Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

Regresión por mínima mediana de cuadrados (LMS) y mínimos cuadrados recortados (LTS)

Daniel Fernández López

Dirigido por Raúl Pérez Fernández

UNIVERSIDAD DE OVIEDO
Facultad de Ciencias
Grado en Matemáticas

Febrero de 2021

Índice general

1. Introducción **3**
- 1.1. Motivación . . . 3
- 1.2. Objetivos . . . 3
- 1.3. Organización del trabajo . . . 4

2. Regresión lineal simple por mínimos cuadrados (LS) **6**
- 2.1. Contexto histórico . . . 6
- 2.2. Introducción . . . 7
- 2.3. Estimación mínimo cuadrática de los parámetros (LS) . . . 8
  - 2.3.1. Estimación de β_0 y β_1 . . . 8
  - 2.3.2. Propiedades de los estimadores . . . 10
  - 2.3.3. Estimación de σ^2 . . . 11
- 2.4. Test de hipótesis sobre los coeficientes de regresión . . . 12
  - 2.4.1. T-test . . . 12
- 2.5. Estimación por intervalos en la regresión lineal simple . . . 14
  - 2.5.1. Intervalos de confianza de β_0, β_1 y σ^2 . . . 14
- 2.6. Predicción de nuevas observaciones . . . 15
- 2.7. Coeficiente de determinación . . . 15
- 2.8. Regresión lineal múltiple . . . 17
- 2.9. Ejemplo . . . 18

3. Residuos, apalancamiento y puntos de influencia **24**
- 3.1. Análisis de los residuos . . . 24
  - 3.1.1. Definición de los residuos . . . 25
  - 3.1.2. Métodos para reescalar los residuos . . . 25
  - 3.1.3. Gráficos de residuos . . . 32
- 3.2. Detección y tratamiento de outliers . . . 38
- 3.3. Diagnóstico del apalancamiento y la influencia . . . 41
  - 3.3.1. Apalancamiento . . . 42
  - 3.3.2. Distancia de Cook . . . 43
  - 3.3.3. Varianza generalizada y COVRATIO . . . 44
  - 3.3.4. Ejemplo de punto de influencia . . . 45
  - 3.3.5. Tratamiento de datos de influencia . . . 48

4. Introducción a la regresión lineal robusta **50**
4.1. Punto de ruptura 51
4.2. M-estimadores 53

5. Regresión lineal por mínima mediana de cuadrados (LMS) **56**
5.1. Definición del estimador LMS 56
5.2. Propiedades generales 58
5.2.1. Existencia de solución 58
5.2.2. Propiedades de equivarianza 60
5.2.3. Punto de ruptura del método LMS 62
5.2.4. Propiedades asintóticas 66
5.3. Consideraciones finales 66

6. Regresión lineal por mínimos cuadrados recortados (LTS) **67**
6.1. Definición del estadístico 67
6.2. Propiedades generales 68
6.2.1. Propiedades de equivarianza 69
6.2.2. Punto de ruptura del método LTS 70
6.2.3. Propiedades asintóticas 72
6.3. Consideraciones finales 73

7. Caso unidimensional: Problema de localización **74**
7.1. Introducción 74
7.2. Estimador de localización LS 75
7.3. Estimador de localización LMS 75
7.4. Estimador de localización LTS 77
7.5. Ejemplo numérico 78

8. Ejemplo: Jugadores de la NBA **80**
8.1. Resolución del problema 80

9. Conclusión **93**

Capítulo 1

Introducción

1.1. Motivación

Desde que comenzó mi formación académica en las matemáticas, tanto primaria como secundaria y universitaria; la cuestión que como estudiante me he planteado a mí mismo o al propio docente es: ¿Qué aplicaciones tiene esto?. La respuesta a esta pregunta me la he ido dando a medida que iban pasando los cursos y he ido indagando a título personal en los conocimientos desarrollados en el Grado. Con este trabajo, en cierta medida, he intentado dar una contestación válida a esa cuestión. Si bien es cierto que no creo que las distintas ramas de las matemáticas sean ajenas entre ellas, es más, pienso que buscan explicar lo mismo desde distintos puntos de vista; la estadística siempre ha despertado un mayor interés en mí que el resto de áreas. Mi principal motivación era la de unir la parte teórica de la estadística con la parte práctica, para demostrar que conociendo los fundamentos matemáticos, siempre se obtendrá un resultado más preciso y correcto sobre cualquier problema que con la mera intuición que una persona de otro campo científico podría tener.

En particular, cuando empecé a tratar con bases de datos, se despertó en mí una curiosidad sobre el hecho de que los datos atípicos generalmente sean eliminados porque 'molestan'. Nunca me pareció del todo correcto pasar por alto estas observaciones, como si su existencia fuera ficticia, y me interesé en ellas; en ver cómo afectan a un análisis o qué se puede hacer en lugar de obviarlas. La regresión lineal es una técnica asentada en todos los cursos de estadística y creí que, por su simplicidad e importancia, era la mejor opción para estudiar el efecto de estos valores atípicos.

1.2. Objetivos

El trabajo tiene como propósito principal desarollar la teoría de la regresión lineal desde la base, e ir exponiendo los distintos problemas que van

apareciendo en el método clásico de mínimos cuadrados. Los valores atípicos u outliers son uno de los principales contratiempos que nos podemos encontrar a la hora de aplicar ese procedimiento. Se busca ilustrar al lector con diversas alternativas robustas que eviten estos problemas y no desvirtúen el trabajo realizado por la comunidad estadística hasta el momento. Para ello, se han expuesto de forma teórica y práctica dos de los métodos menos sensibles frente a la presencia de valores atípicos, la regresión por mínima mediana de cuadrados y por mínimos cuadrados recortados. La finalidad no es otra que la de dotar al lector de distintos procedimientos y despertar su curiosidad en ellos, no asumiendo siempre como correcto o válido lo que se estudia de forma clásica. La mejor manera de mostrarlo gráficamente es mediante la regresión lineal simple, y es por ello que el trabajo se ha centrado en ese caso particular.

1.3. Organización del trabajo

La estructura del trabajo es la siguiente:

- **Capítulo 1. Regresión lineal simple por mínimos cuadrados (LS)**

  Se introducen los principales conceptos de la regresión lineal simple por mínimos cuadrados, desarollando la teoría desde la base y explicando aquellos parámetros de interés para el trabajo.

- **Capítulo 2. Residuos, apalancamiento y puntos de influencia**

  Es el punto de inflexión del trabajo, donde se redactan los distintos tipos de residuos con los que podemos trabajar y las diversas técnicas que nos permiten detectar puntos con influencia en el ajuste de mínimos cuadrados. Esencialmente se ilustran los problemas que pueden aparecer distorsionando por completo un análisis clásico de regresión lineal.

- **Capítulo 4. Introducción a la regresión lineal robusta**

  Se presentan brevemente los conceptos ensenciales para medir la robustez de un método para la regresión lineal, centrándonos en el punto de ruptura. Su objetivo es el de ir introduciendo alternativas robustas al método de mínimos cuadrados hasta llegar a las dos estimaciones que dan título al trabajo.

- **Capítulo 5. Regresión lineal por mínima mediana de cuadrados (LMS)**

  Una vez detectados los problemas causados por los outliers, se ilustra la regresión LMS a través de sus propiedades y diversos ejemplos gráficos.

- **Capítulo 6. Regresión lineal por mínimos cuadrados recortados (LTS)**

  La finalidad es la de introducir la regresión LTS como otra alternativa a las expuestas anteriormente y que, además, soluciona los problemas comentados previamente.

- **Capítulo 7. Caso unidimensional: Problema de localización**

  Se busca exponer el problema de localización de forma sencilla y tratarlo brevemente para los ajustes hechos previamente. La finalidad es la de recoger el único caso no estudiado en los capítulos anteriores.

- **Capítulo 8. Ejemplo: Jugadores de la NBA**

  Se aplican a la práctica la mayoría de conocimientos descritos anteriormente. El fin es el de dar una visión global sobre el trabajo y recalcar su importancia con un ejmplo de aplicación al mundo real.

- **Capítulo 9. Conclusión**

  Valoración propia sobre lo expuesto a lo largo de los capítulos anteriores.

Capítulo 2

Regresión lineal simple por mínimos cuadrados (LS)

2.1. Contexto histórico

Desde los orígenes del pensamiento humano, la búsqueda sobre la relación que puede existir entre dos sucesos ha sido una de nuestras principales motivaciones. Crocker [8] resumió con sus palabras los fundamentos de la investigación humana: 'Conocer si los sucesos se relacionan y, con qué intensidad lo hacen, facilita a las personas explicar el pasado, controlar el presente y predecir el futuro'. De estas palabras podemos extraer la importancia que tiene para la sociedad el dominio de los conceptos de correlación y regresión.

El origen de estas nociones proviene, en gran parte, de diversos estudios en la rama biológica de la ciencia. La idea actual de correlación y regresión se debe en gran medida a Galton [13], quien se casó con una prima de Charles Darwin y, a partir de su parentesco con el científico, desarrolló un gran interés en los estudios sobre la herencia. Podemos considerar a Galton como un ingenioso amateur, ya que, sin conocer los métodos estadísticos de la época, estudia la variabilidad de características humanas. Motivado por Darwin, diseñó un experimento con semillas de guisantes, cuya finalidad era la de estudiar su peso en dos generaciones distintas. Repartió las semillas entre 7 amigos, que a su vez las cultivaron y le enviaron las semillas cosechadas. Sus conclusiones entre las semillas madre y las semillas hija fueron interesantes. Notó que el peso medio de las semillas cosechadas era función lineal del peso de las semillas cultivadas con una pendiente menor que la unidad. Desarrollando su estudio, no tardó en darse cuenta que si el grado de asociación entre dos características se mantenía constante, entonces la pendiente de la recta de regresión podría ser calculada si se conoce la variabilidad de ambas medidas. Todas sus conclusiones llegaron gracias a representaciones gráficas y experimentos empíricos, por lo que los errores se sucedieron en sus estudios, aunque asentó las bases de la regresión moderna.

Fue Pearson quien formalizó la teoría matemática sobre la regresión y la correlación, continuando con el trabajo previo de Galton debido a su interés en la Biometría [29].

Finalmente, las ideas modernas sobre regresión se originan en los trabajos de Legendre y Gauss, sobre el método de mínimos cuadrados, con la finalidad de ajustar los datos sobre las órbitas de cuerpos celestes. Para una introducción más completa a la historia del estudio de la regresión, se refiere el artículo [33].

2.2. Introducción

La regresión lineal simple es una técnica estadística cuya finalidad es la búsqueda de una relación lineal entre dos variables de estudio. En este caso, dispondremos de una variable explicativa o regresor x relacionada con una variable respuesta o dependiente y a través de una línea recta. El modelo de la regresión lineal simple se ajusta a la siguiente expresión:

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x} + \epsilon \tag{2.1}$$

donde β_0 y β_1 son constantes desconocidas (coeficientes de regresión). Intuitivamente se observa que las constantes β_0 y β_1 representan, respectivamente, el punto de corte y la pendiente de la recta descrita. Por otra parte, el vector ϵ hace referencia a la componente aleatoria del error del modelo, siendo cada componente el desvío frente al ajuste lineal asociado a cada observación. Para simplificar el procedimiento, dicho error se supone de media cero y varianza desconocida σ^2. De forma adicional, generalmente asumiremos estos errores como incorrelados entre sí, es decir, no existe dependencia lineal entre los valores de cada componente del vector error.

A lo largo del trabajo también será utilizada indistintamente la notación:

$$\mathrm{y} = \mathrm{X}\vec{\beta} + \epsilon$$

donde

$$\mathrm{X} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix} = (\vec{1}_n | \mathrm{x}_1)$$

y $\vec{\beta} = (\beta_0, \beta_1)$, ya que algunas demostraciones son válidas también para el caso múltiple y trabajar con la forma matricial facilita su entendimiento.

Es conveniente partir de un estudio donde la variable explicativa x esté controlada por el analista y medida con un error despeciable, mientras que la variable respuesta y sea aleatoria. En otras palabras, para cada valor x

de x podemos encontrar la distribución de probabilidad de y condicionada por dicho valor x. La esperanza y la varianza de esta distribución son:

$$\mathrm{E(y|x=}\, x) = \beta_0 + \beta_1 x \tag{2.2}$$

$$\mathrm{Var(y|x=}\, x) = \mathrm{Var}(\beta_0 + \beta_1 x + \epsilon) = \sigma^2 \tag{2.3}$$

Obsérvese que por ser los errores incorrelados, las respuestas también lo son.

A lo largo de este Capítulo se ha tomado como referencia el libro [27], para seguir una estructura y una notación coherente durante el mismo. El caso múltiple se desarrolla en el Capítulo 6 de [25], el cual ha sido consultado para una mejor comprensión.

2.3. Estimación mínimo cuadrática de los parámetros (LS)

Los coeficientes de regresión han de ser estimados a partir de la muestra dada. Para ello, supóngase que partimos de un conjunto de n-pares de datos representada por $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$.

2.3.1. Estimación de β_0 y β_1

Utilizaremos el método de mínimos cuadrados, es decir, estimaremos los valores de β_0 y β_1 de forma que la diferencia cuadrática entre las observaciones y_i y la recta obtenida sea mínima, en otras palabras, el método de mínimos cuadrados intenta que la distancia vertical del punto a la recta se minimice. De manera similar a la Ecuación (2.1) podemos escribir:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2.4}$$

Nos podemos referir a la Ecuación (2.1) como modelo poblacional de regresión, a su vez que la Ecuación (2.4) puede interpretarse como el modelo muestral. El criterio de mínimos cuadrados consiste en obtener los valores que minimizan la expresión:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.5}$$

Denotando por $\hat{\beta}_0$ y $\hat{\beta}_1$ a las estimaciones mínimo cuadradas correspondientes, estas han de satisfacer:

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

y

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

Simplificando ambas ecuaciones llegamos a que:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned} \tag{2.6}$$

Estas últimas ecuaciones son llamadas ecuaciones normales de mínimos cuadrados. De la Ecuación (2.6) llegamos a la solución:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{2.7}$$

y

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{2.8}$$

donde

$$\overline{y} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} y_i \quad \text{y} \quad \overline{x} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i$$

son las medias muestrales de y y x respectivamente. Las estimaciones mínimo cuadradas obtenidas en las Ecuaciones (2.7) y (2.8) son las correspondientes al punto de corte β_0 y la pendiente β_1 de la recta de regresión. La última igualdad puede ser expresada de una forma más compacta

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{2.9}$$

haciendo uso de las identidades correpondientes a la suma de cuadrados de x y la suma de cuadrados de x e y:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

y

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} = \sum_{i=1}^{n} y_i(x_i - \overline{x})$$

Así, el modelo de regresión lineal ajustado es:

$$\hat{\text{y}} = \hat{\beta}_0 + \hat{\beta}_1 \text{x} \tag{2.10}$$

En este momento aparece un concepto vital en el desarrollo de la regresión lineal, el residuo. Este se define como la diferencia entre el valor observado y_i y el valor ajustado obtenido $\hat{y}_i$, y, además, sirve como estimación del error. Matemáticamente, es representado por:

$$e_i = y_i - \hat{y}_i \quad \text{con } i = 1, 2, \ldots, n \tag{2.11}$$

2.3.2. Propiedades de los estimadores

Los estimadores $\hat{\beta_0}$ y $\hat{\beta_1}$ hallados presentan varias propiedes remarcables. Antes de entrar en detalle con estas propiedades destáquese que, por las Ecuaciones (2.7) y (2.8), las estimaciones obtenidas son combinaciones de todas las observaciones y_i.

La primera propiedad que es conveniente comprobar es que tanto $\hat{\beta_0}$ como $\hat{\beta_1}$ son estimadores insesgados de β_0 y β_1 respectivamente.

Probémoslo para $\hat{\beta_1}$, siendo $c_i = (x_i - \overline{x})/S_{xx}$ para $i = 1, 2, \ldots, n$. Teniendo en cuenta este coeficiente c_i, $\hat{\beta_1}$ puede escribirse como $\sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i + \epsilon_i)$, aplicando directamente el modelo de regresión formulado en la Ecuación (2.4). Esta expresión se puede simplificar aún mas como sigue

$$\sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i \epsilon_i = \beta_1 + \sum_{i=1}^{n} c_i \epsilon_i$$

De aplicar que los coeficientes c_i verfican $\sum_{i=1}^{n} c_i = 0$ y $\sum_{i=1}^{n} c_i x_i = 1$ llegamos a que

$$E(\hat{\beta_1}) = E\left(\beta_1 + \sum_{i=1}^{n} c_i \epsilon_i\right) = E(\beta_1) + \sum_{i=1}^{n} c_i E(\epsilon_i) = \beta_1$$

es decir, $\hat{\beta_1}$ es un estimador insesgado de β_1.

Para demostrar que $\hat{\beta_0}$ es un estimador insesgado de β_0 demostraremos que $E(\hat{\beta_1}) = \beta_1$.

$$E(\hat{\beta_0}) = E(\overline{y}) - E(\hat{\beta_1}\overline{x}) = \frac{1}{n}\left(\sum_{i=1}^{n} E(y_i) - E(\hat{\beta_1}) \sum_{i=1}^{n} E(x_i)\right)$$
$$= \frac{1}{n}\left(\sum_{i=1}^{n} (\beta_0 + \beta_1 E(x_i) - \beta_1 E(x_i))\right) = \beta_0$$

Ahora veamos cual es la varianza de $\hat{\beta_1}$

$$\mathrm{Var}(\hat{\beta_1}) = \mathrm{Var}(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i^2 \ \mathrm{Var}(y_i) \tag{2.12}$$

porque, como ya se ha indicado previamente, las observaciones y_i son incorreladas, y por tanto se puede aplicar que la varianza de la suma es la suma de las varianzas. Se había supuesto que $\mathrm{Var}(y_i) = \sigma^2$, consecuentemente,

$$\mathrm{Var}(\hat{\beta_1}) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \overline{x})^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}} \tag{2.13}$$

Por otro lado, la varianza de $\hat{\beta}_0$ se calcula

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\overline{y} - \hat{\beta}_1\overline{x}) = \text{Var}(\overline{y}) + \overline{x}^2\ \text{Var}(\hat{\beta}_1) - 2\overline{x}\ \text{Cov}(\overline{y}, \hat{\beta}_1) =$$
$$= \text{Var}(\overline{y}) + \overline{x}^2\text{Var}(\hat{\beta}_1) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)$$

Hemos usado que $\text{Cov}(\overline{\text{y}}, \hat{\beta}_1) = 0$ y $\text{Var}(\overline{\text{y}}) = \sigma^2/n$. Para el modelo de regresión planteado en (2.1) con las hipótesis establecidas, $E(\epsilon) = 0$, $\text{Var}(\epsilon) = \sigma^2$, y los errores incorrelados; los estimadores obtenidos por el método de mínimos cuadrados son los que tienen varianza mínima entre todos aquellos que son insesgados expresados mediante combinaciones lineales de y_i. Este resultado se recoge en el Teorema de Gauss-Markov.

A continuación, vamos a mencionar algunas propiedades adicionales que pueden resultar útiles:

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

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

2. La recta de regresión calculada por el método de mínimos cuadrados contiene al punto $(\overline{y}, \overline{x})$ de los datos (el centroide).

3. La suma de los residuos ponderados por el valor correspondiente de la variable explicativa es cero.

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

4. La suma de los residuos ponderados por el valor correspondiente de los variable ajustados es cero.

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

2.3.3. Estimación de σ^2

Para poder construir intervalos de confianza y test de hipótesis en el modelo de regresión, es necesario estimar el parámetro σ^2. En un caso ideal nos gustaría poder obtener la estimación sin depender de cómo se adecua el modelo ajustado. Sin embargo, esta situación solo es posible cuando hay varias observaciones de y para un mismo valor de x, o por otro lado cuando disponemos de información previa sobre el parámetro. Cuando no nos encontramos en ninguna de las dos situaciones anteriores, entonces tenemos que hacer uso del residuo para estimar σ^2.

$$SS_{Res} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \tag{2.14}$$

Sustituyendo en la ecuación $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, llegamos a

$$SS_{Res} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 - \hat{\beta}_1 S_{xy} \tag{2.15}$$

Haciendo uso de que

$$\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = \sum_{i=1}^{n} (y_i - n\overline{y})^2 \equiv SS_T$$

llegamos al resultado siguiente

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy} \tag{2.16}$$

La suma cuadrada de los residuos tiene $n-2$ grados de libertad, ya que dos grados de libertad están asociados a las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$, necesarios para obtener $\hat{y}_i$. El valor esperado de SS_{Res} es $E(SS_{Res}) = (n-2)\sigma^2$, luego un estimador insesgado de σ^2 es

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res} \tag{2.17}$$

llamado media cuadrada residual. A la raíz cuadrada de $\hat{\sigma}^2$ algunas veces se le llama error típico de regresión, y tiene las mismas unidades que la variable respuesta y. Diremos que $\hat{\sigma}^2$ es una estimación de σ^2 que depende del modelo, porque es calculada a partir de los residuos del modelo de regresión.

2.4. Test de hipótesis sobre los coeficientes de regresión

Una vez encontradas las estimaciones de los parámetros por el método LS (mínimos cuadrados), aparece la necesidad de construir test de hipótesis sobre dichos coeficientes. Para poder plantear el problema necesitamos asumir que los errores del modelo ϵ_i se distribuyen según una normal. Teniendo lo último en cuenta, necesitamos suponer que los errores se distribuyen de forma normal e independiente con media 0 y varianza σ^2.

2.4.1. T-test

Supongamos que queremos contrastar que la pendiente del modelo de regresión β_1 equivale a una constante, denotada β_{10}. Matemáticamente se escribe:

$$H_0 : \beta_1 = \beta_{10},\ H_1 : \beta_1 \neq \beta_{10} \tag{2.18}$$

Obtendremos una región de rechazo con dos colas, por lo que es importante remarcar que estamos planteando un test bilateral. Como hemos supuesto

que los errores ϵ_i son independientes y siguen una $\mathcal{N}(0, \sigma^2)$, luego las observaciones y_i siguen $\mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ y mantienen la independencia. Sabemos que $\hat{\beta}_1$ es una combinación lineal de las observaciones, así sigue una distribución normal de media β_1 y varianza σ^2/S_{xx} por los valores hallados en la sección previa. Planteamos entonces el estadístico de contraste

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}}$$

cuya distribución es $\mathcal{N}(0, 1)$ bajo la hipótesis nula, por el TCL. Si el valor de σ^2 es conocido, podemos usar el estadístico Z_0 para el test (2.18). Sin embargo, esto raramente ocurre. Aparece entonces el estimador insesgado, MS_{Res}, de σ^2. Se conoce que $(n-2)MS_{Res}/\sigma^2$ sigue una distrución χ^2_{n-2}, y además $\hat{\beta}_1$ y MS_{Res} son independientes (covarianza nula). Trabajaremos entonces con el estadístico

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \tag{2.19}$$

con distribución de probabilidad t_{n-2} si la hipótesis nula es cierta. Por tanto, haciendo uso de este estadístico de contraste,la hipótesis nula será rechazada para aquellos valores que estén en la región crítica

$$RC = \{t_0 \in \mathbb{R} \mid |t_0| > t_{\alpha/2, n-2}\}$$

donde α es el nivel de significación establecido por el analista.
Al denominador del estadístico t_0 se le conoce por error estándar de la pendiente. Esto es,

$$se(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}}$$

Podemos contrastar el valor del parámetro β_0 de manera similar, planteando el test de hipótesis

$$H_0 : \beta_0 = \beta_{00},\ H_1 : \beta_0 \neq \beta_{00} \tag{2.20}$$

y haciendo uso del estadístico de contraste

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(1/n + \overline{x}^2/S_{xx})}}$$

La región crítica obtenida vuelve a ser

$$RC = \{t_0 \in \mathbb{R} \mid |t_0| > t_{\alpha/2, n-2}\}$$

De nuevo, el denominador de t_0 recibe el nombre de error estándar del punto de corte, al que denotamos por $se(\hat{\beta}_0)$.

Un caso digno de ser mencionado es el test de hipótesis de la Ecuación (2.18), donde

$$H_0 : \beta_1 = 0 \ , \ H_1 : \beta_1 \neq 0$$

Esta hipótesis contrasta la dependencia lineal entre las variables x e y, donde si no se rechaza H_0 significará que ambas variables no tienen ningún tipo de dependencia lineal. Esto no significa que necesariamente sean independientes una de la otra, simplemente que no podemos encontrar una relación lineal. Por otro lado, si se rechaza la hipótesis nula podemos asegurar que existe dependencia lineal, aunque no tiene por qué ser la mejor. Puede ocurrir que un polinomio en x con un grado mayor que 1 se ajuste mejor a la relación observada.

Otra situación comúnmente encontrada, es aquella en la que se trata con datos donde el mejor ajuste por regresión lineal conlleva el uso de una recta que pase por el origen. Se plantearía entonces el siguiente test de hipótesis:

$$H_0 : \beta_0 = 0 \ , \ H_1 : \beta_0 \neq 0$$

Podemos referirnos a este caso concreto como modelo de regresión sin intercepto. Más formalmente, el modelo sin intercepto es

$$\mathrm{y} = \beta_1 \mathrm{x} + \epsilon$$

2.5. Estimación por intervalos en la regresión lineal simple

A lo largo de esta sección trataremos de encotrar los intervalos de confianza adecuados para los parámetros de la regresión.

2.5.1. Intervalos de confianza de β_0, β_1 y σ^2

En la Sección anterior, hemos deducido que, si se cumple que los errores se distribuyen de forma normal e independiente, los estadísticos pivote $\frac{\hat{\beta_1}-\beta_1}{se(\hat{\beta_1})}$ y $\frac{\hat{\beta_0}-\beta_0}{se(\hat{\beta_0})}$ siguen una distribución t_{n-2}. Por tanto, un intervalo con un nivel de confianza $100(1-\alpha)$ para el parámetro β_1 viene dado por

$$\hat{\beta_1} - t_{\alpha/2,n-2} se(\hat{\beta_1}) \leq \beta_1 \leq \hat{\beta_1} + t_{\alpha/2,n-2} se(\hat{\beta_1}) \tag{2.21}$$

Análogamente, un intervalo de confianza (IC) para β_0 con el mismo nivel será

$$\hat{\beta_0} - t_{\alpha/2,n-2} se(\hat{\beta_0}) \leq \beta_0 \leq \hat{\beta_0} + t_{\alpha/2,n-2} se(\hat{\beta_0}) \tag{2.22}$$

De nuevo, bajo la hipótesis previa requerida de que los errores verifiquen que se distribuyen de manera normal e independiente, podremos plantear una expresión para un IC del parámetro σ^2. Trabajando con la expresión ($n-$

$2)MS_{Res}/\sigma^2$, la cual es conocido que sigue una distribución de probabilidad χ^2_{n-2}, podremos plantear el intervalo

$$\frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}} \tag{2.23}$$

2.6. Predicción de nuevas observaciones

El modelo de regresión tiene varias aplicaciones, entre las que destaca la predicción de nuevas observaciones. El objetivo principal será estimar nuevos valores de la variable y en función de los valores que toma la variable x. Sea x_0 un valor de la variable de interés, entonces podemos escribir el modelo de regresión

$$\hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{2.24}$$

donde el valor estimado y_0 corresponde a la nueva observación.

Una vez obtenida esta estimación puntual, veamos cómo establecer un intervalo de predicción para una observación futura y_0.

Escribimos la variable aleatoria con distribución normal y media 0.

$$\psi = y_0 - \hat{y_0}$$

Calculemos su varianza para poder estimar el intervalo que buscamos.

$$\text{Var}(\psi) = \text{Var}(y_0 - \hat{y_0}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]$$

ya que la observación futura y_0 es independiente de $\hat{y_0}$. Por tanto, llegamos a que el intervalo de predicción con un nivel de confianza $100(1-\alpha)$ para una observación futura en x_0 es

$$\begin{aligned} &\hat{y_0} - t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)} \\ &\leq y_0 \leq \hat{y_0} + t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)} \end{aligned} \tag{2.25}$$

Como podemos observar en la expresión obtenida, el rango de este intervalo será mínimo cuando $x_0 = \overline{x}$, y a medida que la expresión $(x_0 - \overline{x})^2$ sea mayor, el rango del intervalo aumentará.

2.7. Coeficiente de determinación

A la cantidad

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \tag{2.26}$$

la llamamos coeficiente de determinación, el cual representa la proporción de variabilidad explicada por la variable x. Esto se debe a que la expresión tiene en cuenta los términos SS_T, que explica la variabilidad en y sin tener en cuenta el efecto causado por la variable x, y SS_{Res}, que mide la variabilidad restante en y después de haberse considerado x. Como $0 \leq SS_{Res} \leq SS_T = SS_R + SS_{Res}$, se obtiene que $0 \leq R^2 \leq 1$. Cuanto más cerca de 1 esté el valor de R^2, mayor será la variabilidad explicada en y por el modelo de regresión.

A pesar de esta interpretación, tener un valor elevado de R^2 no significa necesariamente que el modelo de regresión adoptado sea un ajuste preciso, ya que siempre es posible hacerlo más grande si se añaden más regresores. Aunque R^2 no puede hacerse más pequeño si añadimos regresores al modelo, esto no necesariamente significa que el nuevo modelo sea superior al antiguo. El coste computacional y teórico de trabajar con una variable explicativa adicional puede no 'compensar' a la hora de realizar el ajuste, ya que la variabilidad explicada puede ser ligeramente superior pero nada reseñable. Todo modelo debe de seguir el principio de la navaja de Ockham; 'en igualdad de condiciones, la explicación más sencilla suele ser la más probable'. En el caso múltiple se realiza la llamada 'Regresión por etapas', donde se van añadiendo sucesivamente más variables al ajuste hasta encontrar la aproximación óptima. Para paliar el efecto causado por añadir una variable explicativa más que aporte poca información adicional, se define el coeficiente de determinación ajustado R_a^2, donde k es el número de regresores considerados. En la regresión lineal simple $k = 1$.

$$R_a^2 = 1 - \frac{n-1}{n-k-1}\frac{SS_{Res}}{SS_T}$$

Normalmente, en los casos prácticos, nos encotraremos con valores elevados para este coeficiente. En caso de obtener una cantidad pobre, puede significar que disponemos de un modelo poco explicativo, con lo cual bastaría con añadir una o varias variables de predicción o regresión.

Es importante comentar la relación existente entre el coeficiente de determinación y el coeficiente de correlación de Pearson. Este último se denota por ρ y mide el grado de asociación lineal entre dos variables, mientras que el coeficiente de determinación mide la proporción de variación explicada de la variable dependiente por la variable explicativa. El coeficiente de correlación de Pearson entre dos variables aleatorias x e y, para una muestra dada, se define

$$\rho = \frac{E(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

donde el término σ_{xy} es la covarianza de x e y.

Es común contrastar si $\rho = 0$ para poder deducir si existe dependencia lineal entre ambas variables.

Se puede calcular este coeficiente sobre una muestra, llamado coeficiente

de correlación muestral

$$r = \frac{\sum_{i=1}^{n} y_i(x_i - \overline{x})}{[\ \sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2\ ]^{1/2}} = \frac{S_{xy}}{[S_{xx}SS_T]^{1/2}}$$

Nótese que

$$\hat{\beta}_1 = \left(\frac{SS_T}{S_{xx}}\right)^{1/2} r$$

Aunque los parámetros $\hat{\beta}_1$ y r estén relacionados, proporcionan información diferente. Mientras que r mide la asociación lineal entre x e y, el coeficiente $\hat{\beta}_1$ muestra el cambio de media en y por cada unidad de cambio de x.

En el caso de regresión lineal simple, existe una relación entre el coeficiente R^2 y r

$$r^2 = \hat{\beta}_1^2 \left(\frac{S_{xx}}{SS_T}\right) = \frac{\hat{\beta}_1 S_{xy}}{SS_T} = \frac{SS_R}{SS_T} = R^2$$

2.8. Regresión lineal múltiple

A lo largo del trabajo nos basaremos principalmente en el caso simple, es decir, donde solamente tenemos una variable explicativa x_1. Esta elección facilita considerablemente el estudio de los distintos tipos de ajuste y, además, nos permite exponer gráficamente de una manera mucho más sencilla las diferencias entre las estimaciones robustas de los Capítulos 5 y 6 frente al ajuste mínimo cuadrático clásico. Aún así, se introducirá la notación para el caso múltiple, ya que las propiedades estudiadas en adelante serán probadas para el caso p-dimensional, donde buscaremos predecir una variable explicada y a través de la siguiente expresión

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \ldots + \beta_{p-1}\mathrm{x}_{p-1} + \epsilon \tag{2.27}$$

En forma matricial, se escribirá el problema (2.27) como

$$\mathrm{y} = \mathrm{X}\vec{\beta} + \vec{\epsilon} \tag{2.28}$$

- y es un vector columna con n observaciones de la variable aleatoria explicada y.
- X (matriz de regresión) es una matriz de dimensión $n \times p$, la cual recoge las n observaciones de las $p-1$ variables explicativas (x_1, ..., x_{p-1}) que se han observado, y una columna formada por el vector de unos con n componentes. En general es de rango p.
- $\vec{\beta}$ es el vector de parámetros de dimensión p.
- $\vec{\epsilon}$ es el vector columna n-dimensional de valores de las perturbaciones aleatorias.

El modelo de regresión estimado se escribe $\hat{\mathrm{y}} = \mathrm{X}\hat{\beta}$, donde la estimación $\hat{\beta}$ viene de minimizar la expresión $g(\vec{\beta}) = (y - \mathrm{X}\vec{\beta})'(y - \mathrm{X}\vec{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\vec{\beta})^2$. Se obtiene así que

$$\hat{\beta} = (\mathrm{X'X})^{-1}\mathrm{X'y} \Rightarrow \hat{\mathrm{y}} = \mathrm{X}\hat{\beta} = \mathrm{Hy}$$

La matriz H=X(X'X)$^{-1}$X' recibe el nombre de matriz de influencia o 'hat matrix'.

2.9. Ejemplo

Una vez hecho el desarrollo teórico de los conceptos necesarios de la regresión lineal, vamos a ilustrar estos resultados expuestos a través de un ejemplo. Para ello se ha construido una base de datos a partir de dos vectores de tamaño 20 generados por R. El primer vector, al que nos referiremos como la variable aleatoria x, corresponde a 20 valores de una distribución $\mathcal{N}(2, 3)$ generados de forma aleatoria. La segunda variable aleatoria, y, se ha construido a través de una transfromación lineal de x ($3.5x$+10) y sumándole valores aleatorios de una $\mathcal{N}(0, 2)$.

Por tanto, el conjunto de datos activo con el que se trabaja queda representado en la Tabla 2.1.

Una vez disponemos de los datos, lo más conveniente es representarlos mediante un diagrama de dispersión, mostrado en la Figura 2.1. Claramente se observa una relación importante entre la variable x y la variable y. Es más, podemos considerar que se ajusta a un modelo lineal simple $\mathrm{y} = \beta_0 + \beta_1\mathrm{x} + \epsilon$ de una forma bastante razonable.

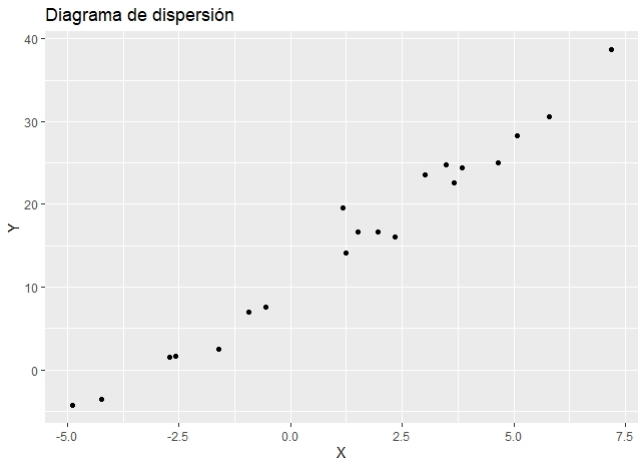


Figura 2.1: Diagrama de dispersión de la variable x frente a la variable y

ID	x	y
1	-4.88	-4.31
2	2.34	16.09
3	3.84	24.43
4	1.16	19.61
5	-0.94	6.97
6	1.95	16.61
7	1.51	16.65
8	-2.70	1.54
9	-1.60	2.53
10	3.01	23.54
11	4.64	25.00
12	3.49	24.79
13	-4.24	-3.60
14	3.65	22.62
15	5.07	28.34
16	7.19	38.72
17	-0.56	7.57
18	1.25	14.18
19	-2.57	1.63
20	5.79	30.59

Tabla 2.1: Conjunto de datos

En primer lugar vamos a calcular los parámetros del modelo, la suma cuadrática de x y la suma cuadrática de x e y respectivamente.

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 255.6004 - \frac{751.1011}{20} = 218.0453$$

y

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = 1197.73 - \frac{8591.748}{20} = 768.1422$$

Una vez obtenidos estos valores, ya podemos estimar los coeficientes de regresión con las relaciones descritas en la teoría.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 3.522856$$

y

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 15.67481 - 3.522856 \times 1.370311 = 10.8474$$

De forma intuitiva, y como los datos del problema están completamente especificados, vemos que el valor tomado por los coeficientes se ajusta bastante al real. Conocer con esta precisión la construcción de un problema es

remotamente improbable, aún así el objetivo de este ejemplo es ilustrar la utilidad de la regresión lineal simple para un tipo concreto de problemas.

En la Tabla 2.2 se recogen los valores observados (y_i), los valores estimados ($\hat{y}_i$) y el residuo (e_i).

y	$\hat{y}_i$	e_i
-4.31	-6.36	2.05
16.09	19.07	-2.98
24.43	24.37	0.06
19.61	14.94	4.67
6.97	7.53	-0.56
16.61	17.72	-1.11
16.65	16.18	0.47
1.54	1.34	0.20
2.53	5.21	-2.68
23.54	21.46	2.08
25.00	27.21	-2.21
24.79	23.14	1.65
-3.60	-4.07	0.47
22.62	23.71	-1.09
28.34	28.70	-0.36
38.72	36.17	2.56
7.57	8.89	-1.32
14.18	15.24	-1.06
1.63	1.80	-0.17
30.59	31.23	-0.64
$\sum y_i = 313.4962$	$\sum \hat{y}_i = 313.4962$	$\sum e_i \approx 0$

Tabla 2.2: Valores observados, estimados y residuo

El ajuste por mínimos cuadrados viene dado por

$$\hat{\mathrm{y}} = 10.8474 + 3.522856 \cdot \mathrm{x}$$

Habiendo obtenido la ecuación del modelo, surgen varias preguntas de interés sobre este.

1. ¿Cómo se ajusta a los datos reales?
2. ¿Es un modelo bueno para predecir datos futuros?
3. ¿Alguna de las hipótesis previas mencionadas no se cumple?

Estas cuestiones han de ser tratatadas antes de utilizar el modelo para ajustar y predecir datos. Para evaluar el modelo, los residuos juegan un papel

fundamental. En el siguiente capítulo se tratará cómo se adecua un modelo de regresión lineal simple a un conjunto de datos determinado.

Una vez obtenidas las aproximaciones de los coeficientes de regresión, se van a estudiar sus propiedades. Para estimar σ^2 en el ejemplo tratado primero calculamos

$$SS_T = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 7687.302 - \frac{98279.88}{20} = 2773.308$$

A partir de este valor se calcula SS_{Res}

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy} = 2773.308 - 3.522856 \times 768.1422 = 67.25393$$

Por tanto, la estimación de σ^2 se obtiene de

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = \frac{67.25393}{18} = 3.736329$$

El error estándar de la pendiente es

$$se(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}} = \sqrt{\frac{3.736329}{218.0453}} = 0.1309029$$

Así, el estadístico del t-test es

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{3.522856}{0.1309029} = 26.91198$$

Si elegimos el nivel de significación $\alpha = 0.05$, el valor crítico de t es $t_{0.025,18} = -2.100922$. Por tanto, se rechaza la hipóteis $H_0 : \beta_1 = 0$ y concluimos que existe una relación lineal entre las variables x e y.

Ahora se construirán los intervalos de confianza para los parámetros $\hat{\beta}_1$, $\hat{\beta}_0$ y σ^2 con un nivel de confianza del 95 % . Por la teoría se ha visto que

$$\hat{\beta}_1 + t_{0.025,18} se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 - t_{0.025,18} se(\hat{\beta}_1)$$

obteniéndose el intervalo $(3.247839, 3.797873)$.

Análogamente, el intervalo de confianza para β_0 es

$$\hat{\beta}_0 + t_{0.025,18} se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 - t_{0.025,18} se(\hat{\beta}_0)$$

Haciendo de nuevo los cálculos al igual que para el caso anterior, llegamos al intervalo de confianza $(9.864241, 11.83056)$,donde

$$se(\hat{\beta}_0) = \sqrt{MS_{Res}(1/n + \overline{x}^2/S_{xx})} = 0.4679667$$

El valor α prefijado es el que marca el rango del intervalo, si lo cambiamos también lo harían los extremos de este.

Para el parámetro σ^2, manteniendo el mismo valor de α se calcula como

$$\frac{(18)MS_{Res}}{\chi^2_{0.975,18}} \leq \sigma^2 \leq \frac{(18)MS_{Res}}{\chi^2_{0.025,18}}$$

y se llega al intervalo de confianza $(2.133259, 8.171061)$.

Para encontrar un intervalo de predicción de futuras observaciones con nivel de confianza al 95 %, usaremos las ecuaciones desarrolladas en la teoría.

$$\hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}$$
$$\leq y_0 \leq \hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}$$

Suponemos un valor $x_0 = 5$ para realizar la predicción.

$$28.46168 + (-2.100922)\sqrt{3.736329\left(1 + \frac{1}{20} + \frac{(5 - 1.370311)^2}{218.0453}\right)}$$
$$\leq y_0 \leq 28.46168 - (-2.100922)\sqrt{3.736329\left(1 + \frac{1}{20} + \frac{(5 - 1.370311)^2}{218.0453}\right)}$$

Calculando, el intervalo es

$$y_0 \in (24.18234, 32.74102)$$

Podemos hacer una representación del intervalo de confianza calculado a partir de la fórmula descrita para el ejemplo, representado en la Figura 2.2.

Finalmente, se van a calcular los coeficientes R^2 y r con los datos mostrados en la Tabla 1.

$$R^2 = 1 - \frac{SS_{Res}}{SS_T} = 1 - \frac{67.25393}{2773.308} = 0.9757496$$

y

$$r = \frac{S_{xy}}{\sqrt{S_{xx}SS_T}} = \frac{768.1422}{\sqrt{218.0453 \times 2773.308}} = 0.9878004$$

Así, la proporción de variabilidad explicada por el regresor x es casi toda (97.57496 %), y existe una dependencia lineal clara entre las variables x e y.

Por tanto, el ajuste obtenido por el método de mínimos cuadrados queda representado en la Figura 2.3.

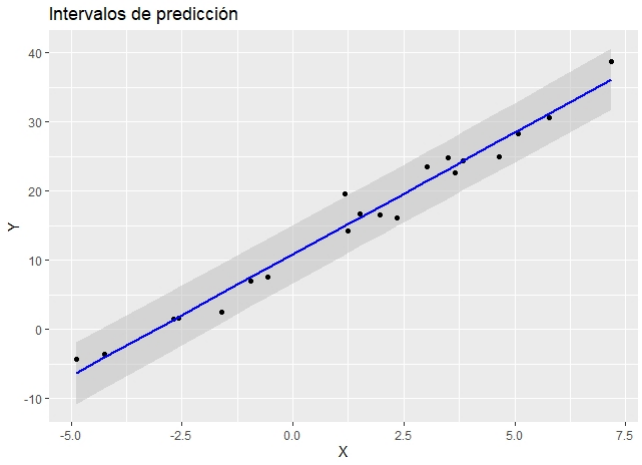


Figura 2.2: El intervalo con nivel de confianza del 95 % para los datos del ejemplo

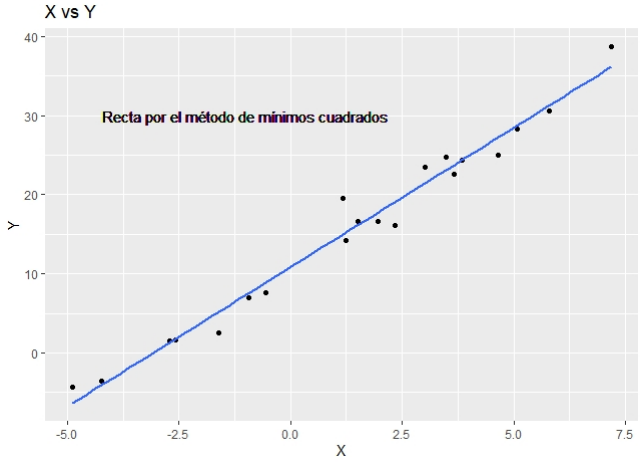


Figura 2.3: Recta de regresión para el conjunto de datos del ejemplo

Capítulo 3

Residuos, apalancamiento y puntos de influencia

En las secciones anteriores se han ido requiriendo ciertas hipótesis previas que han sido mencionadas. Recapitulando, estas son las siguientes:

1. La relación entre las variables x e y es aproximadamente lineal.
2. El error ϵ tiene media cero y varianza constante σ^2.
3. Los errores son incorrelados.
4. Los errores siguen una distribución normal.

Es importante confirmar siempre que estas hipótesis se cumplen para el modelo en estudio, ya que si se producen alteraciones graves podrían hacer que se comporte de manera inestable. Es decir, una muestra distinta del estudio podría conducirnos a un modelo del que extraigamos conclusiones totalmente opuestas. Anteriormente se han ido exponiendo los distintos estadísticos resumen del modelo de regresión. Estos son solo indicadores globales, por lo que no debemos centrar el estudio de la adecuación del modelo en ellos. En este capítulo se presentarán distintos métodos útiles para encontrar violaciones en las hipótesis previas.

A lo largo de este capítulo se han tomado como referencias los libros [27, 10], para seguir una estructura y una notación coherente durante el mismo.

3.1. Análisis de los residuos

Los procedimientos de diagnóstico se basan principalmente en el estudio de los residuos. A través de ellos podremos detectar problemas en el ajuste obtenido por el modelo, o intuir la aparición de ciertos datos que tienen una influencia importante sobre la recta construida. Mediante la representación

de los residuos y sus diferentes tipos de escalas se intentará comprobar si las hipótesis previas descritas se verifican o no.

3.1.1. Definición de los residuos

En el Capítulo 2 hemos definido los residuos como

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

donde y_i es el valor observado e $\hat{y}_i$ es su correspondiente valor aproximado. Por tanto, una interpretación intuitiva, a la par que correcta, es entender el concepto de residuo como las desviación entre los datos observados y los valores ajustados. Además, es una manera de medir la variabilidad no explicada por el modelo de regresión de la variable respuesta.
Cualquier variación en las hipótesis previas requeridas a los errores puede detectarse a través de los residuos, ya que, a parte de las interpretaciones mencionadas, es conveniente pensar en ellos como los valores observados de los errores en el modelo. Una manera muy efectiva y utilizada es representar los residuos para verificar la lista de hipótesis y comprobar la adecuación del modelo a los datos.
Los residuos tienen media cero, y su varianza media aproximada se estima como

$$\frac{\sum_{i=1}^{n}(e_i - \overline{e})}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Res}$$

En el caso simple tenemos $n-2$ grados de libertad y los residuos son no independientes.

3.1.2. Métodos para reescalar los residuos

En algunas ocasiones es cómodo trabajar con residuos en distintas escalas. En esta sección trataremos varios procedimientos que nos lo permitan. Son especialmente útiles para encontrar outliers dentro de nuestro conjunto de observaciones, es decir, observaciones que distan considerablemente del resto de datos.

- **Residuos estandarizados** Es un reescalado comúnmente utilizado en varios ámbitos, y como disponemos de la media ($\overline{e}_i = 0$) y la varianza (MS_{Res}), entonces podemos denotar los residuos estandarizados como

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}, \quad i = 1, 2, \ldots, n$$

  Los residuos d_i tienen media 0 y aproximadamente varianza 1. Consecuentemente, un residuo estandarizado con un valor grande ($|d_i| > 3$) indicará un outlier o valor atípico.

- **Residuos estudentizados** Para poder tratar de forma apropiada este método, es necesario recurrir al concepto de la matriz de influencia (hat matrix) H. Esta matriz es de dimensión $n \times n$ y se expresa como H = X(X'X)$^{-1}$X'; donde X es una matriz de dimensión $n \times p$, con $p = 2$ en el caso de la regresión lineal simple. En ese caso, escribiremos X como

$$\mathrm{X} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix} = (\vec{1}_n | \mathrm{x}_1)$$

En el caso de tener más variables explicativas basta con añadir las $p-2$ columnas necesarias a la matriz X.
La matriz H nos permite trabajar con los valores observados de y en lugar de los valores de las variables en X, ya que

$$\hat{\mathrm{y}} = \mathrm{X}\hat{\beta} = \mathrm{H}\mathrm{y}$$

siendo $\hat{\beta}$ el vector de los coeficientes de regresión estimados. Esta matriz H tiene diversas propiedades; es idempotente (HH=H) y es simétrica (H'=H), al igual que la matriz I − H, siendo I la matriz identidad. Una vez introducida la matriz de influencia, podemos expresar los residuos en notación matricial como

$$e = \mathrm{y} - \mathrm{H}\mathrm{y} = (\mathrm{I} - \mathrm{H})\mathrm{y} \tag{3.1}$$

Habiendo establecido esta notación, ya podemos tratar el tema que nos ocupa. Usando MS_{Res} como la varianza del i-ésimo residuo, e_i es tan solo una aproximación del residuo i. Podemos mejorar la escala del residuo e_i dividiéndolo por la desviación típica exacta.
Haciendo uso de la notación alternativa para los residuos en la Ecuación (3.1) y sustituyendo $\mathrm{y} = \mathrm{X}\beta + \epsilon$, llegamos a

$$\begin{aligned} e = (\mathrm{I} - \mathrm{H})(\mathrm{X}\beta + \epsilon) = \mathrm{X}\beta - \mathrm{H}\mathrm{X}\beta + (\mathrm{I} - \mathrm{H})\epsilon = \\ = \mathrm{X}\beta - \mathrm{X}(\mathrm{X'X})^{-1}\mathrm{X'X}\beta + (\mathrm{I} - \mathrm{H})\epsilon = (\mathrm{I} - \mathrm{H})\epsilon \end{aligned} \tag{3.2}$$

Hemos visto así que los residuos se expresan como combinación lineal de y o de los errores ϵ de la misma forma.

La matriz de varianzas-covarianzas de los residuos se calcula

$$\mathrm{Var}(e) = \mathrm{Var}\left[(\mathrm{I} - \mathrm{H})\epsilon\right] = (\mathrm{I} - \mathrm{H})\mathrm{Var}(\epsilon)(\mathrm{I} - \mathrm{H})' = \sigma^2(\mathrm{I} - \mathrm{H}) \tag{3.3}$$

ya que $\mathrm{Var}(\epsilon) = \sigma^2\mathrm{I}$ e I−H es una matriz simétrica e idempotente. Generalmente I−H no es diagonal, por lo que los residuos presentan varianzas distintas y están correlados entre ellos.

Denotaremos la varianza del i-residuo como

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$$

donde h_{ii} es el elemento diagonal de la matriz H y verifica $0 < h_{ii} < 1$. Siguiendo la misma notación, la covarianza entre dos residuos cualesquiera será

$$\mathrm{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$$

Como h_{ii} es una forma de medir la localización del punto x_i en el espacio asociado a x, la varianza del residuo e_i será distinta en función de donde se encuentre dicho punto x_i. Normalmente, los puntos cercanos al centro del espacio asociado a x tendrán un ajuste de mínimos cuadrados más pobre que los que estén mas alejados. Sin embargo, los problemas en las hipótesis establecidas por el modelo serán más propensos a darse en dichos puntos remotos, lo que hace que sean difíciles de detectar usando los residuos ordinarios o los estadarizados, ya que normalmente serán mas pequeños debido al método utilizado.

Sea y_n el valor observado en la variable respuesta del n-ésimo punto, y tomemos x_n los valores específicos para las variables explicativas o regresores. Los valores aproximados para la variable respuesta por la recta de regresión en función de los $n-1$ primeros puntos serán denotados $\hat{y}_n^*$. Consideramos el valor $\delta = y_n - \hat{y}_n^*$ la diferencia entre el n-ésimo valor observado y el calculado a partir de los $n-1$ primeros. Si un punto es remoto en términos de la variable explicativa y $|\delta|$ toma un valor elevado, entonces tenemos un punto influyente. Usando todos los datos para predecir el valor, lo denotamos ahora $\hat{y}_n$. Se tiene, por la información que aporta la matriz H, que

$$\hat{y}_n = \hat{y}_n^* + h_{nn}\delta$$

donde h_{nn} es el n-ésimo elemento de la diagonal de H. Si el punto es remoto en términos del espacio definido por los valores de los regresores, entonces h_{nn} se aproxima a 1, e $\hat{y}_n$ tiende a y_n consecuentemente. Por este motivo la varianza de los residuos de estos puntos es pequeña, lo que complica su análisis.

Un procedimiento lógico será examinar los residuos estudentizados

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}}, \quad i = 1, 2, \ldots, n \tag{3.4}$$

Estos residuos tienen varianza constante $\mathrm{Var}(r_i) = 1$, independientemente de la localización del punto x_i cuando el modelo es correcto. En muchas ocasiones la varianza de los residuos se acaba estabilizando, particularmente para conjuntos de datos grandes. En estos casos

la información proporcionada es equivalente a la de los residuos estandarizados. Sin embargo, como se ha visto que un punto con un residuo grande y un valor de h_{ii} elevado es potencialmente influyente en el ajuste por mínimos cuadrados, es conveniente el procedimiento estudentizado.

- **Residuos PRESS** En los dos métodos anteriores se han ilustrado dos alternativas que detectan los outliers o puntos remotos de forma efectiva. En este apartado, se explica un nuevo procedimiento basado en la examinación de las cantidades obtenidas de operar $y_i - \hat{y}_{(i)}$, donde $\hat{y}_{(i)}$ es el i-ésimo valor ajustado para la variable respuesta basado en todas las observaciones salvo la de la posición i. Haciendo un razonamiento similar al caso estudentizado, la lógica nos lleva a pensar que si la observación y_i es realmente inusual entonces ejercerá una inluencia considerable en el modelo de regresión. En esta situación el valor ajustado $\hat{y}_i$ será similar al valor observado y_i y, por ende, el residuo ordinario e_i será pequeño. Sin embargo, al eliminar la i-ésima observación, el valor $\hat{y}_{(i)}$ no puede ser influenciado por esta y, por ello, el residuo resultante debería indicar la presencia de un outlier. Si eliminamos la observación y_i, aproximaremos el modelo de regresión a través de las $n-1$ restantes, y tendremos los llamados residuos PRESS

$$e_{(i)} = y_i - \hat{y}_{(i)}, \quad i = 1, 2, \ldots, n$$

Puede resultar algo complejo ya que parece necesario calcular estos residuos para n hiperplanos (rectas en el caso simple) de regresión distintos. A pesar de esto, es posible calcular los residuos PRESS con solo un ajuste de regresión para las n observaciones, debido a la siguiente igualdad

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}, i = 1, 2, \ldots, n \tag{3.5}$$

La equivalencia se demuestra como sigue.

Demostración : Sea $\hat{\beta}_{(i)}$ el vector de coeficientes de regresión obtenidos sin tener en cuenta la i-ésima observación. Entonces

$$\hat{\beta}_{(i)} = \left[\mathrm{X'}_{(i)}\mathrm{X}_{(i)}\right]^{-1} \mathrm{X'}_{(i)}\mathrm{y}_{(i)}$$

donde $\mathrm{X}_{(i)}$ e y_i son la matriz X y el vector y sin la observación i-ésima. Por tanto, el residuo PRESS en la posición i puede escribirse como

$$\begin{aligned} e_{(i)} &= y_i - \hat{y}_{(i)} = y_i - \mathrm{x'}_i\hat{\beta}_{(i)} \\ &= y_i - \mathrm{x'}_i \left(\mathrm{X'}_{(i)}\mathrm{X}_{(i)}\right)^{-1} \mathrm{X'}_{(i)}\mathrm{y}_{(i)} \end{aligned}$$

donde las matrices $(\mathrm{X'X})^{-1}$ y $[\mathrm{X'}_{(i)}\mathrm{X}_{(i)}]^{-1}$ guardan la relación siguiente, llamada identidad de Sherman-Morrison-Woodbury.

$$\left[\mathrm{X'}_{(i)}\mathrm{X}_{(i)}^{-1}\right] = (\mathrm{X'X})^{-1} + \frac{(\mathrm{X'X})^{-1}\,\mathrm{x}_i\mathrm{x}_i'\,(\mathrm{X'X})^{-1}}{1 - h_{ii}}$$

y $h_{ii} = \mathrm{x'}_i\,(\mathrm{X'X})^{-1}\,\mathrm{x}_i$. Usando la anterior relación entre las matrices

$$
\begin{aligned}
e_{(i)} &= \mathrm{y}_i - \mathrm{x'}_i \left[(\mathrm{X'X})^{-1} + \frac{(\mathrm{X'X})^{-1}\,\mathrm{x}_i\mathrm{x'}_i\,(\mathrm{X'X})^{-1}}{1-h_{ii}}\right] \mathrm{X'}_{(i)}\mathrm{y}_{(i)} \\
&= \mathrm{y}_i - \mathrm{x'}_i\,(\mathrm{X'X})^{-1}\,\mathrm{X'}_{(i)}\mathrm{y}_{(i)} - \frac{\mathrm{x'}_i\,(\mathrm{X'X})^{-1}\,\mathrm{x}_i\mathrm{x'}_i\,(\mathrm{X'X})^{-1}\,\mathrm{X'}_{(i)}\mathrm{y}_{(i)}}{1-h_{ii}} \\
&\frac{(1-h_{ii})\mathrm{y}_i - (1-h_{ii})\mathrm{x'}_i\,(\mathrm{X'X})^{-1}\,\mathrm{X'}_{(i)}\mathrm{y}_{(i)} - h_{ii}\mathrm{x'}_i\,(\mathrm{X'X})^{-1}\,\mathrm{X'}_{(i)}\mathrm{y}_{(i)}}{1-h_{ii}} \\
&\frac{(1-h_{ii})\mathrm{y}_i - \mathrm{x'}_i\,(\mathrm{X'X})^{-1}\,\mathrm{X'}_{(i)}\mathrm{y}_{(i)}}{1-h_{ii}}
\end{aligned}
$$

Finalmente, ya que $\mathrm{X'y} = \mathrm{X'}_{(i)}\mathrm{y}_i + \mathrm{x}_i\,y_i$, llegamos a la ecuación

$$
\begin{aligned}
e_{(i)} &= \frac{(1-h_{ii})\mathrm{y}_i - \mathrm{x'}_i\,(\mathrm{X'X})^{-1}\,(\mathrm{X'y} - \mathrm{x}_i y_i)}{1-h_{ii}} \\
&= \frac{(1-h_{ii})y_i - \mathrm{x'}_i\,(\mathrm{X'X})^{-1}\,\mathrm{X'y} + \mathrm{x'}_i\,(\mathrm{X'X})\,\mathrm{x}_i y_i}{1-h_{ii}} \\
&= \frac{(1-h_{ii})y_i - \mathrm{x'}_i\hat{\beta} + h_{ii}y_i}{1-h_{ii}} \\
&= \frac{y_i - \mathrm{x'}_i\hat{\beta}}{1-h_{ii}}
\end{aligned}
$$

donde el numerador se corresponde con el valor del residuo e_i y, en consecuencia, tenemos la igualdad.

□

De la Ecuación (3.5) es fácil ver que el residuo PRESS es tan solo un residuo ordinario ponderado en función de los elementos diagonales de la matriz H. Los puntos altamente influyentes serán aquellos con un valor elevado de h_{ii}, luego un residuo PRESS pequeño. Generalmente, una gran diferencia con el residuo ordinario significará que es un punto donde el modelo se ajusta bien.

La varianza del residuo PRESS ($e_{(i)}$) es

$$\mathrm{Var}[e_{(i)}] = \mathrm{Var}\left[\frac{e_i}{1-h_{ii}}\right] = \frac{1}{(1-h_{ii})^2}[\sigma^2(1-h_{ii})] = \frac{\sigma^2}{1-h_{ii}}$$

luego el residuo estandarizado será

$$\frac{e_{(i)}}{\sqrt{\mathrm{Var}[e_{(i)}]}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

que, utilizando MS_{Res} como estimación de σ^2, se corresponde con el residuo estudentizado.

A partir de los residuos PRESS, se define el llamado estadístico PRESS mediante la expresión

$$PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 \tag{3.6}$$

Este estadístico se utiliza principalmente como medida de evaluación del modelo, es decir, nos indica cómo de precisas seran las predicciones futuras hechas por el modelo de regresión. Un modelo con un valor pequeño del estadístico PRESS es preferible a uno donde sea grande.

- **R-Student** En los residuos estudentizados hemos utilizado MS_{Res} como estimador de σ^2 a la hora de calcular los residuos r_i. En este caso, vamos a estimar el parámetro de la varianza basándonos en todos los datos salvo la observación i. Denotamos a la aproximación de σ^2 por $S^2_{(i)}$, y toma el valor

$$S^2_{(i)} = \frac{(n-p)MS_{Res} - e_i^2/(1-h_{ii})}{n-p-1} \tag{3.7}$$

Demostración : A lo largo de la demostración haremos uso de la identidad de Sherman-Morrison-Woodbury

$$\left[\mathrm{X'}_{(i)} \mathrm{X}^{-1}_{(i)} \right] = (\mathrm{X'X})^{-1} + \frac{(\mathrm{X'X})^{-1} \mathrm{x}_i \mathrm{x'}_i (\mathrm{X'X})^{-1}}{1 - h_{ii}}$$

Si multiplicamos ambos lados de la igualdad por $\mathrm{X'y} - \mathrm{x}_i \mathrm{y}_i$, obtenemos

$$\hat{\beta}_{(i)} = \hat{\beta} - (\mathrm{X'X})^{-1} \mathrm{x}_i y_i + \frac{(\mathrm{X'X})^{-1} \mathrm{x}_i \mathrm{x'}_i (\mathrm{X'X})^{-1} (\mathrm{X'y} - \mathrm{x}_i y_i)}{1 - h_{ii}}$$

que se reduce a

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{(\mathrm{X'X})^{-1} \mathrm{x}_i e_i}{1 - h_{ii}} \tag{3.8}$$

Ahora

$$(n-p-1)S^2_{(i)} = \sum_{j \neq i} (y_j - \mathrm{x'}_j \hat{\beta}_{(i)})^2 \tag{3.9}$$

y usando la ecuación (3.8), esto se convierte en

$$\sum_{j \neq i} \left(y_j - \mathrm{x'}_j \hat{\beta}_{(i)} \right)^2 = \sum_{j=1}^{n} \left(y_j - \mathrm{x'}_j \hat{\beta} + \frac{\mathrm{x'}_j (\mathrm{X'X})^{-1} \mathrm{x}_i e_i}{1 - h_{ii}} \right)^2 - \left(y_i - \mathrm{x'}_i \hat{\beta} + \frac{h_{ii} e_i}{1 - h_{ii}} \right)^2$$
$$= \sum_{j=1}^{n} \left(e_j + \frac{h_{ii} e_i}{1 - h_{ii}} \right)^2 - \frac{e_i^2}{(1 - h_{ii})^2}$$

Desarrollando el cuadrado del primer término

$$\sum_{j=1}^{n}\left(e_j+\frac{h_{ii}e_i}{1-h_{ii}}\right)^2=\sum_{j=1}^{n}e_j^2+\frac{2e_i}{1-h_{ii}}\sum_{j=1}^{n}e_jh_{ij}-\frac{e_i^2}{(1-h_{ii})^2}\sum_{j=1}^{n}h_{ij}^2$$

Sin embargo, ya que Hy=Hŷ, $\sum_{j=1}^{n}e_jh_{ij}=0$, H es idempotente y se cumple $\sum_{j=1}^{n}h_{ij}^2=h_{ii}$, podemos escribir (3.9) como

$$\begin{aligned}(n-p-1)S_{(i)}^2&=\sum_{j=1}^{n}e_j^2+\frac{h_{ii}e_i^2}{(1-h_{ii})^2}-\frac{e_i^2}{(1-h_{ii})^2}\\&=\sum_{j=1}^{n}e_j^2-\frac{e_i^2}{(1-h_{ii})}\\&=(n-p)MS_{Res}-\frac{e_i^2}{(1-h_{ii})}\end{aligned}$$

Llegando finalmente a la ecuación que buscábamos

$$S_{(i)}^2=\frac{(n-p)MS_{Res}-e_i^2/(1-h_{ii})}{n-p-1}$$

□

Se utiliza para estimar σ^2 el parámetro obtenido en (3.7) en lugar de MS_{Res}, con la finalidad de producir un residuo estudentizado de forma externa, el cual es llamado R-student. La expresión de dicho residuo es

$$t_i=\frac{e_i}{\sqrt{S_{(i)}^2(1-h_{ii})}},\quad i=1,2,\ldots,n \tag{3.10}$$

En muchas ocasiones difiere poco del residuo r_i, aunque cuando la observación i es influyente, el estimador $S_{(i)}^2$ será considerablemente distinto de MS_{Res}. Por este motivo, el estadístico R-student es más sensible a estos puntos.

Datos auto-mpg 2.1

Para ilustrar la diferencia producida entre las distintas escalas de residuos, se ha implementado en el programa R el conjunto de datos denominado 'auto-mpg' [11]. En este conjunto de datos se resumen 9 variables sobre 398 automóviles tomados como muestra. De estas 9 variables, 4 de ellas son variables factor ('cylinders', 'model year', 'origin', 'car name'), y las 5 restantes son numéricas ('mpg', 'displacement', 'horsepower', 'weight', 'acceleration'). El objetivo del modelo de regresión lineal múltiple planteado es predecir las millas recorridas por galón ('mpg') de cada automóvil en función de las 4 variables cuantitativas restantes.

El modelo de regresión obtenido es $\vec{y} = \mathrm{X}\vec{\beta} + \vec{\epsilon}$, donde $\vec{y}$ es un vector columna de longitud $n = 398$; X es la submatriz del conjunto de datos inicial que presenta los valores de las 4 variables explicativas, concatenada a un vector de unos, y con dimensión total $n \times p$, donde $p = 5$. Por último tenemos el vector columna de coeficientes de regresión $\vec{\beta} = (\mathrm{X'X})^{-1}\mathrm{X'}\vec{y}$ de longitud p. El vector $\vec{\epsilon}$ es el llamado vector de errores aleatorios.

Para el conjunto de datos 'auto-mpg' se han calculado las distintas escalas de residuos comentadas previamente, a partir de las expresiones obtenidas en cada apartado. Los 25 primeros valores (de los 398 en cada columna), se pueden observar en la Tabla 3.1.

	e_i	d_i	r_i	$e_{(i)}$	t_i
1	-0.22	-0.05	-0.05	-0.22	-0.01
2	-9.06	-2.12	-2.15	-9.24	-0.51
3	6.60	1.55	1.55	6.65	0.37
4	-0.24	-0.06	-0.06	-0.24	-0.01
5	-4.82	-1.13	-1.13	-4.86	-0.27
6	-2.88	-0.67	-0.68	-2.91	-0.16
7	-1.40	-0.33	-0.33	-1.42	-0.08
8	6.83	1.60	1.61	6.89	0.38
9	-2.10	-0.49	-0.50	-2.13	-0.12
10	-3.65	-0.86	-0.86	-3.68	-0.20
11	4.41	1.03	1.04	4.49	0.24
12	6.82	1.60	1.61	6.92	0.38
13	-3.95	-0.93	-0.93	-3.98	-0.22
14	3.25	0.76	0.76	3.27	0.18
15	2.08	0.49	0.49	2.12	0.12
16	-1.35	-0.32	-0.32	-1.36	-0.07
17	-2.16	-0.51	-0.51	-2.18	-0.12
18	2.23	0.52	0.53	2.26	0.12
19	1.61	0.38	0.38	1.65	0.09
20	1.90	0.45	0.45	1.91	0.10
21	-0.49	-0.12	-0.12	-0.50	-0.03
22	0.20	0.05	0.05	0.20	0.01
23	-7.35	-1.72	-1.74	-7.48	-0.41
24	-6.68	-1.57	-1.57	-6.75	-0.37
25	1.40	0.33	0.33	1.41	0.08

Tabla 3.1: Diferentes escalas de residuos para los datos 'auto-mpg'

3.1.3. Gráficos de residuos

La representación gráfica de los residuos es una manera efectiva de investigar la precisión del ajuste de un modelo de regresión y de verificar las

hipótesis establecidas al principio del Capítulo 3. En este apartado se introducen los gráficos más comunes, los cuales serán representados mediante R. En su representación se trabaja con los residuos estandarizados d_i.

- **P-P Plot y Q-Q Plot** El P-P Plot se basa en la construcción de la probabilidad normal teórica en relación a los residuos. Es una gráfico diseñado de tal forma que la distribución normal acumulada se representa como una línea recta. Sean $d_{[1]} < d_{[2]} < \ldots < d_{[n]}$ los residuos estandarizados ordenados de menor a mayor valor. Si representamos $d_{[i]}$ frente a la probabilidad acumulada $P_i = (i - \frac{1}{2})/n, \quad i = 1, 2, \ldots, n$ en el gráfico de probabilidad normal, los puntos deberán estar en torno a dicha línea recta. Esta recta normalmente se determina de forma visual haciendo énfasis en los valores centrales antes que en los extremos. Si los puntos representados se alejan de la recta substancialmente, entonces la distribución de los residuos no es normal. En el software R, en lugar de representar las probabilidades se representan los valores esperados de una normal estándar Φ. Esto se deduce de $E(d_{[i]}) \approx \Phi^{-1}\left[(i - \frac{1}{2})/n\right]$. Este tipo de gráfico recibe el nombre de 'Q-Q Plot'

  Para ejemplificarlo se hace uso de los datos 'auto-mpg', donde el gráfico 'Normal Q-Q' es el de la Figura 3.1

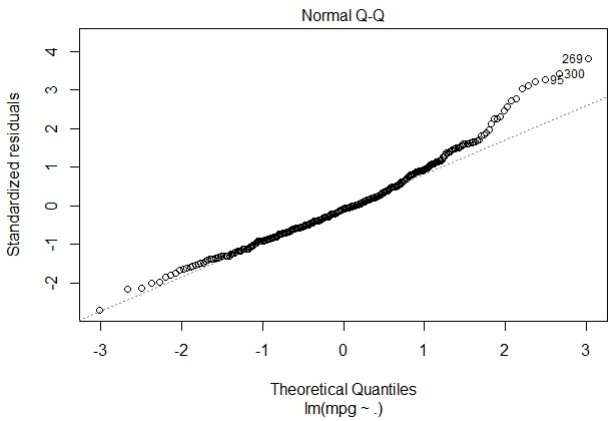


Figura 3.1: Gráfico normal de los residuos estandarizados

Claramente podemos observar que los residuos se ajustan bastante bien a la línea recta, salvo en los datos del extremo derecho. Por tanto, podemos suponer que siguen una distribución normal. Los datos del extremo derecho presentan un indicio de outlier para esos valores. De manera alternativa, una forma bastante rutinaria de intuir si un conjunto de valores se ajusta a una distribución normal es represen-

tar dichos datos en un diagrama de cajas como el de la Figura 3.2. En el 'boxplot' se visualiza claramente una distribución simétrica en torno al 0, a excepción de esos últimos residuos que toman valores más elevados, como pasaba en la Figura 3.1.

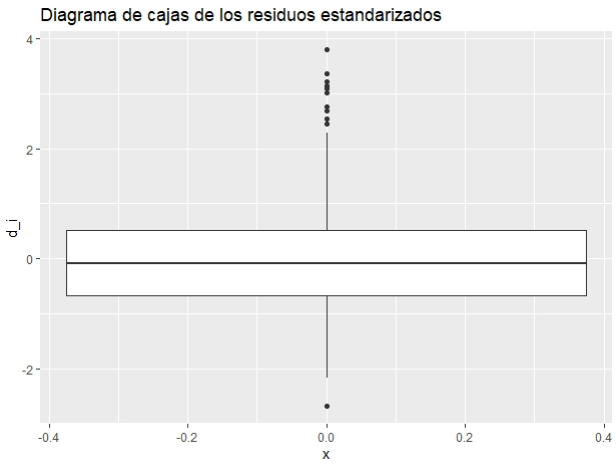


Figura 3.2: Diagrama de cajas de los residuos d_i

En casos generales, cuando disponemos de una muestra pequeña, generalmente se producen gráficos que se desvían bastante de la linealidad. Por el contrario, para tamaños de muestra grandes, los gráficos se comportan mucho mejor. Si tenemos que establecer algún tipo de criterio para el tamaño de muestra, se puede considerar que a partir de 20-30 puntos son necesarios para producir un gráfico suficientemente estable para ser interpretado.

En los 'Q-Q Plot' es difícil detectar anomalías incluso aunque los errores ϵ_i no sigan una distribución normal. Este problema tiene su origen en que los residuos no son una muestra aleatoria simple, si no que son restos de la estimación de los parámetros. De hecho, los residuos son combinaciones lineales de los errores del modelo, en consecuencia ajustar los parámetros tiende a destruir los indicios de no normalidad en los residuos.

- **Gráfico de los residuos frente a los valores ajustados** Una representación los residuos (e_i) frente a los correspondientes valores ajustados $\hat{y}_i$ es útil para detectar varios desajustes comunes. Si los residuos representados pueden acotarse en una banda horizontal, entonces no hay defectos obvios en el modelo. Si, por el contrario, observamos una tendencia clara de los residuos, esto puede indicar que la relación estudiada entre las variables no es la indicada (por ejemplo podría ser mejor un modelo polinomial) o que los residuos no son incorrelados. También puede darse el caso en el que los datos aumenten (o

disminuyan) su dispersión en funcion de y. Algo que combina las dos posibilidades de este último caso puede darse cuando y es una proporción entre 0 y 1, como pasa en el modelo binomial donde la varianza es mayor en 0.5 que en 1.

Estos gráficos también nos muestran outliers potenciales. Antes de considerarlos como puntos remotos es recomendable estudiar si puede deberse a problemas en la varianza de los datos, o a la no linealidad entre las variables respuesta y regresores.

Los residuos frente a los valores ajustados presentan una representación para el conjunto de datos 'auto-mpg' de la forma mostrada en la Figura 3.3.

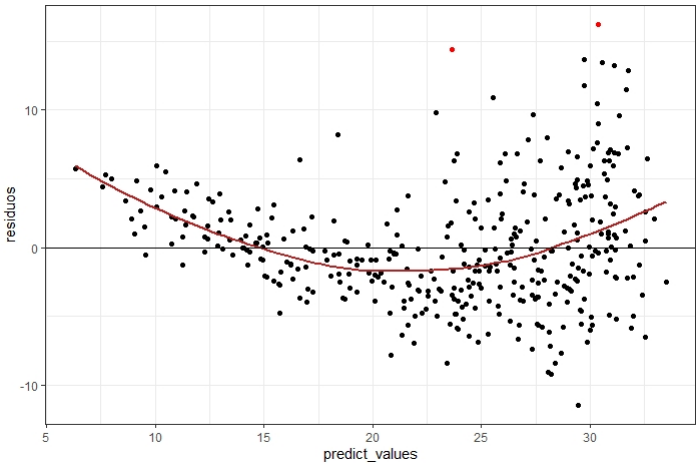


Figura 3.3: Residuos e_i frente a los valores ajustados

A simple vista no se observan problemas en tratar el modelo de regresión lineal que se ha supuesto. Es más, en la gráfica aparecen 2 individuos, el 300 y el 269 coloreados en rojo, que puede intuirse que son outliers, ya que en el 'Q-Q Plot' también presentaban anomalías.

En el gráfico anterior se han utilizado los residuos originales e_i; pero el software R también nos da la posibilidad de mostrar por pantalla este tipo de gráficos para la raíz cuadrada de los residuos estandarizados $\sqrt{d_i}$ frente a los valores ajustados. Se utiliza principalmente para comprobar que la homocedasticidad de los datos se verifica. Diremos que un modelo de regresión lineal es homocedástico si la varianza de los residuos se mantiene constante.

En la Figura 3.4, como no se observa ningún patrón definido por los residuos entonces podemos suponer que tienen la misma varianza.

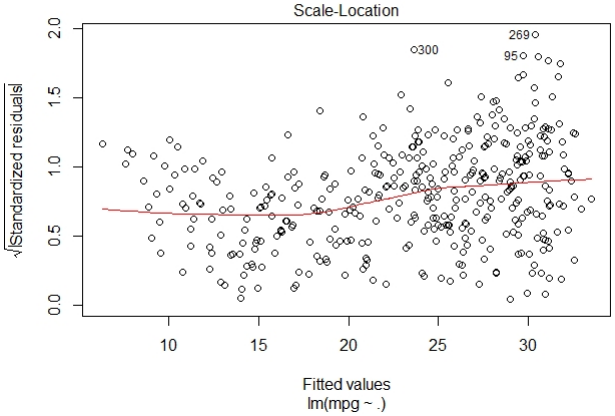


Figura 3.4: Residuos $\sqrt{d_i}$ frente a los valores ajustados

- **Gráfico de residuos frente al apalancamiento**

  En este tipo de gráficos juega un papel importante la llamada distancia de Cook. Esta distancia mide cómo cambia el vector de estimadores $\hat{\beta}$ cuando se elimina cada observación.

  El objetivo es detectar puntos de apalancamiento, luego una expresión útil para la distancia de Cook donde se hace uso de los residuos t_i y de la matriz de influencia H es

  $$D_i = \left(\frac{1}{p}\right) t_i^2 \frac{h_{ii}}{1 - h_{ii}}$$

  Esta definición permite de forma intuitiva interpretar la distancia de Cook. Si el valor t_i es muy grande, hará que la distancia de Cook aumente siempre que no tengamos un valor cercano a 0 de h_{ii}. Por tanto, nos proporciona información sobre los posibles sujetos influyentes sobre el modelo de regresión. La distancia de Cook se explica de manera más detallada en la Subsección 3.3.2 de este mismo capítulo.

  Este tipo de gráficos contienen los residuos estandarizados y se utilizan para interpretar qué datos pueden perturbar el análisis. Si visualizamos algún dato por fuera del valor 0.5 en la distancia de Cook (línea discontinua), podemos hacer el análisis sin ese punto y veremos que el modelo se verá alterado. Es reconocido el criterio en el que si la distancia de Cook $D_i > 0.5$, entonces el i-ésimo punto tiene gran influencia en el modelo, mientras que cuando $D_i > 1$ el punto se considera outlier.

  En el gráfico de apalancamiento de la Figura 3.5 para el conjunto de datos 'auto-mpg', no aparece ningún sujeto fuera de la zona determinada por la distancia de Cook. Aún así, después de haber representado

los diferentes gráficos de residuos llaman la atención los individuos 300, 269 y 95 que podrían ser posibles puntos influyentes.

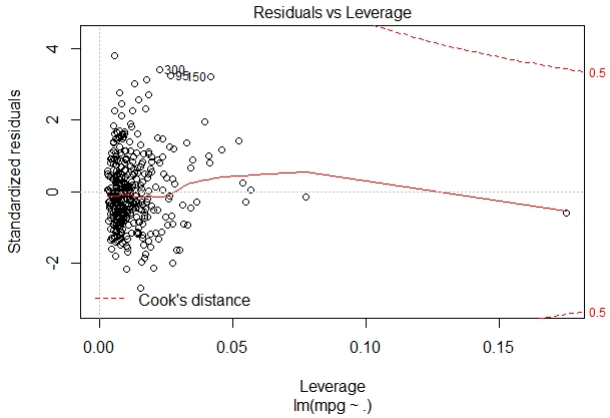


Figura 3.5: Residuos $\sqrt{d_i}$ frente al apalancamiento

- **Gráficos de regresión parcial** Son una variación de los gráficos que enfrentan los residuos con los predictores, lo que implica que son una manera mejorada de estudiar la relación marginal de un regresor. Este tipo de gráficos resulta útil para evaluar si hemos expresado la relación entre las variables explicativas y la explicada de una forma correcta. Además, nos permite conocer si alguno de los regresores no aporta apenas información al modelo.

  Los gráficos de regresión parciales consideran la aportación marginal de un regresor x_j mediante el estudio del resto de las variables explicativas. En este tipo de representaciones, la variable respuesta y se toma junto al regresor x_j para contruir el modelo de regresión a partir del resto de variables explicativas. Repitiendo este procedimiento para cada variable explicativa del modelo, obtenemos información sobre la naturaleza de la relación marginal para cada regresor x_j tomado.

  Para ilustrarlo matemáticamente, supongamos que tenemos el modelo de regresión $\mathrm{y} = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \epsilon$. Supongamos que queremos verificar si la relación entre x_1 e y está correctamente especificada. Para ello haremos la regresión de y con x_2 y obtendremos unos valores ajustados y unos residuos:

$$\hat{y}_i(x_2) = \hat{\theta}_0 + \hat{\theta}_1 x_{i2}$$
$$e_i(y|x_2) = y_i - \hat{y}_i(x_2), \quad i = 1, 2, \ldots, n$$

Ahora apliquemos la regresión a x_1 y x_2 y calculemos los residuos:

$$\hat{x}_{i1}(x_2) = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2}$$
$$e_i(x_1|x_2) = x_{i1} - \hat{x}_{i1}(x_2), \quad i = 1, 2, \ldots, n$$

El gráfico de regresión parcial del regresor x_1 se obtiene de representar los residuos $e_i(y|x_2)$ frente a $e_i(x_1|x_2)$. Si la variable x_1 aporta una relación lineal al modelo, el gráfico deberá mostrar una recta con pendiente distinta de 0. Esta pendiente se corresponderá con el coeficiente de regresión de x_1 en el modelo múltiple. Si por el contrario el gráfico muestra una curva de otro tipo, entonces se necesitarán términos de mayor orden de x_1 o una transformación de esta variable. Cuando los datos representados aparecen a lo largo de una banda horizontal en el gráfico de residuos parcial, nos indica que el regresor x_1 no añade información adicional para predecir los valores de y.

- **Gráfico de residuos parciales** Está relacionado de manera estrecha con el gráfico de regresión parcial, ya que también está diseñado para mostrar la relación entre la variable respuesta y los regresores. Suponiendo que tenemos x_1, x_2,..., x_k regresores, se definen los residuos parciales de x_j como

$$e_i^*(y|x_j) = e_i + \hat{\beta}_j x_{ij}, \quad i = 1, 2, \ldots, n$$

donde e_i son los residuos ordinarios con todos los regresores en el modelo. Cuando se representan estos residuos parciales frente a x_{ij} el gráfico obtenido es una recta con pendiente $\hat{\beta}_j$, el coeficiente de regresión asociado a x_j en el modelo. La interpretación de estos gráficos es similar a la hecha con los gráficos de regresión parcial.

3.2. Detección y tratamiento de outliers

Un outlier es una observación que toma valores extremos, es decir, que es considerablemente distinta al resto de datos. Aquellos residuos que tengan un valor absoluto bastante elevado, indican un outlier potencial en el espacio y. Dependiendo de su localización en el espacio x, los outliers pueden tener una influencia moderada o severa en el modelo de regresión. Para identificarlos son de gran utilidad los gráficos de residuos expuestos previamente, especialmente el 'Q-Q Plot' y el diagrama que muestra los residuos frente al apalancamiento. Es conveniente utilizar los residuos escalados, como los estudentizados y los R-student, ya que facilitan la detección de outliers potenciales.

Los outliers deben de ser investigados con cierto cuidado para detectar la razón de su comportamiento tan extraño. Algunas veces puede tratarse de

un error en la medida o un evento inusual pero hasta cierto punto razonable. Si la presencia de estos sujetos es debida a un error en el instrumento de medida o un fallo en la recolección de datos, el outlier debería de ser corregido si es posible o, en otro caso, eliminado del conjunto. Es necesario tener una evidencia no estadística fuerte de que el outlier es un valor erróneo antes de que sea descartado, ya que eliminarlos del proceso es siempre deseable para ajustar de una manera más precisa el resto de valores. Omitir un outlier que pueda ser perfectamente plausible para mejorar el modelo de regresión puede acabar siendo peligroso, ya que puede dar al intérprete una falsa realidad de precisión en la estimación. Los outliers también pueden indicar una insuficiencia en el modelo para ajustar ciertos datos de una región concreta.

Hay diversos test de hipótesis para contrastar si un punto determinado es un outlier. A continuación se hará el desarrollo teórico de un test basado en R-Student [27].

Supóngase que el modelo con el que se trabaja es $\mathrm{y} = \mathrm{X}\beta + \epsilon$, cuando en realidad el modelo correcto es

$$\mathrm{y} = \mathrm{X}\beta + \delta + \epsilon \tag{3.11}$$

donde δ es un vector columna de ceros de longitud n, a excepción de la observación en la posición u, que toma el valor δ_u. Así,

$$\delta = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \delta_u \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Para ambos modelos se supone que $E(\epsilon) \sim N(o, \sigma^2\mathrm{I})$. El objetivo es encontrar un test apropiado para la hipótesis

$$H_0 : \delta_u = 0, \quad H_1 : \delta_u \neq 0$$

Este procedimiento asume que el interés recae esencialmente sobre la observación u, ya que se dispone información sobre la presencia de un outlier en dicho dato.

El primer paso es encontrar una estimación apropiada de δ_u. Un candidato lógico es el residuo e_u. Sea $e = [\mathrm{I} - \mathrm{X}(\mathrm{X'X})^{-1}\mathrm{X'}]\mathrm{y}$ el vector de residuos $n \times 1$. El valor esperado de e es

$$\begin{aligned} &E(e) = E([\mathrm{I} - \mathrm{X}(\mathrm{X'X})^{-1}\mathrm{X'}]\mathrm{y}) = [\mathrm{I} - \mathrm{X}(\mathrm{X'X})^{-1}\mathrm{X'}]E(\mathrm{y}) \\ &= [\mathrm{I} - \mathrm{X}(\mathrm{X'X})^{-1}\mathrm{X'}][\mathrm{X}\beta + \delta] = [\mathrm{I} - \mathrm{X}(\mathrm{X'X})^{-1}\mathrm{X'}]\mathrm{X}\beta + [\mathrm{I} - \mathrm{X}(\mathrm{X'X})^{-1}\mathrm{X'}]\delta \\ &= [\mathrm{X\text{-}X}]\beta + [\mathrm{I} - \mathrm{X}(\mathrm{X'X})^{-1}\mathrm{X'}]\delta = [\mathrm{I} - \mathrm{X}(\mathrm{X'X})^{-1}\mathrm{X'}]\delta \end{aligned}$$

Luego,

$$E(e_u) = (-h_{uu})\delta_u$$

donde h_{uu} es el u-ésimo valor de la diagonal de la matriz de influencia H. De esta forma, el estimador insesgado de δ_u es

$$\hat{\delta}_u = \frac{e_u}{1 - h_{uu}}$$

La expresión de $\hat{\delta}_u$ coincide justamente con el sumando u del estadístico PRESS en (3.6). El siguiente paso es determinar la varianza del estimador. Nótese que

$$\begin{aligned}
&\text{Var}(e) = \text{Var}([\text{I} - \text{X}(\text{X'X})^{-1}\text{X'}]\text{y}) \\
&= [\text{I} - \text{X}(\text{X'X})^{-1}\text{X'}]\sigma^2 I[\text{I} - \text{X}(\text{X'X})^{-1}\text{X'}]' \\
&= \sigma^2[\text{I} - \text{X}(\text{X'X})^{-1}\text{X'}][\text{I} - \text{X}(\text{X'X})^{-1}\text{X'}] \\
&= \sigma^2[\text{I} - \text{X}(\text{X'X})^{-1}\text{X'}]
\end{aligned}$$

Luego, $\text{Var}(e_u) = (1 - h_{uu})\sigma^2$. La varianza de $\hat{\delta}_u$ es por tanto

$$\text{Var}(\frac{e_u}{1 - h_{uu}}) = \frac{1}{(1 - h_{uu})^2}\text{Var}(e_u) = \frac{(1 - h_{uu})\sigma^2}{(1 - h_{uu})^2} = \frac{\sigma^2}{1 - h_{uu}}$$

Nótese que e es una combinación lineal de Y. En consecuencia, e es combinación lineal de variables aleatorias distribuidas según una distribución normal. De aquí se sigue que e sigue una distribución normal de la misma forma que $\hat{\delta}_u$ también. Consecuentemente, bajo la hipótesis nula $H_0 : \delta_u = 0$,

$$\frac{e_u/(1 - h_{uu})}{\sigma/(\sqrt{1 - h_{uu}})} = \frac{e_u}{\sigma\sqrt{1 - h_{uu}}}$$

sigue una distribución normal estándar. En general, σ^2 es desconocido, así que se utiliza MS_{Res} como estimador insesgado. Es más, ya hemos visto que $\frac{MS_{Res}}{\sigma^2}$ es una variable aleatoria que, dividida por sus grados de libertad, se comporta según una χ^2. De esta manera, el candidato a estadístico del test es

$$\frac{e_u}{\sqrt{MS_{Res}(1 - h_{uu})}}$$

que sigue una distribución t siempre que $e = \text{I} - [\text{X}(\text{X'X})^{-1}\text{X'}]\text{y}$, y $SS_{Res} = \text{y}'[\text{I} - \text{X}(\text{X'X})^{-1}\text{X'}]\text{y}$ sean independientes. Podemos demostrar que e y SS_{Res} son independientes si

$$[\text{I} - \text{X}(\text{X'X})^{-1}\text{X'}]\sigma^2 I[\text{I} - \text{X}(\text{X'X})^{-1}\text{X'}]' = 0$$

Pero, desafortunadamente, esta igualdad no se da. El problema reside en

$$SS_{Res} = e'e = \sum_{i=1}^{n} e_i^2$$

lo que significa que SS_{Res} está correlado con cada componente del residuo, porque el cuadrado de cada componente residual es un sumando de SS_{Res}. En la ecuación (3.7) habíamos desarrollado un estimador de σ^2 donde la observación u es eliminada. Esta estimación es independiente de e_u debido a la hipótesis de independencia de los errores aleatorios. Como resultado, un estadístico apropiado para el modelo definido es

$$\frac{e_u}{S_{(u)}\sqrt{1-h_{uu}}}$$

que coincide con el residuo t_i o R-student. Bajo la hipótesis $H_0 : \delta_u = 0$, este estadísitico sigue una distribución t_{n-p-1}, y bajo $H_1 : \delta_u \neq 0$, el estadístico sigue una distribución $t'_{n-p-1,\gamma}$, donde

$$\gamma = \frac{\delta_u}{\sigma/(\sqrt{1-h_{uu}})} = \frac{\delta_u\sqrt{1-h_{uu}}}{\sigma}$$

Es conveniente notar que la importancia del test recae en h_{uu}.

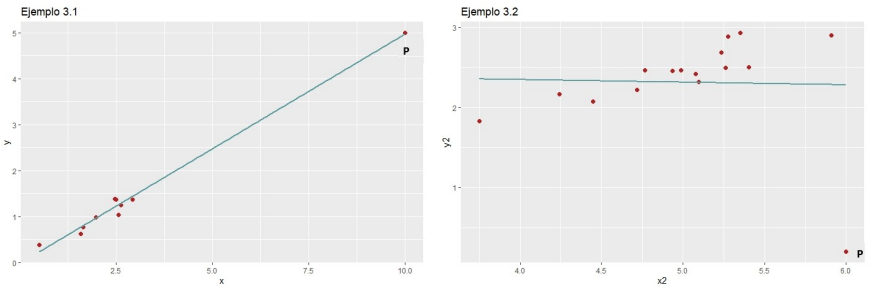
3.3. Diagnóstico del apalancamiento y la influencia

"All models are wrong; some models are useful" ("Todos los modelos son erróneos; algunos son útiles"), es la cita atribuida al famoso estadístico británico George Box. Esta sentencia resume a la perfección lo que cabe esperar de un modelo de regresión. Cuando decidimos construir una recta de regresión que resuma n observaciones estamos renunciando a $n-2$ grados de libertad para estimar el error, cuando en su lugar podríamos haber elegido un modelo polinomial de mayor orden. La simplicidad de una recta es lo que lleva al analista a elegir este tipo de modelos y, por tanto, el ajuste lineal tiene que resumir la mayor cantidad de información posible si queremos que resulte fiable. De esta manera, resulta vital detectar los puntos influyentes que causan desajustes en el modelo de regresión. En esta sección trataremos de diagnosticar la influencia que ejercen los outliers. Una vez tratado este tema, ya se habrá establecido una base más que suficiente para introducir los procedimientos de la regresión lineal robusta, creados con la intención de solucionar los problemas de normalidad y de desajustes del modelo de mínimos cuadrados.

Considérense las situaciones de la Figura 3.6.

En la situación ilustrada en la Figura 3.6 (a), el punto etiquetado P de esta figura está considerablemente alejado del resto de la muestra pero, en cambio, está situado prácticamente en la misma recta que pasa por el resto de observaciones. Este dato es un ejemplo de punto de apalancamiento; es decir, tiene un valor inusual en el eje x pero no afecta en ningún momento la

estimación de los coeficientes de regresión, aunque tendrá un efecto considerable en algunos estadísticos como el R^2. Fijándonos ahora en la situación representada en la Figura 3.6 (b), el punto etiquetado P toma un valor algo extraño en la coodernadas x e y. Esta observación recibe el nombre de punto influyente, es decir, tiene un impacto notable sobre los coeficientes de regresión y 'empuja' la recta hacia su posición.



(a) Punto de apalancamiento P (b) Observación influyente P

Figura 3.6: Ejemplos de outliers

En algunas situaciones puede aparecer un subconjunto de datos que ejerzan una influencia desproporcionada sobre el modelo y sus propiedades. En casos extremos, es posible que los parámetros estimados dependan más del subconjunto de puntos influyentes que de la mayoría de datos. Obviamente, es una situación nada deseable ya que buscamos que el modelo de regresión sea representativo de todos los datos, no solo de unos pocos. Por tanto, el objetivo es encontrar los puntos influyentes y evaluar su impacto sobre el modelo. Si estas observaciones son erróneas, bastará con eliminarlas del conjunto de datos; en caso contrario, puede hacernos replantear si el modelo lineal es adecuado o si el método de mínimos cuadrados aplicado está siendo informativo.

En esta sección se presentarán distintos diagnósticos sobre el apalancamiento y la influencia, los cuales han de ser aplicados junto a las técnicas de análisis de residuos presentadas en la Sección 3.1.

3.3.1. Apalancamiento

En los ejemplos anteriores se ha visto que la situación de los puntos en el eje x es importante para determinar las propiedades del modelo de regresión. En particular, los outliers tienden a causar variaciones en las estimaciones, los errores, los valores predichos y los estadísticos resumen. La matriz de influencia

$$\mathrm{H} = \mathrm{X(X'X)^{-1}X'}$$

juega un papel esencial a la hora de detectar observaciones influyentes. Como se ha explicado previamente, H determina las varianzas y covarianzas de $\hat{y}$

y los residuos, ya que $\mathrm{Var}(\hat{\mathrm{y}}) = \sigma^2\mathrm{H}$ y $\mathrm{Var}(\mathrm{e}) = \sigma^2(\mathrm{I}-\mathrm{H})$. Los elementos h_{ij} de la matriz H se interpretan como la cantidad de apalancamiento ejercido por la observación y_i sobre el valor ajustado $\hat{y}_j$, es decir, un valor 'grande' significará que la influencia ejercida es 'inusual' con respecto al resto de la muestra.

Generalmente se hace hincapié en los valores de la diagonal h_{ii} de la matriz H, que se escriben

$$h_{ii} = \mathrm{x'}_i(\mathrm{X'X})^{-1}\mathrm{x}_i$$

donde $\mathrm{x'}_i$ es la i-ésima fila de la matriz X. La diagonal de la matriz de influencia H es una medida de distancia estandarizada de la observación i al centro del espacio x. Así, los valores diagonales elevados revelan influencias potenciales debidos a su posición en el eje x respecto al resto de la muestra. El rango de la matriz H es p ya que coincide con el de la matriz X, por lo que $\overline{h} = p/n$ donde n es el número de columnas de H. Se asume como criterio para detectar puntos remotos que cualquier observación que exceda el valor diagonal de la matriz H en más de $2p/n$, serán un punto de apalancamiento. Este criterio será válido siempre que $2p/n < 1$.

En la Figura 3.6 (a) se ve un ejemplo de punto de apalancamiento que no ejerce influencia sobre la regresión luego, a pesar de que tendrá un valor diagonal elevado, esto no es criterio para deducir si un punto es influyente o no. Debido a que los elementos diagonales de H solo proporcionan información sobre la localización del datos en el espacio x, algunos analistas estudian de forma parelela los residuos estudentizados o los R-student; ya que observaciones con valores altos tanto en los residuos como en la diagonal son candidatos a ser influyentes.

3.3.2. Distancia de Cook

Cook [5, 6] sugirió una manera de considerar tanto la localización de los puntos en el eje x como la variable respuesta para medir la influencia. Hizo uso de la distancia al cuadrado entre los estimadores por mínimos cuadrados con todas las observaciones $\hat{\beta}$ y los estimadores obtenidos si se elimina el punto i, denotado $\hat{\beta}_{(i)}$. La fórmula de la distancia se escribe

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'\mathrm{M}(\hat{\beta}_{(i)} - \hat{\beta})}{c}, \quad i = 1, 2, \ldots, n \tag{3.12}$$

donde $\mathrm{M} = \mathrm{X'X}$ y $c = pMS_{Res}$, luego un manera análoga de escribir la distancia de Cook es

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'\mathrm{X'X}(\hat{\beta}_{(i)} - \hat{\beta})}{pMS_{Res}}, \quad i = 1, 2, \ldots, n$$

Aquellos puntos con valores grandes de D_i tienen una influencia considerable en la estimación mínimo cuadrada $\hat{\beta}$.

La magnitud tomada por D_i generalmente se evalua comparándola con una $F_{\alpha,p,n-p}$; luego si $D_i = F_{0.5,p,n-p}$, eliminar el punto i desplazará $\hat{\beta}_{(i)}$ hasta el límite de aproximadamente un 50 % de la región de confianza de β basada en el conjunto de datos total. Es un desplazamiento notable que indica que la estimación mínimo cuadrada es sensible al i-ésimo punto. Ya que $F_{0.5,p,n-p} \approx 1$, normalmente consideramos que un punto que cumple que $D_i > 1$ es influyente. Lo deseable sería que cada estimación $\hat{\beta}_{(i)}$ se desplazará a lo sumo un 10 % o 20 % de la región de confianza.

Una manera alternativa de escribir la distancia de Cook es

$$D_i = \frac{r_i^2}{p} \frac{\text{Var}(\hat{y}_i)}{\text{Var}(e_i)} = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}, \quad i = 1, 2, \ldots, n$$

Así, D_i depende del cuadrado del residuo estudentizado r_i y de los valores diagonales de la matriz H. Puede interpretarse como la distancia del vector x_i al centroide de los datos restantes. Luego, D_i es combinación de una componente que refleja cómo ajusta el modelo la observación y_i, y una componente que mide cómo de lejos está ese punto del resto de datos.

Como $\text{X}\hat{\beta}_{(i)} - \text{X}\hat{\beta} = \hat{\text{y}}_{(i)} - \hat{\text{y}}$, otra forma de escribir la distancia de Cook es

$$D_i = \frac{(\hat{\text{y}}_{(i)} - \hat{\text{y}})'(\hat{\text{y}}_{(i)} - \hat{\text{y}})}{pMS_{Res}}, \quad i = 1, 2, \ldots, n$$

Por tanto, una última manera de interpretar la distancia de Cook es como la distancia Euclídea entre los valores ajustados sin la i-ésima observación, y los valores ajustados totales, todo ello dividido por el coeficiente pMS_{Res}.

Lo anterior ha sido desarrollado para el caso en el que se diagnostique la influencia de un único valor sobre el modelo, pero pueden aparecer grupos de outliers que ejerzan un impacto indebido sobre la regresión.

3.3.3. Varianza generalizada y COVRATIO

El diagnóstico D_i proporciona un punto de vista sobre el efecto de la observaciones sobre las estimaciones de los coeficientes $\hat{\beta}_j$ y los valores ajustados $\hat{y}_i$. Sin embargo, no se extrae ninguna información general sobre la precisión de la estimación. En este punto es cuando aparece el concepto de varianza generalizada, definida como el determinante de la matriz de varianzas-covarianzas, y la cual se usa para cuantificar la precisión. Se define la varianza generalizada de $\hat{\beta}$ como

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

Para cuantificar la influencia ejercida por la i-ésima observación en la precisión de las estimaciones, se define

$$COVRATIO_i = \frac{\left|(\text{X'}_{(i)}\text{X}_{(i)})^{-1} S^2_{(i)}\right|}{|(\text{X'X})^{-1} MS_{Res}|}, \quad i = 1, 2, \ldots, n$$

Claramente si $COVRATIO_i > 1$, la observación i mejora la precisión de la estimación, mientras que si $COVRATIO_i < 1$, el i-ésimo punto empeora el modelo en cuanto a su precisión. Computacionalmente,

$$COVRATIO_i = \frac{(S^2_{(i)})^p}{MS^p_{Res}} \left(\frac{1}{1 - h_{ii}} \right)$$

Nótese que $[1/(1 - h_{ii})]$ se corresponde con el cociente de $\left|(\mathrm{X'}_{(i)}\mathrm{X}_{(i)})^{-1}\right|$ por $\left|(\mathrm{X'X})^{-1}\right|$, por lo que un punto de apalancamiento elevado hará que $COVRATIO_i$ también lo sea. Resulta lógico, ya que un punto de apalancamiento elevado siempre ayudará a que el modelo explique mejor la variación de los datos, salvo que sea un outlier en el eje y por lo que empeoraría el ajuste del modelo de la regresión. Si se da este caso, la cantidad $\frac{\left(S^2_{(i)}\right)^p}{MS^p_{Res}}$ será mucho más pequeña que una unidad.

Definir algún tipo de cota que nos haga distinguir si un punto es influyente en la precisión del modelo no es siempre recomendable, tan solo si se dispone de una muestra de tamaño grande. Belsley, Kuh, y Welsch [2] sugirieron que si $COVRATIO_i > 1 + 3p/n$ o si $COVRATIO_i < 1 - 3p/n$, entonces el punto debe de considerarse influyente. La cota inferior solamente es válida cuando $n > 3p$.

3.3.4. Ejemplo de punto de influencia

Sea el conjunto de datos de la Tabla 3.2, formado por dos variables numéricas 'Extraction' y 'Titration' con 20 observaciones muestrales [31].

Si se representan los puntos en un diagrama de dispersión y se ajustan mediante el método de mínimos cuadrados obtendremos la gráfica de la Figura 3.7.

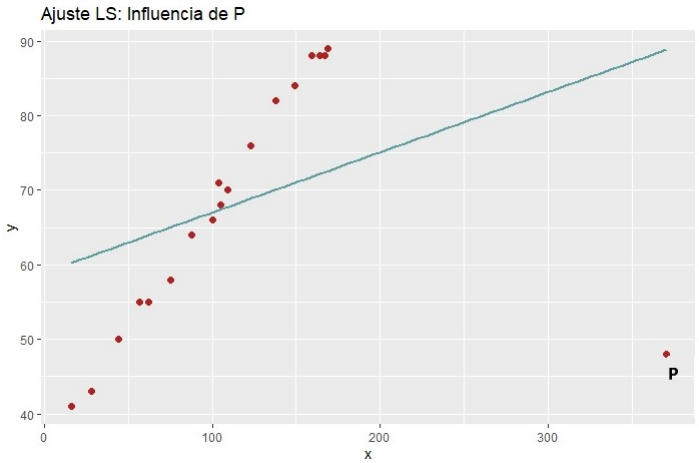


Figura 3.7: Ajuste por mínimos cuadrados: Ejemplo de punto influyente P

ID	Extraction (x)	Titration (y)
1	123.00	76.00
2	109.00	70.00
3	62.00	55.00
4	104.00	71.00
5	57.00	55.00
6	370.00	48.00
7	44.00	50.00
8	100.00	66.00
9	16.00	41.00
10	28.00	43.00
11	138.00	82.00
12	105.00	68.00
13	159.00	88.00
14	75.00	58.00
15	88.00	64.00
16	164.00	88.00
17	169.00	89.00
18	167.00	88.00
19	149.00	84.00
20	167.00	88.00

Tabla 3.2: Conjunto de datos

A simple vista el ajuste obtenido está lejos de ser el deseable, ya que el punto P ejerce una influencia notable sobre el método de mínimos cuadrados. Utilicemos los conceptos de esta sección para realizar un diagnóstico del punto.

Calculamos los coeficientes y estadísticos del modelos de mínimos cuadrados con el dato de influencia y sin él, mostrados en la Tabla 3.3.

Dataframe	$\hat{\beta}_0$	$\hat{\beta}_1$	MS_{Res}	R^2
Todas las observaciones	58.94	0.08	243.32	0.14
Sin la observación 6	35.32	0.32	1.57	0.99

Tabla 3.3: Variación en los coeficientes y estadísticos según se considere el punto de influencia o no

La diferencia entre los valores con la observación 6 y sin ella son notables. Eliminando el dato se consigue explicar prácticamente toda la variabilidad de los datos ya que la estimación de σ^2 se reduce considerablemente.

Si se introduce en R el código

```
model <- lm(y ~ x, data = Datos)
```

```
summary(model)
print(influence.measures(model))
```

podemos extraer los datos diagonales de la matriz de influencia y la distancia de Cook de cada observación. La Tabla 3.4 recoge la información.

ID	cook.d(D_i)	hat(h_{ii})
1	0.01	0.05
2	0.00	0.05
3	0.02	0.08
4	0.00	0.05
5	0.02	0.09
6	14.44	0.62
7	0.04	0.10
8	0.00	0.05
9	0.15	0.15
10	0.11	0.13
11	0.02	0.05
12	0.00	0.05
13	0.04	0.06
14	0.01	0.07
15	0.00	0.06
16	0.04	0.07
17	0.05	0.07
18	0.04	0.07
19	0.02	0.06
20	0.04	0.07

Tabla 3.4: Estadísticos para detectar observaciones de influencia

Los valores correspondientes al dato 6 son mucho mayores que el resto. Se ha comentado que cualquier observación que cumpla que $h_{ii} > 2p/n$ se puede considerar un punto de apalancamiento. En este caso $2p/n = 0.2$, con lo que como $h_{66} = 0.62 \gg 0.2$ tenemos que es un punto de apalancamiento. Por otro lado, el valor que toma D_i en la observación 6 es 14.44, por lo que utilizando el criterio en el que se considera que un dato es influyente si $D_i > 1$, se concluye que efectivamente la observación 6 causa un impacto en el modelo.

Por último calculemos el valor de $COVRATIO_6$

$$\begin{aligned} COVRATIO_6 &= \frac{\left(S^2_{(6)}\right)^p}{MS^p_{Res}} \left(\frac{1}{1-h_{66}}\right) \\ &= \frac{(1.573661)^2}{(243.3163)^2} \left(\frac{1}{1-0.62}\right) = 0.0001093699 \end{aligned}$$

que, en efecto, es mucho más pequeño que 1 y, entonces, incluir la observación 6 degradará la precisión. En lugar de hacer la cuenta a través de los parámetros, se puede introducir en R

```
covratio(model)
```

y veremos que el valor coincide a la perfección.

Siguiendo el criterio establecido por Belsley, Kuh, y Welsch [2] deducimos que la observación 6 es influyente sobre el modelo de regresión, porque $0.0001093699 \ll 0.7 = 1 - 3p/n$.

El ajuste por mínimos cuadrados de los datos considerados en la Tabla 3.2 sin tomar la observación 6 mejora considerablemente, como se puede observar en la gráfica representada en la Figura 3.8.

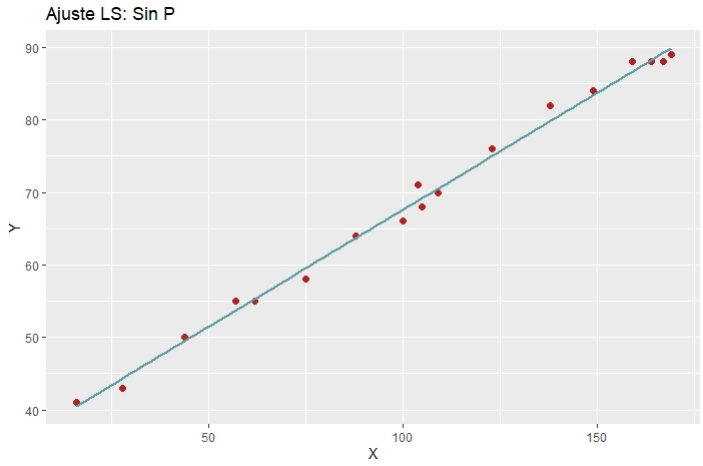


Figura 3.8: Ajuste por mínimos cuadrados: Ajuste sin el punto P en los datos

3.3.5. Tratamiento de datos de influencia

Diagnosticar el apalancamiento y la influencia es una parte vital a la hora de construir el modelo de regresión. La técnicas expuestas en este capítulo pretenden ayudar al analista a detectar los puntos problemáticos y razonar cuáles son desechables, o cuáles merecen un tratamiento más profundo. Como norma general, todo punto 'extraño' producto de un error de medida o de interpretación ha de ser eliminado en caso de no poder ser corregido. Sin embargo, si mediante estos análisis previos llegamos a descubrir la existencia de un punto influyente y perfectamente razonable dentro de los datos, no hay justificación para que sea eliminado.

Es en este punto cuando aparece el tema principal del trabajo, la regresión lineal robusta y los distintos modelos que la componen. La finalidad de las estimaciones robustas radica en darle menos importancia a las observaciones en proporción a las magnitudes residuales o su influencia, por lo que

se consiguen modelos lineales menos alterados por los puntos influyentes que en el caso de mínimos cuadrados. Utilizaremos el 'punto de ruptura' como criterio para evaluar nuestros modelos en los próximos capítulos.

Capítulo 4

Introducción a la regresión lineal robusta

El análisis de regresión es una herramienta estadística que se aplica de forma rutinaria en la mayoría de las ciencias. Debido a su tradición y la facilidad en su computación, el método de mínimos cuadrados (LS) ha sido el más utilizado en todos los ámbitos científicos, llegando a 'eclipsar' a las diversas alternativas que han ido apareciendo a lo largo de la historia. Sin embargo, como se ha estudiado en los capítulos previos, este procedimiento produce errores considerables cuando aparecen outliers en el análisis de los datos. Tanto los puntos de influencia referentes a las variables explicativas como a la explicada, pueden distorsionar el estudio de la regresión por mínimos cuadrados, haciendo que el método sea completamente inservible.

Para remediar estos problemas se han ido desarrollando nuevas técnicas estadísticas, llamados métodos robustos. Existe una corriente generalizada en creer que estos procedimientos alternativos simplemente 'esconden' los outliers, es decir, los pasan por alto. Esta afirmación no podría estar más lejos de la realidad, ya que gracias al ajuste robusto podremos detectar los outliers de una forma mucho más clara, siendo los residuos mucho mayores que los obtenidos por el método de mínimos cuadrados. El objetivo final de los siguientes capítulos será el de ensalzar la utilidad de la regresión robusta a la hora de detectar outliers.

Una alternativa a los métodos robustos generalmente utilizada son los diagnósticos de apalancamiento e influencia, ya comentados en la Sección 3.3. La finalidad de ambos procedimientos es la misma, sin embargo la manera de trabajar es opuesta. En el análisis de diagnóstico, el analista pretende identificar los outliers para poder ajustar los datos de la mejor manera posible aplicando la forma clásica. En cambio, la aproximación por métodos robustos primero se ajusta a los datos haciendo justicia a la mayoría de ellos, y a continuación detecta los outliers como aquellos puntos con un residuo elevado. Ambos caminos pueden llevarnos a un resultado final bastante si-

milar, o incluso el mismo. Elegir entre uno u otro se basa principalmente en el gusto del analista y su familiaridad con las técnicas aplicadas. Aún así, el hecho de detectar los outliers para después ajustar la recta de mínimos cuadrados a los datos de la mejor forma posible no deja de ser un procedimiento robusto.

En los dos primeros capítulos se ha desarrollado la teoría relativa al ajuste por mínimos cuadrados, centrando el estudio de este a través de sus residuos y la hipótesis previa de normalidad de los errores, vital en el desarrollo del procedimiento. A diferencia del método de mínimos cuadrados, la regresión robusta permite asignar de forma desigual la importancia de cada observación, haciendo que los outliers tengan menos peso en la estimación. En aquellos casos donde los errores no se comporten de acuerdo a una distribución normal, ya sea, por ejemplo, por tener unas colas de distribución mucho más pronunciadas, resulta útil aplicar la regresión lineal robusta.

En este capítulo, se expondrá el concepto de punto de ruptura para poder comparar los estadísticos robustos más básicos a modo de introducción, puesto que el objetivo final del trabajo es el de estudiar la regresión lineal por mínima mediana de cuadrados y por mínimos cuadrados recortados.

Además de los desarrollos teóricos expuestos a continuación, un estudio más detallado con R se explica en [26, 34].

4.1. Punto de ruptura

En la Figura 3.7, hemos visto que el punto P altera por completo la recta de mínimos cuadrados. Si calculásemos el valor de los residuos para dicho ajuste, veríamos que aparecen varios outliers que realmente son fruto de la influencia ejercida por P. Por esto, es muy peligroso deshacerse de aquellos datos que presenten un residuo elevado ya que, como en este caso, estaríamos haciendo un análisis completamente erróneo. Queda así demostrado empíricamente que con un solo punto se puede distorsionar por completo la estimación por mínimos cuadrados. Por otro lado, veremos que existen estimadores que pueden tratar con datos que cuenten con un cierto porcentaje de outliers. Para formalizar este aspecto, se introdujo el concepto de punto de ruptura.

El punto de ruptura de un estimador $\hat{\theta}$ de θ puede ser definido, de una manera coloquial, como la mayor cantidad de contaminación (proporción de outliers) que el conjunto de datos puede contener de forma que $\hat{\theta}$ siga aportando información sobre θ.

La definición formal que usaremos será la dada por Donoho y Huber [9, 19], ya que se trata de una versión referida a muestras aleatorias simples finitas.

Definición 4.1. *Sea una muestra de n datos cualquiera, dada por*

$$Z = \{(x_{10}, \ldots, x_{1p-1}, y_1), \ldots, (x_{n0}, \ldots, x_{np-1}, y_n)\}$$

y sea T el estimador de la regresión. Esto significa que, aplicando dicho estimador T a la muestra Z, llegamos al vector de coeficientes de regresión

$$T(Z) = \hat{\beta}$$

Ahora, considérense todas las muestras alteradas Z', las cuales son obtenidas de reemplazar m puntos de la muestra original Z por valores arbitrarios. Denotamos por bias$(m; T, Z)$ a la cantidad máxima de distorsión causada por cada contaminación de los datos

$$bias(m; T, Z) = sup_{Z'} \|T(Z') - T(Z)\|$$

donde el supremo se calcula respecto a todas las muestras posibles Z'. Si la cantidad bias$(m; T, Z)$ es infinita, esto significa que m outliers pueden tener un efecto arbitrariamente grande sobre T, a lo que puede referirse como que el estimador 'se rompe'. Por lo tanto, el punto de ruptura del estimador T para una muestra finita Z se define como

$$\epsilon_n^*(T, Z) = \text{mín} \left\{ \frac{m}{n} \middle| \ bias(m; T, Z) \ es \ infinito \right\} \tag{4.1}$$

En otras palabras, es la menor fracción de contaminación que puede hacer que el estimador T tome valores arbitrariamente lejanos de $T(Z)$.

Para el caso del método de mínimos cuadrados, hemos comprobado que un solo outlier puede perturbar tanto como quiera la estimación obtenida. De esta forma, el punto de ruptura equivale a

$$\epsilon_n^*(T, Z) = \frac{1}{n}$$

que tiende a cero a medida que el tamaño muestral aumenta. Por tanto, se dice que el ajuste por mínimos cuadrados tiene un punto de ruptura del 0 %. Este parámetro vuelve a reflejar la sensibilidad extrema del ajuste LS a los outliers.

Es en este punto cuando empiezan a surgir alternativas robustas al método de mínimos cuadrados. Como se ha indicado al principio del trabajo, la idea fundamental del ajuste LS es la de minimizar la cantidad

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 \tag{4.2}$$

es decir, optimizar el ajuste minimizando los residuos todo lo posible.

En 1887, Edgeworth [12], mejorando una propuesta de Boscovich, dio el primer paso para encontrar un estimador de regresión más robusto. Boscovich argumentó la influencia desmesurada de los outliers sobre el ajuste de

mínimos cuadrados se debía al cuadrado considerado en la Ecuación (4.2). En consecuencia, propuso un estimador de 'mínimos valores absolutos', dado por

$$\min_{\hat{\beta}} \sum_{i=1}^{n} |e_i| \tag{4.3}$$

Esta técnica puede denotarse como regresión L_1, mientras que la regresión por mínimos cuadrados sería L_2 (por la norma que inducen respectivamente). A pesar de este cambio de definición, el punto de ruptura sigue siendo del 0 %; ya que, si bien la estimación es robusta frente a los outliers en el eje y, la distorsión producida por un punto alejado en el eje x es significativa.

4.2. M-estimadores

Los M-estimadores se basan en el reemplazo de los residuos e_i^2 en la Ecuación (4.2) por otra función dada por

$$\min_{\hat{\beta}} \sum_{i=1}^{n} \rho(e_i) \tag{4.4}$$

donde ρ es una función par con un único mínimo en cero. Derivando esta expresión con respecto a los coeficientes $\hat{\beta}_j$ llegamos a

$$\sum_{i=1}^{n} \psi(e_i)\mathrm{x}_i = \mathbf{0} \tag{4.5}$$

donde ψ es la derivada de ρ, y x_i es el vector fila de valores de la i-ésima observación

$$\mathrm{x}_i = (x_{i0}, \ldots, x_{ip-1})$$
$$\mathbf{0} = (0, \ldots, 0)$$

Tenemos entonces en (4.5) un sistema de p-ecuaciones, no siendo siempre fácil encontrar esta solución. En la práctica se utilizan esquemas iterativos basados en el ajuste de mínimos cuadrados ponderado o el llamado H-algoritmo, en referencia a Huber [20]. A diferencia de SS_{Res} y la expresión en la Ecuación (4.3), la solución de la Ecuación (4.5) no es equivariante respecto al aumento del eje y. Por ello, los residuos tienen que estandarizarse mediante una estimación σ, obteniendo

$$\sum_{i=1}^{n} \psi(e_i/\hat{\sigma})\mathrm{x}_i = \mathbf{0} \tag{4.6}$$

donde el valor de $\hat{\sigma}$ ha de estimarse simultáneamente. Huber propuso usar la función

$$\psi(t) = \min(c, max(t, -c))$$

ya que hace que los M-estimadores con esta función sean más eficientes que los obtenidos por la regresión L_1, a la par que son robustos frente a outliers en el eje y. Sin embargo, el punto de ruptura continúa siendo $1/n$ ya que en el eje x siguen siendo sensibles ante los outliers.

La solución a este problema fueron los GM-estimadores (M-estimadores generalizados), con el propósito de corregir la influencia en el eje x mediante la ponderación por una función peso w. Mallows [24] propuso en 1975 la expresión

$$\sum_{i=1}^{n} w(x_i)\psi(e_i/\hat{\sigma})\mathrm{x}_i = \mathbf{0} \tag{4.7}$$

mientras que acorde a [16] Schweppe sugirió en 1977 el uso de

$$\sum_{i=1}^{n} w(x_i)\psi(e_i/(w(x_i)\hat{\sigma}))\mathrm{x}_i = \mathbf{0} \tag{4.8}$$

Estos estimadores fueron construidos con la esperanza de paliar la influencia de un solo outlier, pudiendo medirse esta a través de la llamada función de influencia. Basándose en este criterio, se buscaron las mejores opciones para w y ψ, recibiendo los GM-estimadores el nombre de estimadores de influencia recortada. Muy a nuestro pesar, el punto de ruptura de los GM-estimadores más utilizados está acotado superiormente por $\frac{1}{p+1}$, donde p es el número de coeficientes de regresión. Este hecho conlleva a puntos de ruptura bajos incluso para problemas con pocas variables explicativas. Es más, no se sabe qué GM-estimador puede utilizarse para alcanzar el punto de ruptura óptimo.

A lo largo de la historia han ido apareciendo muchos otros estimadores, pudiendo citarse los L-estimadores (Bickel [3], Koenker y Bassett [23]) o los R-estimadores (Jureckovà [22], Jaeckel [21]) como los más reconocidos actualmente. Desafortunadamente, en el caso simple, ninguno de los métodos alcanza un punto de ruptura por encima del 30 %, es más, alguno de ellos no se pueden extender para el caso con $p > 2$.

Es en este punto cuando cabe preguntarse si podremos encontrar un estimador robusto con un punto de ruptura suficientemente elevado. La respuesta la dio Siegel en 1982 [32], quien propuso el estimador de 'medianas repetidas' con un punto de ruptura del 50 %. Es más, este resultado es el mejor que podemos esperar, ya que cuando el conjunto de datos está muy contaminado, resulta imposible distinguir entre aquellos puntos 'buenos' y 'malos'. El estimador de Siegel se define como sigue:

Para cualesquiera n observaciones

$$(x_{i1}, y_1), \ldots, (x_{in}, y_n)$$

se calcula el vector parámetro que ajusta estos puntos de forma exacta. La j-ésima coordenada de este vector es denotada por $\beta_j(i_1, \ldots, i_n)$. El estimador

de medianas repetidas de la regresión se define coordenada a coordenada como

$$\hat{\beta}_j = \text{med}_{i1}(\ldots(\text{med}_{in-1}(\text{med}_{in}\beta_j(i_1,\ldots,i_n)))\ldots)$$

Este estimador se puede computar explícitamente, pero requiere considerar todos los subconjuntos de n observaciones, tarea que puede ser costosa. Es por esto que para pocas variables explicativas se aplica con facilidad y de forma exitosa. Este estimador, al contrario de otros, no es equivariante a transformaciones lineales de x_i, debido a su construcción por coordenadas. La idea de considerar la mediana a la hora de calcular el estimador es la base del ajuste por mínima mediana de cuadrados, el cual será tratado en el próximo capítulo.

Capítulo 5

Regresión lineal por mínima mediana de cuadrados (LMS)

5.1. Definición del estimador LMS

En el anterior capítulo hemos visto que a lo largo de la historia han ido apareciendo diferentes estimadores robustos con el objetivo de elevar el punto de ruptura, es decir, de disminuir su sensibilidad frente a la presencia de outliers. Si prestamos atención a los diferentes caminos que se han ido tomando para mejorar el ajuste por mínimos cuadrados, veremos que solo se ha considerado importante la influencia del cuadrado en la operación, obviando el efecto causado por la operación suma. De hecho, el nombre abreviado del ajuste por mínimos cuadrados ('least sum of squares') es LS, pasando por alto la palabra 'sum', como si la única opción sensata con n números positivos fuese la de sumarlos. Sin embargo, el estimador por mínima mediana de cuadrados reemplaza la operación aditiva por la mediana, dando lugar al estimador

$$\text{mín}_{\hat{\beta}}\text{med}_i(e_i^2(\beta)) = \text{mín}_{\hat{\beta}}\text{med}_i(\text{y}_i - \text{x}_i'\beta)^2 \tag{5.1}$$

definido por Rousseeuw en 1984 [30, 31].

Más adelante, comprobaremos que el punto de ruptura es del 50 %, el valor posible más alto. El estimador LMS es claramente equivariante frente a transformaciones lineales en las variables explicativas, ya que en la Ecuación (5.1) solo se tienen en cuenta los residuos.

La solución del ajuste por LMS de la regresión simple con un punto de intersección β_0 viene dada por

$$\text{mín}_{\hat{\beta}}\text{med}_i(\text{y}_i - \hat{\beta}_1 \cdot \text{x}_i - \hat{\beta}_0)$$

siendo $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$.

Una manera alternativa a la Ecuación (5.1) para definir el problema es la siguiente

$$\text{mín}(e_i^2(\beta))_{(h)} \tag{5.2}$$

donde los residuos son ordenados de manera creciente y lo que se busca es minimizar el h-ésimo residuo al cuadrado. Entonces el problema de regresión LMS es el problema de encontrar $\beta \in \mathbb{R}^p$ tal que en caso de que n sea impar el h-ésimo residuo ordenado sea mínimo con $h = [n/2] + 1$, y en caso de que n sea par la suma del h-ésimo y el $(h+1)$-ésimo residuos ordenados sea mínima con $h = n/2$. El propio Rousseeuw ha estudiado el uso de otros valores de h, bajo el nombre de Least Quantile Regression (LQS).

La idea geométrica subyacente a este método es la de encontrar la 'banda más fina' que cubra la mitad de las observaciones; tomando como mitad al valor [n/2] + 1, y siendo [n/2] la parte entera de este mismo valor, puesto que para un número par de n observaciones la mediana no siempre es única. La recta de regresión obtenida por LMS se corresponde de manera exacta con la mitad de esa banda. Para ilustrar la idea de ajuste por mínima mediana de cuadrados y mostrar la diferencia con la estimación mínimo cuadrática, se ha considerado el mismo ejemplo que en la Figura 3.7. La recta de regresión calculada por la estimación LMS se ilustra en la Figura 5.1, comparándola con la aproximación lineal por mínimos cuadrados (en gris).

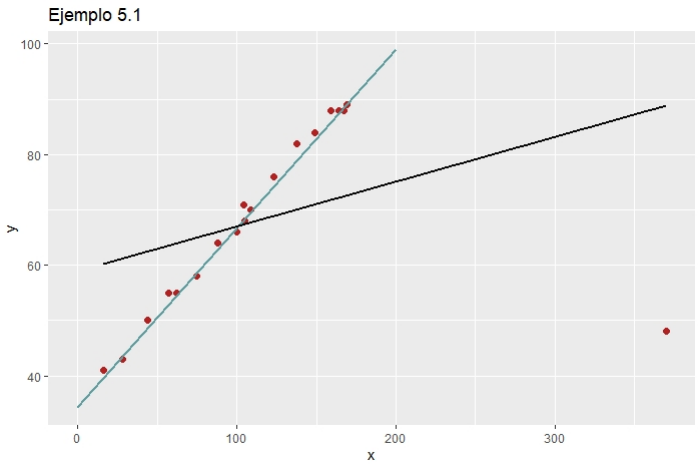


Figura 5.1: Ajuste por mínima mediana de cuadrados

La diferencia es notable. Mientras que en el ajuste por mínimos cuadrados el outlier alteraba completamente la estimación, utilizando este procedimiento robusto no hay ningún tipo de distorsión en la recta de regresión. Es más, la estimación es bastante similar a la mostrada en la Figura 3.8, donde el outlier había sido eliminado del análisis.

A diferencia de la regresión lineal por mínimos cuadrados, no existe una fórmula explícita para estimar los coeficientes de la regresión en el caso LMS.

Es por eso que se hará un estudio de sus propiedades esenciales, y para su aplicación se utilizará el algoritmo ya implementado en R, a través de la función 'lqs' del paquete 'MASS'.

5.2. Propiedades generales

Se presentarán las principales propiedades del estimador LMS para el caso múltiple, tomando una muestra de n observaciones y $p-1$ variables explicativas que tratarán de predecir el valor de la variable independiente y, a través del modelo lineal $\mathrm{y} = \mathrm{X}\beta + \epsilon$. X es la matriz de observaciones muestrales de las $p-1$ variables explicativas $\mathrm{x}_{.1}, \ldots, \mathrm{x}_{.p-1}$ concatenadas al vector independiente de unos ($\mathrm{x}_{.0}$); β el vector p dimensional de coeficientes de regresión desconocido; ϵ el vector de n-componentes del error.

A lo largo de esta sección supondremos que todas las observaciones toman valores no nulos ($x_i \neq \mathbf{0}$). En caso contrario, esa observación no aportaría información alguna acerca de β. Esto se verifica automáticamente si el modelo tiene intersección (β_0) puesto que la primera columna de X tomaría el valor 1. Es más, se asume que en el espacio $(p+1)$ dimensional de (x_i, y_i) no existen hiperplanos verticales pasando por el cero que contengan más de la mitad de las observaciones. Tal hiperplano vertical se entiende como el subespacio p-dimensional que contiene a $(0, \ldots, 0)$ y $(0, \ldots, 0, 1)$.

5.2.1. Existencia de solución

Proposición 5.1. *Siempre existe solución para el problema*

$$\text{mín}_{\hat{\beta}} med_i(e_i^2(\beta))$$

Demostración: Vamos a trabajar en el espacio $(p+1)$-dimensional denotado por E, de las observaciones (x_i, y_i). El espacio definido por x_i es el hiperplano horizontal a través del origen ($y = 0$). Se consideran dos casos:

- **Caso 1** Supongamos un caso particular en el que existe un subespacio V $(p-1)$-dimensional del hiperplano $y = 0$, que pasa por el 0 y contiene al menos $[n/2]+1$ de x_i. Es así que las observaciones (x_i, y_i) correspondientes a cada x_i generan el subespacio vectorial S de E, que como máximo alcanza dimensión p. Como hemos asumido que E no contiene hiperplanos verticales con $[n/2]+1$ observaciones (hipótesis previas), se sigue que el subespacio S no puede contener al hiperplano $(0, \ldots, 0, 1)$; por lo que la dimensión de S es como máximo $p-1$. Esto implica que existe un hiperplano no vertical H dado por una determinada ecuación $\mathrm{y}=\mathrm{x}\beta$ que contiene a S. Para este valor de β claramente $\mathrm{med}_i(e_i^2) = 0$, que justamente coincide con el valor mínimo que puede tomar.

- **Caso 2** Ahora nos encontramos en el caso general, donde no tiene por qué existir el subespacio V. El resto de la demostración se enfocará en probar la existencia de una bola centrada en el origen en el espacio de todos los β posibles, donde podremos encontrar el mínimo de $\text{med}_i(e_i^2)(\beta)$. Como la función a minimizar es continua en β en un compacto, la existencia del mínimo está garantizada (Teorema de Weierstrass). Tomamos:

$$\delta = \frac{1}{2}\text{inf}\{\tau > 0; \text{existe un subespacio } V \text{ de dimensión } p-1 \text{ de } y = 0 \text{ de tal forma que } V^\tau \text{contiene como mínimo} [n/2]+1 \text{ observaciones } x_i\}$$

donde V^τ es el conjunto de todos los puntos x con distancia a V menor que τ. El Caso 1 se corresponde con $\delta = 0$ (V contenía por lo menos a la mitad de las observaciones, luego la distancia era $\tau = 0$). Denotemos por $M := \max_i |y_i|$. Ahora nos enfocaremos tan solo en el dominio delimitado por la bola cerrada centrada en el origen de radio $(\sqrt{2}+1)\frac{M}{\delta}$. En realidad, para cualquier β con $||\beta|| > (\sqrt{2}+1)\frac{M}{\delta}$, se comprobará que

$$\text{med}_i(e_i^2)(\beta) > \text{med}_i(y_i^2) = \text{med}_i(e_i^2)(0)$$

por lo que será imposible encontrar funciones que hagan mínima la expresión fuera de la bola. Razonando como en el Caso 1, tenemos que β determina un hiperplano no vertical H dado por y=xβ. Por la fórmula de las dimensiones de Grassmann, propia del ágebra lineal:

$$\dim(H \cap (\text{y} = 0)) = \dim(H) + \dim(\text{y} = 0) - \dim(H + (\text{y} = 0))$$
$$= p + p - (p+1) = p - 1$$

H depende del valor de β, por lo que para valores $||\beta|| > (\sqrt{2}+1)\frac{M}{\delta}$, se comprueba que $H \neq$(y= 0). Es por ello que $(H \cap (\text{y} = 0))^\delta$ contiene a lo sumo $[n/2]$ de las observaciones x_i. Para cada una de las observaciones restantes (x_i, y_i) no pertenecientes a H, construimos el plano vertical 2-dimensional P_i que contiene al punto en cuestión y es ortogonal a $(H \cap (\text{y} = 0))$. Para que el plano pueda ser considerado vertical ha de pasar por los puntos (x_i, y_i) y $(x_i, y_i + 1)$, ya que no contiene al cero. Vemos que

$$|e_i| = |x_i\beta - y_i| \geq |x_i\beta| - |y_i| \tag{5.3}$$

con $|x_i\beta| > \delta|\tan(\alpha)|$, siendo $\alpha \in (-\pi/2, \pi/2)$ el ángulo formado por H y la línea horizontal en P_i, por lo tanto, de aplicar el teorema de Pitágoras

$$|\alpha| = \arccos\left\{\frac{1}{\sqrt{1+||\beta||^2}}\right\}$$

luego

$$|\tan(\alpha)| = \tan\left(\arccos\left\{\frac{1}{\sqrt{1+||\beta||^2}}\right\}\right)$$

y, aplicando que $\tan(x) = \frac{\text{sen}(x)}{\cos(x)} = \frac{\sqrt{1-\cos^2(x)}}{\cos(x)}$, basta con sustituir por el valor de $|\alpha|$ en la expresión. Obtenemos de esta forma que $|\tan(\alpha)| = ||\beta||$.

Habíamos supuesto que $||\beta|| > (\sqrt{2}+1)\frac{M}{\delta}$, de forma que

$$|x_i\beta| > \delta||\beta|| > M \geq |y_i|$$

de manera que, por (5.3)

$$|e_i| = |x_i\beta - y_i| > (\delta||\beta|| - |y_i|)$$

sustituyendo por la cota inferior de $||\beta||$ y elevando al cuadrado

$$e_i^2 > ((\sqrt{2}+1)M - |y_i|)^2 > ((\sqrt{2}+1)M - M)^2 > 2M^2$$

para al menos $n - [n/2]$ de las observaciones. Por ello

$$\text{med}_i(e_i^2)(\beta) > M^2 \geq \text{med}_i(y_i^2)$$

Luego la función objetivo asociada a esos valores de β es mayor que aquella donde $\beta = 0$. Por lo que bastaría con buscar una solución β en la bola $B(0, (\sqrt{2}+1)\frac{M}{\delta})$. Como la función es continua, por el Teorema de Weiersstrass se alcanzá mínimo dentro de la bola y, finalmente, la existencia de la solución al problema planteado.

□

5.2.2. Propiedades de equivarianza

Para los estimadores de la regresión, se pueden considerar tres tipos de equivarianza, ordenadas de mayor a menor importancia: regresión, escalado y afín. Las propiedades de equivarianza permiten al analista controlar la solución cuando se hacen transformaciones del problema. Dispondremos de una muestra $Z = \{(x_{10}, \ldots, x_{1p-1}, y_1), \ldots, (x_{n0}, \ldots, x_{np-1}, y_n)\}$ y el estimador T de la regresión obtenido por el método LMS para la muestra dada.

Definición 5.1. *Diremos que un estimador T de la regresión es equivariante por regresión si*

$$T(\{(x_i, y_i + x_i \cdot v); i = 1, \ldots, n\}) = T(\{(x_i, y_i); i = 1, \ldots, n\}) + v$$

siendo v un vector columna de longitud p.

Esta propiedad va implícita en el concepto de estimador de regresión. Nótese que un coeficiente tan representativo a la hora de estudiar la regresión lineal, como es el coeficiente de determinación (R^2), no es equivariante por regresión, debido a que depende de la pendiente de la superficie de regresión [1].

Definición 5.2. *Diremos que un estimador T de la regresión es equivariante por escalado si*

$$T\left(\{(x_i, c \cdot y_i); i = 1, \ldots, n\}\right) = c \cdot T\left(\{(x_i, y_i); i = 1, \ldots, n\}\right)$$

para cualquier constante real c arbitraria.

En otras palabras, el ajuste es independiente de la unidad de medida elegida para la variable explicada o respuesta y.

Definición 5.3. *Diremos que un estimador T de la regresión es equivariante afín si*

$$T\left(\{(x_i A, y_i); i = 1, \ldots, n\}\right) = A^{-1} T\left(\{(x_i, y_i); i = 1, \ldots, n\}\right)$$

para cualquier matriz cuadrada no singular A.

Es decir, la equivarianza afín significa que cualquier transformación lineal de x_i transforma equivalentemente al estimador T. Si se verifica esta propiedad, podremos variar el sistema de coordenadas de las variables explicativas o regresores sin que afecten a la estimación.

Proposición 5.2. *El estimador LMS verifica las tres propiedades de equivarianza; es decir, es equivariante por regresión, equivariante por escalado y equivariante afín.*

Demostración: Supongamos para toda la demostración que $\hat{\mathrm{y}}_i = \mathrm{x}_i\beta$. Por lo que $e_i^2 = (y_i - x_i\beta)^2$. Por otro lado, recordemos que el estimador LMS consiste en minimizar la expresión $\mathrm{med}_i(e_i^2)$, así que basta con estudiar los tres tipos de equivarianza para la mediana de los residuos al cuadrado.

- La equivarianza por regresión se comprueba directamente de

  $$\mathrm{med}_i\left(\{y_i + x_i \cdot v\} - \{x_i\beta + x_i \cdot v\}\right)^2 = \mathrm{med}_i\left(y_i - x_i\beta\right)^2 = \mathrm{med}_i(e_i^2)$$

  siendo v un vector columna de longitud p.

- La equivarianza por escalado se comprueba siguiendo la misma idea

  $$\mathrm{med}_i\left(c \cdot y_i - c \cdot x_i\beta\right)^2 = c^2 \mathrm{med}_i\left(y_i - x_i\beta\right)^2 = c^2 \mathrm{med}_i(e_i^2)$$

  donde c es una constante real arbitraria.

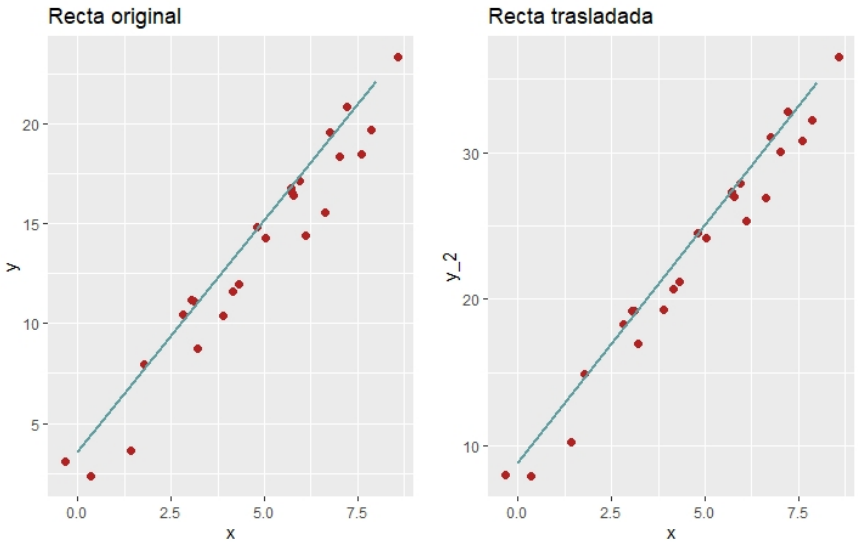
- La equivarianza afín se obtiene de

$$\text{med}_i \left(y_i - \{x_i A\}\{A^{-1}\beta\}\right)^2 = \text{med}_i \left(y_i - x_i\beta\right)^2 = \text{med}_i(e_i^2)$$

para cualquier matriz cuadrada A no singular.

□

Para ilustrar gráficamente que las propiedades de equivarianza se verifican por el estimador LMS, se genera una muestra aleatoria simple de una variable x que sigue una distribución $\mathcal{N}(5, 2)$, y a partir de esta se calcula una variable y que se obtiene de y= 2.25·x+3, más una cantidad llamada 'ruido' que altere ligeramente los datos. Calculamos la recta de regresión por el método LMS, $\hat{y} = 2.320 \cdot x + 3.554$ (Figura 5.2a). A partir de la primera muestra realizamos una transformación de los datos $y_2 = y + \lambda_0 + \lambda_1 \cdot x$, con $\lambda_0 = 5.275$ y $\lambda_1 = 0.91955$, y calculamos de nuevo la recta de regresión para esta muestra de datos. La estimación obtenida por LMS es $\hat{y}_2 = 3.240 \cdot x + 8.829$ (Figura 5.2b). Justamente se verifica que esta última estimación es $(2.320 + \lambda_1) \cdot x + (3.554 + \lambda_0)$. En la Figura 5.2 se muestran ambas gráficas.



(a) Recta original (b) Recta trasladada

Figura 5.2: Aplicación de la propiedad de equivarianza por regresión en LMS

5.2.3. Punto de ruptura del método LMS

Como fuimos introduciendo a lo largo del Capítulo 4, ha habido una constante evolución en la búsqueda de estimadores de la regresión con puntos de ruptura elevados. Estudiaremos en lo que sigue las propiedades de

ruptura del ajuste LMS, usando el concepto dado en la Definición 4.1. Para los siguientes resultados, tomamos una muestra aleatoria de tamaño n $(x_1, y_1), \ldots, (x_n, y_n)$ denotada por Z, y un estimador de la regresión T. Esto significa que aplicando el estimador T a la muestra Z tendremos una estimación del coeficiente de regresión $\hat{\beta}$. Recordemos que la expresión del punto de ruptura de T en Z es

$$\epsilon_n^*(T, Z) = \text{mín} \left\{ \frac{m}{n} \middle| \text{ bias}(m; T, Z) \text{ es infinito} \right\}$$

m representa la cantidad de puntos de la muestra original que son reemplazados en cada muestra.

Si hacemos uso de la definición en la Ecuación (5.1) utilizada hasta ahora para los anteriores resultados, tendríamos las siguientes propiedades:

Proposición 5.3. *Si el número de variables explicativas es $p > 1$ y disponemos de una muestra con n observaciones, entonces el punto de ruptura del método LMS es*

$$([n/2] - p + 2)/n$$

La demostración se puede consultar en la página 118 del libro [31].

Cuando en el Capítulo 4 hablamos del punto de ruptura, establecimos que la cota máxima a alcanzar era del 50 %, y justamente el estimador de las medianas repetidas lo verificaba. Sin embargo, el punto de ruptura expuesto en la Proposición 5.3 es ligeramente inferior al 50 %. Para solucionar esto, Rousseeuw sugiere la definición alternativa del estimador a partir de la ecuación (5.2) en el que se trata de minimizar el h-ésimo residuo ordenado donde $h = [n/2] + [(p+1)/2]$ en lugar del h-ésimo residuo ordenado asociado con la mediana.

Proposición 5.4. *Cualquier estimador equivariante por regresión T satisface*

$$\epsilon_n^*(T, Z) \leq ([(n-p)/2] + 1)/n$$

Demostración: Vamos a probar el resultado por reducción al absurdo. Supongamos que el punto de ruptura es estrictamente mayor que $([(n-p)/2] + 1)/n$. Esto significa que existe una constante finita r de forma que $T(Z')$ está contenido en la bola $B(T(Z), r)$ para todas las muestras Z' que compartan al menos $m = n - [(n-p)/2] - 1$ puntos con Z. Construimos un vector columna p-dimensional $v \neq \mathbf{0}$, de manera que $\mathrm{x}_{.1}v = 0, \ldots, \mathrm{x}_{.p-1}v = 0$. La cantidad $2m-(p-1) \leq n$. Por tanto, los $2m-(p-1)$ primeros puntos de Z pueden reemplazarse por

$$(x_1, y_1), \ldots, (x_{p-1}, y_{p-1}), (x_p, y_p), \ldots, (x_m, y_m),$$
$$(x_p, y_p + x_p\lambda v), \ldots, (x_m, y_m + x_m\lambda v)$$

para cualquier $\lambda > 0$. La estimación $T(Z')$ de esta nueva muestra Z' pertenece a la bola $B(T(Z), r)$. $T(Z')$ también puede escribirse como $T(Z'') + \lambda v$, con $T(Z'')$ perteneciente a la misma bola. Por tanto tenemos que la estimación $T(Z')$ está en la bola $B(T(Z) + \lambda v, r)$, lo que supone una contradicción. Para valores suficientemente grandes de λ, la intersección de ambas bolas es vacía.

□

Todo método de regresión robusto debe verificar que en el caso en que la mayoría de observaciones se distribuyan según una relación lineal exacta, entonces el ajuste de la regresión robusta debe de ser esta ecuación.

Proposición 5.5. *Si el número de variables explicativas es uno o más, y existe β de tal forma que al menos $n - [n/2] + p - 1$ observaciones satisfacen $y_i = x_i\beta$ de forma exacta, entonces la solución obtenida por LMS es exactamente β sin importar el resto de observaciones.*

Demostración: Por hipótesis, existe β de tal forma que al menos $n - [n/2] + p - 1$ de las observaciones se encuentran en el hiperplano H definido por la ecuación y=xβ. Por tanto β es solución al problema (5.1) ya que $\text{med}_i(e_i^2(\beta)) = 0$.

Supongamos que existe otra solución $\beta' \neq \beta$, que define al plano H' produciendo los residuos $e_i^2(\beta')$. Si intersecamos ambos planos $H \cap H'$, por la fórmula de las dimensiones de Grassmann

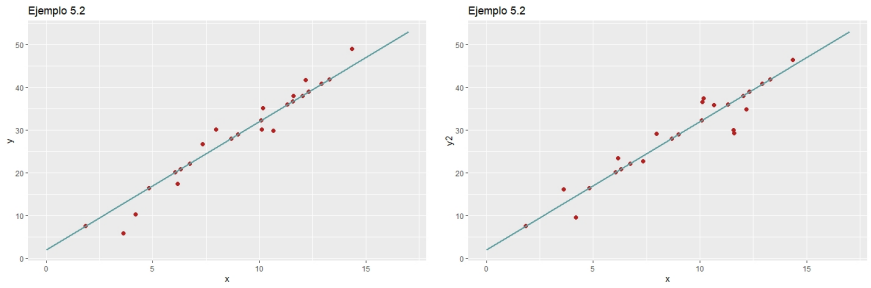
$$\dim(H \cap H') = \dim(H) + \dim(H') - \dim(H + H') = p - 1$$

por lo que la intersección contiene a lo sumo $p - 1$ observaciones. Para los $n - [n/2]$ puntos de H que no están en la intersección se tiene que $e_i^2(\beta') > 0$, por tanto β' no puede ser una solución.

□

Para ilustrar este resultado gráficamente, se han simulado 24 observaciones de una distribución $\mathcal{N}(9, 3)$. A partir de esta variable aleatoria se ha construido una variable y 3x+ 2. Por la Proposición 5.5, si al menos $n - [n/2] + p - 1 = 25 - 12 + 2 - 1 = 14$ se encuentran en el hiperplano (recta en dimensión 2) definido por la ecuación y = 3x+ 2, entonces la solución obtenida por LMS es exactamente la recta definida sin importar el resto de observaciones. El valor de p es 2 ya que tendríamos la columna de observaciones x_0 asociada al término independiente, y la variable x_1 definida por la normal. En la Figura 5.3 aparecen dos casos. En el primer gráfico, se ha añadido ruido a 11 observaciones de la muestra de y con el objetivo de desplazarlas de su disposición lineal en la recta. Por la Proposición 5.5, el máximo permitido de datos perturbados es 11 para que no altere la ecuación; mientras que en el segundo gráfico se han variado 12 observaciones, 1 más de lo permitido. Es obvio que en el primer caso la recta se ajusta de forma

exacta a los 14 puntos originales. Sin embargo, en el segundo caso el ajuste no se ve alterado tampoco, cuando por la teoría debería de desplazarse ligeramente.



(a) 11 observaciones varían (b) 12 observaciones varían

Figura 5.3: Aplicación de la Proposición 5.5

Al estar trabajando con el estimador definido por la mediana, siempre que tengamos $[n/2]+1$ puntos alineados la mediana de los residuos al cuadrado será 0. Es decir, la recta de regresión definida por la estimación LMS se ajustará perfectamente a los datos. Por tanto, en nuestro caso, al tener 13 puntos alineados nos aseguramos que la recta obtenida es una solución a la regresión LMS. Sin embargo, puede no ser única. Imaginémonos una situación muy sencilla, donde el conjunto de datos cuenta con tres observaciones. Si estas tres observaciones están alineadas, la recta obtenida por el LMS será única y se ajustará a los puntos perfectamente. En cambio, si movemos uno de eso puntos fuera de esa línea, podremos trazar tres rectas distintas que sean solución LMS y ajusten perfectamente dos de los puntos. Por la Proposición 5.5 necesitaríamos $n-[n/2]+p-1=3-1+2-1=3$ puntos para que no se perturbe el ajuste y, en cambio, cualquier recta que contenga a dos observaciones será solución al estimador LMS.

Gracias al resultado de la Proposición 5.5 podemos formalizar la definición de punto de ajuste exacto como

$$\delta_n^*(T,Z)=\text{mín}\{m/n;\text{existe } Z' \text{ tal que } T(Z')\neq\beta\}$$

donde Z es la muestra tal que $\mathrm{y}_i=\mathrm{x}_i\beta$ para todo i, y Z' es una muestra que varía en m puntos arbitrarios de Z. Por ejemplo, en el ejemplo expuesto en la Figura 5.3 el punto ajuste exacto sería 12/25. De forma intuitiva, diremos que el punto de ajuste exacto es la fracción mínima de contaminación que puede hacer que $T(Z')\neq\beta$. Es más, si el estimador T es equivariante por regresión y por escalado, como ocurre en el caso de LMS, se cumple:

$$\delta_n^*(T,Z)\geq\epsilon_n^*(T,Z)$$

5.2.4. Propiedades asintóticas

El problema del método LMS reside en su falta de eficiencia por tener una velocidad de convergencia $n^{-1/3}$. Existen varias posibilidades para acelerar la convergencia del estimador, como puede ser combinar su uso con el de los M-estimadores. La manera que nos interesa desarrollar en este trabajo para paliar la falta de eficacia en la convergencia del LMS, será definiendo el estimador de la regresión por mínimos cuadrados recortados (LTS). Su velocidad de convergencia es $n^{-1/2}$.

5.3. Consideraciones finales

En el último capítulo se ilustrará el ajuste obtenido por la estimación LMS, comparándolo con las estimaciones LTS y LS.

Como conclusión de este capítulo, podemos decir que el estimador LMS es muy robusto frente a la presencia de outliers, mejorando considerablemente el ajuste por mínimos cuadrados en este caso. Más que como modelo de predicción, tambien podemos utilizarlo para detectar observaciones atípicas. Cuando nos encontramos con un problema de esta índole, se pueden aplicar tanto los métodos LS como LMS para obtener el ajuste de regresión correspondiente. Si el resultado de ambos procedimientos es similar, la regresión por mínimos cuadrados resulta fiable. En contrapartida, si hay diferencia significativa entre ambas salidas, podremos detectar qué observaciones la causan fijándonos en los residuos de la estimación LMS. En la Figura 5.1 hemos trabajado con 20 datos, una muestra bastante pequeña para las bases de datos que aparecen en la realidad, y el ajuste por mínima mediana de cuadrados detectó el outlier presente en la observación 6, estimando la recta de regresión que mejor se adecúa al resto de puntos. Este ejemplo indica que además es aplicable a conjuntos de datos con pocas observaciones.

Capítulo 6

Regresión lineal por mínimos cuadrados recortados (LTS)

En el Capítulo 5, hemos llegado a que el problema de la regresión por mínima mediana de cuadrados reside en su velocidad de covergencia. La intención es la de mejorar la convergencia sin empeorar otras propiedades, como pueden ser las propiedades de equivarianza o del punto del ruptura al 50 %.

Se han tomado como referencia [30, 31], además de [14, 17], donde se estudian sus propiedades desde una visión más computacional.

6.1. Definición del estadístico

Como solución al problema de convergencia del ajuste LMS, Rousseeuw propuso en 1983-1984 [30, 31] el estimador por mínimos cuadrados recortados (LTS), dado por la expresión

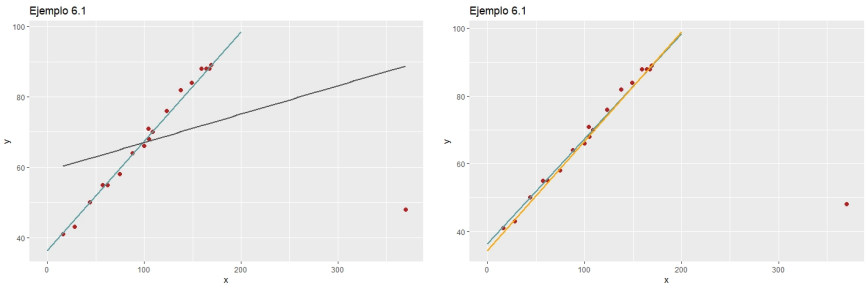
$$\text{mín}_{\hat{\beta}} \sum_{i=1}^{h} (e^2)_{(i)} \tag{6.1}$$

donde

$$(e^2)_{(1)} \leq \ldots \leq (e^2)_{(n)}$$

son los residuos al cuadrado ordenados. Nótese que primero se elevan al cuadrado y luego se ordenan, es decir, el orden sería equivalente a hacerlo en valor absoluto. La definición dada en la Ecuación (6.1) es muy similar a la del estimador por mínimos cuadrados, con la única diferencia de no tener en cuenta los $n - h$ residuos al cuadrado más grandes en la suma. También difiere de la regresión por mínima mediana de cuadrados, donde no se consideran solamente los residuos más pequeños. Esta es justamente la razón por la que se añade el término 'recortados'.

Tomando $h = [n/2] + 1$, el estimador LTS alcanza el mismo punto de ruptura que el estimador LMS en la Proposición 5.3, mientras que para el valor $h = [n/2] + [(p+1)/2]$ el LTS alcanza el máximo valor posible del 50 %. Al igual que para el ajuste LMS, se puede dibujar la recta de regresión para la estimación LTS del ejemplo de la Figura 3.7, comparándola con la recta de regresión gris obtenida a partir del método LS. Tanto esta representación, como la comparación del LMS con el LTS aparecen la Figura 6.1. La línea en color anaranjado es la recta de regresión calculada por el ajuste LMS.



(a) Estimación LTS y LS (b) Estimaciones LMS y LTS

Figura 6.1: Ejemplo de estimación LTS frente a la estimación LMS

Para la ejecución práctica de la regresión lineal por mínimos cuadrados recortados se usará el algoritmo implementado en la función 'lqs' en R del paquete 'MASS', ya que tampoco podemos calcular los coeficientes de forma explícita.

6.2. Propiedades generales

De la misma forma que se hizo en el anterior capítulo, se citarán las principales propiedades del estimador LTS para el caso múltiple, tomando una muestra de n observaciones y $p-1$ variables explicativas, que buscarán predecir el valor de la variable independiente y, a través del modelo lineal $\mathrm{y} = \mathrm{X}\beta + \epsilon$. X es, de nuevo, la matriz de observaciones muestrales de las $p-1$ variables explicativas $\mathrm{x}_{.1}, \ldots, \mathrm{x}_{.p-1}$, concatenadas al vector independiente de unos ($\mathrm{x}_{.0}$); β el vector p dimensional de coeficientes de regresión desconocido; ϵ el vector de n-componentes del error.

A lo largo de esta sección supondremos que todas las observaciones toman valores no nulos ($x_i \neq \mathbf{0}$). En caso contrario, esa observación no aportaría información alguna acerca de β. Esto se verifica automáticamente si el modelo tiene intersección (β_0) puesto que la primera columna de X tomaría el valor 1. Es más, se asume que en el espacio $(p+1)$ dimensional de (x_i, y_i) no existen hiperplanos verticales pasando por el cero que contengan más de la mitad de las observaciones. Tal hiperplano vertical se entiende como el subespacio p-dimensional que contiene a $(0, \ldots, 0)$ y $(0, \ldots, 0, 1)$.

6.2.1. Propiedades de equivarianza

Conserva las mismas propiedades que el estimador LMS.

Proposición 6.1. *El estimador por mínimos cuadrados recortados (LTS) es equivariante por regresión, equivariante por escalado y equivariante afín.*

Demostración: Supongamos para toda la demostración que $\hat{\mathrm{y}}_i = \mathrm{x}_i\beta$. Por lo que $e_i^2 = (y_i - x_i\beta)^2$. Por otro lado, recordemos que el estimador LTS consiste en minimizar la expresión $\sum_{i=1}^{h}(e^2)(i)$, así que basta con estudiar los tres tipos de equivarianza para el sumatorio.

- La equivarianza por regresión se comprueba directamente de

$$\sum_{i=1}^{h}((\{y_i + x_i \cdot v\} - \{x_i\beta + x_i \cdot v\})^2)_{(i)} = \sum_{i=1}^{h}((y_i - x_i\beta)^2)_{(i)} = \sum_{i=1}^{h}(e^2)_{(i)}$$

  siendo v un vector columna de longitud p.

- La equivarianza por escalado se comprueba siguiendo la misma idea

$$\sum_{i=1}^{h}((c \cdot y_i - c \cdot x_i\beta)^2)_{(i)} = c^2\sum_{i=1}^{h}((y_i - x_i\beta)^2)_{(i)} = c^2\sum_{i=1}^{h}(e^2)_{(i)}$$

  donde c es una constante real arbitraria.

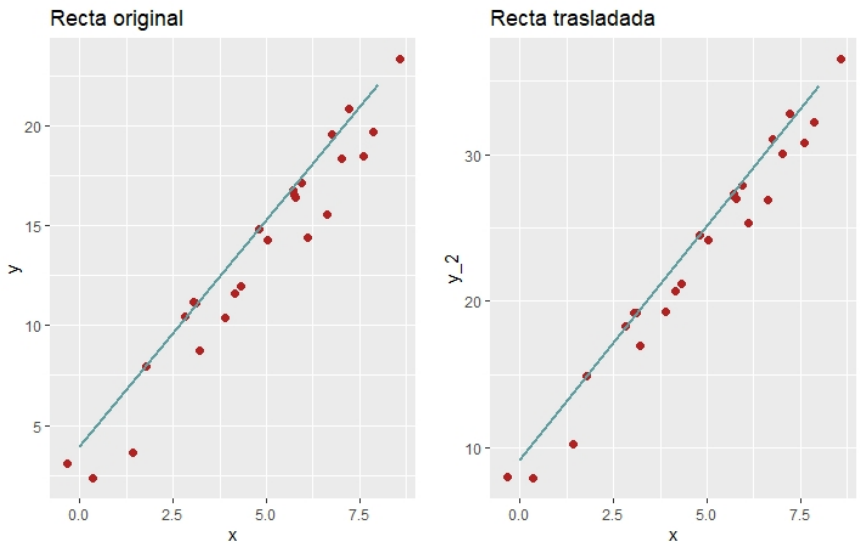
- La equivarianza afín se obtiene de

$$\sum_{i=1}^{h}(\left(y_i - \{x_iA\}\{A^{-1}\beta\}\right)^2)_{(i)} = \sum_{i=1}^{h}((y_i - x_i\beta)^2)_{(i)} = \sum_{i=1}^{h}(e^2)_{(i)}$$

  para cualquier matriz cuadrada A no singular.

□

Para ilustrar gráficamente que las propiedades de equivarianza se verifican por el estimador LTS, se toma la misma muestra aleatoria simple que en la Figura 5.2, de una variable x que sigue una distribución $\mathcal{N}(5, 2)$, y a partir de esta se calcula una variable y que se obtiene de y= 2.25·x+3, más una cantidad llamada 'ruido' que altere ligeramente los datos. Calculamos la recta de regresión por el método LTS, $\hat{y} = 2.271 \cdot x + 3.909$ (Figura 6.2a). A partir de la primera muestra realizamos una transformación de los datos $y_2 = y + \lambda_0 + \lambda_1 \cdot x$, con $\lambda_0 = 5.275$ y $\lambda_1 = 0.91955$, y calculamos de nuevo la recta de regresión para esta muestra de datos. La estimación obtenida por LTS es $\hat{y}_2 = 3.190 \cdot x + 9.184$ (Figura 6.2b). Justamente se verifica que esta última estimación es $(2.271 + \lambda_1) \cdot x + (3.909 + \lambda_0)$. En la Figura 6.2 se muestran ambas gráficas.



(a) Recta original (b) Recta trasladada

Figura 6.2: Aplicación de las propiedades de equivarianza al LTS

6.2.2. Punto de ruptura del método LTS

Estudiaremos en lo que sigue las propiedades de ruptura del ajuste LTS, al igual que se hizo en el Capítulo 5 para el estimador LMS, y usando el concepto dado en la Definición 4.1. Para los siguientes resultados, tomamos una muestra aleatoria de tamaño n $(x_1, y_1), \ldots, (x_n, y_n)$ denotada por Z, y un estimador de la regresión T. Esto significa que aplicando el estimador T a la muestra Z tendremos una estimación del coeficiente de regresión $\hat{\beta}$.

Proposición 6.2. *El punto de ruptura del método LTS definido por la Ecuación (6.1) con* $h = [n/2] + [(p+1)/2]$ *es*

$$\epsilon_n^*(T, Z) = ([(n-p)/2] + 1)/n$$

Demostración: Para poder demostrar este resultado, asumimos de nuevo que todas las observaciones con $(x_{i0}, \ldots), x_{ip} = \mathbf{0}$ han sido eliminadas. Vamos a probar el resultado por doble desigualdad.

- La desigualdad $\epsilon_n^*(T, Z) \leq ([(n-p)/2] + 1)/n$ se sigue directamente de aplicar la Proposición 5.4 y las propiedades de equivarianza que satisface el estimador LTS, en particular la equivarianza por regresión.
- Para demostrar la desigualdad $\epsilon_n^*(T, Z) \geq ([(n-p)/2]+1)/n$ se necesita algo más de trabajo. Como la muestra $Z = \{(x_i, y_i); i = 1, \ldots, n\}$ está

formada por n observaciones en el espacio p-dimensional, se sigue que

$$\rho = \frac{1}{2}\text{inf}\{\tau > 0; \text{existe un subespacio } (p-1) - \text{dimensional}$$
$$V \text{ de } (\text{y} = 0) \text{ tal que } V^{\tau} \text{ cubre al menos } p \text{ de las } x_i\}$$

es estrictamente positivo. Supongamos que β minimiza (6.1) para la muestra Z, y denotamos por H al hiperplano correspondiente dado por la ecuación y=xβ. Denotamos por $M = \text{máx}_i|e_i|$, donde $e_i = y_i - x_i\beta$. Construimos a partir de $n - ([(n-p)/2])$ observaciones de Z una muestra contaminada Z'. Es suficiente con comprobar que $||\beta - \beta'||$ está acotado, donde β' se corresponde con Z'. Sin perdida de generalidad, asumiremos que $\beta' \neq \beta$, luego el hiperplano asociado H' será diferente al hiperplano H. Por el teorema de las dimensiones de Grassmann $dim(H' \cap H)$ tiene dimensión $p-1$. Definimos A como el conjunto de observaciones buenas restantes, conteniendo por lo menos $n - ([(n-p)/2]+1)/n-(p-1)$ puntos. Consideramos un punto en A cualquiera, denotado por (x_a, y_a), cuyos residuos son $e_a = y_a - x_a\beta$ y $e'_a = y_a - x_a\beta'$. Construimos un plano P_a dos dimensional que pasa por el punto (x_a, y_a) y el plano ortogonal a la proyección de $H' \cap H$ sobre (y=0). Se sigue de aquí que

$$|e'_a - e_a| = |x_a\beta' - x_a\beta| > \rho(||\beta' - \beta|| - 2||\beta||)$$

La suma de los h primeros residuos cuadrados de la nueva muestra tomada Z', con al menos $n - ([(n-p)/2]) \geq h$ residuos siendo los mismos que para la muestra Z. Como β' se corresponde con Z' se tiene también que

$$\sum_{i=1}^{h}((y'_i - x'_i\beta')^2)(i) \leq hM^2 \tag{6.2}$$

Si ahora asumimos que

$$|\beta' - \beta| \geq 2||\beta|| + M(1+\sqrt{h})/\rho$$

luego para todo a en A que se cumple

$$|e'_a - e_a| > \rho(||\beta' - \beta|| - 2||\beta||) \geq M(1+\sqrt{h})$$

luego

$$|e'_a| \geq |e'_a - e_a| - |e_a| > M(1+\sqrt{h}) - M = M\sqrt{h}$$

Notemos que $n - |A| \leq h-1$. Por tanto, cualquiero conjunto de tamaño h de (x'_i, y'_i) deben contener al menos una observación del tipo (x_a, y_a), luego

$$\sum_{i=1}^{h}((y'_i - x'_i\beta')^2)(i) \leq (e'_a)^2 > hM^2$$

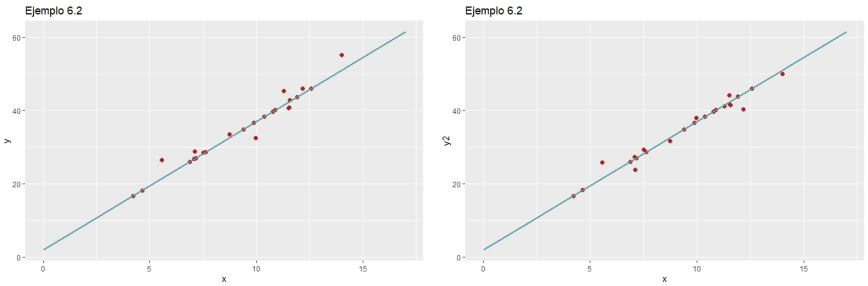
que supone una contradicción con (6.2)

□

Como ocurría en el caso del ajuste de regresión LMS, el estimador LTS también verifica la propiedad de ajuste exacto.

Proposición 6.3. *Si existe algún valor β para el que más de $\frac{1}{2}(n+p-1)$ observaciones satisfacen $y_i = x_i\beta$ de manera exacta, entonces la solución al estimador LTS es β sin importar el resto de observaciones.*

Al igual que se hizo en el Capítulo 5, vamos a comprobar empíricamente este resultado. Para ello, volvemos a suponer 24 observaciones de una distribución $\mathcal{N}(9,3)$. A partir de esta variable aleatoria se construye una variable y = 3.5x + 2. Por la Proposición 6.3, si más de $\frac{1}{2}(n+p-1) = \frac{1}{2}(24+2-1) = 12.5$ observaciones se encuentran en el hiperplano (recta en dimensión 2) definido por la ecuación y = 3.5x + 2 , entonces la solución obtenida por LMS es exactamente la recta original sin importar el resto de observaciones. El valor $p = 2$ ya que tendríamos la columna de observaciones x_0 asociada al término independiente, y la variable x_1 definida por la distribución normal.



(a) 11 observaciones varían (b) 12 observaciones varían

Figura 6.3: Aplicación de la Proposición 6.3

En la Figura 6.3 (a) la recta se ajusta de forma exacta a y= 3.5x +2, donde se han desplazado 11 observaciones, el máximo permitido según la Proposición 6.3. Alterando el valor de una observación más, comprobamos que efectivamente la recta se desplaza ligeramente, ya que el punto de intersección con el eje y es 1.973 en este caso. Queda así verificado experimentalmente el resultado y, además, la robustez del método.

6.2.3. Propiedades asintóticas

El método LTS surgió con la intención de acelerar la velocidad de convergencia alcanzada por el método LMS. De hecho converge como $n^{-1/2}$, lo cual mejora considerablemente al estimador LMS.

La principal desventaja del ajuste por mínimos cuadrados recortados es la cantidad de operaciones que tiene que realizar el algoritmo, teniendo que ordenar n residuos al cuadrado, lo que toma $O(n\log n)$ operaciones comparadas con las $O(n)$ operaciones de la mediana.

6.3. Consideraciones finales

Al igual que ocurre con el ajuste por mínima mediana de cuadrados, la estimación obtenida por LTS es mucho más robusta frente a outliers que la de mínimos cuadrados. El procedimiento para su aplicación es el mismo que el del LMS. En primer lugar, aplicaremos la estimación LTS y LS para obtener el ajuste de regresión correspondiente. Con los residuos del ajuste LTS detectaremos los outliers de una forma mucho más clara, pudiendo elegir si eliminarlos por tratarse de un error de medición, por ejemplo, o mantenerlos ya que aportan información al modelo. Es más, se ha vuelto a encontrar la recta de regresión en la Figura 6.1 para los mismos 20 puntos que en el ejemplo expuesto en el Capítulo 5, demostrando así que puede ser aplicado para conjuntos de datos con pocas observaciones.

Si hubiese que elegir entre uno de los dos métodos robustos estudiados durante todo el trabajo, bajo mi percepción, eligiría la estimación por mínimos cuadrados recortados. Mantiene las propiedades del ajuste LMS y, además, mejora su velocidad de convergencia. Mi elección coincide con la del precursor de ambos métodos, Peter J. Rousseeuw, quien se decanta por la estimación LTS antes que la LMS [31].

Capítulo 7

Caso unidimensional: Problema de localización

7.1. Introducción

Hasta ahora, hemos trabajado siempre con la presencia de variables explicativas ($p > 1$), es decir, la variable a predecir y se estimaba a partir de un modelo lineal dependiente de una serie de variables x_i elegidas por el analista. Un caso especial de la regresión, es el unidimensional ($p = 1$); donde tan solo disponemos del vector de parámetros $\beta = \beta_0$, es decir, el vector de valores independientes. Por tanto, el modelo subyacente de regresión para el caso unidimensional, también llamado problema de localización, para una muestra $(y_i)_{i=1,\ldots,n}$ es

$$y_i = \beta + \epsilon_i \tag{7.1}$$

siendo ϵ_i el vector de errores con las hipótesis de normalidad supuestas durante todo el trabajo. En esta situación, el estimador T de β del modelo de regresión, recibe el nombre de estimador de localización univariante. A simple vista puede parecer un problema trivial, sin embargo, los estimadores de localización han sido muy estudiados en la literatura.

Las propiedades de equivarianza descritas en los Capítulos 5 y 6, se verifican para un estimador de localización T obtenido por las estimaciones descritas. Siendo más precisos, para este tipo de estimadores tenemos dos tipos de propiedades de equivarianza, por traslación y por escalado.

- Diremos que un estimador de localización T, es equivariante por traslación si verifica

  $$T(y_1 + v, \ldots, y_n + v) = T(y_1, \ldots, y_n) + v$$

  para cualquier constante real v.

- Diremos que T un estimador de localización, es equivariante por escalado si verifica

$$T(c \cdot y_1, \ldots, c \cdot y_n) = c \cdot T(y_1, \ldots, y_n)$$

para cualquier constante real c distinta de cero.

7.2. Estimador de localización LS

El estimador obtenido por mínimos cuadrados estudiado en el Capítulo 2 se reduce a minimizar la expresión

$$\sum_{i=1}^{n} (\mathrm{y}_i - \hat{\beta}_0)^2 \tag{7.2}$$

El resultado es la ya conocida media aritmética. Al igual que se argumentó en el Capítulo 4, este estimador es altamente sensible frente a la presencia de outliers. Para disminuir el impacto causado por estos valores atípicos, Huber [18] propuso reemplazar el cuadrado en la Ecuación (7.2) por una función peso adecuada ρ. Esto lleva, de nuevo, a los M-estimadores y, en este caso particular, a M-estimadores de localización. Se define esta alternativa como

$$\min_{\hat{\beta}_0} \sum_{i=1}^{n} \rho(\mathrm{y}_i - \hat{\beta}_0)$$

satisfaciendo la condición necesaria

$$\sum_{i=1}^{n} \psi(\mathrm{y}_i - \hat{\beta}_0) = 0$$

donde ψ es la derivada de ρ.

El estudio de este estimador se escapa de los objetivos del trabajo, para más información se pueden consultar los libros [15, 18].

7.3. Estimador de localización LMS

El estimador de localización LMS se corresponde con minimizar

$$\mathrm{med}_i(\mathrm{y}_i - \beta_0)^2$$

La existencia de esta estimación para una muestra $(y_1, \ldots, y_n)$ está garantizada por la Proposición 5.1, por lo que podremos demostrar el siguiente resultado.

Proposición 7.1. *Sea $p = 1$ y todos las observaciones $x_{i0} = 1$, por lo que la muestra se reduce a $(y_i)_{i=1,\ldots,n}$. Si se cumple que*

$$m_T^2 := med_i(e_i^2) = med_i(y_i - T)^2 = \text{mín}_\beta med_i(y_i - \beta_0)^2$$

entonces los valores $T - m_T$ y $T + m_T$ son observaciones en la muestra.

Demostración:

- Caso 1. En primer lugar supongamos que n es impar, con $n = 2k - 1$. Por tanto, $med(e_i^2)$ se alcanza en el k-ésimo término. Por consiguiente, al menos uno de los puntos $T - m_T$ ó $T + m_T$ es una observación; sin pérdida de generalidad, sea $T - m_T$ una observación y $T + m_T$ no. Existe una partición de e_i^2 en $k-1$ términos $\geq m_T^2$, 1 término $= m_T$, y $k - 1$ términos $\leq m_T^2$. Tomamos la observación más grande y_j, la cual es más pequeña que $T + m_T$ (si hay varias iguales se coge solo una), y definimos

  $$T' = \frac{1}{2}((T - m_T) + y_j)$$

  y

  $$m^2 = (\frac{1}{2}|(T - m_T) - y_j|)^2 < m_T^2$$

  Definiendo los residuos $e_i'^2 = (y_i - T')^2$, encontramos de forma analoga a la anterior una partición en $k - 1$ términos $\leq m^2$ (los mismos que en el caso de e_i^2), $k - 2$ términos $\geq m^2$ (los mismos que en el caso anterior salvo y_j), y 2 términos $= m^2$ (y_j y $T - m_T$). Finalmente, $\text{med}(e_i'^2) = m^2 < m_T^2$. Llegamos a una contradicción de suponer que $T + m_T$ no es una observación de la muestra.

- Caso 2. Supongamos que n es par, con $n = 2k$. Denotando por $e_{(1)}^2 \leq \ldots \leq e_{(n)}^2$ a los residuos al cuadrado ordenados, entonces

  $$m_T^2 = \frac{1}{2}(e_{(k)}^2 + e_{(k+1)}^2)$$

  Existe una partición de los residuos al cuadrado en $k - 1$ términos $\leq e_{(k)}^2$, ($e_{(k)}^2$ y $e_{(k+1)}^2$), y $k - 1$ términos $\geq e_{(k+1)}^2$. Si $T - m_T$ es una observación y $T + m_T$ no lo es (o al revés), bastaría con repetir el razonamiento del Caso 1.

  Supongamos ahora que tanto $T - m_T$ y $T + m_T$ no son observaciones de la muestra, lo que fuerza que ($e_{(k)}^2 < e_{(k+1)}^2$), ya que en caso contrario ($e_{(k)}^2 = m_T^2 = e_{(k+1)}^2$). Así, por lo menos el residuo cuadrado $e_{(k+1)}^2 > 0$.

  1. Asumamos que $e_{(k)}^2 = 0$. En ese caso, T coincide exactamente con k observaciones (todos los $k - 1$ residuos anteriores también

valen 0). La siguiente observación más cercana a la estimación, denotada por y_d, estará a una distancia $|e_{(k+1)}|$. Tomando $T' = \frac{1}{2}(T + y_d)$, llegamos a

$$\text{med}(y_i - T')^2 = \frac{1}{2}\left(\left(\frac{1}{2}e_{(k+1)}\right)^2 + \left(\frac{1}{2}e_{(k+1)}\right)^2\right) = \frac{1}{4}e^2_{(k+1)} = m_T^2$$

Llegamos a una contradicción de nuevo.

2. Asumamos que $e^2_{(k)} > 0$. Denotamos por y_j a la observación asociada al residuo $e^2_{(k)}$ y por y_d a la asociada al residuo $e^2_{(k+1)}$. Si las observaciones que causan $e^2_{(k)}$ y $e^2_{(k+1)}$ son todas mayores o menores que las estimadas por T, se puede repetir el argumento del Caso 1. Por tanto, se asumirá, sin pérdida de generalidad, que $y_j < T < y_d$. Tomando $T' = \frac{1}{2}(y_j + y_d)$, llegamos a

$$\text{med}(y_i - T')^2 = \frac{1}{2}((y_j - T')^2 + (y_d - T')^2) < \frac{1}{2}((y_j - T')^2 + (y_d - T')^2)$$

a causa de que la función $12((y_j - T')^2 + (y_d - T')^2)$ alcanza su mínimo en $\frac{1}{2}(y_j + y_d)$. Llegamos de nuevo a una contradicción.

Por tanto, tenemos que tanto $T - m_T$ como $T + m_T$ son observaciones muestrales.

□

La idea subyacente a este resultado es la de encontrar la 'banda' más fina que cubra la mitad de las observaciones. En el caso múltiple esta banda será la región delimitada por dos hiperplanos paralelos.

7.4. Estimador de localización LTS

El estimador en el caso unidimensional de ($\mathrm{y}_i = \hat{\beta}_0$) por el método LTS viene dado por

$$\text{mín}_{\hat{\beta}} \sum_{i=1}^{h} (e^2)_{(i)} \tag{7.3}$$

donde $h = [n/2] + 1$ y $(e^2)_{(i)}$ son los residuos al cuadrado ordenados. Lo que se pretende con el caso unidimensional, tanto para el estimador LMS como el LTS, es determinar la localización de la estimación. En el caso LMS esta era la banda más pequeña que contenía a la mitad de las observaciones. Para determinar la localización de la estimación LTS, tenemos que considerar $n - h + 1$ submuestras de la muestra original de tamaño n:

$$\{y_{(1)}, \ldots, y_{(h)}\}, \{y_{(2)}, \ldots, y_{(h+1)}\}, \ldots, \{y_{(n-h+1)}, \ldots, y_{(n)}\}$$

Cada una de estas submuestras tiene h observaciones, y reciben el nombre de mitades contiguas. Para cada una de las mitades, calculamos su media muestral

$$\overline{y}^1 = \frac{1}{h}\sum_{i=1}^{h} y_{(i)}$$
$$\vdots$$
$$\overline{y}^{n-h+1} = \frac{1}{h}\sum_{i=n-h+1}^{n} y_{(i)}$$

y la correspondiente suma de cuadrados de cada una de ellas

$$SQ^{(1)} = \sum_{i=1}^{h}\{y_{(i)} - \overline{y}^{(1)}\}^2$$
$$\vdots$$
$$SQ^{(n-h+1)} = \sum_{i=n-h+1}^{n}\{y_{(i)} - \overline{y}^{(n-h+1)}\}^2$$

La solución al estimador LTS se corresponde con la media $\overline{y}^{(j)}$ asociada a la suma de cuadrados $SQ^{(j)}$ más pequeña.

7.5. Ejemplo numérico

Se ha generado aleatoriamente la siguiente muestra aleatoria simple de tamaño 20:

$$y_i : (-2.12,\ -0.30,\ 0.52,\ 0.33,\ -0.98,\ -0.55,\ -2.11,\ 0.02,\ -2.72,\ 1.24,\ 0.10,\ 0.71,\ 1.10,\ -0.20,\ 0.26,\ 2.00,\ -0.07,\ -0.57,\ 1.10,\ 1.98)$$

correspondiente a una variable aleatoria y con distribución $\mathcal{N}(0,1)$. Las tres estimaciones de localización para la muestra generada son:

- **Estimación LS**: $\hat{y} = -0.0138$, que justamente es la media aritmética.
- **Estimación LMS**: $\hat{y} = -0.0252$.
- **Estimación LTS**: $\hat{y} = 0.0212$.

Cambiemos ahora manualmente un valor de la muestra y_i, por ejemplo tomamos $y_1' = 1000 \times y_1$. Lo que debería ocurrir es que el ajuste por mínimos cuadrados varíe drásticamente. Las estimaciones obtenidas en este caso son:

- **Estimación LS**: $\hat{y} = -105.7$, que justamente es la media aritmética.

- **Estimación LMS**: $\hat{y} = -0.0252$.
- **Estimación LTS**: $\hat{y} = 0.0212$.

Para la muestra tomada, los dos métodos robustos estudiados se mantienen exactamente iguales. Por lo visto en la teoría, no sorprende la alta sensibilidad de la estimación LS. Tomando otra muestra cualquiera lo que ocurriría es que las estimaciones cambian ligeramente para los métodos LMS y LTS, mientras que para el método LS sería igual a la media muestral.

Capítulo 8

Ejemplo: Jugadores de la NBA

Durante todo el trabajo se han ido exponiendo los conceptos necesarios para aplicar la regresión por mínimos cuadrados (LS), por mínima mediana de cuadrados (LMS) y por mínimos cuadrados recortados (LTS). Recapitulando las ideas más importantes; el ajuste por mínimos cuadrados es sencillo en su compresión y fácilmente computable, sin embargo, es muy fácil distorsionar su estimación con outliers. Para paliar estos datos comúnmente presentes en todos los estudios, ya sea por errores de medición, cambios de escala, o simplemente situaciones atípicas, aparecen las estimaciones robustas. El trabajo se ha centrado en dos de ellas, la estimación LMS y LTS.

En este capítulo buscaremos mostrar las diferencias entre los tres procedimientos a través de un conjunto de datos real. No se pretende decidir si un método es mejor o peor, o si conviene aplicar uno u otro. Esta decisión recae en el analista y cuantas más herramientas tenga a su alcance más preciso será su análisis. Por tanto, la finalidad es comparar, ilustrar los resultados y concluir desde el punto de vista personal qué resulta más correcto para el problema en cuestión.

Para su elaboración se han tenido en cuenta diversos artículos [4, 7, 28], con el fin de estructurar el estudio de una forma coherente y comprensible.

8.1. Resolución del problema

Se ha elegido una base de datos referente a las estadísticas de 651 jugadores de baloncesto de la NBA ('National Basketball Association') en la temporada regular 2019-2020, obtenidas de basketball-reference.com [35]. Dicha base de datos contiene 29 variables (3 cualitativas y 26 numéricas). El conjunto de datos ha sido previamente filtrado y tratado para facilitar su manejo con R y obtener resultados más claros. Para nuestro problema nos enfocaremos en 2 variables numéricas:

- MP (minutos por partido)
- PTS (puntos por partido)

El problema en cuestión es el de aplicar los tres métodos de regresión estudiados, con la finalidad de predecir los puntos por partido a partir del número de minutos disputados. Se ha escogido un caso de regresión lineal simple ya que su interpretación gráfica es mucho más intuitiva, pudiendo comparar a simple vista los distintos procedimientos aplicados.

El coeficiente de correlación lineal de Pearson entre las dos variables de estudio es $\rho = 0.8812139$. Normalmente, por encima de $|\rho| = 0.7$ se suele decir que la dependencia lineal entre las variables es fuerte. En este caso se verifica que la dependencia lineal es fuerte y, además, es positiva (creciente).

El primer paso es conocer la distribución de los datos en el plano, resumir los datos numéricamente y extraer conclusiones previas. Este análisis previo puede ayudarnos a enfocar el problema en una dirección u otra. El resumen numérico se muestra en la Figura 8.1, y el correspondiente diagrama de dispersión en la Figura 8.2.

```
       MP              PTS        
 Min.   : 1.00   Min.   : 0.000  
 1st Qu.:12.00   1st Qu.: 3.800  
 Median :18.50   Median : 6.800  
 Mean   :19.18   Mean   : 8.461  
 3rd Qu.:26.70   3rd Qu.:11.650  
 Max.   :37.50   Max.   :34.300  
```

Figura 8.1: Resumen numérico de las variables 'MP' y 'PTS'

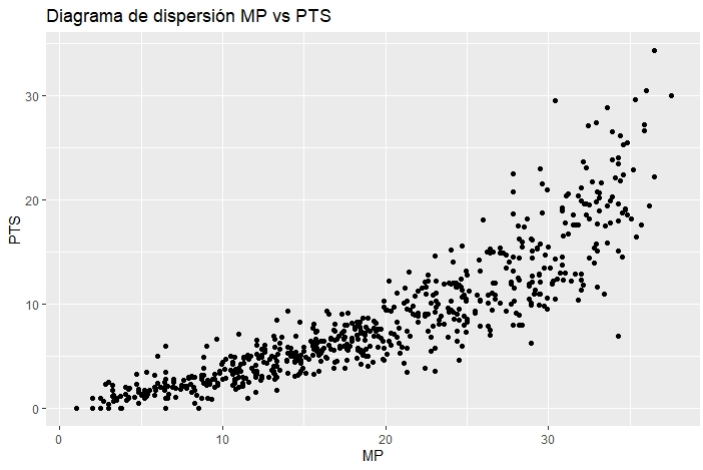
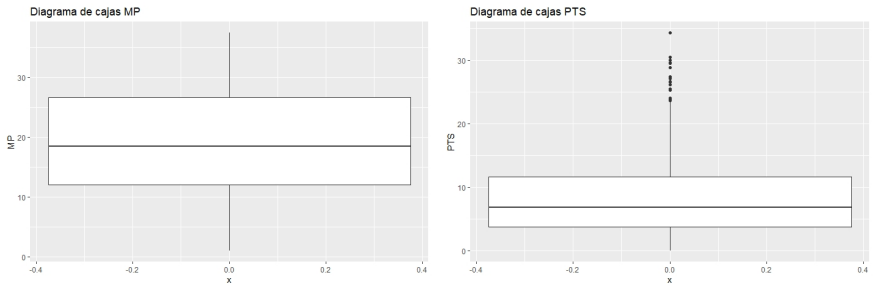


Figura 8.2: Diagrama de dispersión

A simple vista hay varios aspectos que pueden llamar nuestra atención. La variable 'PTS' tiene un máximo muy por encima de los valores acumu-

lados en el tercer cuartil, lo que indica que hay al menos un jugador en especial que marca mucho más puntos que el resto en función de los minutos disputados, es decir, un posible outlier. Tendremos que estudiar si aquellos jugadores con anotaciones muy altas (por ejemplo, más de 30 puntos), han disputado una cantidad de minutos de acuerdo con estos valores o, por el contrario, son valores atípicos. Lo mismo pasará en el caso contrario, es decir, aquellos jugadores con unos valores de anotación muy bajos para el tiempo de juego disputado. Por otro lado, media y mediana son bastante próximas en ambas variables, teniendo en cuenta el rango de valores en los que estamos trabajando, por lo que indicaría una simetría de los datos. Una manera de representar la distribución de los datos de ambas variables es mediante los diagramas de caja de la Figura 8.3. En el diagrama de caja 8.3(b) se identifican ciertos valores que pueden resultar ser outliers. En ambas cajas se muestran el 50 % de los valores observados, siendo relativamente simétricas las dos representaciones.



(a) Distribución de las observaciones en la variable MP

(b) Distribución de las observaciones en la variable PTS

Figura 8.3: Diagramas de caja de las variables MP y PTS

Para llegar al objetivo del problema, que recordemos es el de aplicar regresión lineal al conjunto de datos para explicar la variable 'PTS' a partir de la variable 'MP', se necesita construir un modelo definido por la ecuación lineal:

$$\text{PTS} = \beta_0 + \beta_1 \cdot \text{MP} + \epsilon \tag{8.1}$$

siendo ϵ la componente aleatoria del error.

Una vez definido el modelo, el siguiente paso es estimarlo, para lo que tendremos que estimar el vector de parámetros $\beta = (\beta_0, \beta_1)$ mediante $\hat{\beta} = (\hat{\beta_0}, \hat{\beta_1})$. Es de esta forma que el modelo estimado de (8.1) pasa a ser:

$$\hat{\text{PTS}} = \hat{\beta_0} + \hat{\beta_1} \cdot \text{MP} \tag{8.2}$$

Vamos a abordar el modelo de regresión mediante 3 procedimietos distintos:

1. Estimación de los parámetros de la regresión por mínimos cuadrados (LS), estudio de los residuos y apalancamiento, y conclusión sobre outliers.

2. Estimación de los parámetros de la regresión por mínima mediana de cuadrados (LMS) y cáculo de los residuos.

3. Estimación de los parámetros de regresión por mínimos cuadrados recortados (LTS) y cáculo de los residuos.

Una vez desarrollados los tres procedimientos, los compararemos y extraeremos conclusiones.

Comencemos con la regresión lineal por mínimos cuadrados. Se utilizará la función 'lm' de R para obtener los valores de los parámetros de la regresión, y los distintos estadísticos resumen como es el coeficiente de determinación. La salida que nos proporciona R es la mostrada en la Figura 8.4.

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.6744  -1.6591  -0.0807   1.3702  15.3996

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.09850    0.26917  -11.51   <2e-16 ***
MP           0.60271    0.01269   47.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.932 on 649 degrees of freedom
Multiple R-squared:  0.7765,    Adjusted R-squared:  0.7762
F-statistic:  2255 on 1 and 649 DF,  p-value: < 2.2e-16
```

Figura 8.4: Salida de R: Regresión lineal por mínimos cuadrados (LS)

En la columna 'Estimate' nos indica el valor estimado de $\hat{\beta}$ siendo este $(\hat{\beta}_0, \hat{\beta}_1) = (-3.09850, 0.60271)$. De esta forma podemos escribir el modelo estimado como

$$\hat{\text{PTS}} = -3.09850 + 0.60271 \cdot \text{MP} \tag{8.3}$$

En la columna 'Std. Error' nos muestra la estimación de la desviación típica de $\hat{\beta}_0$ y $\hat{\beta}_1$. En la Subsección 2.3.2 hemos visto las expresiones de los valores de $\text{Var}(\hat{\beta}_0)$ y $\text{Var}(\hat{\beta}_1)$, por lo que la estimación de cada desviación típica será su raíz cuadrada.

En las dos columnas restantes, se realizan los contrastes

$$H_0 : \beta_0 = 0,\ H_1 : \beta_0 \neq 0$$
$$H_0 : \beta_1 = 0,\ H_1 : \beta_1 \neq 0$$

Estos contrastes son los explicados en el T-test de la Subsección 2.4.1. En la columna 't-value' se muestra el valor observado del estadístico t_0 y en

la columna Pr($> |t|$) su p-valor asociado. Ambos p-valores son ≈ 0, por lo que se rechazaría la hipótesis nula y, por ello, tanto la variable MP como el coeficiente β_0 aportan información significativa al modelo.

En la parte de la salida 'Residuals', nos muestra información sobre la estimación de los errores $e_i = y_i - \hat{y}_i$. Teniendo en cuenta que la hipótesis teórica dice que los errores han de ser normales y centrados en 0, el valor muestral de la mediana debería de ser próximo a 0. Esto se cumple ya que es -0.0807.

En la última parte de la salida, bajo el nombre de 'Residual standard error' tenemos la estimación σ, dada por $\sqrt{MS_{Res}}$. Los grados de libertad son precisamente $n - p = 651 - 2 = 649$, ya que estamos trabajando con dos columnas de lo que sería la matriz X de la regresión, la de valores independientes y la asociada a la variable 'MP'. Justo debajo, nos econtramos con el coeficiente de determinación R^2 y el coeficiente de determinación ajustado R_a^2. Este último valor nos indica que el 77.62 % de la variabilidad de los puntos anotados en un partido por los jugadores, se explica mediante el modelo lineal utilizando como variable predictora los minutos disputados.

La última parte de la salida contrasta si el modelo es informativo, es decir si

$$H_0 : \beta_1 = 0 \tag{8.4}$$

lo cual se rechaza, por ser el p-valor próximo a 0. Es más, coincide con el p-valor del T-test ya que nos encontramos en el caso simple. Si estuviésemos en el caso múltiple, el contraste a realizar sería el siguiente

$$H_0 : \beta_1 = \ldots = \beta_{p-1} = 0 \tag{8.5}$$

La recta de regresión dada por la Ecuación (8.3) se dibuja en la Figura 8.5.

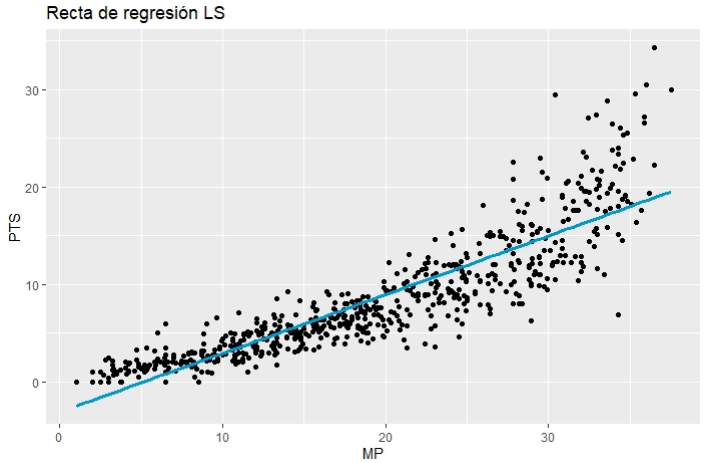


Figura 8.5: Ajuste lineal por mínimos cuadrados (LS)

A simple vista no parece una mala aproximación, aunque sí podemos pensar que aquellos jugadores con puntuaciones muy altas ejercen influencia sobre la predicción. Estudiemos ahora los residuos e_i con las herramientas estudiadas a lo largo del Capítulo 3.

El modelo teórico de regresión parte de la hipótesis de que los residuos (o errores) se distribuyen de forma normal centrados en 0, homocedásticos e independientes. Para que el modelo estimado sea adecuado para hacer las estimaciones, es necesario comprobar que estas condiciones se cumplen. Para ello vamos a realizar una exploración gráfica, mediante los distintos diagramas de residuos estudiados.

En primer lugar, analicemos el Q-Q Plot asociado al modelo lineal por mínimos cuadrados de la Figura 8.6

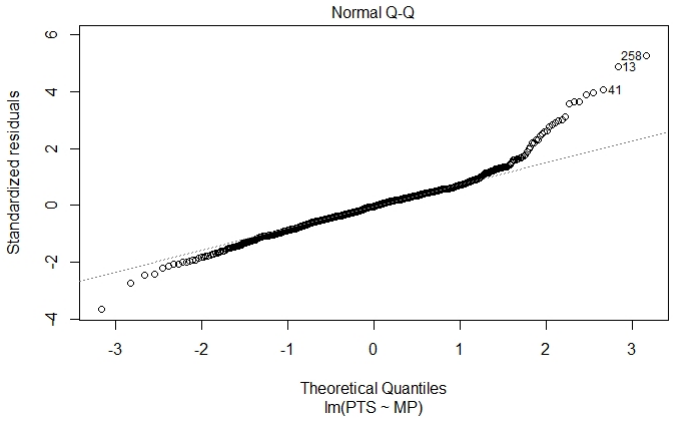


Figura 8.6: Grafico normal de los residuos estandarizados

En el gráfico se representan los cuantiles teóricos de una distribución $\mathcal{N}(0,1)$ frente a los cuantiles muestrales de los valores $d_{[i]}$. Para asumir normalidad los puntos representados deberían de estar cerca de la bisectriz del primer cuadrante (representada en la línea discontinua). En el problema que nos ocupa, vemos que para los valores menores de 2 aproximadamente, los puntos sí se ajustan a la bisectriz del primer cuadrante, salvo unos puntos cercanos a -3. Sin embargo, para valores más altos, los datos se desvían notablemente de la recta. Llaman especialmente la atención los puntos correspondientes a los individuos 13 y 258. Su cuantil teórico es aproximadamente 2, mientras que su cuantil observado está por encima de 4. Lo mismo ocurre para una serie de individuos cercanos a estos otros, como es el 41. Esto nos hace sospechar que la condición de normalidad podría no cumplirse. Para comprobar la normalidad de los residuos tenemos dos opciones, generar sus valores y representarlos en un diagrama de cajas, por ejemplo, o aplicar un test de normalidad a estos. Haremos ambas alternativas utilizando el test

de Shapiro-Wilk para comprobar la normalidad. El diagrama de caja es el mostrado en la Figura 8.7. En el gráfico vemos una simetría latente de los residuos, a pesar de que como pasaba en el Q-Q Plot la normalidad presenta serias dudas. El test de normalidad Shapiro-Wilk nos proporciona el p-valor $= 2.737e-15 \approx 0$, por lo que se rechaza la hipótesis de normalidad, y, en consecuencia, las condiciones teóricas sobre el modelo no se cumplen.

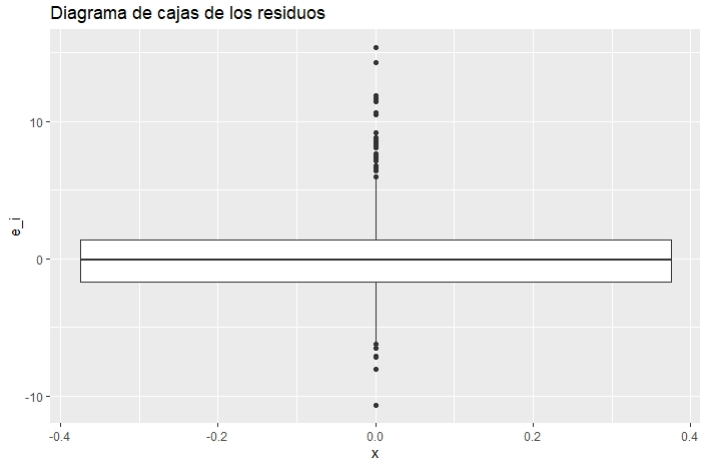


Figura 8.7: Diagrama de caja de los residuos e_i

El siguiente tipo de gráfico estudiado será el de los residuos e_i frente a los valores ajustados. El objetivo es el de determinar si el modelo lineal aplicado es correcto y no sería necesario otro tipo de ajuste (polinomial, exponencial, logarítmico...). Fijémonos ahora en la Figura 8.8.

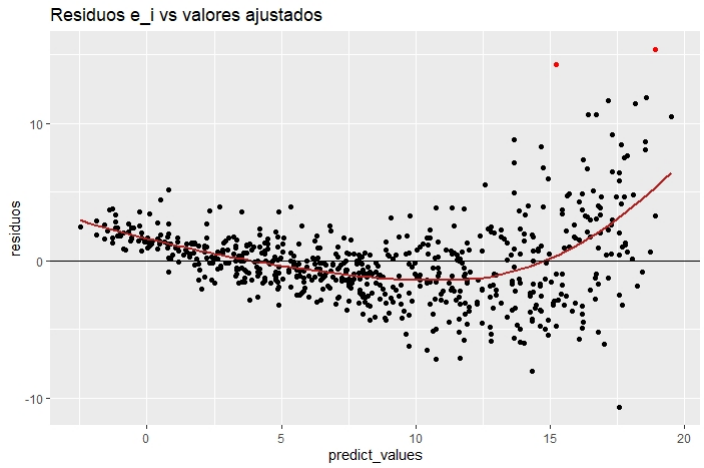


Figura 8.8: Residuos e_i frente a los valores ajustados $\hat{y}_i$

En rojo aparecen los residuos de los individuos 13 y 258, que se habían

identificado previamente como posibles 'outliers'.

- La media de los errores debería de ser próxima a cero. Esto se verifica si la línea roja está cerca del cero. Vemos que justamente por esos individuos más alejados en la 'cola' de la derecha, esto no ocurre, haciendo se desvíe la curva.

- Los residuos deberían ser homocedásticos, es decir, la varianza de los residuos debería de ser constante. Para observar la homocedasticidad, el intervalo de valores en el que se mueven los residuos debería de ser el mismo independientemente del valor ajustado. El problema lo causan principalmente individuos con valores altos de la variable explicativa como son, por ejemplo, los individuos 13 y 258.

- Los residuos ser linealmente independientes, es decir, en esta gráfica no deberíamos de observar ninguna forma establecida entre los residuos. En este caso no aparece ninguna relación lineal clara, aunque sí se puede observar una ligera relación entre los residuos.

Finalmente, representaremos el gráfico de residuos frente al apalancamiento en la Figura 8.9.

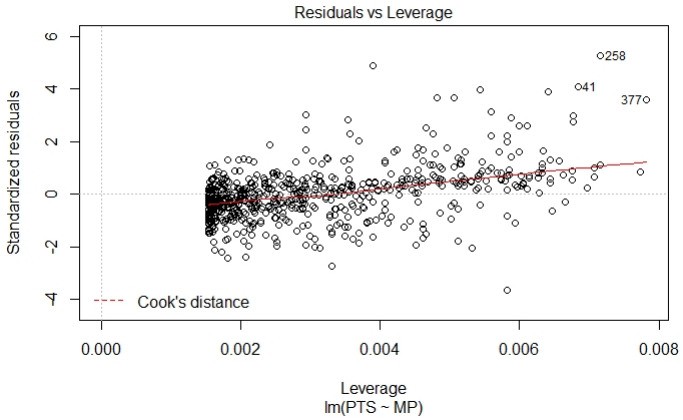


Figura 8.9: Residuos $\sqrt{d_i}$ frente al apalancamiento

En la representación aparecen algunas observaciones atípicas, pero ninguna fuera de la zona determinada por la distancia de Cook. Por ello, por este camino será inviable asegurar que un sujeto es un outlier, aunque sí que nos alerta sobre algunos candidatos como son el 41, 258 y 377.

El último paso, de acuerdo a lo estudiado en el Capítulo 3, es estudiar las medidas de influencia. Utilizaremos los criterios descritos en la Sección 3.3, detectaremos los outliers, y construiremos el modelo LS sin estas observaciones. Extraemos del conjunto de datos aquellas observaciones que verifican

$COVRATIO_i > 1 + 3p/n$ y $COVRATIO_i < 1 - 3p/n$, que era el criterio que habíamos considerado para detectar puntos influyentes sobre el ajuste. En la Figura 8.10 observamos los residuos frente a los valores ajustados de nuevo, pero esta vez aparecen en rojo aquellos puntos que hemos detectado como influyentes.

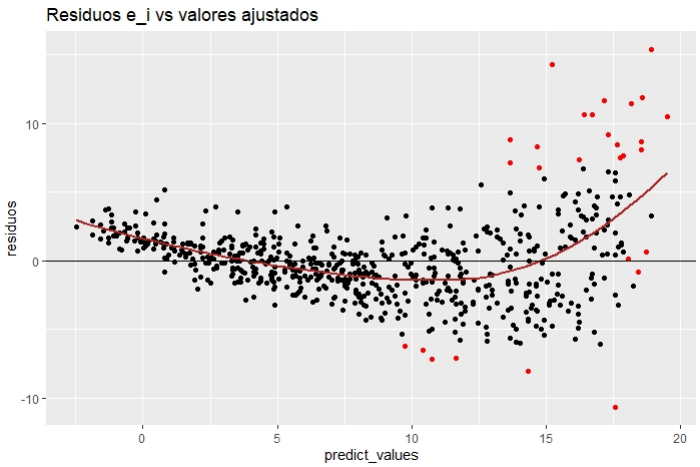


Figura 8.10: Residuos e_i frente a los valores ajustados $\hat{y}_i$ mostrando outliers

En esta gráfica vuelven a coincidir los individuos 13 y 258, ya detectados previamente. Podemos buscar en la base de datos original quiénes son y por qué resultan ser valores atípicos. Por un lado, la observación 13 es Giannis Antetokounmpo, quien obtuvo 29.5 puntos de media en 30.4 minutos. Por otro lado, la observación 258 se corresponde con James Harden, quien juega 36.5 minutos por partido de media y anota unos 34.3 puntos. La conclusión general que podemos extraer de estos datos atípicos es que se corresponden en su mayoría con jugadores AllStar, es decir, los mejores durante la temporada regular de la NBA. Dentro de esos puntos rojos también se encuentran jugadores reconocidos en la liga como LeBron James o Kawhi Leonard. Este criterio es tan solo una generalización, ya que también hay jugadores con muy pocas anotaciones en relación a los minutos disputados. Por ejemplo, Kenrich Williams, que anota 3.5 puntos de media en los 21.3 minutos que juega.

A continuación, eliminaremos estas observaciones y estimaremos de nuevo el modelo de regresión lineal, mostrados en la Tabla 8.1.

Claramente la variabilidad explicada es mayor en el modelo que no tiene en cuenta los outliers, y el resto de parámetros también sufren un cambio. Vamos a representar ambas rectas de regresión en el diagrama de dispersión para visualizar las diferencias. La línea naranja es la estimación del modelo sin tener en consideración los outlier detectados, mientras que la azul es la original. La Figura 8.11 recoge lo explicado.

Dataframe	$\hat{\beta}_0$	$\hat{\beta}_1$	MS_{Res}	R^2
Todas las observaciones	-3.09850	0.60271	8.6142	0.7765
Sin las observaciones influyentes	-2.40459	0.55345	5.1121	0.8227

Tabla 8.1: Variación en los coeficientes y estadísticos según se consideren los puntos de influencia o no

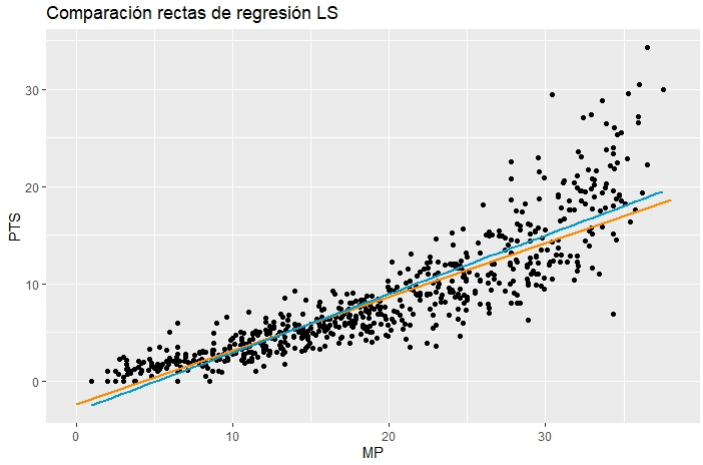


Figura 8.11: Ajuste por mínimos cuadrados de ambos modelos de regresión lineal

No hay una gran diferencia entre ambas rectas ya que no exite un outlier tan influyente como ocurría en el ejemplo estudiado en la Subsección 3.3.4. Aún así, sí que nos damos cuenta que esos jugadores de baloncesto con mucha cantidad de minutos y pocos puntos o con una anotación alta en relación al tiempo de juego, empeoraban el ajuste por mínimos cuadrados.

Es en este momento cuando aplicaremos los estimadores robustos por mínima mediana de cuadrados (LMS), y mínimos cuadrados recortados (LTS). Cabe esperar que la recta de regresión tenga menor pendiente que la recta naranja incluso. Una vez calculadas ambas rectas, obtendremos los residuos y evaluaremos aquellos que sean más altos.

Haciendo uso de la función 'lqs' del paquete 'MASS' de R, llegamos a las estimaciones del modelo (8.1), mediante el método LMS

$$\hat{\text{PTS}} = -0.71 + 0.40 \cdot \text{MP} \tag{8.6}$$

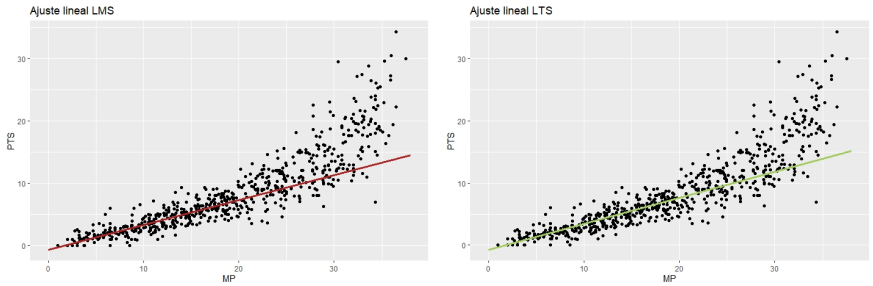
y el método LTS

$$\hat{\text{PTS}} = -0.7519 + 0.4184 \cdot \text{MP} \tag{8.7}$$

Ambas pendientes son menores que en el ajuste por mínimos cuadrados, y el punto de corte con el eje y se eleva. De esta forma, ambas rectas se ven

menos influenciadas por la presencia de outliers. En la Figura 8.12 aparecen dibujadas las rectas obtenidas de aplicar los métodos robustos.

Calculando el coeficiente de determinación R^2 para el ajuste por mínima mediana de cuadrados obtenemos $R^2_{LMS} = 0.63$. Del mismo modo, $R^2_{LTS} = 0.667$ para el ajuste por mínimos cuadrados recortados. No sorprende que la variabilidad explicada sea menor en estos dos casos, ya que la estimación por mínimos cuadrados es la que maximiza este valor.



(a) Recta de regresión por el método LMS

(b) Recta de regresión por el método LTS

Figura 8.12: Ajustes lineales LMS y LTS

Finalmente, se muestra en la Figura 8.13 un diagrama en el que aparecen las 4 rectas de regresión calculadas durante el problema.

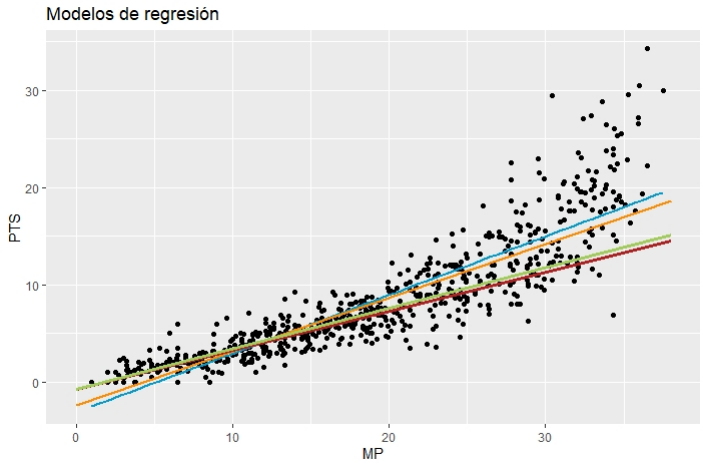


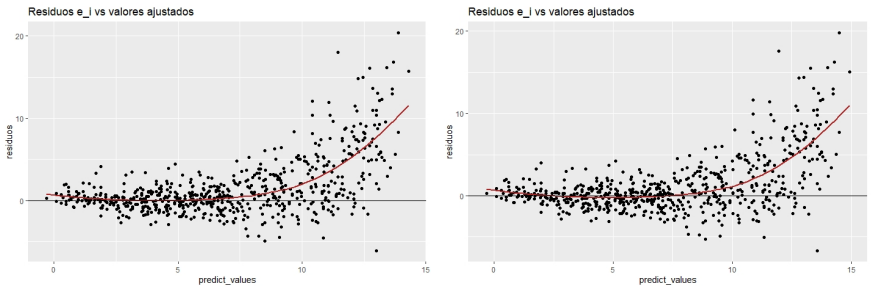
Figura 8.13: Rectas de regresión obtenidas por los distintos métodos estudiados

Resulta obvio que las rectas calculadas por los métodos robustos se ven menos influenciadas por las observaciones atípicas, haciendo más sencillo detectar qué jugadores se alejan de lo esperado. Es fácil deducir que los jugadores con una puntuación media por partido muy alta (cola derecha de

la gráfica) son los que distorsionan principalmente el análisis, y, por ello, toman los valores residuales más elevados en ambos métodos robustos.

Al igual que hicimos para la estimación proporcionada por el método de mínimos cuadrados, es posible representar los residuos de ambos métodos robustos frente a los valores ajustados. La Figura 8.14 recoge lo descrito.

De nuevo, si bien mejoramos el ajuste clásico de mínimos cuadrados, se intuye una relación lineal entre los residuos. Esto se debe a los outliers. En el ejemplo tratado aparece una serie de valores atípicos que distorsionan el análisis, los cuales, mayoritariamente, se encuentran por encima de los 20 puntos por partido. Por esta razón, lo más adecuado sería distinguir aquellos jugadores que anotan menos de esta puntuación media, y elaborar un modelo de regresión lineal. Con los datos restantes, se ajustaría su tendencia mediante otro modelo de regresión lineal. Es decir, lo más correcto sería hacer dos grupos de observaciones y estimar su puntuación por separado. Otra posible opción sería hacer uso de regresión no lineal.



(a) Residuos e_i frente a los valores ajustados $\hat{y}_i$ por LMS

(b) Residuos e_i frente a los valores ajustados $\hat{y}_i$ por LTS

Figura 8.14: Diagramas de residuos e_i frente a los valores ajustados por métodos robustos

Normalmente, el conocimiento matemático es suficiente para realizar un estudio sobre una base de datos, sin embargo, la ausencia de conocimientos específicos sobre el tema puede hacer que se nos escapen detalles importantes. En este caso, si utilizamos el conocimiento que podamos tener sobre la NBA, resultará lógico pensar que esos jugadores con tantos puntos por partido son los que disputan el AllStar. Es por eso que, escogiendo aquellos datos referentes a estos jugadores AllStar, podemos distinguirlos en el diagrama de dispersión y ver que, en efecto, son una de las causas principales de las diferencias con respecto a la recta LS. En la Figura 8.15 aparecen señalados estos jugadores mediante una estrella.

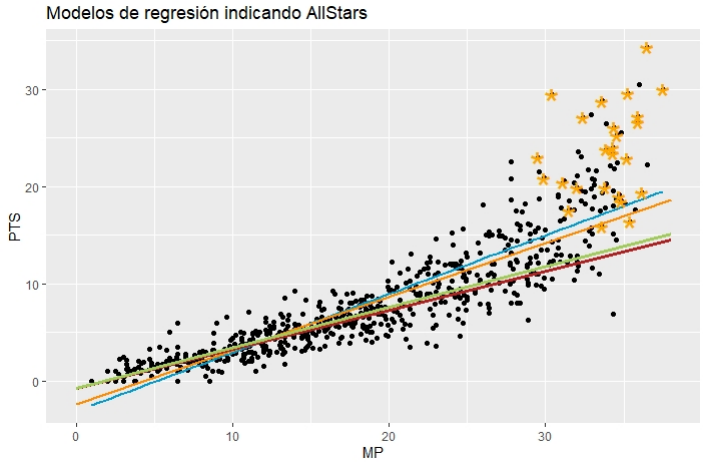


Figura 8.15: Rectas de regresión indicando los jugadores AllStar

Capítulo 9

Conclusión

En el ejemplo final hemos utilizado la mayor parte de la teoría expuesta a lo largo del trabajo. El propósito de aplicar los métodos robustos y el ajuste clásico de mínimos cuadrados no era otro que el de indicar las distintas alternativas que ha de barajar todo estadístico que trabaje con una base de datos de esta índole. En este ejemplo, las diferencias son notorias entre los tres procedimientos. Es más, aunque detectemos los posibles puntos que distorsionan el análisis, la estimación mínimo cuadrática continuará siendo considerablemente distinta a las obtenidas por LMS y LTS. No se busca convencer al lector de que una alternativa sea mejor o peor que la otra, de hecho, la simplicidad suele ser lo más cómodo, pero no siempre lo más correcto. Es por eso que resulta vital realizar un análisis exploratorio de los datos que nos ponga en situación sobre el problema a tratar, y en función de este estudio previo elegir un procedimiento u otro, sin olvidarnos de las desventajas o problemas que puede tener nuestra elección. Con la aproximación lineal de este conjunto de datos último con estas variables, queda probado que aplicar distintos métodos puede llevarnos a un análisis más preciso y correcto.

Si bien la finalidad de este trabajo es mostrar de forma aplicada las diferencias patentes entre los tres métodos, ha sido necesario el desarrollo teórico expuesto a lo largo de los capítulos. En mi opinión, a pesar de que siempre se ha de partir de una intuición fundada en los gráficos y los análisis previos para tomar el camino adecuado; en algunas ocasiones esto puede resultar tedioso, ya sea por un volumen cuantioso de datos o la dificultad en su representación. Ante estas situaciones, es de importancia seguir con rigor todos y cada uno de los pasos descritos, ya que cualquier hipótesis previa necesaria fallida o mala interpretación, puede hacernos llegar a un resultado completamente erróneo.

Los outliers o valores atípicos son comunes en muchas disciplinas, y no tan solo se presentan en modelos lineales. Buscar estimaciones robustas nos previene ante estas situaciones. No quiere decir que nos centremos en ellas

como punto fundamental del análisis, si no que evitemos que distorsionen la aproximación o estimación del resto de observaciones. Por tanto, siendo conservador y adelantándonos ante posibles outliers, podremos despreocuparnos frente a los problemas causados por estos.

Como valoración propia, diría que gracias a los conocimientos adquiridos en la realización de este trabajo, me he dado cuenta de la importancia de los resultados teóricos. Generalmente, en otros ámbitos científicos, se tiende a ignorar el estudio de proposiciones y teoremas para centrarse en su aplicación directa, muchas veces ya implementada directamente en los programas informáticos. Conocer la base matemática de estos algoritmos permite detectar los fallos de manera inmediata en muchas ocasiones, agilizando el proceso de corrección y optimizando todos los procesos. Por esto, pienso que una mezcla de conocimientos teóricos y aplicados es lo que impulsa el conocimiento matemático y justifica su versatilidad.

Bibliografía

[1] J. P. Barret, The coefficient of determination-Some limitations, Am. Stat., 28, 19-20, 1974.

[2] D. A. Belsley , E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, John Wiley & Sons, New York, 1980.

[3] P. J. Bickel, On some analogues to linear combination of order statistics in the linear model, Annals of Statistics, 1, 597–616. 1973.

[4] S. Cankaya , G. T. Kayaalp , L. Sangun , Y. Tahtali & M. Akar, A Comparative Study of Estimation Methods for Parameters in Multiple Linear Regression Model, Journal of Applied Animal Research, 43-47, 2006.

[5] R. D. Cook, Detection of influential observation in linear regression, Technometrics, 19, 15–18. 1977.

[6] R. D. Cook, Influential observations in linear regression, J. Am. Stat. Assoc., 74, 169–174, 1979.

[7] R. D. Cook, S. Weisberg, Regression Diagnostics: Robust versus Least Squares Residuals, Department of Applied Statistics, University of Minnesota, Technical Report Number 538, 1990.

[8] J. Crocker, Judgment of covariation by social perceivers. Psychological Bulletin, 272-–292, 1981.

[9] D. L. Donoho, P. J. Huber, The notion of breakdown point, in A Festschrift for Erich Lehmann, edited by P. Bickel, K. Doksum, and J. L. Hodges, Jr., Wadsworth, Belmont, CA, 1983.

[10] N. R. Draper, H. Smith: Applied Regression Analysis, John Wiley & Sons, New York, 1998.

[11] D. Dua, C. Graff, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences (2017).

[12] F. Y. Edgeworth, On observations relating to several quantities, Hermathena, 6, 279–285, 1887.

[13] F. Galton, Natural Inheritance (5th ed.), New York: Macmillan and Company, 1894.

[14] A.Giloni, M.Padberg, Least Trimmed Squares Regression, Least Median Squares Regression, and Mathematical Programming, Mathematical and Computer Modelling, Pergamon, 1043–1060, 2002.

[15] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel, Robust Statistics The Approach Based on Influence Functions, John Wiley & Sons, 1986.

[16] R. W. Hill, Robust Regression When There Are Outliers in the Caniers, unpublished Ph.D. dissertation, Harvard University, Boston, MA, 1977.

[17] M. Hofmann, C. Gatu, E. J. Konthoghiorghes, An Exact Least Trimmed Squares Algorithm for a Range of Coverage Values, Journal of Computational and Graphical Statistics, Volume 19, Number 1, 191–204, 2010.

[18] P. J. Huber, Robust estimation of a location parameter, Ann. Math. Stat., 35, 73–101, 1964.

[19] P. J. Huber, Finite Sample Breakdown of M- and P-Estimators, The Annals of Statistics, Volume 12, 119–126, 1984.

[20] P. J. Huber, R. Dutter, Numerical solutions of robust regression problems, in: COMPSTA T 1974, Proceedings in Computational statistics, edited by G. Bruckmann, P. Verlag, Vienna, 1974.

[21] L. A. Jaeckel, Estimating regression coefficients by minimizing the dispersion of residuals, 5 , 1449–1458, 1972.

[22] J. Jureckovà, Nonparametric estimate of regression coefficients, 42, 1328–1338, 1971.

[23] R. Koenker, G. J. Bassett, Regression quantiles, Econornetrica, 46, 33–50, 1978.

[24] C. L. Mallows, On some topics in robustness, unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ, 1975.

[25] K.Mardia, J.Kent, J. Bibby, Multivariate Analysis, Academic Press, 1979.

[26] R. A. Maronna, R. D. Martin, V. J. Yohai, M. Salibián-Barrera, Robust Statistics Theory and Methods (with R), John Wiley & Sons, 2019.

[27] D. C. Montgomery, E. A. Peck, G. G. Vining, Introduction to Linear Regression Analysis, Wiley, 2012.

[28] P. Morano, G. de Mare, F. Tajani, LMS for Outliers Detection in the Analysis of a Real Estate Segment of Bari. Lecture Notes in Computer Science. 7974. 457–472, 2013.

[29] K. Pearson, F. Galton: A Centenary Appreciation, Cambridge University Press, 1922.

[30] P. J. Rousseeuw, Least median of squares regression, Journal of the American Statistical Association 79, 871–880, 1984.

[31] P. J. Rousseeuw, A. M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, 1987.

[32] A. F. Siegel, Robust regression using repeated medians, Biometrika, 69, 242–244, 1982.

[33] J.M. Stanton , Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors, Journal of Statistics Education, 9, 1–13, 2017.

[34] R. A. Yaffee, Robust Regression Analysis: Some Popular Statistical Package Options, 2002.

[35] `https://www.basketball-reference.com/leagues/NBA_2020_per_game.html` (último acceso: 23 de enero de 2021).