Residual estandarizado

También conocidos como residuales estudentizados de manera interna, se define como:

$$d_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

La media de un residual estandarizado es 0 y su varianza es 1. Además, asumiendo normalidad, valores por encima de 2, o por debajo de -2 son considerados outliers.

```
modelo |> resid() -> r
summary(modelo)$sigma -> s
modelo |> model.matrix() -> X
X %*% solve(t(X) %*% X) %*% t(X) -> H
(r/(s*sqrt(1-diag(H)))) -> res_stand
res_stand |> round(2)
```

matriz hat

OUTLIER

1	2	3	4	5	6	7	8	9	10	11	12
-0.68	-0.42	-0.09	-0.58	-0.38	-0.30	-0.12	3.11	-0.40	-0.46	-0.08	0.10

Nota: se redondea solo con la finalidad de mostrar los residuales en la diapositiva.

Residual estandarizado

También conocidos como residuales estudentizados de manera interna, se define como:

$$d_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

Residual estudentizado

Sin embargo, de haber outliers, los residuales estandarizados podrían no ser normales. Ante ello, se propone redefinir la expresión para la estimación del CME, retirando la i-ésima observación. Así, el residual estudentizado (externamente) sería:

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

(i): al retirar la i-ésima observación

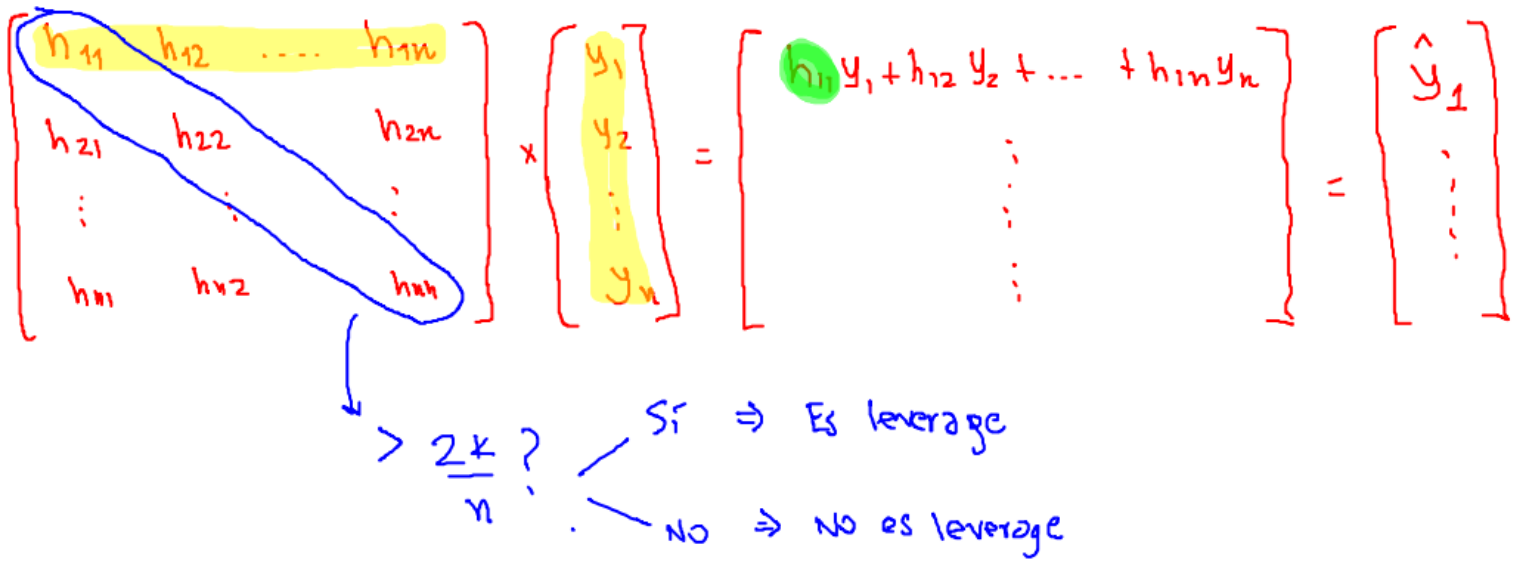
Leverage

Sabemos que:

$$\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Es decir:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n$$



Distancia de Cook

Mide la influencia de cada observación en los coeficientes de regresión estimados

$$D_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)' \mathbf{X}'\mathbf{X} \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{k\hat{\sigma}^2} = \frac{d_i^2 h_{ii}}{k(1-h_{ii})}$$

$\hat{\beta}_{k\times 1}$ $\quad$ $\hat{\beta}_{(i)\,k\times 1}$

$X_{n\times k}$ $\quad$ $X'_{k\times n}$

Obs	D_i	> ?
1	D_1	
2	D_2	
3	D_3	
⋮	⋮	
n	D_n	

$$D_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)' \mathbf{X}'\mathbf{X} \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{k\hat{\sigma}^2} = \frac{d_i^2 h_{ii}}{k(1 - h_{ii})}$$

```
# ¿observación 1 es influyente?
modelo = lm(y~x)
modelo_sin = lm(y[-1]~x[-1])
modelo |> coef() -> beta
modelo_sin |> coef() -> beta_sin
modelo |> model.matrix() -> X
(modelo |> sigma())**2 -> cme
2 -> k
(t(beta-beta_sin)%*%t(X)%*%X%*%(beta-beta_sin))/(k*cme)
```

```
          [,1]
[1,] 0.04252736
```

$= D_1$

```
modelo |> rstandard() -> rsta
modelo |> hatvalues() -> h
(rsta[1]**2*h[1])/(k*(1-h[1]))
```

```
          1
0.04252736
```

$= D_1$

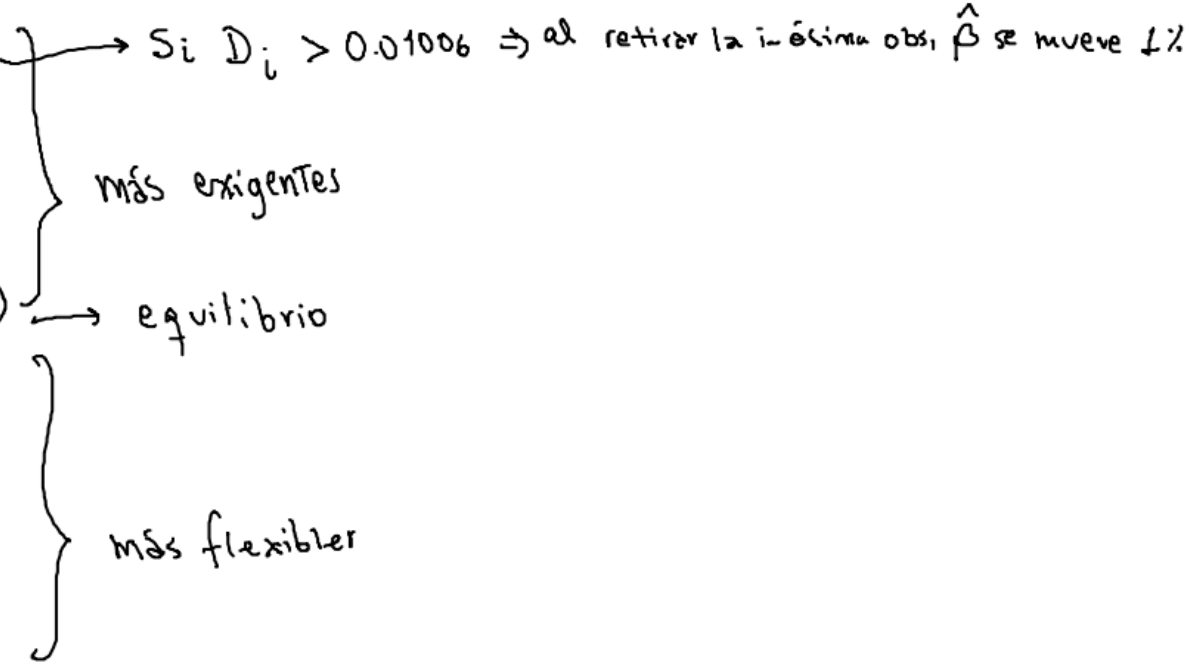
```
modelo |> cooks.distance() |> round(2) |> as.vector()
```

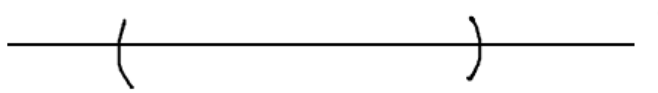
```
 [1] 0.04 0.01 0.00 0.03 0.01 0.00 0.00 0.50 0.01 0.59 0.00 0.00
```

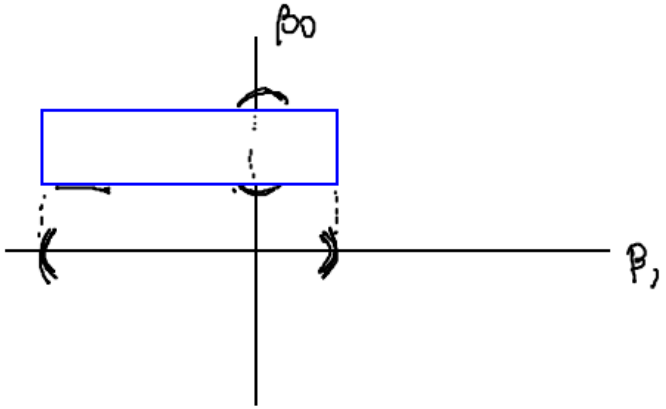
¿Entre qué valores fluctúa F? Interprete los valores de la siguiente tabla:

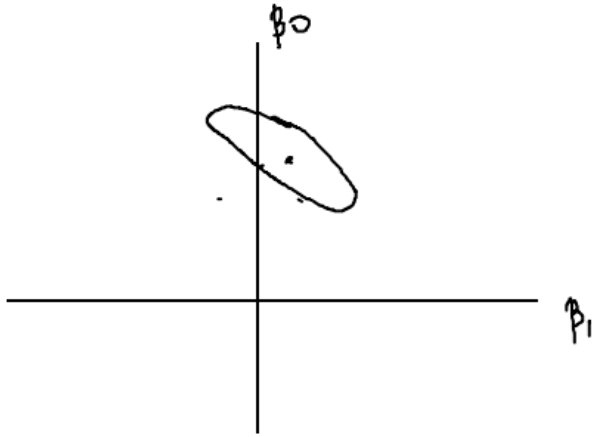
```
alfas = c(0.01,0.05,0.1,0.25,0.5,0.66,0.75,0.90,0.95,0.99)
data.frame(alfas,valorF=qf(alfas,k,n-k))
```

	alfas	valorF
1	0.01	0.01006044
2	0.05	0.05155730
3	0.10	0.10647844
4	0.25	0.29611921
5	0.50	0.74349177
6	0.66	1.20403474
7	0.75	1.59753955
8	0.90	2.92446596
9	0.95	4.10282102
10	0.99	7.55943216

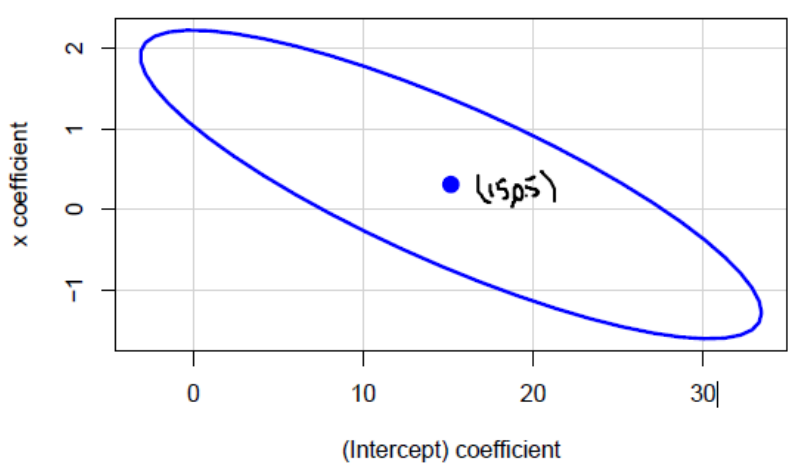


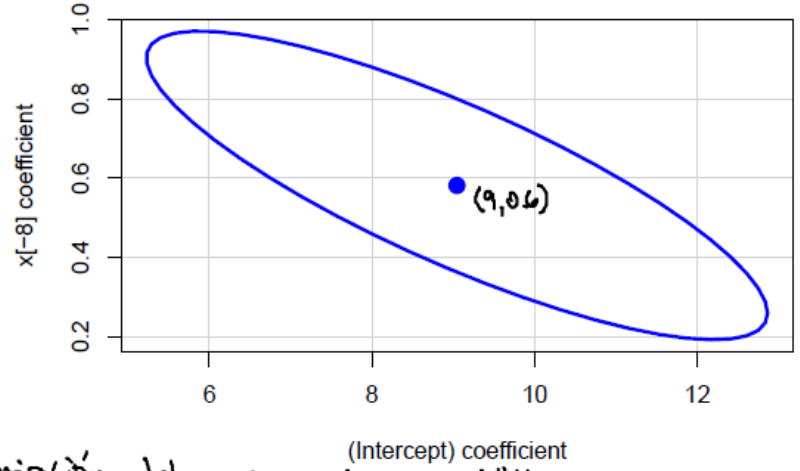






$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$





Variación del centro y la variabilidad