Taller RLM 1

-

3/12/2021

```
## Rows: 1599 Columns: 12

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (12): fixed acidity, volatile acidity, citric acid, residual sugar, chlo...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

##
## Call:
## lm(formula = Calidad ~ Fija + Volatil + Citrico + Azucar + Cloruros,
##     data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3954 -0.3604 -0.1540  0.4216  1.6609
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.972902   0.584358  10.221  < 2e-16 ***
## Fija         0.096393   0.066284   1.454  0.14921
## Volatil     -2.087519   0.494974  -4.217 5.68e-05 ***
## Citrico     -1.686348   0.510522  -3.303  0.00135 **
## Azucar       0.001826   0.045415   0.040  0.96801
## Cloruros     0.786835   0.940631   0.836  0.40500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6124 on 94 degrees of freedom
## Multiple R-squared:  0.1753, Adjusted R-squared:  0.1315
## F-statistic: 3.997 on 5 and 94 DF,  p-value: 0.002482
```

Punto siete. Sumas de cuadrados secuenciales (tipo I) y sumas de cuadrados parciales (tipo II).

```
## Analysis of Variance Table
##
```

```
## Response: Calidad
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Fija       1  0.474  0.4738  1.2632 0.263904
## Volatil    1  2.919  2.9190  7.7829 0.006386 **
## Citrico    1  3.840  3.8399 10.2384 0.001876 **
## Azucar     1  0.000  0.0001  0.0003 0.987118
## Cloruros   1  0.262  0.2624  0.6997 0.404997
## Residuals 94 35.255  0.3751
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos observar en la tabla anterior los menores valores para las sumas de cuadrados de tipo I son: 1. $\text{SS1}_{X_4} = 0.000$ 2. $\text{SS1}_{X_5} = 0.262$ 3. $\text{SS1}_{X_1} = 0.474$ Nuestra tabla anova tambien nos dice que: $\text{SSR}(X_4|X_1, X_2, X_3) = 0.000$ $\text{SSR}(X_5|X_1, X_2, X_3, X_4) = 0.262$ $\text{SSR}(X_1) = 0.474$, lo que quiere decir que las sumas de las diferencias entre la estimación y el valor medio de la variable de respuesta es mínima, por lo que el modelo propuesto no es suficientemente útil, también podemos verlo con el p-value; rechazamos la hipótesis y concluimos que la variable no es significativa para cada modelo planteado.

```
## Anova Table (Type II tests)
##
## Response: Calidad
##           Sum Sq Df F value    Pr(>F)
## Fija       0.793  1  2.1148  0.149211
## Volatil    6.671  1 17.7867 5.685e-05 ***
## Citrico    4.092  1 10.9110  0.001353 **
## Azucar     0.001  1  0.0016  0.968006
## Cloruros   0.262  1  0.6997  0.404997
## Residuals 35.255 94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

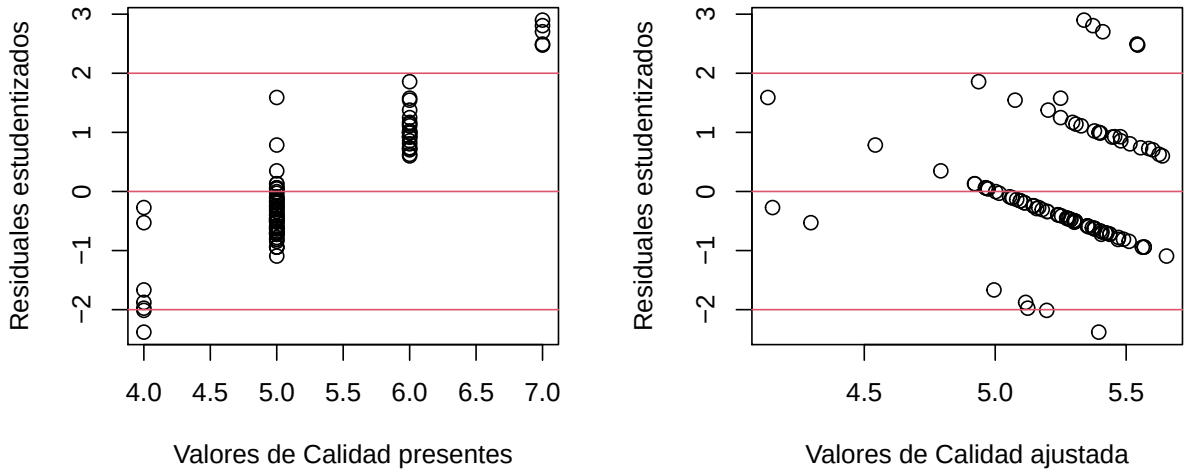
Como podemos observar en la tabla anterior los menores valores para las sumas de cuadrados de tipo II son:

1. $\text{SS2}_{X_4} = 0.001$
2. $\text{SS2}_{X_5} = 0.262$
3. $\text{SS2}_{X_1} = 0.793$

Cada valor nos dice el SSR de cada variable en el modelo completo dadas las demás (ej. $\text{SSR}(X_4|X_1, X_2, X_3, X_4) = 0.001$), lo que quiere decir que las sumas de las diferencias entre la estimación y el valor medio de la variable de respuesta es mínima, por lo que el modelo propuesto no es suficientemente útil, también podemos verlo con el p-value; recordemos que rechazamos la siguiente hipótesis nula cuando el p-value es pequeño, como podemos ver para X_1, X_4, X_5 los p-values son demasiado grandes si fijamos un α de 0.05, por lo que concluimos que estas variables no son significativa para cada el modelo ajustado.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + E_i \text{ , con } E \sim N(0, \sigma^2) \text{ } \mathrm{H}_0 : \beta_j = 0 \text{ vs. } \mathrm{H}_1 : \beta \neq 0$$

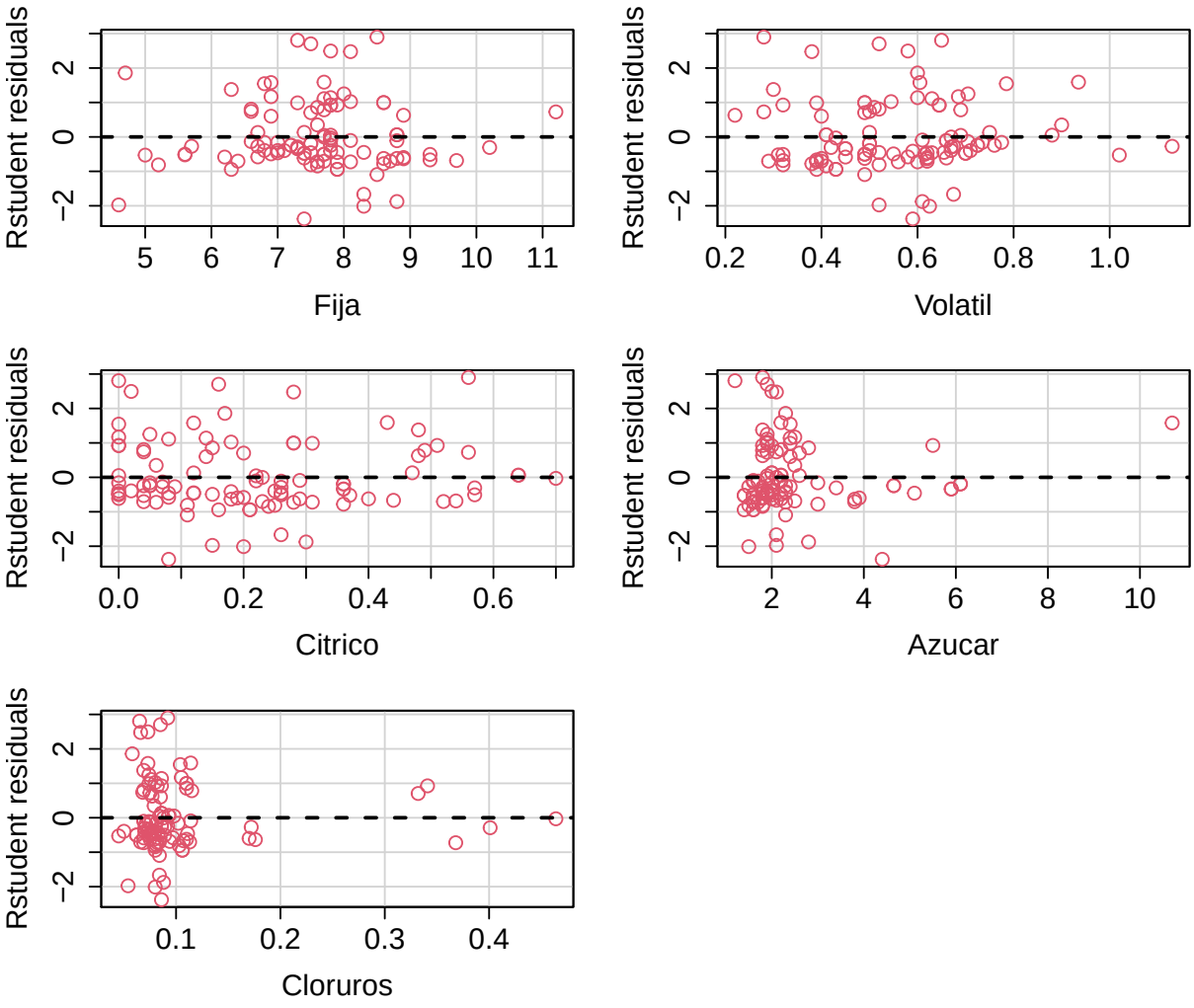
Punto ocho. Gráficos de los residuales estudentizados vs. Valores ajustados y contra las variables de regresión utilizadas.



Como podemos ver en las gráficas anteriores los residuales estudentizados tienen ciertos patrones, en la primera gráfica observamos que entre más alta sea la calidad estos tienden a pasar de negativos a positivos (modelo lineal entre x y y no es adecuado) y, además, que cuando el valor es de calidad es 4 la varianza está mucho más dispersa que cuando el valor de la calidad es 7, haciendo que la varianza no sea constante.

Un motivo de esto puede ser que no se cuenta con un número considerable de observaciones, por lo que el modelo puede ser susceptible a observaciones atípicas o influcienciadoras.

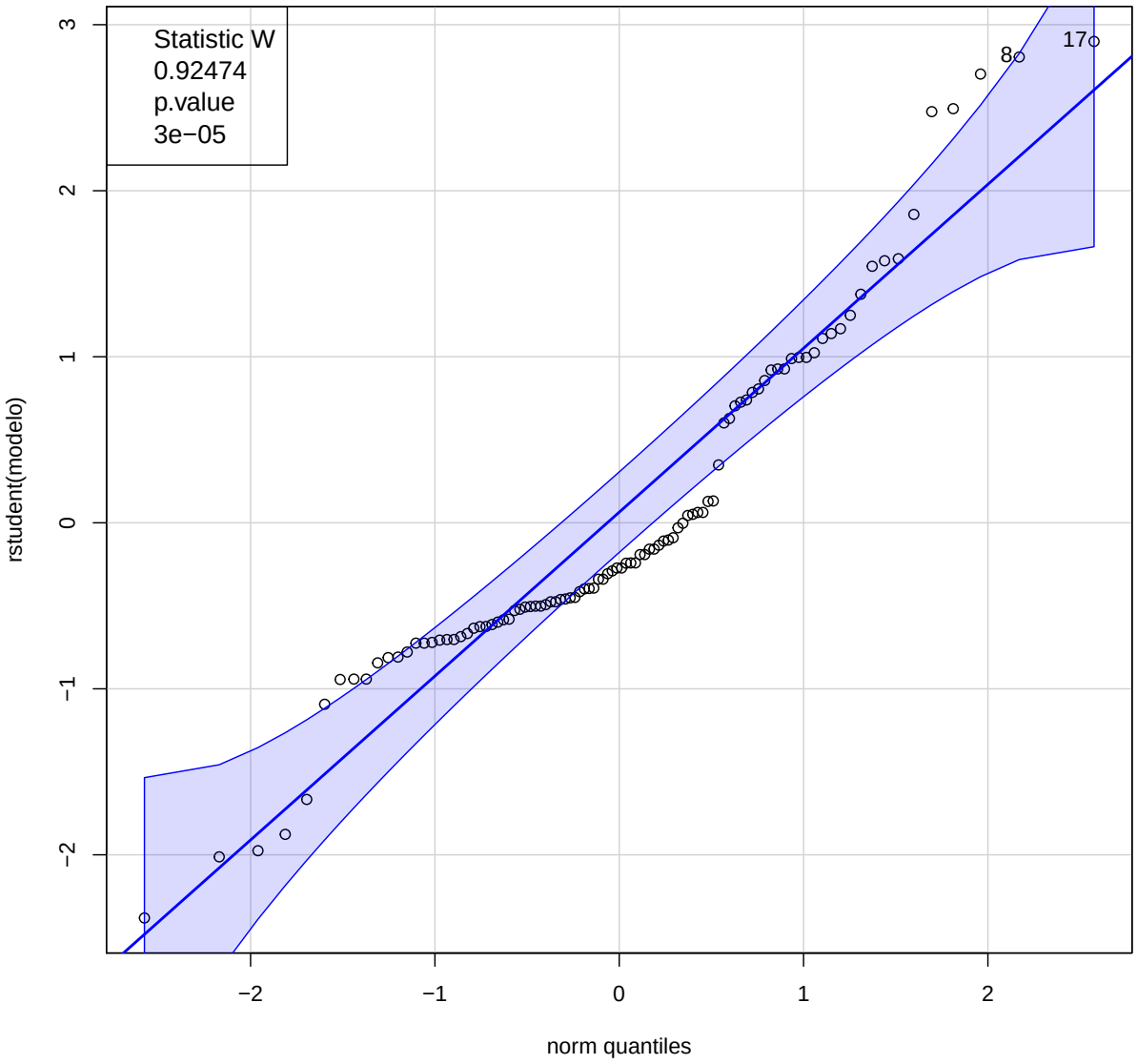
En la gráfica de valores de la calidad ajustada por el modelo vs. residuales estudentizados podemos ver que tiene un efecto similar, pero en este los residuales están un poco más centrados.



En esta gráfica podemos observar el comportamiento de las variables utilizadas para la regresión vs. los residuales, al parecer no hay ningún indicio de que alguna variable afecte el comportamiento de la varianza.

Punto nueve. Gráfica de probabilidad normal para los residuales estudentizados. ¿Existen razones para dudar de la hipótesis de normalidad sobre los errores en este modelo?

```
## [1] 17  8
```



Como podemos ver en el gráfico hay datos que se desvían demasiado de los cuantiles teóricos de la distribución normal, lo cual es una gran señal para dudar de la normalidad de los residuales.

Realizamos el test de Shapiro-Wilk donde la hipótesis nula es que nuestros errores provienen de una distribución normal, podemos ver que el p-value es igual a 0.00003, muchísimo menor a cualquier valor de α que podamos fijar, por lo que rechazamos la hipótesis nula y afirmamos que hay suficiente evidencia para decir que los errores residuales no siguen una distribución normal.

Punto diez. Presencia de observaciones atípicas, de balanceo y/o influencíales.

Como se observa en la tabla anterior los datos influenciables son:

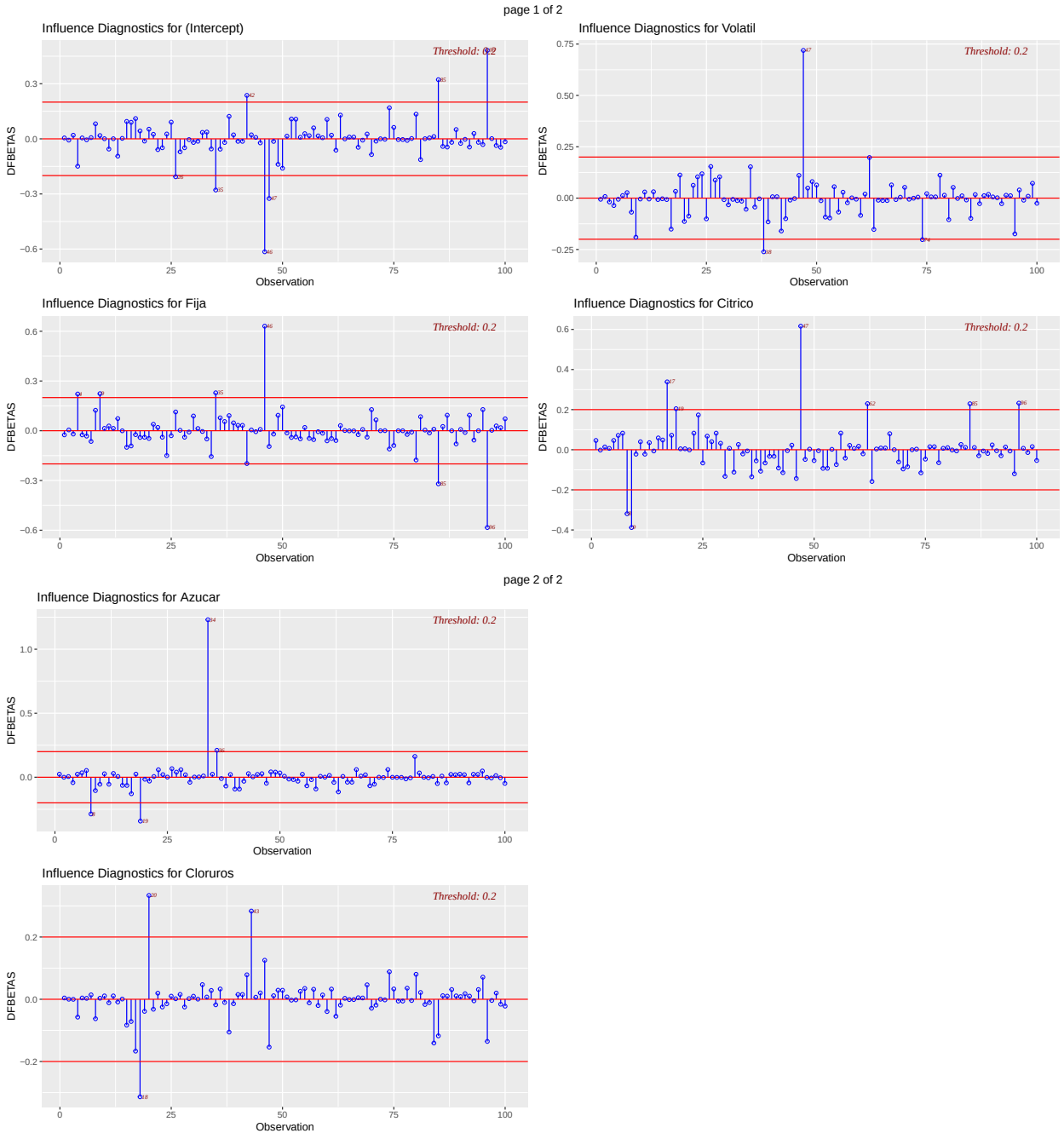
Según la medida DFBetas, los datos influenciables son: 34

Según la medida DFFITS, los datos influenciables son: 34, 47

Según la medida COVRATIO, los datos influenciables son: 4, 8, 9, 17, 18, 19, 20, 34, 38, 39, 43, 63, 82, 84, 95

Según la Distancia de Cook ningún dato es influenciable.

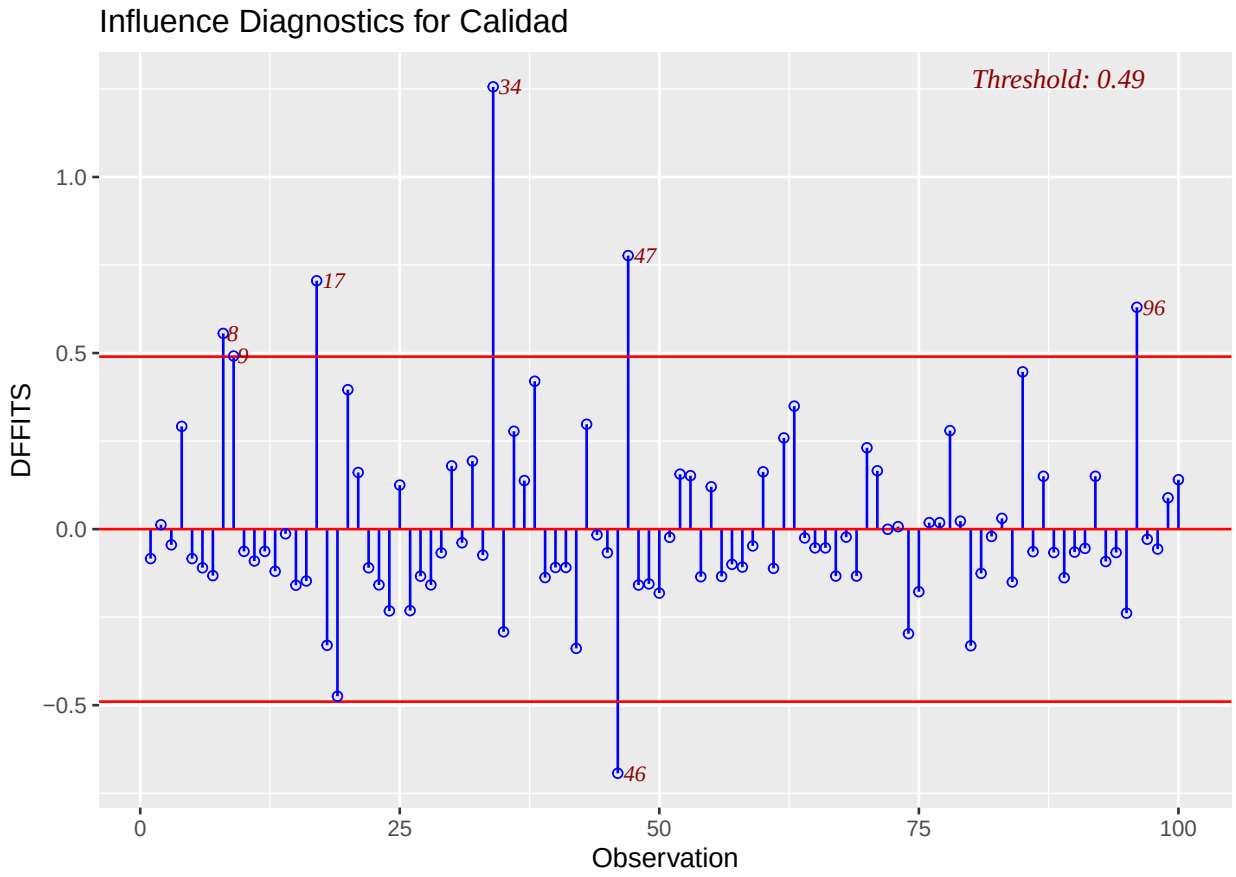
DFBETAS



Como podemos ver el dato numero 34 supera el limite fijado en para los datos de azúcar, recordemos que una observación es candidata a ser influencial mediante este metodo si $|DFBETAS_{j(i)}| > 2/\sqrt{n}$, en este caso nuestro limite es igual a $2/\sqrt{100} = 0.2$

DFFITS

Como podemos ver en la gráfica hay varios datos que superan el limite fijado, recordemos que una observación es candidata a ser influencial si $|DFFITS_{(i)}| > 2\sqrt{\frac{k+1}{n}}$, en este caso nuestro limite es igual a $2\sqrt{\frac{5+1}{100}} \approx 0.49$, con esto en mente, los datos mas potencialmente influencíales de acuerdo a esta medida, en orden, son: 34, 47, 17, 46

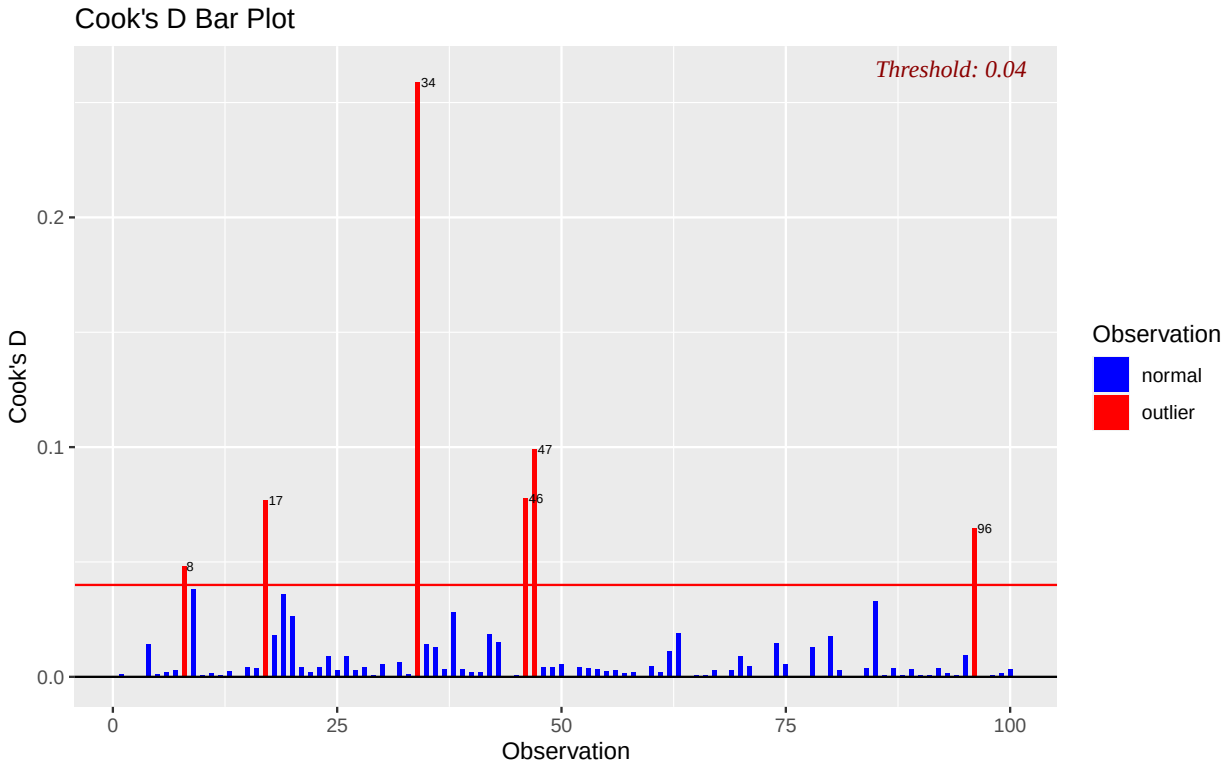


COVRATIO

Como podemos ver los datos numero supera el limite fijado en para los datos de azúcar, recordemos que una observación es candidata a ser influencial si $|COVRATIO_i - 1| > 3(k+1)/n$, en este caso; usamos R y encontramos los datos que cumplen la condición, a continuación los datos potencialmente influencíales y su $COVRATIO$:

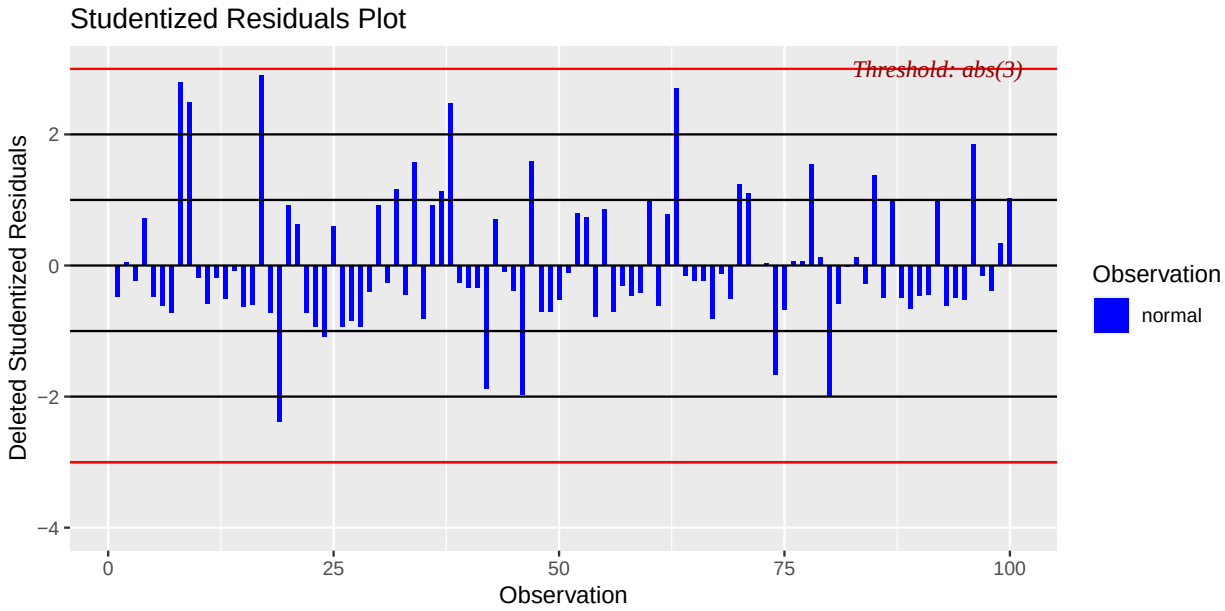
```
##         4         8         9        17        18        19        20        34
## 1.1970381 0.6805927 0.7511249 0.6720391 1.2442410 0.7774960 1.1934652 1.4863231
##        38        39        43        63        82        84        95
## 0.7478104 1.3315360 1.2175042 0.6886982 1.5805205 1.3455928 1.2608571
```

DISTANCIA DE COOK

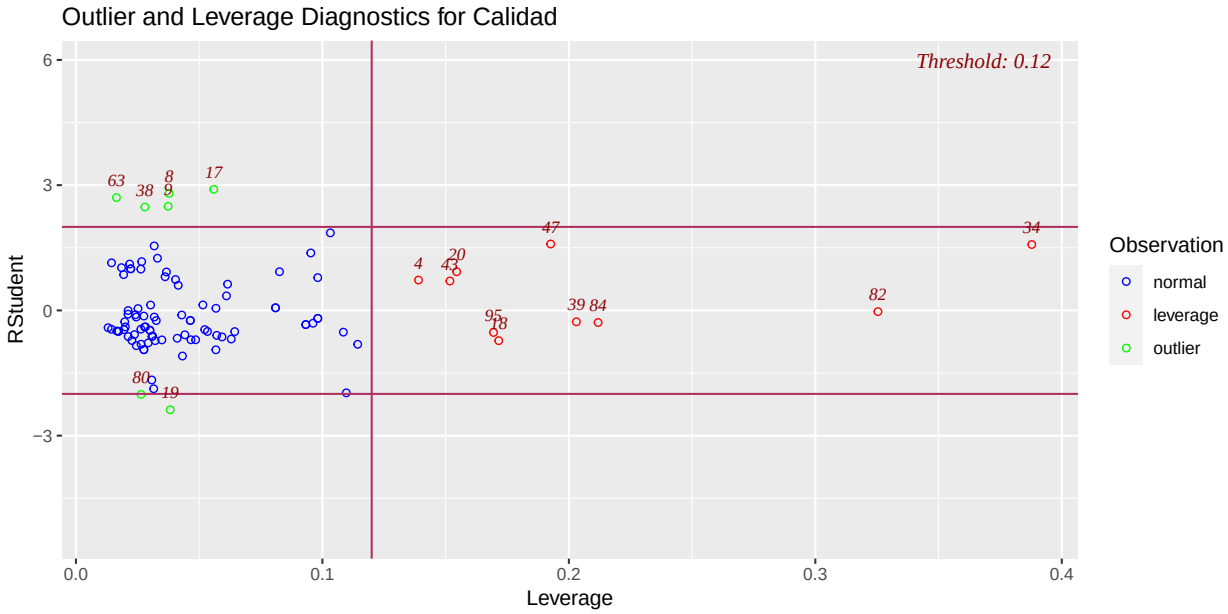


Validamos y encontramos que efectivamente ningún valor sobrepasa 1, pero podemos ver que la observación 34 está demasiado lejos de las demás, por lo que la tendremos en cuenta.

Residuales estudentizados (internamente estudentizados)

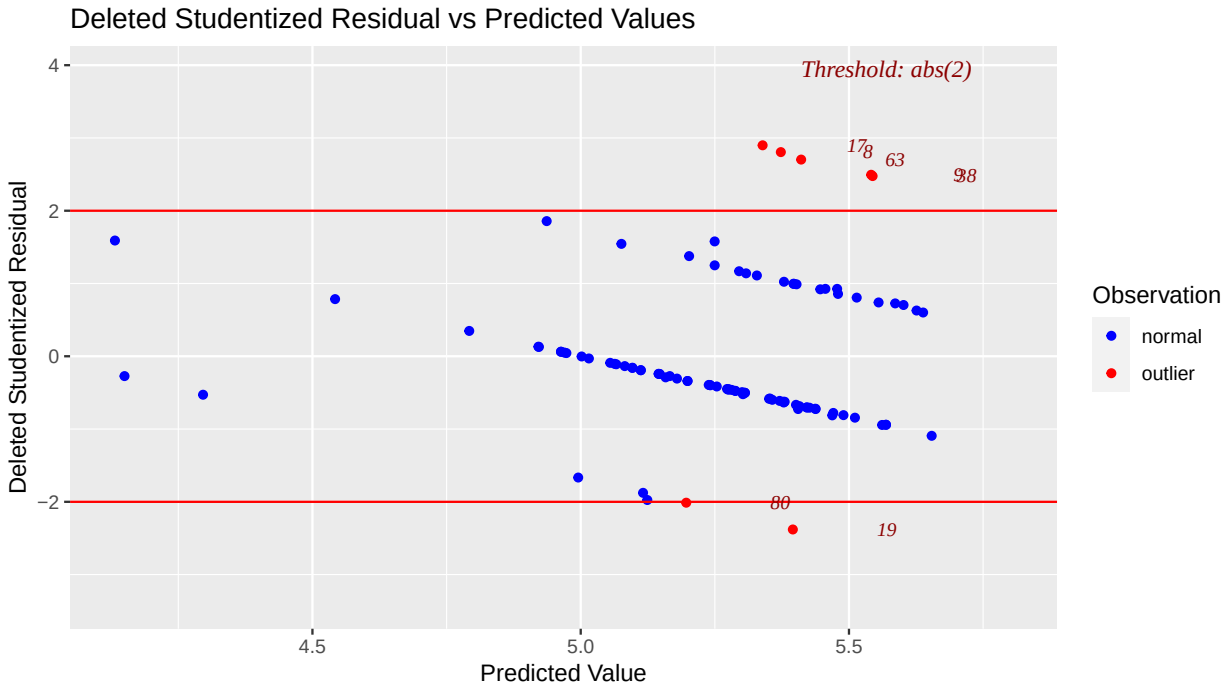


En este gráfico podemos ver que ninguna observación sobrepasa el limite fijado de $|e_i| > 3$, por lo que por este método no encontramos ninguna observación atípica.



En esta gráfica se grafican los hat-values vs. residuales estudentizados, como podemos observar la gráfica usa un limite diferente, este limite es 0.12, segun hemos visto en clase los limite que fijamos son ± 3, con el cual obtendiramos que ningún dato es una observación atípica, pero este se encuentra calculado de una forma diferente, lo cual puede darnos un indicio a cuales podrían ser observaciones atípicas.

En la gráfica anterior tambien podemos observar que los puntos de balanceo pueden ser las observaciones 34, 82, 84, 39, 47, 18, 95, 20, 43, 4 basándonos en su valor h_{ii} y el limite usado $h_{ii} > 2(k+1)/n$, que equivale a $h_{ii} > 0.12$, como se puede observar en la gráfica.



	dfb.1_	dfb.Fija	dfb.Vltl	dfb.Ctrc	dfb.Azcr	dfb.Clrr	dffit	cov.r	cook.d	hat
1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
9	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
10	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
12	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
14	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
15	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
16	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
17	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
18	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
19	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
20	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
21	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
22	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
23	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
24	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
25	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
26	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
27	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
28	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
29	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
30	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
31	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
32	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
33	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
34	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE
35	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
36	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
37	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
38	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
39	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
40	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
41	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
42	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
43	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
44	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
45	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
46	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
47	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
48	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
49	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
50	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
51	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
52	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
53	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
54	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
55	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
56	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
57	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
58	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
59	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE