9

Diagnosis del modelo

En este capítulo se investiga la detección de posibles deficiencias en el modelo por incumplimiento de las hipótesis fijadas en 2.3. Para ello la principal herramienta es el análisis de los residuos que nos permite detectar los siguientes problemas:

1. Algunas de las variables explicativas del modelo tienen una relación no lineal con la variable respuesta.
2. No hay homocedasticidad, es decir, los errores no tienen varianza constante.
3. Los errores no son independientes.
4. Muchas observaciones atípicas.
5. Hay observaciones demasiado influyentes.
6. Los errores no tienen distribución normal

También estudiaremos la consecución del mejor grupo reducido de variables regresoras.

9.1. Residuos

9.1.1. Estandarización interna

Los residuos de un modelo lineal se obtienen como diferencia entre los valores observados de la variable respuesta y las predicciones obtenidas para los mismos datos:

$$\mathbf{e} = (e_1, \ldots, e_n)' = \mathbf{Y} - \widehat{\mathbf{Y}}$$

La media de los residuos es cero

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

y una estimación aproximada de la varianza es

$$\frac{1}{n-k-1}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2 = \mathrm{SCR}/(n-k-1) = \mathrm{ECM}$$

que tiene sólo $n-k-1$ grados de libertad, donde k es el número de variables regresoras, ya que los n residuos no son independientes,

Se llaman *residuos estandarizados* a

$$d_i = \frac{e_i}{\sqrt{\mathrm{ECM}}} \qquad i = 1, \ldots, n$$

que tienen media cero y varianza aproximada uno.

Ahora bien, como el vector de residuos aleatorios es $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y} = (\mathbf{I} - \mathbf{P})\boldsymbol{\epsilon}$, donde $\mathbf{P}$ es la matriz proyección, la matriz de varianzas-covarianzas de los residuos es $\text{var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{P})$ de manera que

$$\text{var}(e_i) = \sigma^2(1 - h_{ii})$$

donde h_{ii} es el i-ésimo elemento[1] de la diagonal de $\mathbf{P}$.

La utilización de los residuos $\mathbf{e}$ como estimaciones de los errores $\boldsymbol{\epsilon}$ requiere que mejoremos la estandarización. Como $0 \leq h_{ii} \leq 1$, utilizar ECM para estimar la varianza $\text{var}(e_i)$ es una sobreestimación:

$$\begin{array}{ccccc} 0 \leq & \text{var}(e_i) & \leq & \sigma^2 \\ 0 \leq & \text{ECM}(1 - h_{ii}) & \leq & \text{ECM} \end{array}$$

De modo que muchos autores recomiendan trabajar con los *residuos studentizados*

$$r_i = \frac{e_i}{[\text{ECM}(1 - h_{ii})]^{1/2}} \qquad i = 1, \ldots, n$$

Además, h_{ii} es una medida de la localización del i-ésimo punto $\mathbf{x}_i$ respecto al punto medio.

En la regresión lineal simple

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{9.1}$$

En el modelo de regresión múltiple

$$h_{ii} = \frac{1}{n}[1 + (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}_{\mathbf{xx}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})] = \frac{1}{n}(1 + D_i^2) \tag{9.2}$$

donde D_i es la llamada distancia de Mahalanobis.

Así, la varianza de un error e_i depende de la posición del punto $\mathbf{x}_i$. Puntos cercanos al punto central $\bar{\mathbf{x}}$ tienen mayor varianza (pobre ajuste MC) que los puntos alejados.

Como las violaciones de las hipótesis del modelo son más probables en los puntos remotos, pero más difíciles de detectar con los residuos e_i (o d_i), porque los residuos son menores, es mejor trabajar con los residuos r_i ya que $\text{var}(r_i) = 1$ constante, desde el punto de vista de la localización de los $\mathbf{x}_i$.

Para n grande se puede trabajar con los d_i o con los r_i. Pero como valores altos de e_i y de h_{ii} pueden indicar un punto de alta influencia en el ajuste MC, se recomienda la utilización de los residuos estudentizados r_i. Estos residuos se utilizarán en el diagnóstico de valores atípicos.

Ejemplo 9.1.1

Si recuperamos el ejemplo de regresión simple propuesto en la sección 1.2 con los datos de tráfico, podemos calcular los residuos studentizados de ese modelo.

Primero calculamos los elementos de la diagonal de la matriz $\mathbf{P}$*, por ejemplo*

$$h_{11} = \frac{1}{24} + \frac{(12.7 - 54.44167)^2}{15257.4383} = 0.155865$$

y con este valor se obtiene el residuo

$$r_1 = \frac{0.528699}{0.2689388(1 - 0.155865)^{1/2}} = 2.13968$$

Los otros residuos se calculan de forma similar, mejor con la ayuda de una hoja de cálculo o con un programa estadístico (ver sección 9.4).

1. En muchos libros escritos en inglés la matriz proyección se llama *hat* y se escribe $\mathbf{H}$.

9.1.2. Estandarización externa

Para calcular los residuos estudentizados r_i en el apartado anterior hemos utilizado ECM como estimador de la varianza σ^2. Nos referiremos a esto como una estimación *interna* puesto que para calcularla se utilizan los n puntos. Otra aproximación consiste en estimar σ^2 con el conjunto de datos sin la i-ésima observación.

Si $s^2_{(i)}$ es la estimación de σ^2 así obtenida, se demuestra que

$$s^2_{(i)} = \frac{(n-k-1)\text{ECM} - e_i^2/(1-h_{ii})}{n-k-2} = \text{ECM}\left(\frac{n-k-1-r_i^2}{n-k-2}\right)$$

Si utilizamos estos estimadores de σ^2 en lugar de ECM, producimos los llamados residuos studentizados externamente o R-Student

$$t_i = \frac{e_i}{[s^2_{(i)}(1-h_{ii})]^{1/2}} \qquad i = 1, \ldots, n \tag{9.3}$$

En la mayoría de situaciones los residuos t_i no diferirán de los residuos studentizados r_i. Sin embargo, si la i-ésima observación es influyente, entonces $s^2_{(i)}$ puede diferir significativamente de ECM y el estadístico t_i será más sensible para este punto. Además, bajo las hipótesis estándar $t_i \sim t_{n-k-2}$, de modo que podemos considerar un procedimiento formal para la detección de valores atípicos mediante el contraste de hipótesis y utilizando algún método múltiple. En la práctica, un diagnóstico "a ojo" es más útil y rápido. En general, se considera que un residuo es atípico o *outlier* si $|t_i| > 2$. Además, la detección de los valores atípicos está ligada a la detección de puntos influyentes.

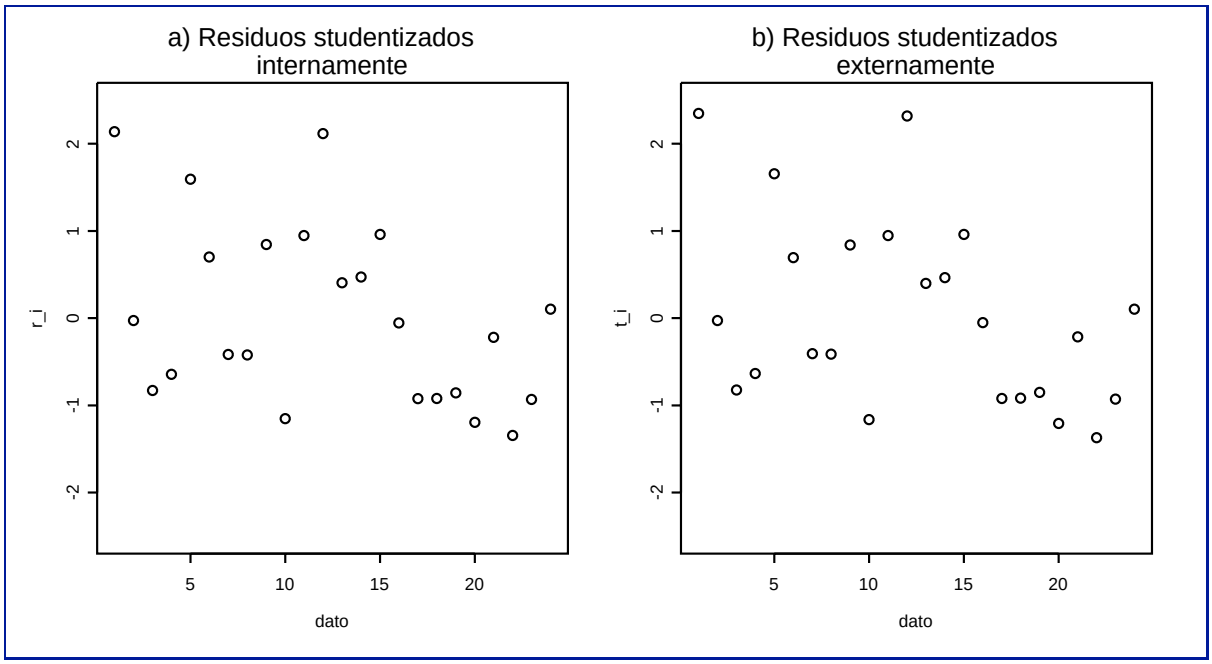


Figura 9.1: Gráficos de los residuos studentizados del ejemplo 9.1.1.

Ejemplo 9.1.2

Vamos a calcular el residuo studentizado externamente t_1 para la primera observación de la regresión simple continuación del ejemplo 9.1.1. Para ello necesitamos el valor del error ECM $= (0.2689388)^2 = 0.072328$ *con el que calculamos*

$$s^2_{(i)} = 0.072328\frac{24-1-1-2.13968^2}{24-1-2} = 0.060004$$

y con esta estimación externa

$$t_1 = \frac{0.528699}{\sqrt{0.060004(1 - 0.155865)}} = 2.349159$$

Siguiendo con la misma idea, también podemos calcular los residuos en función de las predicciones $\hat{y}_{i(i)}$ calculadas con el modelo de regresión sin la i-ésima observación. Sean $e_{(i)} = y_i - \hat{y}_{i(i)}$ los residuos así obtenidos y

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 \tag{9.4}$$

su suma de cuadrados[2]. También algunos autores llaman *error cuadrático de validación* a esta suma de cuadrados por ser una medida externa de precisión del modelo.

Se demuestra que

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \qquad \text{var}(e_{(i)}) = \frac{\sigma^2}{1 - h_{ii}} \tag{9.5}$$

de modo que la estandarización de estos residuos

$$\frac{e_{(i)}}{[\text{var}(e_{(i)})]^{1/2}} = \frac{e_i}{[\sigma^2(1 - h_{ii})]^{1/2}}$$

también depende del estimador que utilicemos para estimar σ^2. Si utilizamos el estimador interno ECM, recuperamos los residuos studentizados r_i y si utilizamos el estimador externo $s_{(i)}^2$ obtenemos los residuos studentizados externamente t_i.

Los residuos asociados con puntos para los que h_{ii} sea grande, tendrán residuos $e_{(i)}$ grandes. Estos puntos serán puntos de alta influencia. Una gran diferencia entre el residuo ordinario e_i y el residuo $e_{(i)}$ indicará un punto en el que el modelo, con ese punto, se ajusta bien a los datos, pero un modelo construido sin ese punto "predice" pobremente.

9.1.3. Gráficos

Algunos gráficos de los residuos nos van a ayudar en el diagnóstico del modelo aplicado.

En primer lugar, el análisis de datos univariante de los residuos y, en particular, los gráficos como histogramas, diagramas de caja, diagramas de tallo y hojas, etc. nos mostrarán algunos detalles. Por ejemplo, en el diagrama de caja podemos estudiar la centralidad, la simetría y la presencia de valores atípicos.

Ejemplo 9.1.3

También con los datos de tráfico del ejemplo de regresión simple propuesto en la sección 1.2 podemos representar algunos gráficos de los residuos sin estandarizar. En la figura 9.2 se muestran dos de los gráficos obtenidos con el programa SPSS. En ellos se observa una cierta asimetría de los residuos, aunque no hay ningún valor atípico.

Otros gráficos adecuados para el análisis de la regresión son:

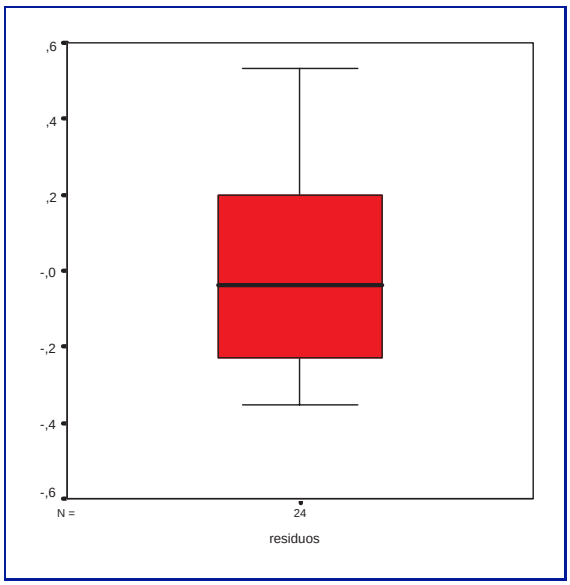
- Gráfico de dispersión de los residuos respecto al índice $i = 1, \ldots, n$.

  Este diagrama puede indicar algún tipo de correlación no deseada entre los residuos o alguna agrupación contraria a la supuesta aleatoriedad (figura 9.3 a).

- Gráfico de los residuos versus los datos de la variable respuesta.

  Permite observar los residuos desde los valores observados de la variable respuesta.

2. *prediction error sum of squares*




```
RESIDUO Stem-and-Leaf Plot

Frequency     Stem &  Leaf

   14,00       -0 .  00011122222333
    8,00        0 .  01112224
    2,00        0 .  55

Stem width:  1,000000
Each leaf:       1 case(s)
```

Figura 9.2: Boxplot y diagrama de tallo y hojas de los residuos en la regresión simple del ejemplo 9.1.3.

- Gráfico de los residuos versus los valores ajustados.

  Este gráfico es muy importante porque debe mostrar una total aleatoriedad. La dispersión horizontal no debe presentar ninguna tendencia. Una curvatura indica la violación del supuesto de linealidad del modelo en el caso de regresión lineal simple (figura 9.3 b). Una forma triangular indica una posible heterogeneidad o violación de la hipótesis de varianza constante de los errores.

- Gráficos de los residuos versus las observaciones de la variable o variables regresoras.

  Sirven para detectar si las variables regresoras o explicativas han de incluirse en el modelo con alguna transformación no lineal.

- Gráfico de los valores observados versus los valores ajustados.

  La proximidad de los puntos a la bisectriz muestra el ajuste de la recta de regresión (figura 9.3 c).

- Gráfico de los cuantiles de la distribución normal o QQ-plot y gráfico de las probabilidades acumuladas de la distribución normal o PP-plot.

  Con estos gráficos se pretende visualizar el ajuste de la distribución muestral de los residuos a la ley normal. En el QQ-plot se dibujan los puntos asociados a los cuantiles de la distribución normal (estándar en R o sin estandarizar como en SPSS). En el PP-plot se dibujan las probabilidades acumuladas estimadas y teóricas para la distribución normal. En ambos casos se dibuja también una recta que representa el ajuste perfecto a la distribución normal. Los desvíos exagerados de dichas rectas indican una posible violación de la hipótesis de normalidad (figura 9.3 d).

  El estudio de la normalidad de los residuos se debe completar con algún contraste de ajuste como la prueba ji-cuadrado o el test de Kolmogorov (ver sección 9.4).

Ejemplo 9.1.4

Como continuación del ejemplo de regresión simple 9.1.3 con los datos de tráfico, podemos representar algunos gráficos como los de la figura 9.3. Entre esos gráficos podemos destacar la no aleatoriedad manifiesta del gráfico (b) que indica un ajuste no lineal entre las variables. Ello justifica la introducción del modelo parabólico (ejercicio 9.1).

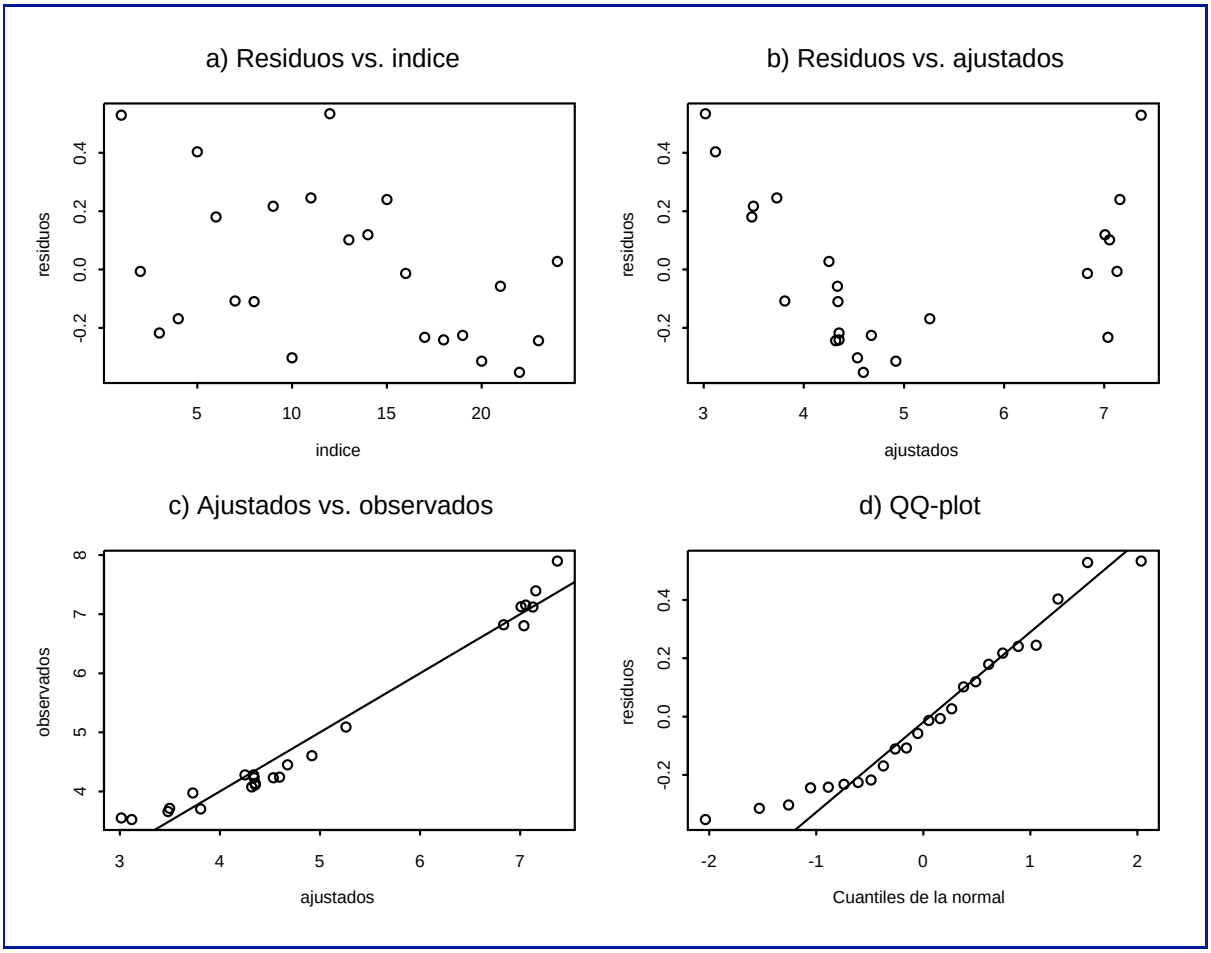


Figura 9.3: Gráficos en el análisis de la regresión simple del ejemplo 9.1.4.

9.2. Diagnóstico de la influencia

Ocasionalmente hallamos que algún dato o un pequeño subconjunto de datos ejerce una desproporcionada influencia en el ajuste del modelo de regresión. Esto es, los estimadores de los parámetros o las predicciones pueden depender más del subconjunto influyente que de la mayoría de los datos. Queremos localizar estos puntos influyentes y medir su impacto en el modelo. Si por alguna razón concreta son puntos "malos" los eliminaremos, pero si no ocurre nada extraño, su estudio puede darnos algunas claves del modelo.

9.2.1. Nivel de un punto

Casi siempre los puntos definidos por las variables regresoras o explicativas forman una nube y están razonablemente repartidos alrededor del punto medio. Sin embargo, alguno de ellos o un pequeño grupo puede aparecer muy alejado del resto. Estos valores son potencialmente peligrosos, puesto que pueden afectar excesivamente al ajuste del modelo. Vamos a definir el concepto de nivel[3] de un punto y señalaremos los que tengan un nivel muy alto (*leverage points*).

El nivel de un punto es una medida de la distancia del punto al centroide del conjunto de datos. Existen varias propuestas pero la más extendida se basa en los elementos h_{ii} de la diagonal de la matriz proyección $\mathbf{P}$. Estos elementos se calculan con las fórmulas 9.1 en el caso de la regresión simple y 9.2 para la regresión múltiple.

3. *leverage*

Como

$$\sum_{i=1}^{n} h_{ii} = \text{traza}(\mathbf{P}) = \text{rango}(\mathbf{P}) = k + 1$$

el tamaño medio de cada h_{ii} es $(k + 1)/n$. Así, cuando un punto verifique $h_{ii} > 2(k + 1)/n$ diremos que dicha observación es un punto de alto nivel. Estos puntos se deben marcar para su posterior estudio ya que son potencialmente influyentes.

Ejemplo 9.2.1

Siguiendo con el ejemplo 9.1.1 los datos con mayor nivel son

dato	*nivel*
1	0.15586452
15	0.13601868
2	0.13354830

Dado que $2(k + 1)/n = (2 \cdot 2)/24 = 0.1666$*, no hay ningún punto de alto nivel.*

9.2.2. Influencia en los coeficientes de regresión

Entre las medidas de influencia sobre los coeficientes de regresión la más empleada es la distancia de Cook (1977,1979)

$$C_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{(k + 1)\text{ECM}} \qquad i = 1, \ldots, n \tag{9.6}$$

donde $\widehat{\boldsymbol{\beta}}$ son las estimaciones MC en el modelo con todos los puntos, mientras que $\widehat{\boldsymbol{\beta}}_{(i)}$ son las estimaciones sin el i-ésimo punto. Esta medida calcula la distancia cuadrática entre $\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\beta}}_{(i)}$, relativa a la geometría fija de $\mathbf{X}'\mathbf{X}$.

Otra versión equivalente de esta distancia es

$$C_i = \frac{(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})'(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})}{(k + 1)\text{ECM}}$$

ya que $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ y $\widehat{\mathbf{Y}}_{(i)} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{(i)}$.

Sin embargo para el cálculo de esta distancia es mejor utilizar la fórmula

$$C_i = \frac{r_i^2}{k + 1} \cdot \frac{h_{ii}}{1 - h_{ii}}$$

donde la primera parte depende del ajuste al modelo de la i-ésima predicción, mientras que el segundo factor es una función de la distancia del punto $\mathbf{x}_i$ al centroide del conjunto de observaciones de las variables explicativas. Una demostración de esta fórmula puede verse en el ejercicio 9.19 del libro de Ugarte y Militino[70].

La búsqueda de puntos influyentes se puede iniciar con la identificación de puntos con distancia de Cook elevada. Sin embargo se desconoce la distribución exacta de este estadístico y no hay reglas fijas para la determinación de los puntos con valor de C_i grande. Los puntos con distancias de Cook grandes pueden ser influyentes y podemos extraerlos del análisis para ver si los cambios son apreciables.

Ejemplo 9.2.2

Con el ejemplo de regresión simple que estamos estudiando desde el ejemplo 9.1.1 se observa que los datos con mayor distancia de Cook son:

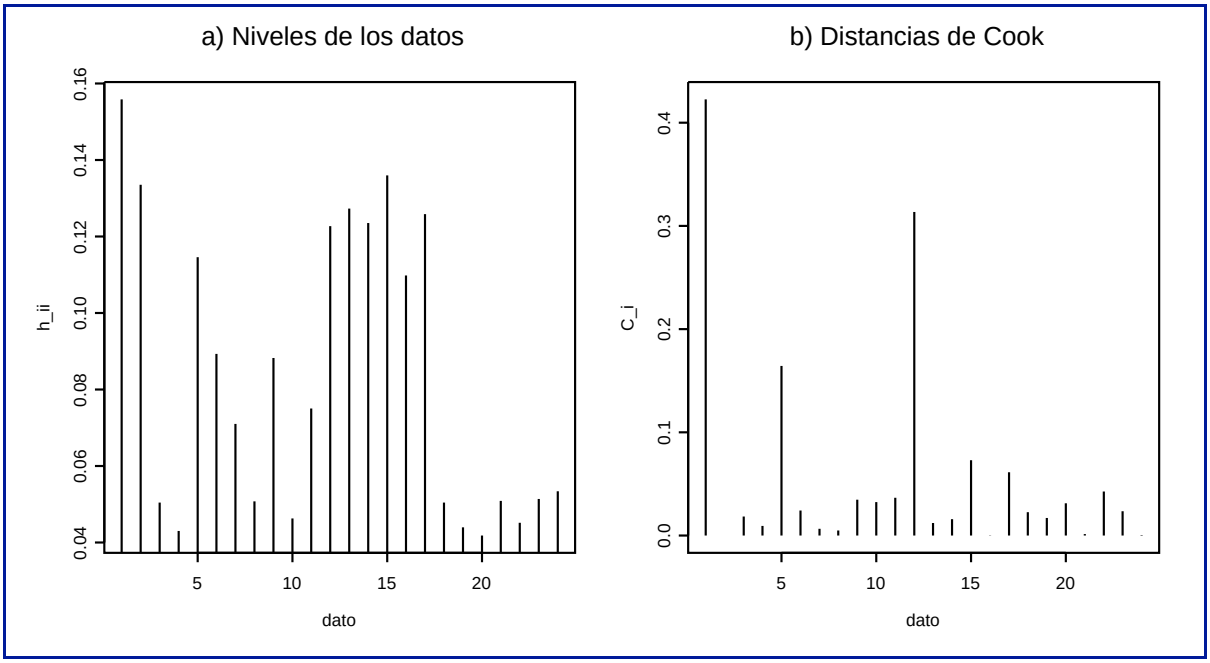


Figura 9.4: Gráficos de los niveles y distancias de Cook de los datos del ejemplo 9.2.2.

dato	h_{ii}	r_i	C_i
1	0.1559	2.1397	0.4227
12	0.1227	2.1178	0.3136

Estos datos son los de mayor influencia debida al gran residuo studentizado (los dos mayores) y a su alto nivel, especialmente el dato 1.

Otra medida de influencia sobre cada coeficiente de regresión por separado fue propuesta por Belsley et al.[6] y consiste en la diferencia estandarizada entre la estimación MC de dicho parámetro con todas las observaciones y la estimación MC del mismo sin la i-ésima:

$$\text{Dfbetas}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s^2_{(i)} c_{jj}}}$$

para $j = 0, 1, \ldots, k$ y $i = 1, \ldots, n$, donde c_{jj} es el j-ésimo elemento de la diagonal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$ y $s^2_{(i)}$ la estimación MC de la varianza σ^2 sin la i-ésima observación. Observemos que $s^2_{(i)} c_{jj}$ es una estimación de la varianza $\text{var}(\hat{\beta}_j) = \sigma^2 c_{jj}$.

Un valor absoluto desmesurado de esta medida indica una gran influencia de la observación i-ésima sobre la estimación del coeficiente β_j. En la práctica se considera una observación influyente cuando |Dfbetas| > 1 para un pequeño conjunto de datos y |Dfbetas| > $2/\sqrt{n}$ en general.

9.2.3. Influencia en las predicciones

Como hemos visto, la distancia de Cook es también una medida de la influencia de un punto sobre el conjunto de predicciones.

Otra medida de influencia de la i-ésima observación sobre la predicción de la propia observación i es el estadístico

$$\text{Dffits}_i = \frac{|\hat{y}_i - \hat{y}_{i(i)}|}{\sqrt{s^2_{(i)} h_{ii}}}$$

donde se estandariza la diferencia entre las predicciones de la i-ésima observación con y sin ella misma.

A partir de las ecuaciones 9.3 y 9.5 se demuestra que (ejercicio 9.3)

$$\text{Dffits}_i = |t_i| \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \tag{9.7}$$

donde t_i son los residuos studentizados externamente.

En general se considera que la influencia es notable si el Dffits es superior a $2\sqrt{(k+1)/n}$, mientras que para un conjunto de datos reducido basta que sea mayor que uno.

Ejemplo 9.2.3

Como continuación del ejemplo 9.2.2 podemos calcular el Dffits_1 *para la primera observación:*

$$\text{Dffits}_1 = |2.349159| \sqrt{\frac{0.155865}{1 - 0.155865}} = 1.009439$$

que supera el valor frontera $2\sqrt{2/24} = 0.577$ *y muestra la alta influencia de esta observación.*

9.3. Selección de variables

Con el objetivo de considerar el mejor modelo de regresión posible, el experimentador debe seleccionar un conjunto de variables regresoras entre las observadas y, si es necesario, entre potencias y productos de las mismas. Una primera decisión fijará el tipo de relación funcional con la variable respuesta pero, en todo caso, la selección de un conjunto reducido de variables explicativas es un problema complicado. Si consideramos un número demasiado pequeño de variables es posible que la potencia del modelo se vea reducida y que las estimaciones obtenidas sean sesgadas, tanto de los coeficientes de regresión, como de las predicciones. Este sesgo se origina ya que los errores calculados con los datos observados pueden contener efectos no aleatorios de las variables desechadas. Por otra parte, un número muy grande de variables explicativas complica la utilidad práctica del modelo y, aunque mejora el ajuste aparente, aumenta la varianza de los estimadores de los parámetros.

Decidir el mejor conjunto de variables es prácticamente un arte, en el que algunas técnicas sirven de apoyo: test t de Student de los coeficientes de regresión, test F de significación de la regresión, estudio de la multicolinealidad, etc. Sin embargo, ya hemos alertado sobre la utilización ciega de los test t parciales para medir la importancia de las variables. Así pues, es preciso añadir algunas técnicas específicas para comparar modelos de regresión que pasamos a detallar.

9.3.1. Coeficiente de determinación ajustado

Esta técnica consiste en calcular los coeficientes de determinación de todos los modelos posibles con la combinación de cualquier número de variables explicativas. Para evitar los problemas que justifican la definición 8.2.1 resulta obvio utilizar el coeficiente ajustado cuando hay muchas variables en juego. El objetivo es reconocer el modelo con mayor coeficiente. Sin embargo, si el número de variables es considerable esta técnica puede tener dificultades de cálculo.

9.3.2. Criterio C_P de Mallows

Con este criterio se debe fijar en primera instancia un número P de parámetros, incluido el término independiente, aunque con posterioridad se podrá variar. Se trata de hallar el mejor modelo con P

variables explicativas, incluida la constante, utilizando el estadístico de Mallows

$$C_P = \frac{\text{SCR}_P}{\hat{\sigma}^2} - (n - 2P)$$

donde SCR_P es la suma de cuadrados residual del modelo particular y $\hat{\sigma}^2$ un estimador de la varianza del modelo que acostumbra a ser el ECM del modelo completo.
Para el modelo completo $P = k + 1$, el estadístico de Mallows es

$$C_{k+1} = \frac{\text{SCR}}{\text{ECM}} - (n - 2(k + 1)) = n - (k + 1) - (n - 2(k + 1)) = k + 1$$

También para todo modelo no completo se puede demostrar que, si el modelo es adecuado, aproximadamente $E(C_P) = P$. En consecuencia parece recomendable elegir los conjuntos para los que C_P sea aproximadamente P.

9.3.3. Selección paso a paso

El procedimiento se puede realizar hacia adelante (forward stepwise) o hacia atrás (backward stepwise), seleccionando las variables una a una e incorporándolas desde el modelo inicial o eliminándolas desde el modelo completo en función de su contribución al modelo. Aunque es el método más utilizado por su facilidad de computación, este sistema tiene el inconveniente de que puede conducir a modelos distintos y no necesariamente óptimos.
En la selección hacia adelante se incorpora como primera variable la de mayor F de significación de la regresión simple. La segunda variable se selecciona por su mayor contribución al modelo que ya contiene la primera variable del paso anterior y así sucesivamente.

9.4. Ejemplos con R

Con los datos de tráfico de la sección 1.2 se calcula la regresión como se explica en la sección 6.9 mediante la instrucción

```
>
recta<-lm(rvel~dens)
```

Para el análisis de los residuos, la función `summary` nos ofrece un resumen de cinco números

```
 Call: lm(formula = rvel ~ dens)
 Residuals:
     Min       1Q   Median       3Q      Max
 -0.3534 -0.2272 -0.03566 0.1894 0.5335
```

También podemos obtener algunos gráficos univariantes como los de la figura 9.5 con las siguientes instrucciones:

```
> par(mfrow=c(1,2))
> par(pty="s")
> hist(residuals(recta),xlab="residuos")
> title("a) Histograma")
> boxplot(residuals(recta))
> title("b) Diagrama de caja")
> stem(residuals(recta))

N = 24   Median = -0.0356607
Quartiles = -0.228869, 0.1987335
```

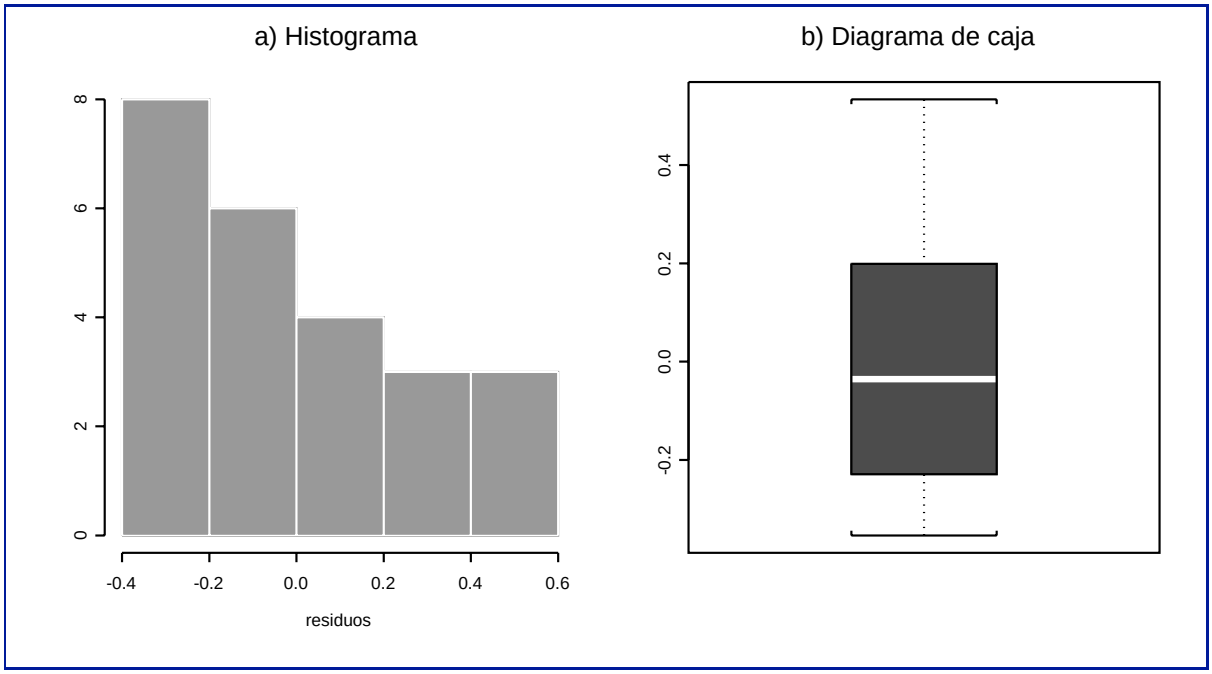


Figura 9.5: Gráficos de los residuos de la regresión simple del ejemplo de la sección 1.2.

```
Decimal point is 1 place to the left of the colon

  -3 : 510
  -2 : 44332
  -1 : 711
  -0 : 611
   0 : 3
   1 : 028
   2 : 245
   3 :
   4 : 0
   5 : 33
```

Para obtener los gráficos de la figura 9.3 se requieren las siguientes instrucciones:

```
> par(mfrow=c(2,2))
> plot(residuals(recta),xlab="indice",ylab="residuos")
> title("a) Residuos vs. indice")
> plot(fitted(recta),residuals(recta),xlab="ajustados",ylab="residuos")
> title("b) Residuos vs. ajustados")
> plot(fitted(recta),rvel,xlab="ajustados",ylab="observados")
> abline(0,1)
> title("c) Ajustados vs. observados")
> qqnorm(residuals(recta),xlab="Cuantiles de la normal",ylab="residuos")
> qqline(residuals(recta))
> title("d) QQ-plot")
```

R también permite obtener 6 gráficos para el análisis de un modelo de regresión lineal de una forma directa, mediante las instrucciones

```
> par(mfrow=c(2,3))
> plot(recta)
```

En cuanto a los contrastes de ajuste a la distribución normal, podemos optar entre el test de Kolmogorov-Smirnov `ks.gof` y la prueba ji-cuadrado `chisq.gof`. En nuestro caso:

```
> ks.gof(residuals(recta), distribution = "normal")

    One sample Kolmogorov-Smirnov Test of Composite Normality

data:  residuals(recta)
ks = 0.129, p-value = 0.5 alternative
hypothesis: True cdf is not the normal distn. with estimated parameters
sample estimates:
     mean of x standard deviation of x
 2.298509e-017             0.2630273
```

También se puede calcular la regresión con la instrucción

```
recta.ls<-lsfit(dens,rvel)
```

que nos proporciona muchos de los elementos para el diagnóstico en la forma:

```
> recta.diag<-ls.diag(recta.ls)
> recta.diag$hat # nivel
...
> recta.diag$std.res # residuos studentizados
...
> recta.diag$stud.res # residuos studentizados externamente
...
> recta.diag$cooks # distancias de Cook
...
> recta.diag$dfits # medidas Dffits
...
```

Los gráficos ...

```
> par(mfrow=c(1,2))
> par(pty="s")
> plot(recta.diag$hat,type="h",xlab="dato",ylab="h_ii")
> title("a) Niveles de los datos")
> plot(recta.diag$cooks,type="h",xlab="dato",ylab="C_i")
> title("b) Distancias de Cook")

> par(mfrow=c(1,2))
> par(pty="s")
> plot(recta.diag$std.res,xlab="dato",ylab="r_i",ylim=c(-2.5,2.5))
> title("a) Residuos studentizados \n internamente")
> plot(recta.diag$stud.res,xlab="dato",ylab="t_i",ylim=c(-2.5,2.5))
> title("b) Residuos studentizados \n externamente")
```

9.5. Ejercicios

Ejercicio 9.1

Realizar el análisis completo de los residuos del modelo de regresión parabólico propuesto en la sección 1.2 con los datos de tráfico.

Ejercicio 9.2

Realizar el análisis completo de los residuos de los modelos de regresión simple y parabólico propuestos en la sección 1.2 con los datos de tráfico, pero tomando como variable respuesta la velocidad (sin raíz cuadrada). Este análisis debe justificar la utilización de la raíz cuadrada de la velocidad como variable dependiente.

Ejercicio 9.3

Probar la relación 9.7 a partir de las ecuaciones 9.3 y 9.5.

Ejercicio 9.4

Se define el coeficiente de robustez como

$$B^2 = \frac{\text{SCR}}{\text{PRESS}}$$

donde PRESS es la suma de cuadrados 9.4. Este coeficiente está entre 0 y 1 y representa una medida de la robustez del modelo.

Calcular el coeficiente de robustez para los cinco conjuntos de datos de la sección 6.8.