7

Una recta resistente

Para ajustar una linea recta de la forma

$$y = a + bx$$

a un conjunto de datos $(x_i, y_i), i = 1, \ldots, n$ se han desarrollado varios métodos a lo largo de la historia. La regresión por mínimos cuadrados que hemos explicado es el método más conocido y más ampliamente utilizado. Es un método que involucra cálculos algebraicamente simples, utiliza la inferencia deducida para la distribución normal y requiere únicamente una derivación matemática sencilla. Desgraciadamente, la recta de regresión mínimo-cuadrática no es resistente. Un dato "salvaje" puede tomar fácilmente el control de la recta ajustada y conducirnos a conclusiones engañosas sobre la relación entre y y x. La llamada *recta resistente de los tres grupos* evita esta dificultad. Así, esta recta es muy útil en la exploración de los datos y-versus-x.

A continuación exponemos las principales ideas en este tema del clásico libro *Understanding Robust and Exploratory Data Analysis* de Hoaglin, Mosteller y Tukey [40].

7.1. Recta resistente de los tres grupos

7.1.1. Formación de los tres grupos

Empezaremos por ordenar los valores x de manera que $x_1 \leq x_2 \leq \cdots \leq x_n$. Entonces, sobre la base de estos valores ordenados, dividiremos los n puntos (x_i, y_i) en tres grupos: un grupo izquierdo, un grupo central y un grupo derecho, de tamaño tan igual como sea posible. Cuando no hay repeticiones en les x_i, el número de puntos en cada uno de los tres grupos depende del residuo de la división de n por 3:

Grupo	$n = 3k$	$n = 3k + 1$	$n = 3k + 2$
Izquierdo	k	k	$k + 1$
Central	k	$k + 1$	k
Derecho	k	k	$k + 1$

Repeticiones de los x_i nos harán estar alerta para formar tres conjuntos que no separen los puntos con igual x en conjuntos diferentes. Un examen detallado del tratamiento de las repeticiones nos puede llevar incluso a formar únicamente dos grupos. Cuando cada uno de los tercios ha sido definitivamente formado, determinaremos las dos coordenadas de unos puntos centrales, uno para cada grupo, con la mediana de los valores de las x y la mediana de los valores de las y, por separado. Etiquetaremos las coordenadas de estos tres puntos centrales con las letras I de izquierda, C de centro i D de derecha:

$$(x_I, y_I), (x_C, y_C), (x_D, y_D)$$

La figura 7.1 muestra los puntos observados y los puntos centrales de un ejemplo hipotético con 9 puntos. Como se ve en este gráfico, ninguno de los puntos centrales coincide con un punto de los

datos, ya que las medianas de les x y de las y se han calculado separadamente. A pesar de ello, los tres podrían ser puntos observados, como ocurre a menudo, cuando las x y las y siguen el mismo orden.

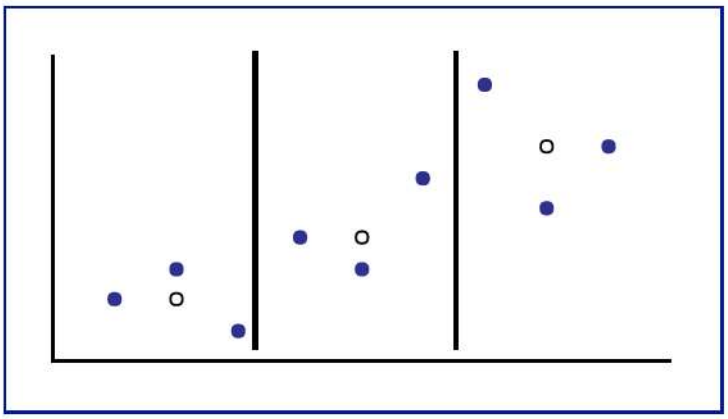

Figura 7.1: Puntos observados y puntos centrales en un ejemplo hipotético.

Este sistema de determinación de los puntos centrales de cada grupo es el que da a la recta que calcularemos su resistencia. Cuanto mayor es el número de puntos observados en cada grupo, la mediana proporciona la resistencia a los valores influyentes de x, y o ambos.

7.1.2. Pendiente e intercepción

Ahora utilizaremos los puntos centrales para calcular la pendiente b y la ordenada en el origen o intercepción a de la recta $y = a + bx$ que ajusta los valores observados y permite la predicción de los valores x_i observados y cualquier otro valor apropiado de x. En este sentido, la pendiente b nos dice cuantas unidades de y cambian por una unidad de x. Es razonable obtener esta información de los datos, en concreto de los puntos centrales de los grupos izquierdo y derecho:

$$b_0 = \frac{y_D - y_I}{x_D - x_I}$$

La utilización de los dos puntos centrales de los grupos extremos nos da la ventaja de medir el cambio de y sobre un intervalo bastante ancho de x, siempre que hayan suficientes puntos observados en estos grupos para asegurar la resistencia.

Cuando tomamos la pendiente b_0 para ajustar el valor y de cada punto central, la diferencia es el valor de la intercepción de una linea con pendiente b_0 que pasa exactamente por este punto. La intercepción ajustada es la media de estos tres valores:

$$a_0 = \frac{1}{3}[(y_I - b_0 x_I) + (y_C - b_0 x_C) + (y_D - b_0 x_D)]$$

De nuevo, como los puntos centrales están basados en la mediana, a_0 es resistente.

El ajuste de una recta en términos de pendiente e intercepción es convencional, pero usualmente artificial. La intercepción, que da el valor de y cuando $x = 0$, puede ser determinada de forma imprecisa, especialmente cuando los valores de x están todos muy alejados del cero y el cero es un valor sin sentido en el rango de las x. Ajustar la recta en términos de pendiente y un valor central de las x, como la media, la mediana o x_C, es mucho más útil. Nosotros escogeremos x_C por conveniencia y entonces la recta inicial es

$$y = a_0^* + b_0(x - x_C)$$

donde b_0 es la de antes y el valor central (también llamado *nivel*) es

$$a_0^* = \frac{1}{3}[(y_I - b_0(x_I - x_C)) + y_C + (y_D - b_0(x_D - x_C))]$$

Como ahora explicaremos, esta recta se toma como punto de partida para ajustar una mejor con iteraciones sucesivas.

7.1.3. Ajuste de los residuos e iteraciones

Una vez que hemos obtenido la pendiente y el nivel de la recta ajustada, el siguiente paso es calcular los residuos para cada punto

$$r_i = y_i - [a^* + b(x_i - x_C)]$$

Los gráficos de los residuos son muy útiles en la evaluación del ajuste y para descubrir patrones de comportamiento inesperados. Pero ahora, de momento, resaltaremos una propiedad general de todo conjunto de residuos, en nuestro problema actual o en situaciones más complejas:

> Si substituimos los valores originales de y por los residuos, es decir, si utilizamos (x_i, r_i) en lugar de (x_i, y_i), $i = 1, \ldots, n$ y repetimos el proceso de ajuste, llegaremos a un ajuste cero.

Para una linea recta esto significa que, con los puntos $(x_i, r_i), i = 1, \ldots, n$ como datos, obtendremos una pendiente cero y un nivel cero. En otras palabras, los residuos no contienen más aportación a la recta ajustada.

Una importante característica de los procedimientos resistentes es que habitualmente requieren iteraciones. Es el caso de la recta resistente de los tres grupos. Los residuos de la recta con la pendiente b_0 y el nivel a_0^* no tienen pendiente y nivel cero cuando hacemos el ajuste de la recta con las mismas x_i, aunque los nuevos valores de pendiente y nivel son substancialmente menores (en magnitud) que b_0 y a_0^*. Por esta razón, pensaremos en b_0 y a_0^* como los valores iniciales de una iteración.

El ajuste a una recta de los residuos obtenidos con la recta inicial da unos valores δ_1 y γ_1 a la pendiente y el nivel, respectivamente. En concreto, utilizaremos los residuos iniciales

$$r_i^{(0)} = y_i - [a_0^* + b_0(x_i - x_C)], \qquad i = 1, \ldots, n$$

en lugar de los y_i y repetiremos los pasos del proceso de ajuste. Como el conjunto de los x_i no ha cambiado, los tres grupos y las medianas de los x en los puntos centrales serán los mismos.

La pendiente y el nivel ajustados son $b_0 + \delta_1$ y $a_0^* + \gamma_1$ y los nuevos residuos

$$r_i^{(1)} = r_i^{(0)} - [\gamma_1 + \delta_1(x_i - x_C)], \qquad i = 1, \ldots, n$$

Ahora podemos avanzar con otra iteración. En general no sabremos si hemos conseguido un conjunto apropiado de residuos, hasta que verifiquemos el ajuste cero. En la práctica continuaremos las iteraciones hasta que el ajuste de la pendiente sea suficientemente pequeño en magnitud, del orden del 1 % o del 0.01 % del tamaño de b_0. Cada iteración añade su pendiente y su nivel a los valores previos

$$b_1 = b_0 + \delta_1, b_2 = b_1 + \delta_2, \ldots$$

y

$$a_1^* = a_0^* + \gamma_1, a_2^* = a_1^* + \gamma_2, \ldots$$

Las iteraciones son normalmente pocas y los cálculos no muy largos.

Tabla 7.1: Edad y altura de unos niños en una escuela privada.

Niño	Edad (meses)	Altura (cm)
1	109	137.6
2	113	147.8
3	115	136.8
4	116	140.7
5	119	132.7
6	120	145.4
7	121	135.0
8	124	133.0
9	126	148.5
10	129	148.3
11	130	147.5
12	133	148.8
13	134	133.2
14	135	148.7
15	137	152.0
16	139	150.6
17	141	165.3
18	142	149.9

Fuente: B.G. Greenberg (1953). "The use of analysis of covariance and balancing in analytical studies", American Journal of Public Health, 43, 692-699 (datos de la tabla 1, pág. 694).

Ejemplo 7.1.1

En una discusión en 1953, Greenberg consideró los datos de edad y altura de dos muestras de niños, una de una escuela privada urbana y la otra de una escuela pública rural. En la tabla 7.1 se reproducen los datos de los 18 niños de la escuela privada.

Aunque los datos no siguen claramente una línea recta, su patrón no es notablemente curvado y el ajuste a una línea puede resumir cómo la altura y crece con la edad x en este grupo de niños. Sólo los niños 13 *y* 17 *tienen puntos muy separados y veremos cómo influyen en el conjunto. Dado que* 18 *es divisible por* 3 *y los datos x no tienen repeticiones, cada grupo contiene seis puntos. Los puntos centrales de cada grupo son*

$$(x_I, y_I) = (115.50, 139.15)$$
$$(x_C, y_C) = (127.50, 147.90)$$
$$(x_D, y_D) = (138.00, 150.25)$$

de forma que el valor inicial de la pendiente es

$$b_0 = \frac{150.25 - 139.15}{138.00 - 115.50} = 0.4933$$

y el valor inicial del nivel

$$a_0^* = \frac{1}{3}[(139.15 - 0.4933(115.5 - 127.5)) + 147.9 + (150.25 - 0.4933(138 - 127.5))] = 146.0133$$

Los datos de la tabla 7.2 están ya ordenados en función de los valores de x = Edad y se han calculado los residuos de la recta inicial.

Para ver cómo van las iteraciones, calcularemos los primeros ajustes de la pendiente y del nivel

$$\delta_1 = \frac{-1.0500 - 0.5367}{138.00 - 115.50} = -0.0705$$
$$\gamma_1 = -0.1519$$

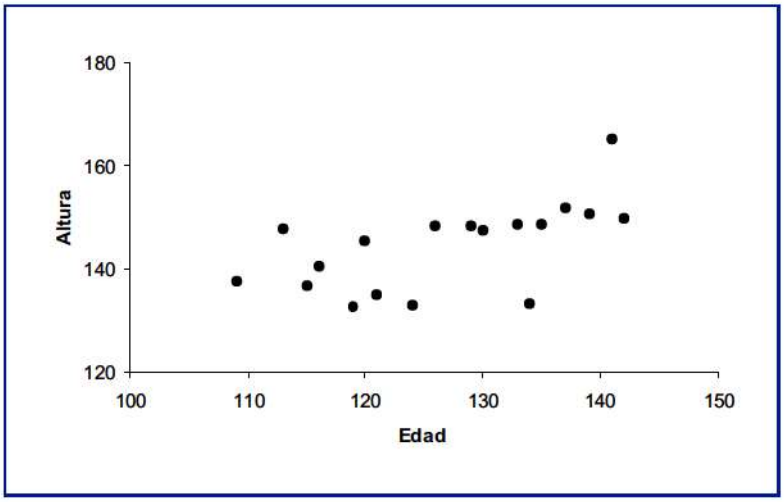

Figura 7.2: Altura versus edad para los niños de una escuela privada.

Tabla 7.2: Edad y altura de los niños en los tres grupos y residuos de la recta inicial

Niño	Edad (meses)	Altura (cm)	Residuo
1	109	137.6	0.7133
2	113	147.8	8.9400
3	115	136.8	−3.0467
4	116	140.7	0.3600
5	119	132.7	−9.1200
6	120	145.4	3.0867
7	121	135.0	−7.8067
8	124	133.0	−11.2867
9	126	148.5	3.2267
10	129	148.3	1.5467
11	130	147.5	0.2533
12	133	148.8	0.0733
13	134	133.2	−16.0200
14	135	148.7	−1.0133
15	137	152.0	1.3000
16	139	150.6	−1.0867
17	141	165.3	12.6267
18	142	149.9	−3.2667

Notemos que δ_1 es sustancialmente menor en magnitud que b_0, pero todavía no es negligible. Dos iteraciones más nos proporcionan unos valores para los que el proceso puede parar: $\delta_3 = -0.0006$ es menor que un 1 % de la pendiente acumulada.

La recta ajustada es

$$y = 145.8643 + 0.4285(x - 127.5)$$

La figura 7.3 representa los residuos de este ajuste. En general, el aspecto global es bastante satisfactorio. Sólo los dos puntos destacados, el del niño 13 y el del niño 17, se separan mucho y son atípicos. También hay tres residuos demasiado negativos para niños que tienen alrededor de 120 meses. Si tuviéramos más información, podríamos estudiar porqué estos niños son demasiado altos o demasiado bajos para su edad. Por ejemplo, podríamos separar los niños de las niñas.

En este ejemplo hemos visto cómo dos puntos, hasta cierto punto inusuales, han tenido muy poco efecto, si han tenido alguno, en el ajuste general de los datos. Una recta ajustada por el método de los mínimos

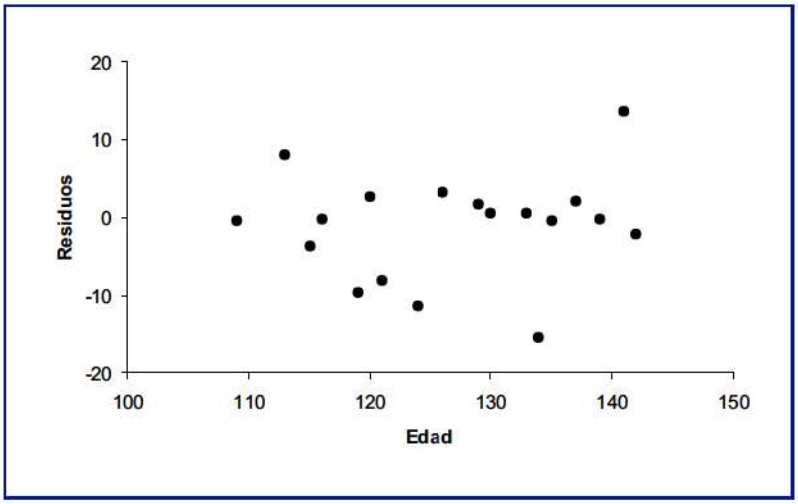

Figura 7.3: Residuos de la altura versus edad, después del ajuste por la recta resistente.

cuadrados corre mucho más riesgo de dejarse influenciar por estos puntos. Para estos datos la recta de regresión mínimo-cuadrática es

$$y = 79.6962 + 0.5113x$$

o

$$y = 144.8853 + 0.5113(x - 127.5)$$

donde observamos cómo los puntos 5, 7, 8 *y* 17 *han torcido la recta. Además, si el valor de y del punto* 13 *no fuera tan bajo, la recta mínimo-cuadrática podría ser más empinada. En todo caso, como la evaluación del ajuste se hace con los residuos, la figura* 7.4 *nos muestra los residuos mínimo-cuadráticos con la edad. Aunque es bastante similar al anterior, este gráfico nos da la sensación de una ligera tendencia a la baja. Es decir, los residuos mínimo-cuadráticos resultarían más horizontales si eliminàramos de ellos una recta con una pendiente ligeramente negativa.*

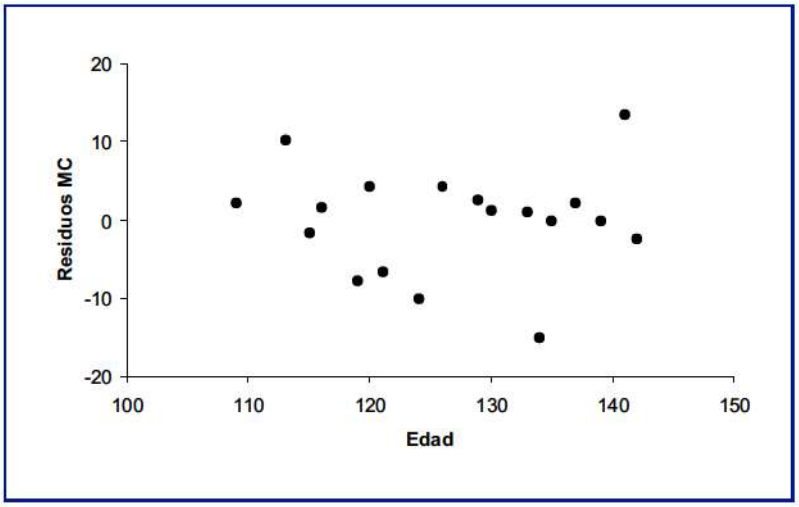

Figura 7.4: Residuos mínimo-cuadráticos versus edad.

En este ejemplo la variabilidad de los residuos merece más atención que la diferencia entre las pendientes de la recta de regresión mínimo-cuadrática y la recta resistente. Por ejemplo, la desviación estándar de los residuos mínimo-cuadráticos es 6.8188 *y el error estándar de la pendiente es* 0.1621, *sobre dos veces la diferencia entre las pendientes.*

Así hemos visto, cualitativamente, cómo algunos datos pueden afectar a la recta mínimo-cuadrática mucho más que la recta resistente. En todo caso, cuando los datos están razonablemente bien dispuestos las dos líneas son parecidas.

7.1.4. Mejora del método de ajuste

Para algunos conjuntos de datos, el procedimiento iterativo explicado para ajustar la recta resistente encuentra dificultades. Los ajustes de la pendiente pueden decrecer muy lentamente o, después de unos pocos pasos, dejar de decrecer y oscilar entre dos valores. Afortunadamente, una modificación elimina completamente estos problemas y permite que el número de iteraciones sea mucho más limitado.

La solución propuesta por Johnstone y Velleman (1982) es un procedimiento iterativo para el cálculo de la pendiente que asegura la convergencia hacia un valor único.

En el cálculo de la pendiente en la $j + 1$ iteración tenemos

$$\delta_{j+1} = \frac{r_D^{(j)} - r_I^{(j)}}{x_D - x_I}$$

y esto será 0 justamente cuando el numerador $r_D^{(j)} - r_I^{(j)} = 0$. Es decir, lo que debemos hacer es hallar el valor de b que proporciona la misma mediana a los residuos del grupo derecho y del grupo izquierdo. Más formalmente

$$\Delta r(b) = r_D(b) - r_I(b)$$

muestra la dependencia funcional de b y prescinde del número de la iteración. Buscamos el valor de b que hace $\Delta r(b) = 0$. Notemos que centraremos el proceso iterativo en b y dejaremos a para el final.

Empezaremos por calcular b_0 como antes y calcularemos $\Delta r(b_0)$ y δ_1 como ya sabemos. A continuación calcularemos $\Delta r(b_0 + \delta_1)$. Generalmente, $\Delta r(b_0)$ y $\Delta r(b_0 + \delta_1)$ tendrán signos opuestos, indicando que el valor deseado de b cae entre b_0 y $b_1 = b_0 + \delta_1$. Si pasa lo contrario, cuando $\Delta r(b_0)$ y $\Delta r(b_0 + \delta_1)$ tienen el mismo signo, hace falta seguir los pasos desde b_0 y $b_1 = b_0 + \delta_1$ hasta que hallamos un b_1 tal que $\Delta r(b_1)$ tiene el signo contrario a $\Delta r(b_0)$.

En este punto tenemos un b_0 con $\Delta r(b_0)$ y un b_1 con $\Delta r(b_1)$ y sabemos que Δr ha de ser 0 para algún valor b entre b_0 y b_1. (Este hecho y que la solución es única requieren una demostración formal que aquí no reproducimos.) Así que podemos continuar por interpolación lineal

$$b_2 = b_1 - \Delta r(b_1)\frac{b_1 - b_0}{\Delta r(b_1) - \Delta r(b_0)}$$

Cuando $\Delta r(b_2)$ no es todavía 0 (o suficientemente cerca de cero), hace falta repetir la interpolación con otro paso. Para hacer esto, consideraremos el intervalo que contiene b utilizando b_2 en lugar de b_1 o de b_0, el que tenga Δr con el mismo signo que $\Delta r(b_2)$. Y así los pasos necesarios.

7.2. Métodos que dividen los datos en grupos

Otras técnicas anteriores al método resistente de los tres grupos fueron propuestas e involucran la división de los datos en grupos. Algunos de estos métodos no pretenden ser una alternativa al método de los mínimos cuadrados y fueron desarrollados para ajustar una recta "cuando ambas variables están sujetas a error".

Método de Wald

Wald (1940) propuso dividir los datos en dos grupos de igual tamaño. Idealmente, los valores teóricos X_i del primer grupo son menores que los del segundo. En la práctica, porque los valores de X_i son desconocidos, agruparemos los puntos en base a los x_i observados.

Supongamos que n es par y sea $m = n/2$. Entonces, si asumimos que los valores de x están ordenados en orden creciente, la pendiente propuesta es

$$b_W = \frac{(y_{m+1} + \cdots + y_n) - (y_1 + \cdots + y_m)}{(x_{m+1} + \cdots + x_n) - (x_1 + \cdots + x_m)}$$

Si $x_{m+1} = x_m$, el método descarta los puntos con repetición en el centro.
El punto de intercepción es

$$a_W = \bar{y} - b_W \bar{x}$$

donde $\bar{y}$ y $\bar{x}$ son las medias totales, de la misma forma que en la recta mínimo-cuadrática.

Método de Nair y Shrivastava

Como una alternativa computacionalmente atractiva respecto al método de los mínimos cuadrados, Nair y Shrivastava (1942) introdujeron el método de las medias por grupo. Si ordenamos las x, podemos considerar un primer grupo con n_I puntos, un segundo grupo con n_D puntos y descartamos los $n - n_I - n_D$ restantes. Los puntos resumen de cada grupo son las medias

$$\bar{x}_I = \frac{x_1 + \cdots + x_{n_I}}{n_I} \qquad \bar{y}_I = \frac{y_1 + \cdots + y_{n_I}}{n_I}$$
$$\bar{x}_D = \frac{x_{n-n_D+1} + \cdots + x_n}{n_D} \qquad \bar{y}_D = \frac{y_{n-n_D+1} + \cdots + y_n}{n_D}$$

y la pendiente y el punto de intercepción resultan de la recta que pasa por $(\bar{x}_I, \bar{y}_I)$ y $(\bar{x}_D, \bar{y}_D)$

$$b_{NS} = \frac{\bar{y}_D - \bar{y}_I}{\bar{x}_D - \bar{x}_I}$$

$$a_{NS} = \bar{y}_I - b_{NS}\bar{x}_I = \bar{y}_D - b_{NS}\bar{x}_D$$

Para formar los grupos se puede tomar $n_I = n_D$ como el entero más próximo a $n/3$.

Método de Bartlett

Bartlett (1949) modificó los dos métodos anteriores con la propuesta

$$b_B = \frac{\bar{y}_D - \bar{y}_I}{\bar{x}_D - \bar{x}_I}$$
$$a_B = \bar{y} - b_B \bar{x}$$

de forma que la recta pasa por el punto $(\bar{x}, \bar{y})$.

Recta de Brown-Mood

La propuesta de Brown y Mood (1951) es un método diferente que utiliza la mediana de dos grupos. La pendiente b_{BM} y el punto de intercepción a_{BM} se calculan de forma que la mediana de los residuos en cada uno de los dos grupos sea cero:

$$\underset{x_i \le M_x}{\text{mediana}}\{y_i - a_{BM} - b_{BM}x_i\} = 0$$
$$\underset{x_i > M_x}{\text{mediana}}\{y_i - a_{BM} - b_{BM}x_i\} = 0$$

La inclusión de la mediana M_x en el primer grupo es arbitraria: el objetivo es que los dos grupos sean muy parecidos en su tamaño.
Para hallar los valores efectivos se propone un método iterativo similar al de las secciones anteriores.

7.3. Métodos que ofrecen resistencia

En la sección anterior hemos visto que la recta resistente de los tres grupos no fue la primera alternativa a la de los mínimos cuadrados. Incluso la última de las rectas propuestas, la recta de Brown-Mood, ofrece también resistencia. Ahora acabaremos esta breve descripción de técnicas con algunas que proporcionan como mínimo un cierto grado de resistencia. Pero primero debemos definir una medida de resistencia.

Una de las atractivas características de la recta resistente de los tres grupos es su habilidad para tolerar puntos "salvajes", es decir, puntos que son inusuales en su valor x o en su valor y o en ambos. Para medir esta resistencia aplicaremos el concepto de colapso (*breakdown*) introducido por Hampel (1971).

Definición 7.3.1

El punto de colapso *(breakdown bound) de un procedimiento para ajustar una recta a n parejas de datos y-versus-x es la proporción k/n, donde k es el mayor número de puntos que pueden ser reemplazados arbitrariamente mientras dejen la pendiente y el punto de intercepción delimitados.*

En la práctica, podemos pensar en enviar puntos al infinito al azar o en direcciones problemáticas hasta que la pendiente y el punto de intercepción no lo puedan tolerar más y se colapsen yendo también ellos hacia el infinito. Nos preguntamos cuan grande debe ser una parte de los datos para que un cambio drástico no afecte de forma considerable la recta ajustada.

Está claro que la recta mínimo-cuadrática tiene punto de colapso cero.

Dado que la recta resistente de los tres grupos usa la mediana dentro de cada grupo, hallaremos su punto de colapso en 1/3 veces el punto de colapso de la mediana de una muestra ordinaria. La mediana es el valor central, entonces su punto de colapso es 1/2, de manera que el punto de colapso de la recta resistente es 1/6. A pesar de las diversas posibilidades de construcción de los tres grupos y el hecho que los puntos salvajes pueden estar repartidos en los tres grupos, la idea es que 1/6 es lo mejor que podemos garantizar en la más desfavorable de las circunstancias.

Residuos mínimo-absolutos

Minimizar la suma de los residuos en valor absoluto tiene una historia casi tan larga como la del método de los mínimos cuadrados. Para ajustar una recta hace falta hallar b_{MA} y a_{MA} que minimicen

$$\sum_{i=1}^{n} |y_i - a_{MA} - b_{MA}x_i|$$

Al contrario que para los mínimos cuadrados, no hay una fórmula para calcular b_{MA} y a_{MA}. De hecho, la pendiente y el punto de intercepción pueden no ser únicos.

Como la mediana es la medida que minimiza

$$\sum_{i=1}^{n} |y_i - t|$$

hace falta esperar que este procedimiento tenga un alto punto de colapso. Desgraciadamente, este colapso es 0. La suma que se minimiza involucra tanto los valores x_i como los y_i y así es posible pensar en un punto (x_i, y_i) que tome el control de la recta.

Mediana de las pendientes por parejas

Otra forma de aplicar la mediana al ajuste de una recta consiste en determinar, para cada pareja de puntos, la pendiente y entonces calcular la mediana de estas pendientes. Con más cuidado, supongamos que los x_i son todos diferentes, definimos

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i} \qquad 1 \leq i < j \leq n$$

que son $n(n-1)/2$ valores. La pendiente ajustada es

$$b_T = \text{Med}\{b_{ij}\}$$

Este método es una propuesta de Theil (1950), mejorada por Sen (1968), para manejar las repeticiones de los x_i.

Para deducir el punto de colapso, supongamos que exactamente k de los n puntos son salvajes. Entonces el número de pendientes salvajes es

$$\frac{k(k-1)}{2} + k(n-k)$$

Si este número es suficientemente grande, b_T quedará descontrolada. Para valores de n grandes, podemos multiplicar el número de pendientes $n(n-1)/2$ por $1/2$, el punto de colapso de la mediana, y igualar con la expresión anterior. Si resolvemos la ecuación planteada para k, obtenemos un valor de k/n aproximadamente de 0.29. Esto quiere decir que el punto de colapso de b_T es 0.29.

Recta con medianas repetidas

Para conseguir un alto punto de colapso, Siegel (1982) ideó el método de las medianas repetidas.

Empezamos con las pendientes por parejas del método anterior, pero ahora tomaremos las medianas en dos pasos, primero en cada punto y después para todos

$$b_{MR} = \underset{i}{\text{Med}}\{\underset{j \neq i}{\text{Med}}\{b_{ij}\}\}$$

En el primer paso se toma la mediana de las pendientes de $n-1$ rectas que pasan por el punto (x_i, y_i) y en el segundo paso se toma la mediana de estas n pendientes.

Para el punto de intercepción calcularemos $a_i = y_i - b_{MR}x_i$ y entonces

$$a_{MR} = \underset{i}{\text{Med}}\{a_i\}$$

Siegel probó que el punto de colapso de la recta con medianas repetidas es esencialmente 1/2.

Discusión

Ahora que tenemos diversos métodos con diferentes puntos de colapso, ¿cómo podemos elegir uno?

Una consideración es el grado de resistencia que una particular aplicación pide. Otro asunto es la precisión relativa de las pendientes estimadas, especialmente en muestras pequeñas. También es evidente que el tiempo de computación es otro de los factores a tener en cuenta.

Finalmente, podemos decir que la recta resistente de los tres grupos tiene un comportamiento suficientemente bueno en los tres aspectos considerados y, por ello, es el método resistente que se ha destacado. Sin embargo, en el capítulo 10 se consideran otros métodos aplicables a la regresión simple o múltiple y que se pueden computar con R o S-PLUS.

7.4. Ejercicios

Ejercicio 7.1

Calcular los parámetros de la recta resistente de los tres grupos con los datos del ejercicio 6.10. Discutir su necesidad frente a la recta mínimocuadrática.

Ejercicio 7.2

Calcular la recta resistente de los tres grupos con los datos (c) de la tabla 6.4. Comparar el resultado con el de la tabla 6.5 y con la recta de regresión mínimocuadrática sin la observación 16.