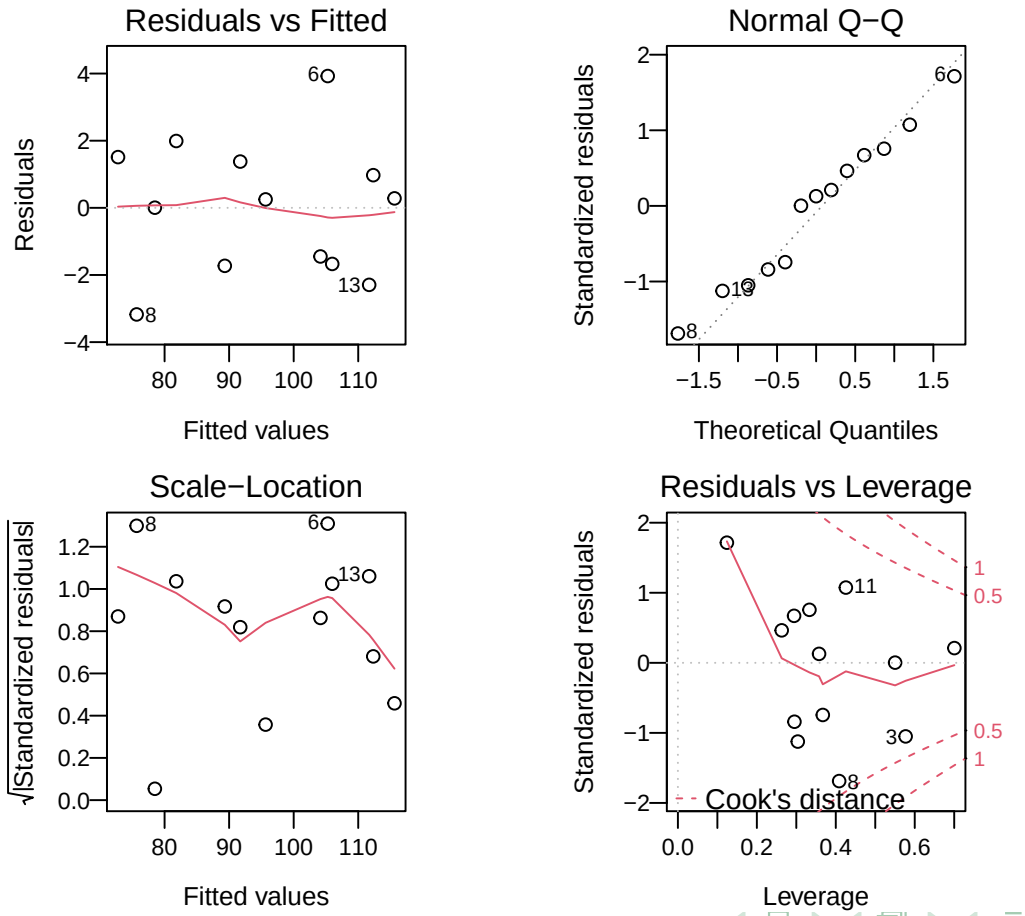
6 Datos influyentes y/o atípicos



Contenido

1. **Introducción**
   - Ejemplo
   - Datos influyentes y atípicos
2. **Datos influyentes y atípicos**
   - Matriz H
   - Definición de residuales
   - Distancia de Cook

Ejemplo: Producción de ácido[1]

i	x_1	x_2	x_3	y	$y - \hat{y}$
1	80	27	89	42	3.235
2	80	27	88	37	−1.917
3	75	25	90	37	4.556
4	62	24	87	28	5.698
5	62	22	87	18	−1.712
6	62	23	87	18	−3.007
7	62	24	93	19	−2.389
8	62	24	93	20	−1.389
9	58	23	87	15	−3.144
10	58	18	80	14	1.267
11	58	18	89	14	2.636
12	58	17	88	13	2.779
13	58	18	82	11	−1.429
14	58	19	93	12	−0.050
15	50	18	89	8	2.361
16	50	18	86	7	0.905
17	50	19	72	8	−1.520
18	50	19	79	8	−0.455
19	50	20	80	9	−0.598
20	56	20	82	15	1.412
21	70	20	91	15	−7.238

El ejemplo son datos de 21 días de operación de una planta de oxidación de amonio en la producción de ácido nítrico.

variable	concepto
i	día
x_1	flujo de aire
x_2	temperatura del agua
x_3	concentración del ácido
y	amonio no convertido

Ajuste:

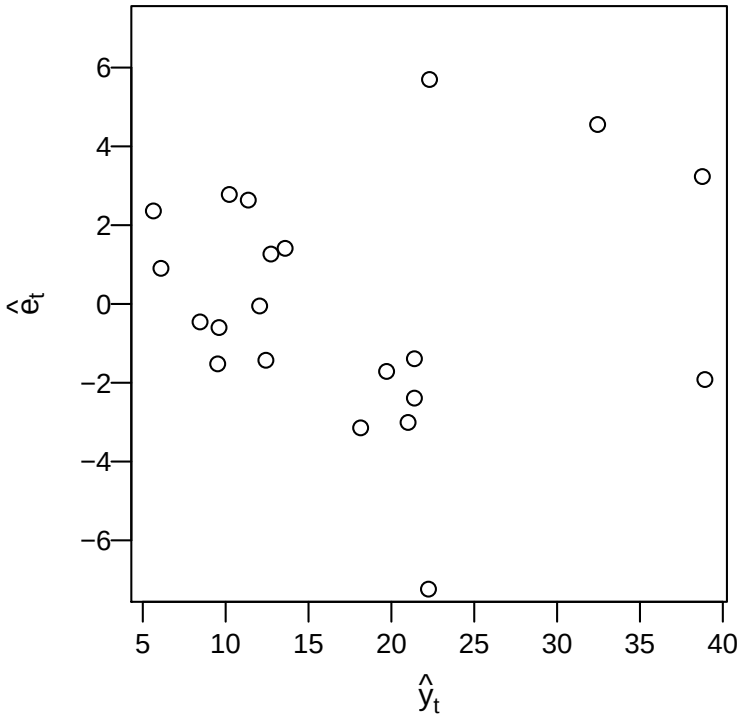
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.9197    11.8960  -3.356  0.00375
x1            0.7156     0.1349   5.307  5.8e-05
x2            1.2953     0.3680   3.520  0.00263
x3           -0.1521     0.1563  -0.973  0.34405

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared: 0.9136,     Adjusted R-squared: 0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

[1] Atkinson (1985).

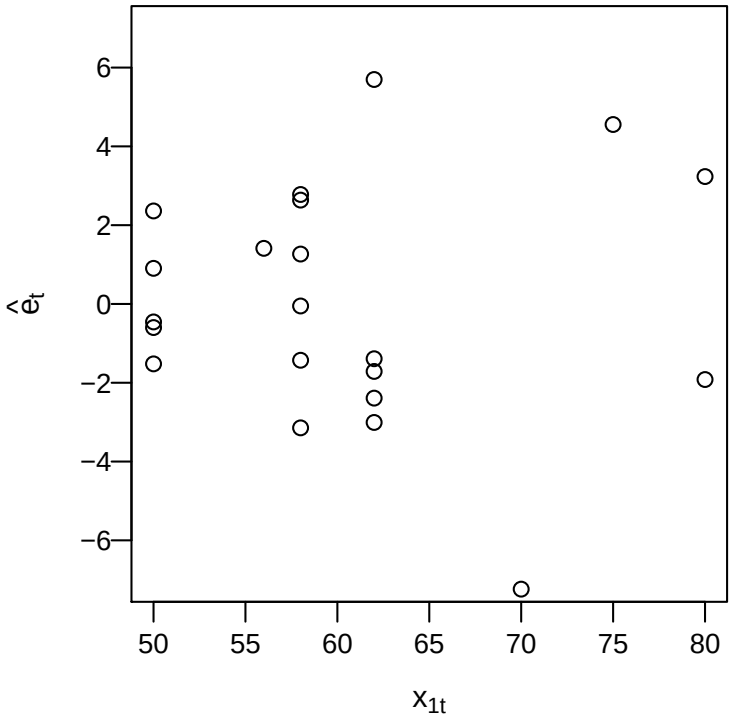
Ejemplo: Producción de ácido (cont.)

Análisis de residuales

a) Residuales vs. respuesta ajustada



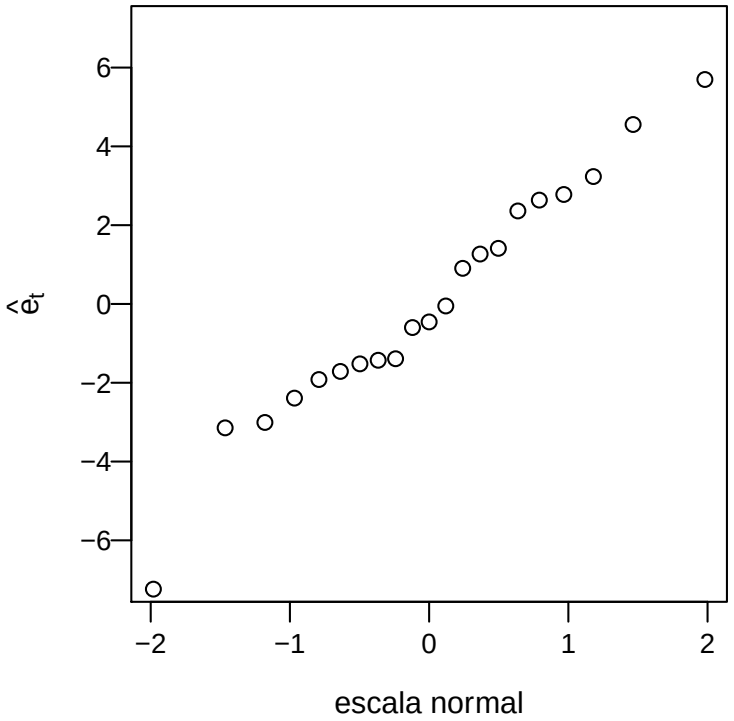
b) Residuales vs. regresor



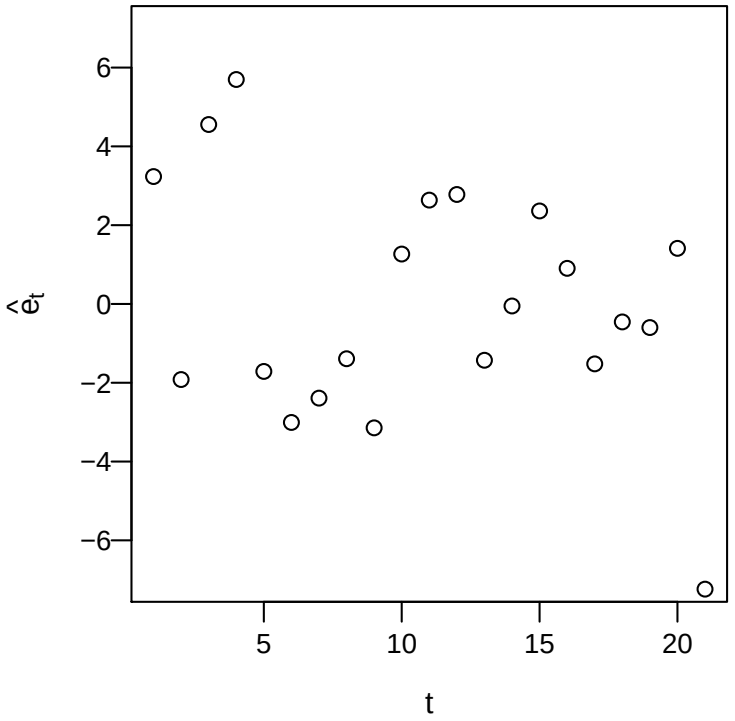
Ejemplo: Producción de ácido (cont.)

Análisis de residuales

c) Residuales en probabilidad normal



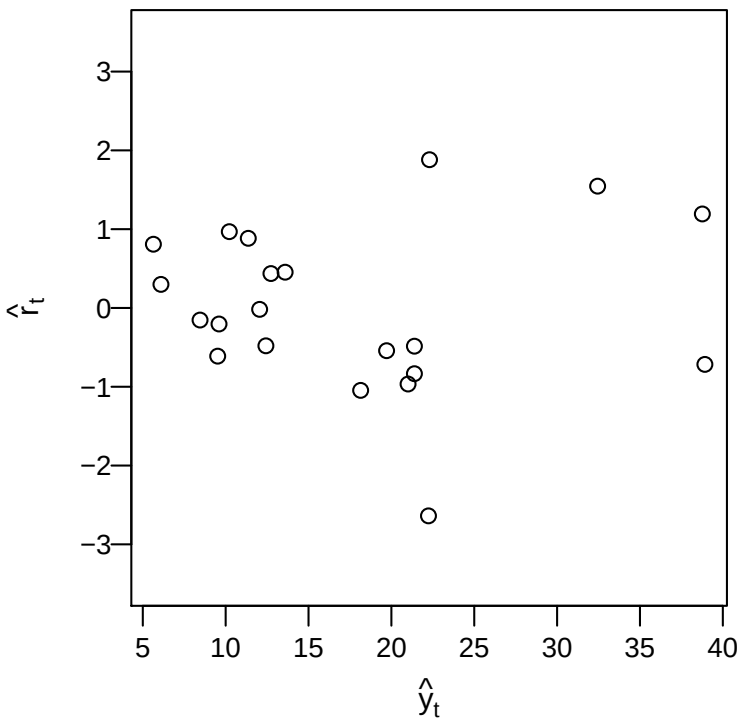
d) Residuales en orden temporal



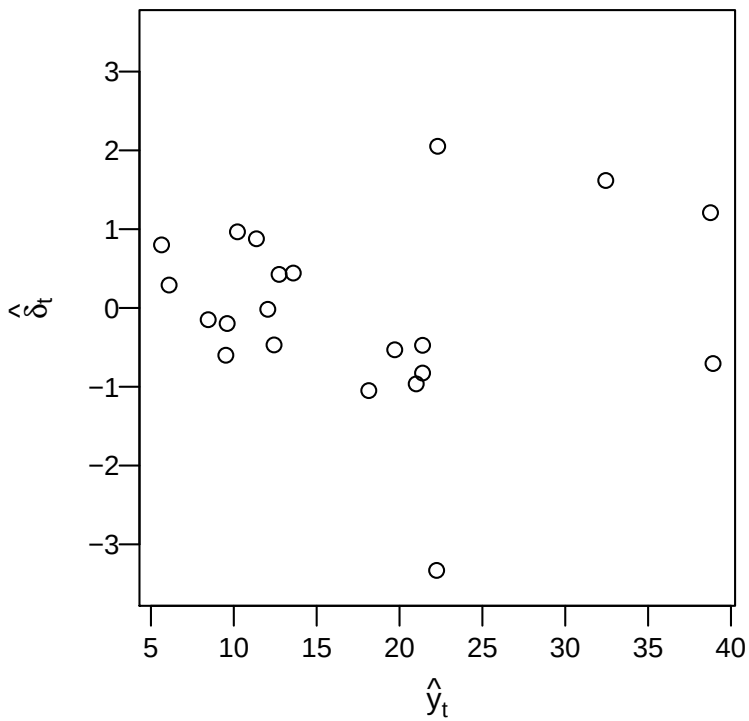
Ejemplo: Producción de ácido (cont.)

Análisis de residuales

e) Residuales estandarizados



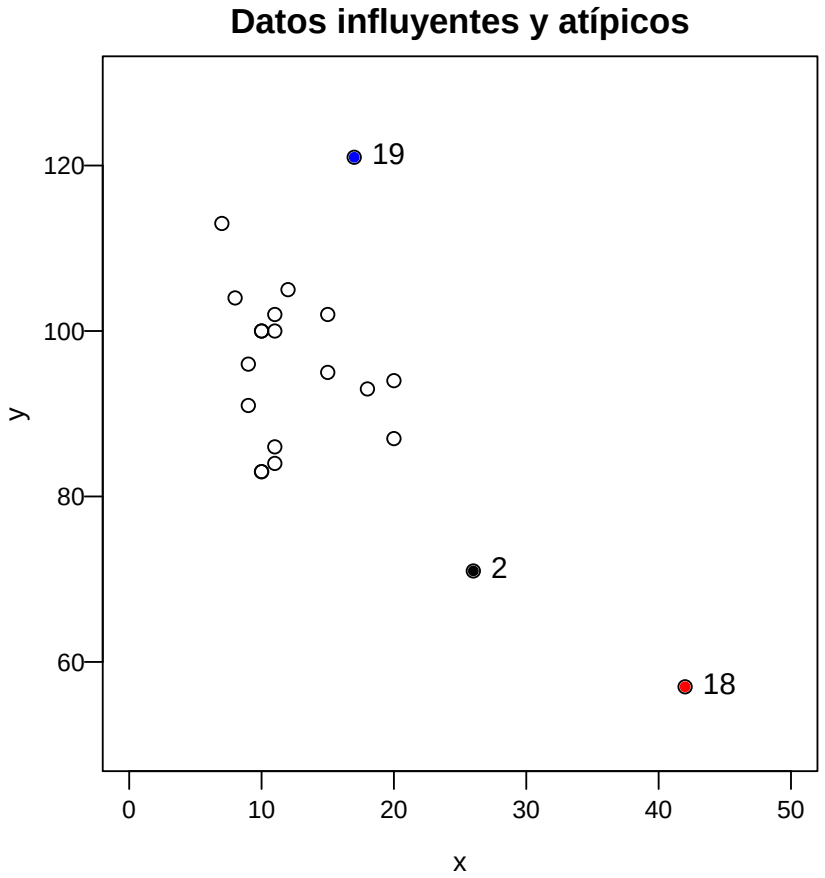
f) Residuales studentizados



Ejemplo: Score de aptitud de Gesell[2]

n: Observación
x: Edad primera palabra (meses)
y: Score de aptitud de Gesell

n	*x*	*y*	*n*	*x*	*y*
1	15	95	11	7	113
2	26	71	12	9	96
3	10	83	13	10	83
4	9	91	14	11	84
5	15	102	15	11	102
6	20	87	16	10	100
7	18	93	17	12	105
8	11	100	18	42	57
9	8	104	19	17	121
10	20	94	20	11	86
			21	10	100



[2] Draper and Smith (1998), p 210.

Considere el modelo

$$y_i = x_i'\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i; \qquad i = 1, \ldots, n$$

Si el ajuste de *mínimos cuadrados* se denota por $\hat{y}_i = x_i'\hat{\beta}$, el correspondiente *residual* está dado por $\hat{e}_i = y_i - \hat{y}_i = y_i - x_i'\hat{\beta}$. Se tiene entonces que

$$\mathbb{E}[\hat{e}_i] = 0; \qquad \text{var}(\hat{e}_i) = \sigma^2(1 - h_{ii}); \qquad i = 1, \ldots, n$$

donde $(h_{ij}) = X(X'X)^{-1}X' = H$ se le conoce como la *matriz "gorro"*.

- h_{ii} se conoce como el *apalancamiento* de la i-ésima observación.

Criterio

Valores con *apalancamiento* "grandes" ($|h_{ii}| > 2p/n$) indica una observación potencialmente influyente.

Asimismo, observaciones con *apalancamiento* "grandes" indican dato atípico, pero no viceversa. (Pues éstos tendrán un valor alejado en el espacio de las x y los datos atípicos tienen un valor alejado en las x, y o en ambas).

−ebz

La matriz H

Propiedades:

- H es simétrica ($H' = H$) e idempotente ($H^2 = H$).
- rango$(H) = q$.
- $\hat{e} = (y - \hat{y}) = (I - H)y = My$.
- $M = I - H$ es simétrica e idempotente.
- cov$(\hat{e}) = \sigma^2 M = \sigma^2(I - H)$.
- corr$(\hat{e}_i, \hat{e}_j) = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}, \quad i \neq j.$

  Esto es, la correlación de los residuales depende exclusivamente de los regresores.

Definición de residuales

⋆ Se definen los *residuales estandarizados* a los residuales divididos por la estimación de la desviación estándar del error: $\hat{r}_i = \hat{e}_i / s$. Luego,

$$\mathbb{E}[\hat{r}_i] = 0, \quad \text{var}(\hat{r}_i) = \frac{1}{s^2}\text{var}(\hat{e}_i) = \frac{1}{s^2}\,\sigma^2(1 - h_{ii}) \approx 1$$

Criterio

Residuales estandarizados *"grandes"* ($|\hat{r}_i| > 3$) indica un potencial valor atípico.

⋆ Se define el i-ésimo *residual studentizado (internamente)* como el residual dividido por la estimación de la desviación estándar del mismo. A saber,

$$\hat{\delta}_i = \frac{\hat{e}_i}{\sqrt{\widehat{\text{var}(\hat{e}_i)}}} = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

Entonces,

$$\mathbb{E}[\hat{\delta}_i] = 0, \quad \widehat{\text{var}(\hat{\delta}_i)} = \frac{1}{(s\sqrt{1 - h_{ii}})^2}\widehat{\text{var}(\hat{e}_i)} = \frac{1}{s^2(1 - h_{ii})}\,s^2(1 - h_{ii}) = 1$$

–ebz

Definición de residuales

$\star$ Denotamos por $s^2_{(i)}$, la estimación de la varianza σ^2 (constante) de los errores *sin* considerar la i-ésima observación. Se puede mostrar que

$$s^2_{(i)} = \frac{(n-q)s^2 - \hat{e}_i^2/(1-h_{ii})}{n-1-q}$$

es una estimación insesgada de σ^2, sin considerar la i-ésima observación.
$\star$ Se define el i-ésimo *residual studentizado (externamente)* al residual dividido por la estimación de la desviación estándar del residual sin incluir precisamente la i-ésima observación. A saber,

$$\hat{t}_i = \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}} \sim t_{(n-1-q)}$$

bajo la hipótesis esférica de los errores ($\epsilon \sim \mathrm{N}(0, \sigma^2)$).

Criterio

Si $\hat{t}_i$ es grande, llama la atención pues ϵ_i no fue considerado en el ajuste.

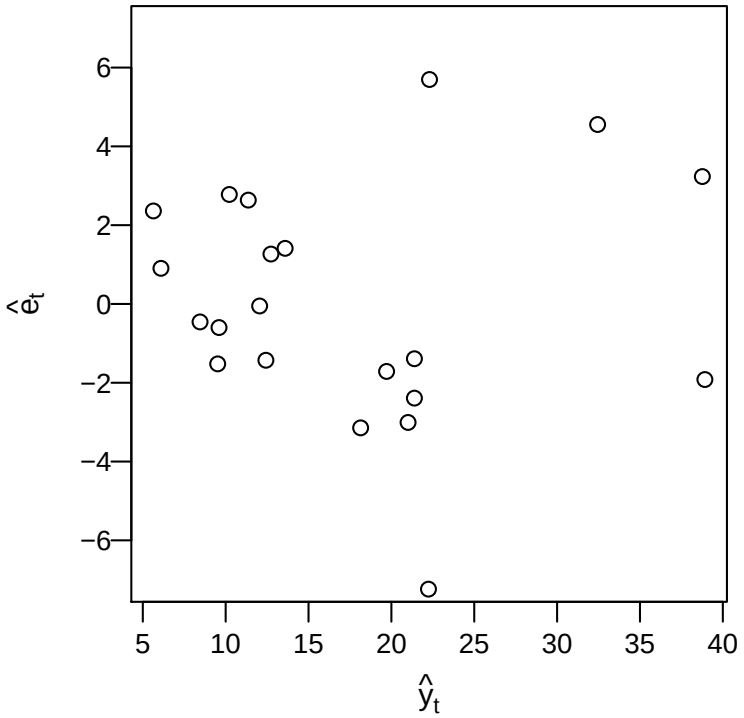
$n < 30$, se recomienda hacer análisis con los residuales studentizados; en otro caso, con los estandarizados.

−ebz

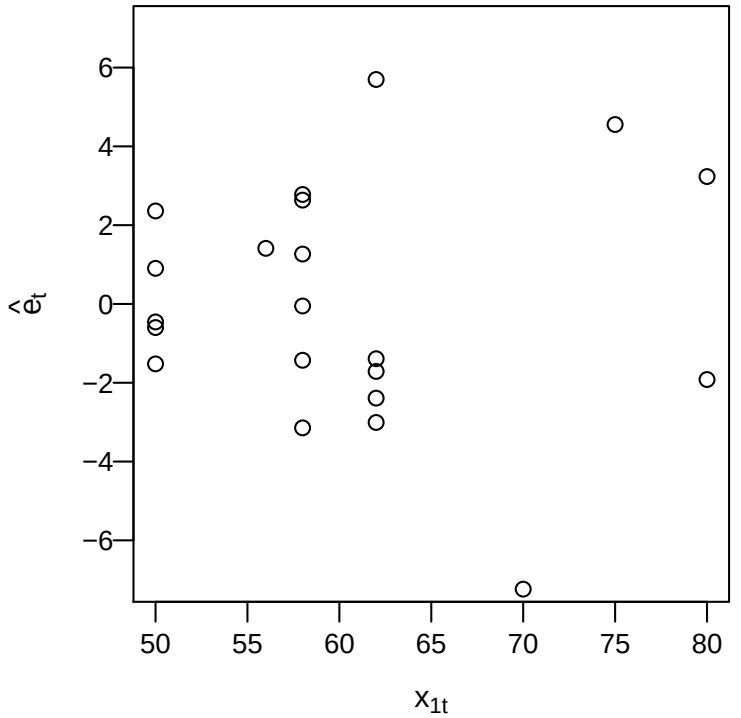
Ejemplo: Producción de ácido (cont.)

Análisis de residuales

a) Residuales vs. respuesta ajustada



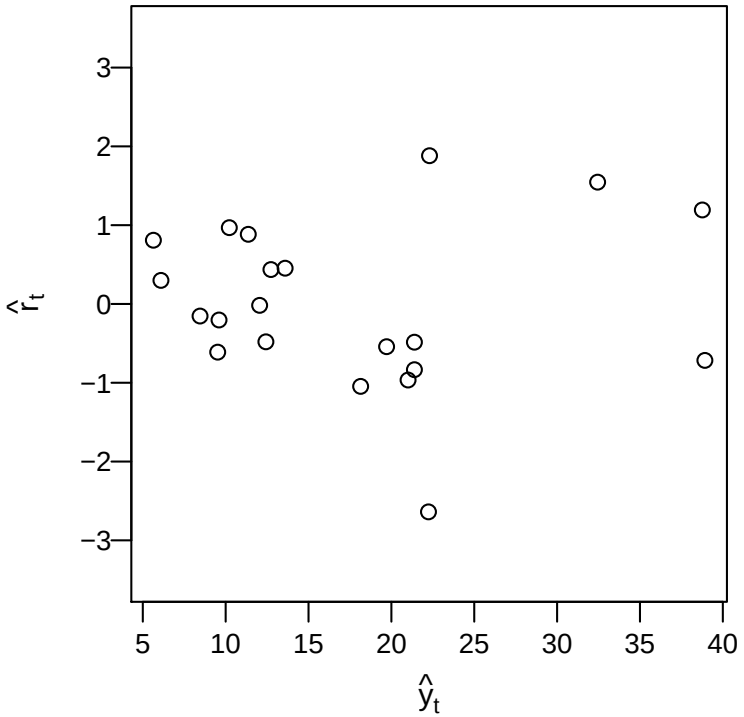
b) Residuales vs. regresor



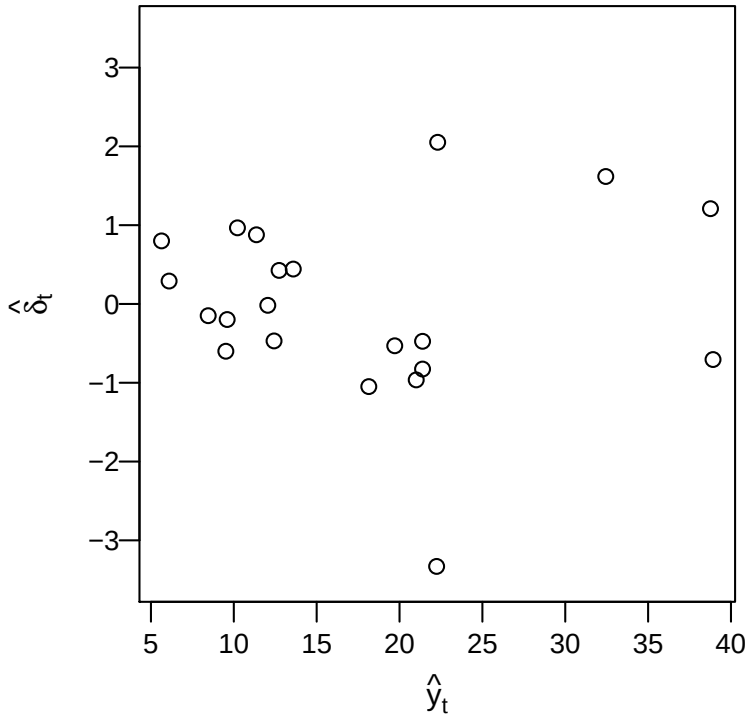
Ejemplo: Producción de ácido (cont.)

Análisis de residuales

e) Residuales estandarizados



f) Residuales studentizados



Residuales y Datos Influyentes

En cualquier juego de datos cuando la estimación de uno o más parámetros dependen "mucho" de unos "pocos" datos, éstos se llaman *datos influyentes* y son síntoma de problemas potenciales. Si es el caso, las conclusiones son muy *sensibles* y posiblemente se necesite de más información. Una medida de la influencia de los datos es la *distancia de Cook*:

$$d_i = \frac{||\hat{y} - \hat{y}_{(i)}||^2}{qs^2}, \quad i = 1, \ldots, n$$

donde $\hat{y}_{(i)}$ es el valor estimado del vector respuesta sin considerar la i-ésima observación en el ajuste.

Si la i-ésima observación no es muy importante, se espera que d_i sea pequeño. Valores grandes de d_i son motivo de cuidado.

$$d_i = \left[\frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}\right]^2 \left[\frac{h_{ii}}{1-h_{ii}}\right] \frac{1}{q} = \left[\hat{\delta}_i\right]^2 \left[\frac{\text{var}(\hat{y}_i)}{\text{var}(\hat{e}_i)}\right] \frac{1}{q}$$

Criterio

Si $d_i > 1$ es posible que la i-ésima observación sea de influencia.

Notas a la distancia de Cook

La distancia de Cook tiene dos componentes:

- El primero, refleja qué tan bien se ajusta el modelo a la i-ésima observación.
- El segundo, nos indica qué tan lejos se encuentra el i-ésimo punto del resto de los datos.

Cada uno de los componentes anteriores (o ambos) contribuyen a una distancia de Cook “grande”.

Extensiones a la distancia de Cook

Belsley, Kuh, and Welsch (1980), proponen varias medidas para estudiar la influencia de la i-ésima observación sobre las estimaciones y ajustes. A saber:

1. *DFBETAS*: El cambio del coeficiente $\hat{\beta}_j$ en unidades de desviaciones estándar. (Qué tanto cambia el j-ésimo coeficiente si se realiza el ajuste sin la i-ésima observación).

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{s_{(i)}\sqrt{C_{jj}}}, \quad j = 1, \ldots, p; \; i = 1, \ldots, n$$

donde C_{jj} es el j-ésimo elemento de la diagonal de la matriz $(X'X)^{-1}$.

Criterio

Si $|DFBETAS_{j,i}| > 2/\sqrt{n}$, entonces la i-ésima observación debe ser examinada.

Extensiones a la distancia de Cook

2. *DFFITS*: Dice cuánto cambia $\hat{y}_i$, en desviaciones estándar, si no se considera la i-ésima observación.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}}, \quad i = 1, \ldots, n$$

Criterio

Si $|DFFITS_i| > 2\sqrt{q/n}$, entonces la i-ésima observación debe ser examinada. (Se considera influyente.)

La distancia de Cook, *DFFITS* y *DFBETAS* no dan una medida en la precisión del modelo.

Extensiones a la distancia de Cook

3. *COVRATIO*: Dice cuánto cambia la precisión de las estimaciones si no se considera la i-ésima observación.

$$COVRATIO_i = \frac{|(X'_{(i)}X_{(i)})^{-1}s^2_{(i)}|}{|(X'X)^{-1}s^2|} = \left[\frac{s^2_{(i)}}{s^2}\right]^q \left(\frac{1}{1-h_{ii}}\right), \quad i = 1, \ldots, n$$

Criterio

Si $COVRATIO_i > 1 + 3q/n$, o bien, $COVRATIO_i < 1 - 3q/n$ $(n > 3q)$ entonces la i-ésima observación debe ser examinada.

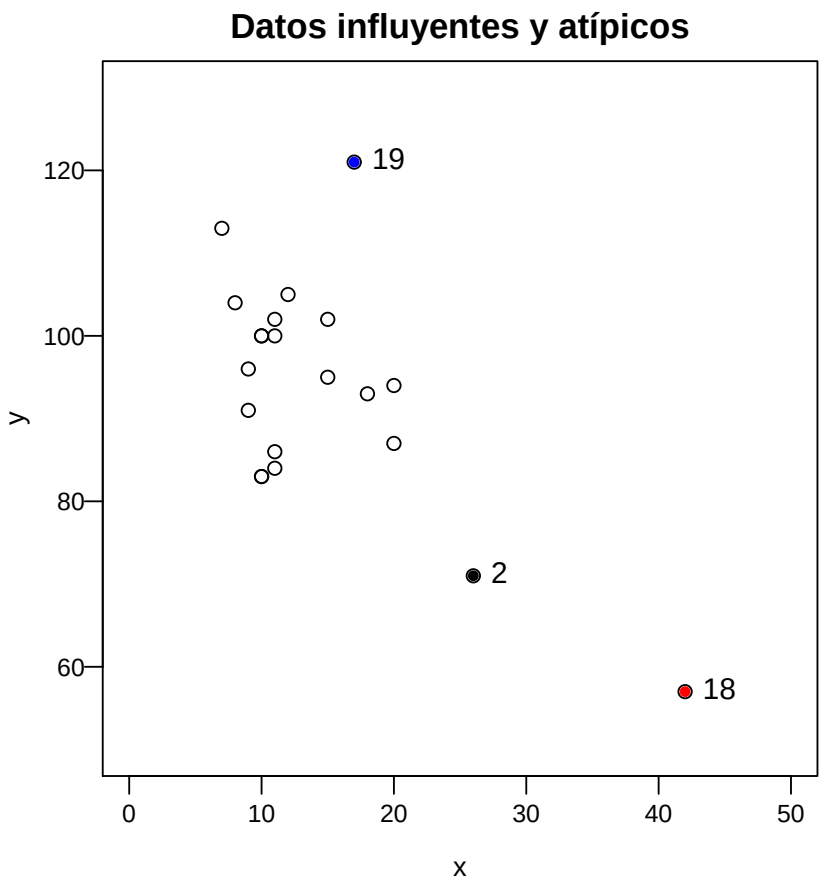
De acuerdo a Montgomery et al. (2001),

- Si $COVRATIO_i < 1$, indica que incluir la i-ésima observación en el modelo *mejora* la precisión en la estimación.
- Si $COVRATIO_i > 1$, indica que incluir la i-ésima observación en el modelo *empeora* la precisión en la estimación.

Ejemplo: Score de aptitud de Gesell (cont.)

n: Observación
x: Edad primera palabra (meses)
y: Score de aptitud de Gesell

n	x	y	n	x	y
1	15	95	11	7	113
2	26	71	12	9	96
3	10	83	13	10	83
4	9	91	14	11	84
5	15	102	15	11	102
6	20	87	16	10	100
7	18	93	17	12	105
8	11	100	18	42	57
9	8	104	19	17	121
10	20	94	20	11	86
			21	10	100

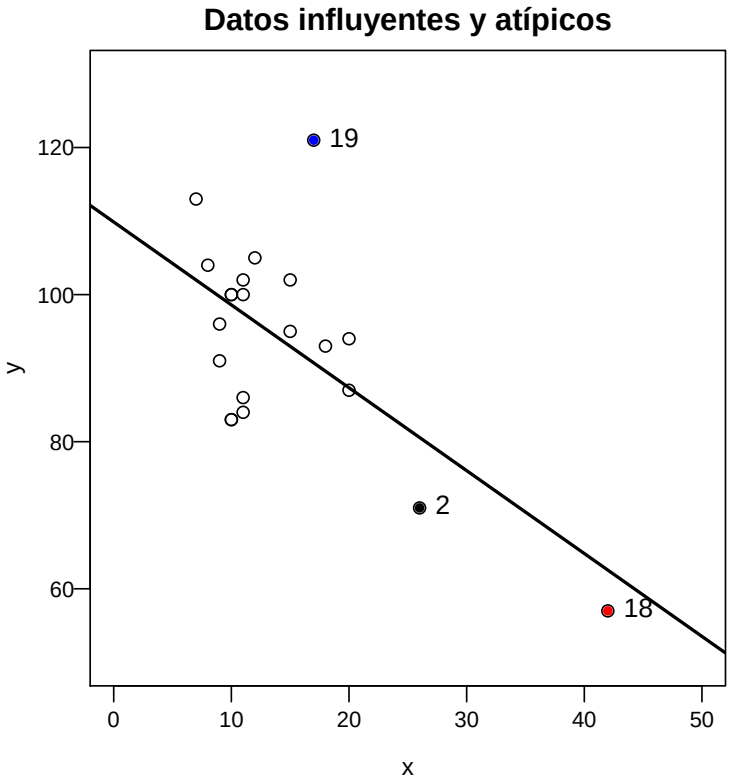


Ejemplo: Score de Aptitud de Gesell (cont.)

Ajuste

Datos completos

```
lm(formula = y ~ x, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.8738     5.0678  21.681 7.31e-15
x            -1.1270     0.3102  -3.633  0.00177

Residual standard error: 11.02 on 19 degrees of freedom
Multiple R-squared:  0.41,      Adjusted R-squared: 0.3789
F-statistic:  13.2 on 1 and 19 DF,  p-value: 0.001769
```



Ejemplo: Score de Aptitud de Gesell (cont.)

Estadístico de Cook y Residuales
Datos completos

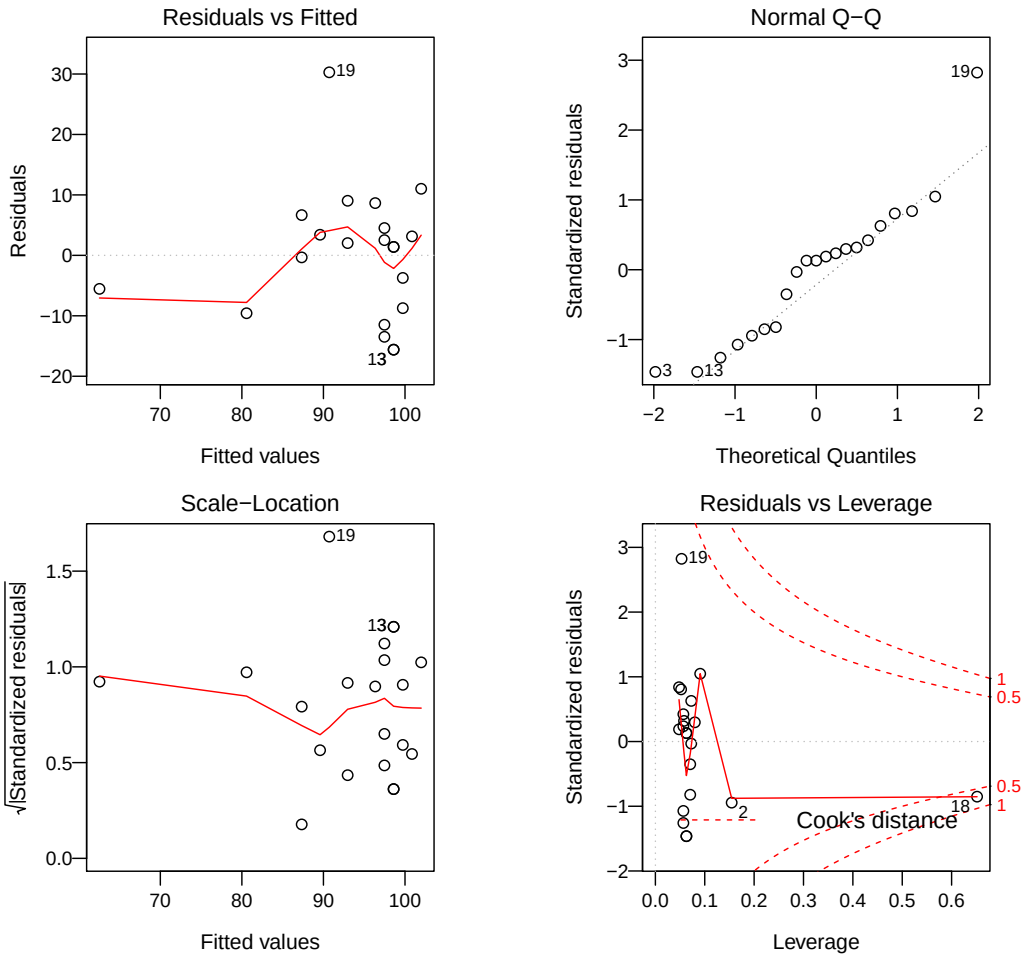
```
> print(influence.measures(lm(y ~ x, data=dat))
Influence measures of lm(formula = y ~ x, data = dat) :

     dfb.1_    dfb.x    dffit cov.r   cook.d    hat inf
1   0.01664  0.00328  0.04127 1.166 8.97e-04 0.0479
2   0.18862 -0.33480 -0.40252 1.197 8.15e-02 0.1545
3  -0.33098  0.19239 -0.39114 0.936 7.17e-02 0.0628
4  -0.20004  0.12788 -0.22433 1.115 2.56e-02 0.0705
5   0.07532  0.01487  0.18686 1.085 1.77e-02 0.0479
6   0.00113 -0.00503 -0.00857 1.201 3.88e-05 0.0726
7   0.00447  0.03266  0.07722 1.170 3.13e-03 0.0580
8   0.04430 -0.02250  0.05630 1.174 1.67e-03 0.0567
9   0.07907 -0.05427  0.08541 1.200 3.83e-03 0.0799
10 -0.02283  0.10141  0.17284 1.152 1.54e-02 0.0726
11  0.31560 -0.22889  0.33200 1.088 5.48e-02 0.0908
12 -0.08422  0.05384 -0.09445 1.183 4.68e-03 0.0705
13 -0.33098  0.19239 -0.39114 0.936 7.17e-02 0.0628
14 -0.24681  0.12536 -0.31367 0.992 4.76e-02 0.0567
15  0.07968 -0.04047  0.10126 1.159 5.36e-03 0.0567
16  0.02791 -0.01622  0.03298 1.187 5.74e-04 0.0628
17  0.13328 -0.05493  0.18717 1.096 1.79e-02 0.0521
18  0.83112 -1.11275 -1.15578 2.959 6.78e-01 0.6516   *
19  0.14348  0.27317  0.85374 0.396 2.23e-01 0.0531   *
20 -0.20761  0.10544 -0.26385 1.043 3.45e-02 0.0567
21  0.02791 -0.01622  0.03298 1.187 5.74e-04 0.0628
```

–ebz

Ejemplo: Score de Aptitud de Gesell (cont.)

Análisis de residuales

Datos completos

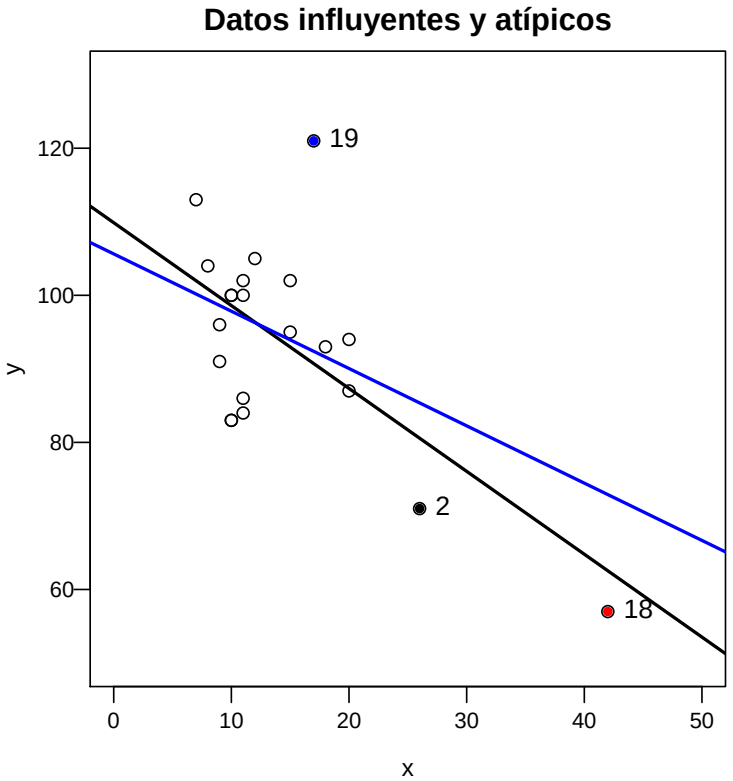


Ejemplo: Score de Aptitud de Gesell (cont.)

Ajuste

Datos incompletos

```
lm(formula = y ~ x, data = dat[-18, ])

Residuals:
    Min      1Q  Median      3Q     Max
-14.838  -8.477   1.779   4.688  28.617

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 105.6299     7.1619  14.749 1.71e-11
x            -0.7792     0.5167  -1.508    0.149

Residual standard error: 11.11 on 18 degrees of freedom
Multiple R-squared: 0.1122,     Adjusted R-squared: 0.06284
F-statistic: 2.274 on 1 and 18 DF,  p-value: 0.1489
```



Ejemplo: Score de Aptitud de Gesell (cont.)

Estadístico de Cook y Residuales
Datos incompletos

```
> print(influence.measures(lm(y ~ x, data=dat[-18,]))
Influence measures of lm(formula = y ~ x, data = dat[-18, ]) :
      dfb.1_    dfb.x   dffit cov.r   cook.d    hat inf
1  -0.000958  0.00916  0.0238 1.190 0.000301 0.0587
2   1.149535 -1.41963 -1.5135 1.357 1.019719 0.4158   *
3  -0.307690  0.20602 -0.3891 0.962 0.071604 0.0695
4  -0.186220  0.13747 -0.2149 1.156 0.023752 0.0846
5  -0.007407  0.07081  0.1843 1.118 0.017423 0.0587
6   0.072145 -0.10309 -0.1250 1.315 0.008237 0.1561
7  -0.019187  0.03173  0.0440 1.249 0.001025 0.1041
8   0.045157 -0.02549  0.0664 1.181 0.002322 0.0587
9   0.133687 -0.10516  0.1459 1.225 0.011144 0.1041
10 -0.093841  0.13409  0.1626 1.306 0.013889 0.1561
11  0.456492 -0.37549  0.4811 1.076 0.112130 0.1279
12 -0.063165  0.04663 -0.0729 1.217 0.002804 0.0846
13 -0.307690  0.20602 -0.3891 0.962 0.071604 0.0695
14 -0.208747  0.11785 -0.3068 1.005 0.045752 0.0587
15  0.076149 -0.04299  0.1119 1.163 0.006552 0.0587
16  0.042432 -0.02841  0.0537 1.199 0.001521 0.0695
17  0.099366 -0.03815  0.1873 1.099 0.017898 0.0522
19 -0.343415  0.65879  1.0298 0.437 0.335261 0.0846   *
20 -0.174596  0.09857 -0.2566 1.056 0.032811 0.0587
21  0.042432 -0.02841  0.0537 1.199 0.001521 0.0695
```

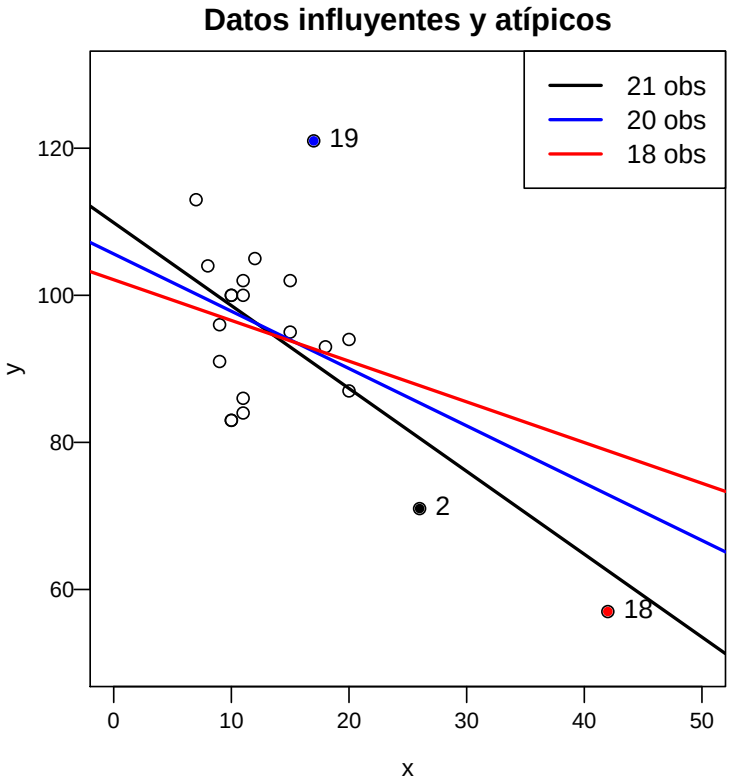
−ebz

Ejemplo: Score de Aptitud de Gesell (cont.)

Ajuste

Datos incompletos

```
lm(formula = y ~ x, data = dat[-c(2, 18, 19), ])

Residuals:
    Min      1Q  Median      3Q     Max
-13.582  -5.614   2.069   5.471  14.757

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.1177     6.6940  15.255 5.93e-11
x            -0.5535     0.5295  -1.045    0.311

Residual standard error: 8.553 on 16 degrees of freedom
Multiple R-squared: 0.06394,    Adjusted R-squared: 0.005439
F-statistic: 1.093 on 1 and 16 DF,  p-value: 0.3114
```



Ejemplo: Score de Aptitud de Gesell (cont.)

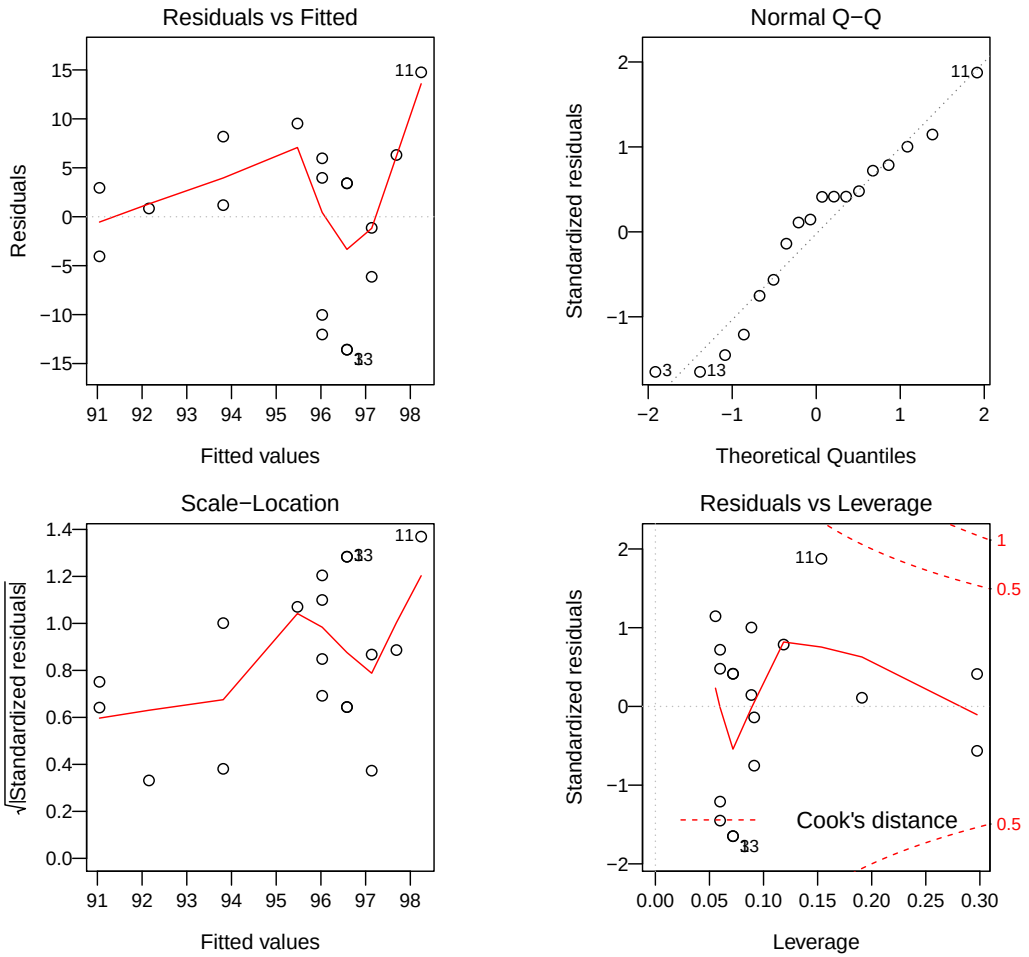
Estadístico de Cook y Residuales
Datos incompletos

```
> print(influence.measures(lm(y ~ x, data=dat[-c(2,18,19),]))
Influence measures of lm(formula = y ~ x, data = dat[-c(2, 18, 19), ]) :

    dfb.1_   dfb.x   dffit cov.r   cook.d    hat inf
1  -0.0152  0.0269  0.0439 1.245 0.001027 0.0888
3  -0.3496  0.2313 -0.4869 0.845 0.104990 0.0717
4  -0.1956  0.1472 -0.2352 1.165 0.028464 0.0913
5  -0.1080  0.1915  0.3130 1.097 0.048964 0.0888
6   0.2621 -0.3239 -0.3592 1.556 0.067444 0.2974   *
7  -0.0332  0.0436  0.0518 1.404 0.001428 0.1910   *
8   0.0642 -0.0315  0.1178 1.176 0.007295 0.0598
9   0.2566 -0.2075  0.2847 1.193 0.041547 0.1186
10 -0.1904  0.2353  0.2609 1.585 0.035916 0.2974   *
11  0.8254 -0.6992  0.8754 0.818 0.318850 0.1535
12 -0.0356  0.0268 -0.0428 1.249 0.000975 0.0913
13 -0.3496  0.2313 -0.4869 0.845 0.104990 0.0717
14 -0.2072  0.1016 -0.3801 0.913 0.066932 0.0598
15  0.0974 -0.0478  0.1788 1.133 0.016494 0.0598
16  0.0806 -0.0533  0.1122 1.200 0.006648 0.0717
17  0.0885 -0.0041  0.2809 1.015 0.038629 0.0556
20 -0.1689  0.0828 -0.3099 0.999 0.046525 0.0598
21  0.0806 -0.0533  0.1122 1.200 0.006648 0.0717
```

Ejemplo: Score de aptitud de Gesell (cont.)

Análisis de residuales

Datos incompletos



Residuales y datos influyentes

- *¿Qué hacer con las observaciones influyentes?*
  El análisis de residuales de apalancamiento de los datos ayuda a conocerlos mejor y a saber qué observaciones hay que examinar más detenidamente.
- *¿Se deben eliminar?*
  De acuerdo a Montgomery (2001), si se sabe que existió un error de captura o que la observación no es parte de la población que se pretende muestrear, sí se debe eliminar. Sin embargo, si es una observación válida no hay justificación para eliminarla.

Referencias

Atkinson, A. C. (1985).
Plots, Transformations, and Regression.
Oxford: Clarendon Press.

Belsley, D. A., E. Kuh, and R. E. Welsch (1980).
Regression Diagnostics.
New York: Wiley.

Draper, N. and H. Smith (1998).
Applied Regression Analysis (3rd ed.).
New York: Wiley.