MAT-266: Análisis de residuos y leverages

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

EX UMBRA IN SOLEM

Suponga el modelo lineal,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

con los Supuestos A1-A4. El vector de residuos es dado por:

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y},$$

con $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, es decir $\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} = \widehat{\mathsf{E}}(\boldsymbol{Y})$.

Bajo el supuesto de normalidad $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, tenemos

$$\boldsymbol{e} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})),$$

es decir

$$\mathsf{E}(e_i) = 0, \quad \mathsf{var}(e_i) = \sigma^2(1 - h_{ii}), \quad \mathsf{Cov}(e_i, e_j) = -\sigma^2 h_{ij}.$$

De ahí que los residuos tienen varianzas diferentes y son correlacionados.

Suponga que σ^2 es conocido, de este modo

$$z_i = \frac{e_i}{\sigma} \sim \mathsf{N}(0,1).$$

Mientras que, el residuo estandarizado es definido como:

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}, \qquad i = 1,\dots,n.$$

Cook y Weisberg (1982)[1] mostraron que

$$\frac{r_i^2}{n-p} \sim \mathsf{Beta}\Big(\frac{1}{2}, \frac{n-p-1}{2}\Big),$$

de este modo

$$\mathsf{E}(r_i) = 0, \quad \mathsf{var}(r_i) = 1, \quad \mathsf{Cov}(r_i, r_j) = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}.$$

[1] Residual and Influence in Regression, Chapman & Hall

Residuos en regresión lineal

Considere el residuo studentizado:

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}, \qquad i = 1, \dots, n.$$

donde

$$s_{(i)}^2 = \frac{1}{n - p - 1} \sum_{j \neq i}^{n} (y_j - \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_{(i)})^2,$$

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1} \boldsymbol{X}_{(i)}^\top \boldsymbol{Y}_{(i)},$$

denotan los estimadores de σ^2 y $\boldsymbol{\beta}$ una vez que la i-ésima observación ha sido eliminada. Es decir,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{(i)} \\ \boldsymbol{x}_i^\top \end{pmatrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{(i)} \\ Y_i \end{pmatrix}$$

EX UMBRA IN SOLEM

Una interpretación interesante de t_i es que corresponde al estadístico t para probar la hipótesis $H_0 : \gamma = 0$ en el modelo de salto en la media, dado por:

$$Y_j = \boldsymbol{x}_j^\top \boldsymbol{\beta} + d_j \gamma + \epsilon_j, \qquad j = 1, \dots, n,$$

donde $d_j = 1$ si $j = i$ y 0 en caso contrario.

El modelo puede ser escrito como:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{d}_i \gamma + \boldsymbol{\epsilon},$$

con $\boldsymbol{d}_i = (\boldsymbol{0}, 1, \boldsymbol{0})^\top$ con un cero en la i-ésima posición.

Lo anterior permite notar que $t_i \sim t(n-p-1)$, y de este modo,

$$\mathsf{E}(t_i) = 0, \qquad \mathsf{var}(t_i) = \frac{n-p-1}{n-p-3} \approx 1$$

QQ-plot en regresión lineal

Objetivo:

Evaluar desvios de normalidad de los residuos studentizados t_i's.[2]

Notación:

Considere $Z_i = t_i$, para $i = 1, \ldots, n$

Idea:

Comparar la CDF muestral para los Z_i's contra la CDF de la N$(0, 1)$.

Asuma que los residuos Z_i están ordenados

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)},$$

los $Z_{(i)}$ son los cuantiles de la CDF muestral, definida como

$$\text{Proportion}(Z \leq Z_{(i)}) = \frac{i}{n}$$

[2] La descripción es válida otras medidas de interés.

Asuma que los residuos Z_i están ordenados

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)},$$

los $Z_{(i)}$ son los cuantiles de la CDF muestral, definida como

$$\text{Proportion}(Z \leq Z_{(i)}) = \frac{i}{n}.$$

Los cuantiles de la distribución teórica, son dados por:

$$q_i^* = \Phi^{-1}\Big(\frac{i}{n}\Big).$$

Si los errores son aproximadamente normales, se debe tener que el gráfico de los pares $(q_1^*, Z_{(1)}), \dots, (q_n^*, Z_{(n)})$ sea a recta identidad.

Se ha sugerido la siguiente aproximación para la esperanza de los estadísticos de orden desde $\mathsf{N}(0,1)$ como:

$$q_i = \Phi^{-1}\Big(\frac{i - 3/8}{n + 1/4}\Big),$$

de este modo, se utilizará el gráfico cuantil-cuantil (QQ-plot) de los pares $(q_i, Z_{(i)})$.

Observación:

- Podemos construir QQ-plots para diversas distribuciones.[3]
- Es difícil chequear visualmente desvios de la distribución de interés.

[3] Por ejemplo, χ^2, t de Student, Poisson, Gama, etc.

QQ-plot con envelopes en regresión lineal

Envelopes simulados son herramientas gráficas para chequear el ajuste de un modelo. Atkinson (1985)[4] sugirió usar el siguiente procedimiento:

- Ajustar un modelo de regresión lineal, calcular residuos y estandarizar para obtener varianza unitaria.
- Generar M (≈ 1000) muestras como respuesta. Para cada muestra ajuste el mismo modelo y calcule los residuos estandarizados
- Ordenar todos los conjuntos de residuos estandarizados.
- El envelope consiste de los cuantiles 2.5% inferior y superior de los residuos estandarizados generados en cada posición.

[4] Plots, Transformations and Regressions, Oxford University Press.

QQ-plot con envelopes en regresión lineal[5]

```
envel.norm <- function(object, nsamples = 1000, alpha = 0.05) {
  x <- model.matrix(object)
  n <- nrow(x); p <- ncol(x)
  H <- x %*% solve(crossprod(x), t(x))
  ti <- rstudent(object)
  Id <- diag(n)
  epsilon <- matrix(0, n, nsamples)
  e <- matrix(0, n, nsamples)
  e1 <- e2 <- numeric(n)
  for (i in 1:nsamples) {
    epsilon[,i] <- rnorm(n)
    e[,i] <- (Id - H) %*% epsilon[,i]
    u <- diag(Id - H)
    e[,i] <- e[,i] / sqrt(u)
    e[,i] <- sort(e[,i])
  }
  for (i in 1:n) {
    eo <- quantile(e[i,], c(alpha / 2, 1 - alpha / 2))
    e1[i] <- eo[1]; e2[i] <- eo[2]
  }
  res <- structure(list(res = sort(ti), elim = cbind(e1,e2)),
                   label = deparse(object$call$formula))
   class(res) <- "envelope"
   res
}
```

[5] Versión preliminar, no recomendable para n "grande".

QQ-plot con envelopes en regresión lineal

```
envelope <- function(object, nsamples = 1000, alpha = 0.05) {
   n <- length(r <- resid(object))
   p <- length(coef(object))
   Y <- scale(qr.resid(qr(model.matrix(object)),
                matrix(rnorm(n * nsamples), n, nsamples)),
                F, T) * sqrt((n - 1) / (n - p))
   Y[,] <- Y[order(col(Y),Y)]
   Y <- matrix(Y[order(row(Y),Y)], nsamples, n)
   x0 <- quantile(1:nsamples, c(alpha / 2, 1 - alpha / 2))
   if (all(x0 %% 1 == 0)) elim <- t(Y[x0,])
   else {
      x1 <- c(floor(x0), ceiling(x0))
      elim <- cbind(Y[x1[1],] + (Y[x1[3],] - Y[x1[1],]) /
                (x1[3] - x1[1]) * (x0[1] - x1[1]),
                Y[x1[2],] + (Y[x1[4],] - Y[x1[2],]) /
                (x1[4] - x1[2]) * (x0[2] - x1[2]))
   }
   res <- sort(r)
   res <- res / sqrt(sum(res^2) / (n - p))
   res <- structure(list(res = res, elim = elim),
                    label = deparse(object$call$formula))
   class(res) <- "envelope"
   res
}
```

QQ-plot con envelopes en regresión lineal

```r
plot.envelope <- function(x, ...) {
   n <- length(x$res)
   ylim <- range(x$res[c(1,n)], x$elim[c(1,n),])
   nscores <- qnorm(ppoints(n))
   oldpar <- par(pty = "s"); on.exit(par(oldpar))
   plot(nscores, x$res, pch = 1, ylim = ylim,
        xlab = "Normal scores", ylab = "Sorted residuals",
        main = attr(x, "label"))
   lines(nscores, x$elim[,1]); lines(nscores, x$elim[,2])
   invisible(x)
}

print.envelope <- function(x, ...) {
   lo <- x$res < x$elim[,1]
   hi <- x$res > x$elim[,2]
   flash <- rep("", length(lo))
   flash[lo] <- "<"; flash[hi] <- ">"
   if (any(lo | hi))
      print(cbind(do.call("data.frame", x), flash = flash)[lo | hi,])
   else cat("All points within envelope\n")
   invisible(x)
}
```

Ejemplo (Herencia de la estatura):

Se recolectó la altura de $n = 1375$ madres en UK (bajo 65 años) y una de sus hijas adultas (sobre 18 años).

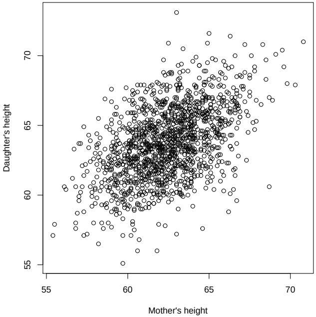
Cargamos el conjunto de datos y hacemos un gráfico:

```
> load("Heights.rda") # carga datos
> plot(dheight ~ mheight, data = Heights)
```

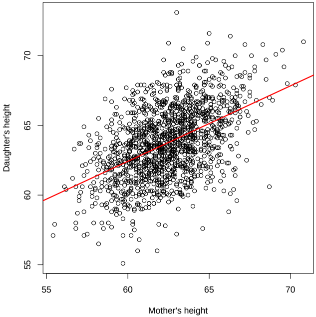
Ajuste de un modelo de regresión lineal simple

```
> fm <- lm(dheight ~ mheight, data = Heights)
> fm

Call:
lm(formula = dheight ~ mheight, data = Heights)

Coefficients:
(Intercept)      mheight
    29.9174       0.5417
```

Herencia de la estatura



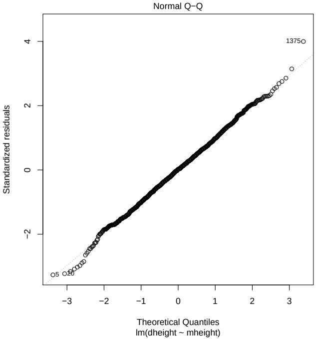
(a) datos estatura



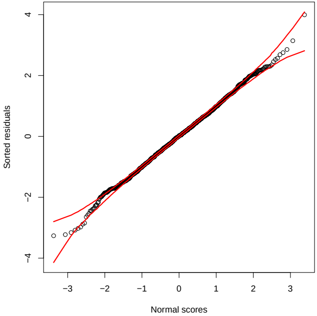
(b) recta ajustada

```
# interpreta script y ejecuta función 'envelope'
> source("envelope.lm.R")
> z <- envelope(fm)

# Salida
> z
            res     elim.1     elim.2 flash
3    -2.9772192 -2.9389051 -2.4155563     <
194  -2.8837417 -2.8506620 -2.3725540     <
83   -2.8469003 -2.7768534 -2.3403934     <
81   -1.9256830 -2.1707912 -1.9326915     >
187  -1.9165152 -2.1591549 -1.9192914     >

...

72   -1.7160276 -1.9010945 -1.7197547     >
997   0.7475798  0.7478225  0.8319662     <
1000  0.7493793  0.7494135  0.8349166     <

# calcula QQ-plots: paneles (a) y (b)
> plot(fm, which = 2)
> plot(z)
```

Herencia de la estatura



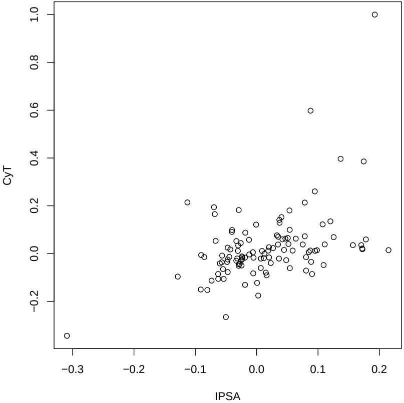
(a) QQ-plot



(b) envelope

Datos de Concha y Toro (Osorio y Galea, 2006)[6]

Rentabilidades mensuales de Concha y Toro vs. IPSA, ajustados por bonos de interés del Banco Central entre marzo/1990 a abril/1999.



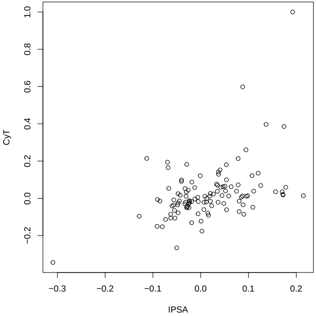
[6] Statistical Papers **47**, 31-38

```
# carga biblioteca 'heavy' y datos 'CyT'
> library(heavy)
> data(cyt)
> fm <- lm(formula = CyT ~ IPSA, data = cyt)
> fm

Call:
lm(formula = CyT ~ IPSA, data = cyt)

Coefficients:
(Intercept)         IPSA
    0.01294      0.88840

# gráfico de CyT con ajuste
> plot(formula = CyT ~ IPSA, data = cyt)
> abline(coef(fm), lwd = 2, lty = 2, col = "red")

# calculo de QQ-plots
> plot(fm, which = 2)
> z <- envelope(fm)
> plot(z)
```

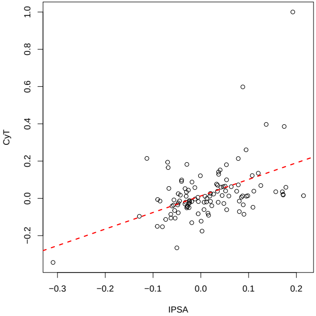
Datos de Concha y Toro

```
# 71 observaciones (de 110) fuera de los límites!
> z
              res       elim.1     elim.2 flash
21   -1.734317225 -3.436526273 -1.9634593     >
27   -1.415056927 -2.788774059 -1.7887230     >
100  -1.404756702 -2.419605631 -1.6748655     >
50   -1.324027912 -2.210958830 -1.5721054     >
54   -1.169203402 -2.054624463 -1.4703409     >
51   -1.153477009 -1.934805526 -1.3798186     >
11   -1.096119396 -1.838246685 -1.3320742     >
47   -1.070881256 -1.742209010 -1.2793535     >
26   -1.007932910 -1.660524800 -1.2222670     >
46   -0.957413175 -1.585646847 -1.1571871     >
22   -0.939792570 -1.516926733 -1.1140743     >
40   -0.936821200 -1.471364845 -1.0692352     >
42   -0.904023872 -1.396710446 -1.0331150     >
64   -0.875677184 -1.345638331 -0.9933418     >
24   -0.866049026 -1.284654537 -0.9480642     >
9    -0.828768419 -1.236817210 -0.9129081     >

...

25    3.757662579  1.787389608  2.7912823     >
12    6.047936818  1.950943630  3.4974732     >
```

Datos de Concha y Toro

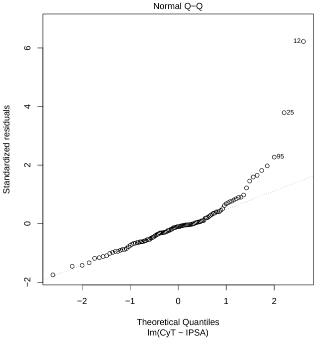


(a) datos Concha y Toro

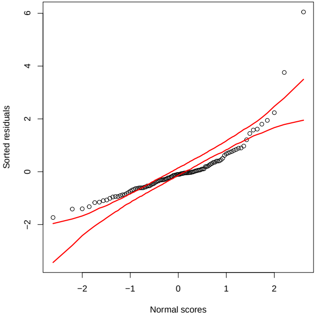


(b) recta ajustada

Datos de Concha y Toro



(a) QQ-plot



(b) envelope

Tenemos que

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}, \qquad \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top. \tag{1}$$

Es fácil notar que,

$$\widehat{Y}_i = \sum_{j=1}^{n} h_{ij}Y_j = h_{ii}Y_i + \sum_{j\neq i} h_{ij}Y_j, \tag{2}$$

es decir el valor predicho es una combinación lineal de las respuestas observadas con pesos dados por los elementos de la matriz de proyección $\boldsymbol{H}$.

La matriz $\boldsymbol{H}$ tiene las siguientes propiedades:

Propiedad 1:

$\boldsymbol{H}$ es simétrica e idempotente con $\operatorname{rg}(\boldsymbol{H}) = \operatorname{tr}(\boldsymbol{H}) = p$.

Propiedad 2:

Los elementos diagonales de $\boldsymbol{H}$ están acotados como:

$$0 \leq h_{ii} \leq 1, \qquad i = 1, \dots, n.$$

En efecto, desde

$$\widehat{\boldsymbol{Y}} \sim \mathsf{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{H}), \qquad \boldsymbol{e} \sim \mathsf{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})).$$

De este modo,

$$\mathsf{var}(\widehat{Y}_i) = \sigma^2 h_{ii}, \qquad \mathsf{var}(e_i) = \sigma^2(1 - h_{ii}),$$

de ahí que obtenemos el resultado.

Para otra demostración, ver Resultado A.6 desde el Apéndice A de las notas de clase.

Propiedad 3:

Tenemos,

$$h_{ii} = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i, \qquad i = 1, \dots, n.$$

Además,

$$\overline{h} = \frac{1}{n} \sum_{i=1}^{n} h_{ii} = \frac{1}{n} \operatorname{tr}(\boldsymbol{H}) = \frac{p}{n}.$$

Propiedad 4:

Sea $\widetilde{\boldsymbol{X}}$ la matriz de datos centrados. En este caso, los elementos diagonales de $\widetilde{\boldsymbol{H}}$ están dados por

$$\widetilde{h}_{ii} = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top (\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}), \qquad i = 1, \dots, n.$$

Luego $\widetilde{h}_{ii}$ es la distancia ponderada desde $\boldsymbol{x}_i$ al centroide $\overline{\boldsymbol{x}}$.

Observación:

Hoaglin y Welsch (1978)[7] sugieren que aquellas observaciones que exceden dos veces su promedio

$$h_{ii} > 2p/n \quad (= 2\overline{h})$$

indican un alto leverage. Mientras que Huber (1981)[8] sugirió identificar observaciones tal que

$$h_{ii} > 0.5,$$

independiente de n o p.

En la práctica se debe prestar atención a casos inusualmente grandes con relación al resto de h_{ii}'s.

[7] The American Statistician **32**, 17-22.

[8] Robust Statistics. Wiley, New York.

Propiedad 5:

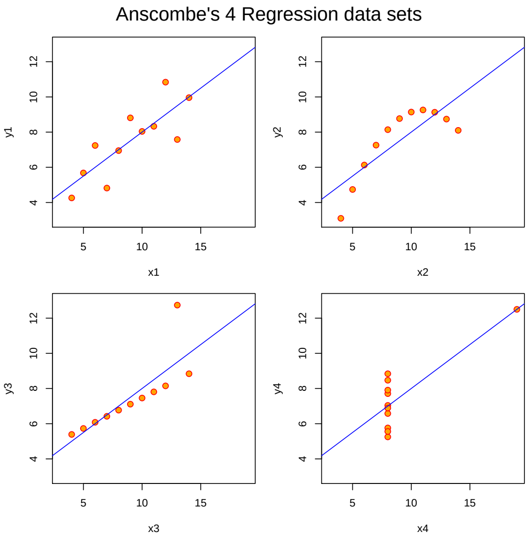
Desde Ecuación (1), sigue que
$$\frac{\partial \widehat{\boldsymbol{Y}}}{\partial \boldsymbol{Y}^\top} = \boldsymbol{H},$$
y en particular $\partial \widehat{Y}_i / \partial Y_i = h_{ii}$, para $i = 1, \dots, n$.

Propiedad 6:

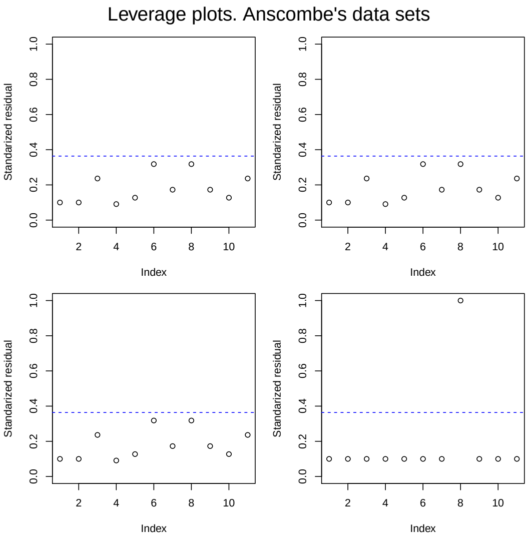
Si el modelo tiene intercepto, entonces $\boldsymbol{H}\boldsymbol{1} = \boldsymbol{1}$.

Considere $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X}_1)$. Sabemos que $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$, y de ahí que
$$\boldsymbol{H}(\boldsymbol{1}, \boldsymbol{X}_1) = (\boldsymbol{1}, \boldsymbol{X}_1),$$
y el resultado sigue.

Cuarteto de regresiones "idénticas" de Anscombe (1973)



Leverages: Datos de Anscombe



Leverages: Datos de Concha y Toro

```
# ajuste de datos 'CyT', almacenando 'x'
> fm <- lm(CyT ~ IPSA, data = cyt, x = TRUE)
> x <-fm$x # extrae 'x'
> z <- influence(fm)
> attributes(z)
$names
[1] "hat"   "coefficients"   "sigma"   "wt.res"

# extrae 'leverages' y calcula punto de corte
> hats <- z$hat
> n <- nrow(x)
> p <- ncol(x)
> cutoff <- 2 * p / n

> which <- hats > cutoff
> idx <- 1:n
> obs <- idx[which]

# gráfico de leverages con punto de corte
> plot(z$hat, ylim = c(0,0.18), ylab = "Leverages")
> abline(h = cutoff, lwd = 2, lty = 2, col = "red")
> text(obs, hat[obs], labels = as.character(obs), pos = 3)

# otros métodos para calcular leverages
> hats <- hatvalues(fm)
> hats <- hat(x, intercept = FALSE)
```

Leverages: Datos de Concha y Toro