.9	0
21	3.1	2.2	0.3	0.6	**61**	1.8	3.2	0.9	0.1
22	0	1.6	0.2	-0.2	**62**	1.8	0.7	0.7	0.7
23	2.3	1.6	2	0	**63**	2.4	3.4	1.5	-0.1
24	0.8	2.9	1.6	0.1	**64**	1.6	2.1	3	-0.3
25	3.1	3.4	2.2	0.4	**65**	0.3	1.5	3.3	-0.9
26	2.6	2.2	1.9	0.9	**66**	0.4	3.4	3	-0.3
27	0.4	3.2	1.9	0.3	**67**	0.9	0.1	0.3	0.6
28	2	2.3	0.8	-0.8	**68**	1.1	2.7	0.2	-0.3
29	1.3	2.3	0.5	0.7	**69**	2.8	3	2.9	-0.5
30	1	0	0.4	-0.3	**70**	2	0.7	2.7	0.6
31	0.9	3.3	2.5	-0.8	**71**	0.2	1.8	0.8	-0.9
32	3.3	2.5	2.9	-0.7	**72**	1.6	2	1.2	-0.7
33	1.8	0.8	2	0.3	**73**	0.1	0	1.1	0.6
34	1.2	0.9	0.8	0.3	**74**	2	0.6	0.3	0.2
35	1.2	0.7	3.4	-0.3	**75**	1	2.2	2.9	0.7
36	3.1	1.4	1	0	**76**	2.2	2.5	2.3	0.2
37	0.5	2.4	0.3	-0.4	**77**	0.6	2	1.5	-0.2
38	1.5	3.1	1.5	-0.6	**78**	0.3	1.7	2.2	0.4
39	0.4	0	0.7	-0.7	**79**	0	2.2	1.6	-0.9
40	3.1	2.4	3	0.3	**80**	0.3	0.4	2.6	0.2

Dado que se trata de un conjunto de datos con más de dos variables explicativas no es posible obtener gáficamente el comportamiento de las observaciones de manera agregada

y lo debemos de visualizar de manera marginal por variable. En la Figura (4.13) se puede observar el comportamiento de los datos y se pueden distinguir claramente los tres grupo de datos atípicos.

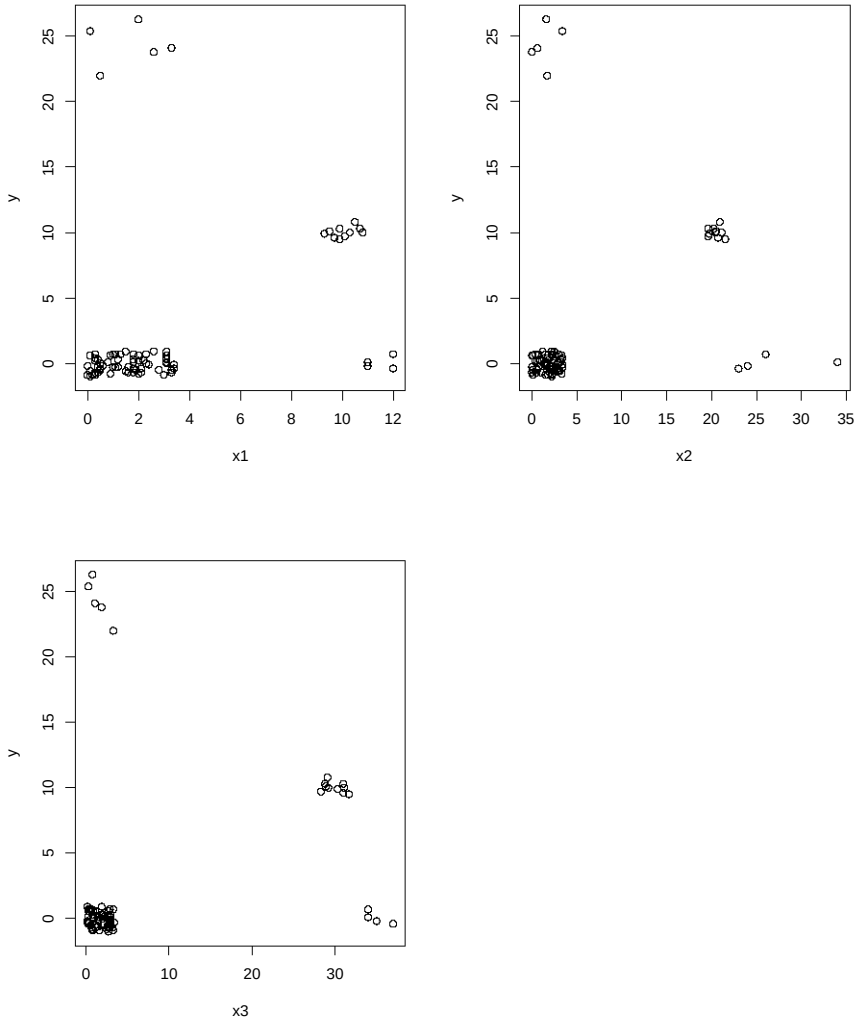


FIGURA 4.13: Comportamiento datos "Artificial"

Realizaremos el análisis de los datos utilizando los sigueintes estimadores: estimador por mínimos cuadrados, M-estimador de Tukey, GM-estimador de Mallows y el MM-estimador (función Biponderada como *ρ-función*).

Realizando el ajuste de los datos basado en los estimadores por mínimos cuadrados obtenemos los siguientes resultados:

```
Call:
lm(formula = artificial_$y ~ artificial_$x1 + artificial_$x2 +
    artificial_$x3)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0565 -2.5104 -1.6271 -0.8488 25.2855

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.2666     1.1241   1.127    0.263
artificial_$x1   0.5541     0.6906   0.802    0.425
artificial_$x2  -0.3776     0.4252  -0.888    0.377
artificial_$x3   0.2548     0.3546   0.718    0.475

Residual standard error: 6.364 on 76 degrees of freedom
Multiple R-squared:  0.09521,        Adjusted R-squared:  0.0595
F-statistic: 2.666 on 3 and 76 DF,  p-value: 0.05374
```

Para identificar el desempeño de los diferentes estimadores, se utilizarán las gráficas de residuales para mostrar si los estimadores logran identificar las observaciones erróneas. En la Figura (4.14) se muestra el comportamiento de los residuales calculados a partir de los valores ajustado por los estimadores por mínimos cuadrados.

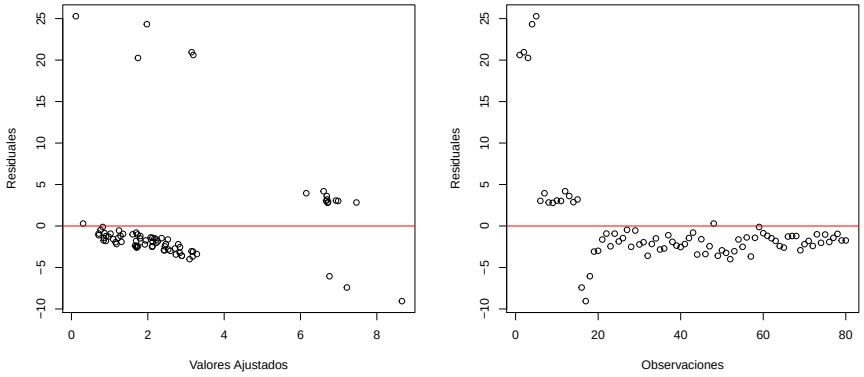


FIGURA 4.14: Gráfica de residuales, estimadores por mínimos cuadrados

Como se esperaba, el ajuste por mínimos cuadrados se distorsiona completamente y no logra explicar el patrón de los datos. Considerando los resultados de R, se observa que ninguna variable es significativa y además, de acuerdo al p-value>0.05, el modelo no tiene ningún sentido ya que no se rechaza la hipótesis nula ($H_0 : \boldsymbol{\beta} = 0$), lo que indica que la variable independiente no está influenciada por las variables regresoras.

Ahora realizaremos el ajuste considerando el M-estimador Biponderado. Lo que se espera observar en las gráficas de residuales en las primeras 15 observaciones sean las que tengan

los residuales más grandes y que los residuales del resto de las observaciones se ubiquen en una banda alrededor del cero.

```
Call: rlm(formula = artificial_$y ~ artificial_$x1 + artificial_$x2 +
    artificial_$x3, psi = "psi.bisquare")
Residuals:
       Min         1Q     Median         3Q        Max
-13.357220  -0.550372  -0.005114   0.594064  26.485121

Coefficients:
               Value   Std. Error t value
(Intercept)    -0.9546  0.1326    -7.1969
artificial_$x1  0.1575  0.0815     1.9331
artificial_$x2  0.1951  0.0502     3.8897
artificial_$x3  0.1778  0.0418     4.2499

Residual standard error: 0.866 on 76 degrees of freedom
```

El comportamiento de los residuales generados con M-estimador de Tukey se muestra en la siguiente figura

```
Robust Regression with Huber Function
Convergence achieved after: 11 iterations
```

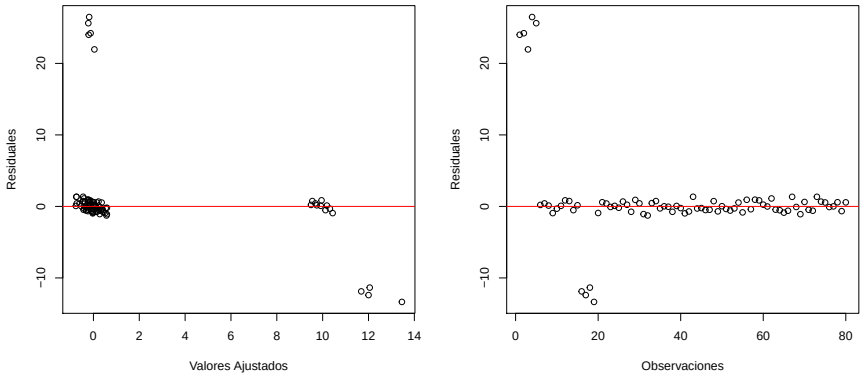


FIGURA 4.15: Gráfica de residuales, M-estimador Biponderado

De acuerdo a lo que se presentó anteriormente, sabemos que los M-estimadores presentan problemas en la presencia de puntos palanca. Considerando las gráficas de la Figura (4.15) podemos observar que el estimador identifica algunos de los datos atípicos pero de manera errónea ya que confunde las observaciones malas con las buenas, en la gráfica de residuales se puede notar como los estimadores se ajustaron a las observaciones 6-15 y 19-80, sin embargo las observaciones 6-15 no siguen el patrón de la mayoría de los datos.

Los resultados obtenidos utilizando el GM-estimador de Mallows son los siguientes:

```
(Intercept) artificial_$x1 artificial_$x2 artificial_$x3
 -0.7072856      0.2028548      0.0981677      0.2046726

Robust Regression with Huber Function
Convergence achieved after: 11 iterations
```

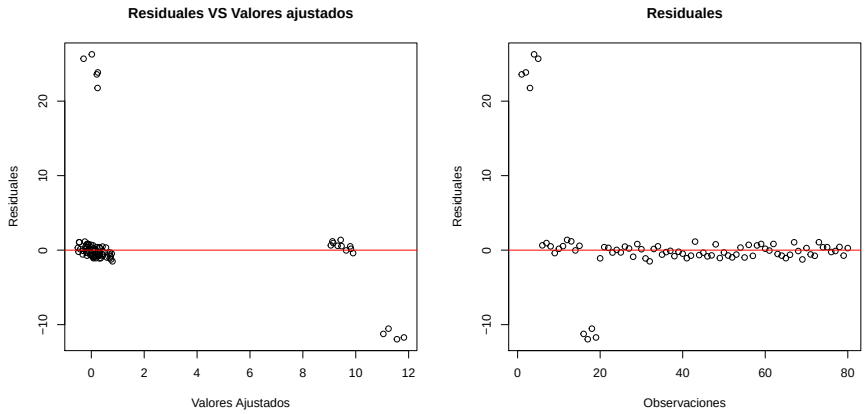


FIGURA 4.16: Gráfica de residuales, GM-estimador

El comportamiento del estimador de influencia acotada es muy similar al del M-estimador, se ajusta a las observaciones malas y pondera a las observaciones que son datos atípicos pero sí siguen el comportamiento de la mayoría de los datos.

Para entender porqué el GM-estimador se comporta igual que el M-estimador observemos a la función de pesos, $w_i(x_i) = \sqrt{1 - h_{ii}}$, utilizada por el estimador

Obs	$w_i(x_i)$	**Obs**	$w_i(x_i)$
1	0.9851	**11**	0.9623
2	0.9675	**12**	0.9657
3	0.9833	**13**	0.9681
4	0.9862	**14**	0.9595
5	0.9738	**15**	0.9561
6	0.9683	**16**	0.9520
7	0.9696	**17**	0.9263
8	0.9568	**18**	0.9442
9	0.9592	**19**	0.6749
10	0.9629		

La tabla anterior muestra que las observaciones que sólo son atípicas respecto a su valor en las variables explicativas tiene un apalancamiento mayor y es por esta razón que son más castigadas por el estimador.

Hasta ahora los diferentes estimadores no han logrado identificar de manera adecuada las observaciones erróneas, pero cómo se comporta el MM-estimador

```
Call: rlm(formula = artificial_$y ~ artificial_$x1 + artificial_$x2 +
    artificial_$x3, data = artificial_, scale.est = "MAD", method = "MM")
Residuals:
    Min      1Q  Median      3Q     Max
-0.9470 -0.3551  0.1504  0.8188 26.2877

Coefficients:
               Value   Std. Error t value
(Intercept)    -0.1974  0.1144    -1.7264
artificial_$x1  0.0951  0.0703     1.3537
artificial_$x2  0.0403  0.0433     0.9306
artificial_$x3 -0.0561  0.0361    -1.5543

Residual standard error: 0.8547 on 76 degrees of freedom
```

La gráfica de residuales son las siguientes:

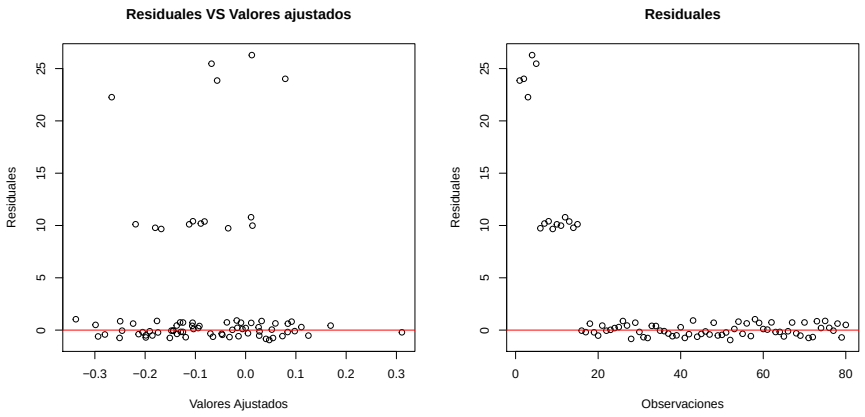


FIGURA 4.17: Gráfica de residuales, MM-estimador

En la Figura (4.17) se puede observar que el MM-estimador identifica de manera correcta a las observaciones atípicas (1-15) y genera un modelo que se ajusta a la mayoría de los datos

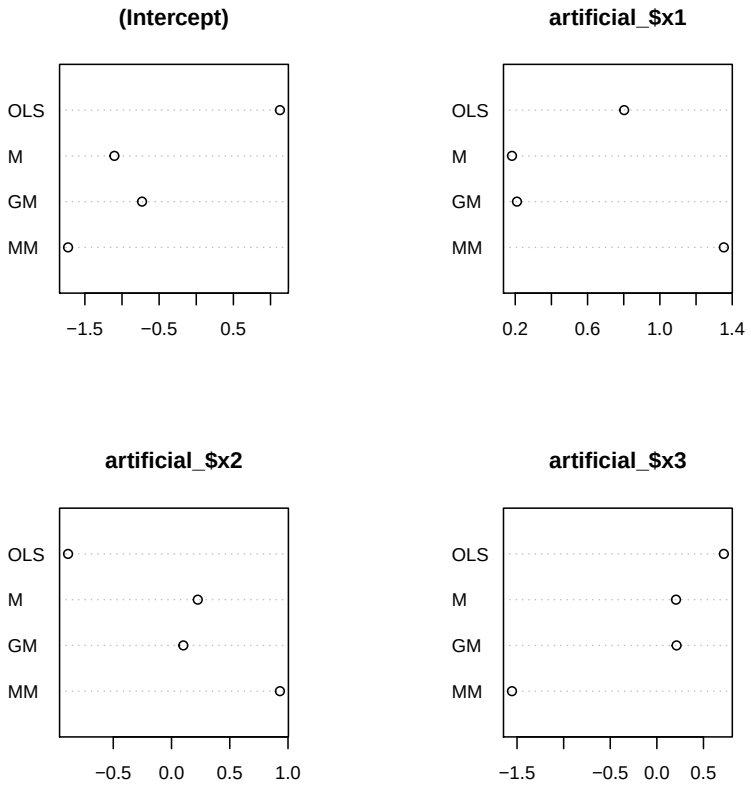


FIGURA 4.18: Comparativo de los diferentes estimadores

Conclusiones

En el desarrollo de este trabajo se presentaron las ventajas de los diferentes métodos de regresión robusta desde los M-estimadores hasta los MM-estimadores, los cuales logran las propiedades deseadas para los estimadores robustos: un elevado punto de ruptura, una alta eficiencia y una influencia acotada. Con ejemplos de datos de estudios reales se mostró las afectacciones que tienen los estimadores por mínimos cuadrados (estimadores óptimos en condiciones idóneas pero que difícilmente se cumplen en la vida real) por la presencia de outliers. Considerando el avance tecnológico, la facilidad en el cómputo de los estimadores como principal argumento para el uso general del método por mínimos cuadrados ha quedado atrás. En el capítulo 3 se ha presentado un algoritmo iterativo para obtener una aproximación de los diferentes estimadores robustos con el fin de mostrar cómo se logra la resistencia a las observaciones atípicas, el grado de eficiencia deseado y la importancia de considerar una medida de escala para obtener un estimador estable. Sin embargo, se puede hacer uso de funciones desarrolladas en diversas paqueterías del software estadístico R.

Además de las ventajas que pueden presentar las técnicas de regresión robusta cuando no se sabe si los datos contienen observaciones atípicas, en los diferentes ejemplos se mostró la utilidad de los estimadores robustos como herramienta para la detección de observaciones atípicas, de tal manera que ayudan a la identificación de aquellas observaciones que tienen un comportamiento diferente a la mayoría de los datos con el fin de que se realice un estudio más detallado sobre estas observaciones. Considerando la fuerza que insertan los estimadores robustos a las técnicas de diagnóstico de regresión, se puede considerar el enfoque robusto o el enfoque clásico de mínimos cuadrados, sin considerar las observaciones identificadas como atípicas, con el fin de obtener el modelo que explique de mejor manera el comportamiento del fenómeno estudiado.

Con el objetivo de resolver los problemas o debilidades los métodos robustos desarrollados desde los años 60's ante diversas condiciones presentadas en fenómenos de la vida real, se desarrollaron los MM-estimadores, estimadores presentados por Victor Yohai con propiedades de robustez deseadas y que no requieren de un gasto computacional

importante para su determinación. En el capítulo 4 se mostraron las ventajas de estos estimadores respecto a la resistencia y la identificación de observaciones atípicas, pero también se mostró que pueden presentar problemas al querer obtener una alta eficiencia respecto a los estimadores por mínimos cuadrados.

Los datos "Artificial" de la paquetería "robustbase" de R, gracias a su construcción, nos permitieron ilustrar el desempeño de los MM-estimadores y compararlo con otros estimadores.

Anexo I: Códigos de R

I.I Método de mínimos cuadrados ponderados - Capítulo 3

```
x<-matrix(ncol=2,nrow=8)
x[,1]<-c(1,1,1,1,1,1,1,1)
x[,2]<-c(1,2,3,4,5,6,7,8)
y<-c(0.8649,0.3878,2.3107,2.2335,4.1564,2.8793,6.0022,-6.0748)

B_est<-matrix(nrow=50,ncol=2)
r<-matrix(nrow=50,ncol=8)
B_est[1,]<-solve(t(x)%*%x)%*%t(x)%*%y #estimador MC
r[1,]<-y-x%*%B_est[1,]
w<-c((ifelse(abs(r[1,])>=1.345,1.345/abs(r[1,]),1))) #Pesos
W<-diag(w)

c<-1
i=1
while (c>0.0001)
{
 i=i+1
 i-1
 B_est[i,]<-solve(t(x)%*%W%*%x)%*%t(x)%*%W%*%y
 r[i,]<-y-x%*%B_est[i,]
 B_est[i,]

 #Convergencia?
 norm<-function(a){sqrt(sum(a^2))}
 c<-norm(B_est[i,]-B_est[i-1,])/norm(B_est[i,])
 ifelse(c<0.0001,"Se cumple el criterio de convergencia?: Si",
                 "Se cumple el criterio de convergencia?: No")
 c<-norm(r[i,]-r[i-1,])/norm(r[i,])
 ifelse(c<0.0001,
        "Se cumple el criterio de convergencia respecto residuales?: Si",
        "Se cumple el criterio de convergencia respecto residuales?: No")
```

```
 w<-c((ifelse(abs(r[i,])>=1.345,1.345/abs(r[i,]),1))) #Pesos
 W<-diag(w)
}
```

I.II Ejemplo: Datos "Phones" - Capítulo 4

```
library(MASS)
data(phones)
attach(phones)
ols<-lm(calls~year)
summary(ols)

fit.hub<-rlm(calls~year,maxit=50)
fit.tuk<-rlm(calls~year,psi="psi.bisquare")

par(mfrow=c(1,1))
plot(year,calls)
abline(ols\$coef,lty=1,col=c("BLACK"))
abline(fit.hub\$coef,lty=4,col=c("Blue"))
abline(fit.tuk\$coef,lty=3,col=c("RED"))
legend(50,200,c("mean","Hubers - MAD","Tukey"),lty=c(1,4,3),
col=c("BLACK","blue","red"))

par(mfrow=c(2,2))
plot(ols,1:2)
plot(ols,4)
h.phones<-hat(model.matrix(ols))
cook.phones<-cooks.distance(ols)
plot(h.phones/(1-h.phones),cook.phones, xlab="h/(1-h)",ylab="Cook Distance")

fit.hub<-rlm(calls~year,maxit=50)
summary(fit.hub)

dia.check.rlm <- function(fit,fitols){
  h.rlm<-hat(model.matrix(fitols))
  cook.d<-cooks.distance(fitols)
  par(mfrow=c(2,2))
obs<-fit\$fit+fit\$res
plot(obs,fit\$fit,xlab="response",ylab="fitted",main="obs vs fitted")
abline(0,1)
plot(obs,fit\$res,xlab="fitted",ylab="residual",main="fitted vs residual")
```

```
qqnorm(fit\$res, main="residuals")
qqline(fit\$res)
plot(fit\$w,ylab="fit weight", main="weights")
par(mfrow=c(1,1))
invisible()
}
dia.check.rlm(fit.hub,ols)

fit.tuk<-rlm(calls~year,psi="psi.bisquare")
summary(fit.tuk,cor=F)
dia.check.rlm(fit.tuk,ols)
```

I.III Ejemplo: Datos "Stars CYG" - Capítulo 4

```
library(MASS)
library(robustbase)
library(mvtnorm)
library(pcaPP)
library(rrcov)

data(starsCYG)
attach(starsCYG)

plot(log.Te,log.light, xlab=c("Luz"),ylab="Temperatura",
      main="Datos \"Stars CYG\"")
ols<-lm(log.light~log.Te)
abline(ols\$coef,lty=1)

fit.hub<-rlm(log.light~log.Te,maxit=50)
summary(fit.hub)

dia.check.rlm <- function(fit,fitols){
  h.rlm<-hat(model.matrix(fitols))
  cook.d<-cooks.distance(fitols)
par(mfrow=c(2,2))
obs<-fit\$fit+fit\$res
plot(obs,fit\$fit,xlab="Valor Observado",
        ylab="Valor Ajustado",main="Observaciones vs Ajuste")
identify(obs,fit\$fit)
abline(0,1)
plot(obs,fit\$res,xlab="Valor Ajustado",
        ylab="Residual",main="Residual vs Ajuste")
```

```
identify(obs,fit\$res)
  qqnorm(fit\$res, main="Residuales")
identify(fit\$res)
qqline(fit\$res)
par(mfrow=c(1,1))
invisible()
}

dia.check.rlm(fit.hub,ols)

fit.tuk<-rlm(log.light~log.Te,psi="psi.bisquare")
summary(fit.tuk,cor=F)
dia.check.rlm(fit.tuk,ols)

plot(fit.tuk\$w,ylab="Peso", xlab="Observaci'on",
      main="Peso asignado",pch=3,col="BLUE")
points(fit.hub\$w,pch=1,col="BLACK")
legend(40,.82,c("Tukey","Huber"),pch=c(3,1),col=c("BLUE","BLACK"))

par(mfrow=c(1,1))
plot(log.Te,log.light, xlab=c("Luz"),ylab="Temperatura")
abline(ols\$coef,lty=1,col=c("BLACK"))
abline(fit.hub\$coef,lty=4,col=c("Blue"))
abline(fit.tuk\$coef,lty=3,col=c("RED"))
legend(3.5,4.5,c("OLS","Huber","Tukey"),lty=c(1,4,3),
col=c("BLACK","blue","red"))
```

I.IV Ejemplo: Datos "Artificial" - Capítulo 4

```
library(MASS)
library(robustbase)
library(mvtnorm)
library(pcaPP)
library(rrcov)
library(circular)
library(wle)
library(rrcov)
library(car)
library(robustreg)

data(artificial)
attach(artificial)
```

```
par(mfrow=c(2,2))
plot(artificial\$x1,artificial\$y, xlab="x1",ylab="y")
plot(artificial\$x2,artificial\$y, xlab="x2",ylab="y")
plot(artificial\$x3,artificial\$y, xlab="x3",ylab="y")
par(mfrow=c(1,1))

fit.ols<-lm(
    artificial\$y~artificial\$x1+artificial\$x2+artificial\$x3)
summary(fit.ols)
par(mfrow=c(1,2))
plot(fit.ols\$fitted.value,fit.ols\$residuals,
     xlab="Valores Ajustados", ylab="Residuales")
abline(h=0,lty=1,col="RED")

plot(x=c(1:80),y=fit.ols\$residuals,
     ylab="Residuales", xlab="Observaciones")
abline(h=0,lty=1,col="RED")

fit.tuk<-rlm(
    artificial\$y~artificial\$x1+artificial\$x2+artificial\$x3,
    psi="psi.bisquare")
summary(fit.tuk)
par(mfrow=c(1,2))
plot(fit.tuk\$fitted.value,fit.tuk\$residuals,
     xlab="Valores Ajustados", ylab="Residuales")
abline(h=0,lty=1,col="RED")
plot(x=c(1:80),y=fit.tuk\$residuals,
     ylab="Residuales", xlab="Observaciones")
abline(h=0,lty=1,col="RED")

fit.gm.hub<-robustRegH(
  artificial\$y~artificial\$x1+artificial\$x2+artificial\$x3,
    data=artificial,m=FALSE)

fit.gm.hub<-robustRegH(
  artificial\$y~artificial\$x1+artificial\$x2+artificial\$x3,
  data=artificial,m=FALSE)
x.mat<-cbind(1,x.artificial_)
fitted.value<-fit.gm.hub\$coef%*%t(x.mat)
fitted.value<-t(fitted.value)
residual.value<-y.artificial-fitted.value

par(mfrow=c(1,2))
```

```
plot(fitted.value,residual.value, main="Residuales VS Valores ajustados",
     xlab="Valores Ajustados", ylab="Residuales")
abline(h=0,lty=1,col="RED")

plot(x=c(1:80),y=residual.value, main="Residuales",
     ylab="Residuales", xlab="Observaciones")
abline(h=0,lty=1,col="RED")

par(mfrow=c(1,1))

fit.mm<-rlm(
  artificial\$y~artificial\$x1+artificial\$x2+artificial\$x3,
  data=artificial, method="MM",scale.est ="MAD")
summary(fit.mm)

par(mfrow=c(1,2))
plot(fit.mm\$fitted.value,fit.mm\$residuals,
    main="Residuales VS Valores ajustados",
     xlab="Valores Ajustados", ylab="Residuales")
abline(h=0,lty=1,col="RED")

plot(x=c(1:80),y=fit.mm\$residuals, main="Residuales",
     ylab="Residuales", xlab="Observaciones")
abline(h=0,lty=1,col="RED")


s0<-1.4826*median(abs(residual.value))
par(mfrow=c(2,2), pty="s", cex=0.8)
for(i in 1:4)
  { vet<-c(
    fit.mm\$coef[i] / sqrt(diag(vcov(fit.mm)))[i],
    fit.gm.hub\$coef[i] / s0,
    fit.tuk\$coef[i] / fit.tuk\$s,
    fit.ols\$coef[i] / sqrt(diag(vcov(fit.ols)))[i])
    names(vet)<-c("MM", "GM", "M", "OLS")
  dotchart(vet, main=names(fit.ols\$coef)[i], cex=.79)}
```

Bibliografía

[1] Andersen, R.: *Modern Methods For Robust Regression.* SAGE Publications, 2008

[2] Bellio, R. y Ventura, L.: *An Introduction to Robust Estimation with R Functions*, Department of Statistics, University of Udine y Department of Statistics, University of Padova, 2005.

[3] Huber, P.: *Robust estimation of a location parameter*, The Annals of Mathematical Statistics, 1964.

[4] Huber, P.: *Robust Statistics*, John Wiley & Sons, Inc, 1981.

[5] Maronna, R.: *Estimación Robusta en el Modelo Lineal*, Sociedad Chilena de Estadística, 1996.

[6] Maronna, R. y Martin, R. and Yohai, V.: *Robust Statistics: Theory and Methods*, John Wiley & Sons, Ltd, 2006.

[7] Martin, D. y College,D.: *Teaching Leverage, Outliers, and Influential Observations in Introductory Statistics Courses.*

[8] Olive, D.: *Applied Robust Statistics*, Department of Mathematics, Southern Illinois University, 2005.

[9] Rousseeuw, P. y Croux, C.: *Alternatives to the Median Absolute Deviation*, American Statistical Association, 1993.

[10] Rousseeuw, P. y Leroy, A.: *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc, 1987.

[11] Rousseeuw, P. and Yohai, V.: *Robust regression by means of S-estimators*, Robust and Nonlinear Time Series, 1984.

[12] Stuart, C.: *Robust Regression*, 2011.

[13] Venables, W. y Ripley, B.:, *Modern Applied Statistics with S*, Springer, 2002.

[14] Zamar, R.: *Estimación Robusta*, Estadistica Española, 1981.