

Taller RLM 1

Sofía Cuartas García Simón Cuartas Rendón Julián Alejandro Úsuga Ortiz
Deivid Zhang Figueroa

Enero de 2022

Punto 1. Descripción de la base de datos.

Para impulsar la industria de vinos y su crecimiento se invierte en tecnología para el proceso de producción y venta.

Los datos fueron recolectados por un sistema computarizado (iLab), que gestiona automáticamente el proceso de elaboración del vino, de las solicitudes de muestreo de pruebas del productor y análisis sensorial y al laboratorio. Las variables que están incluidas en esta base de datos son:

- **Fixed acidity.** Puede traducirse como *acidez fija* y está dado en gramos de ácido tartárico ($C_4H_6O_6$) por decímetro cúbico ($\frac{g[C_4H_6O_6]}{dm^3}$). Es un componente de la acidez total de los vinos que incluye únicamente a los ácidos no volátiles y, en el caso particular del ácido tartárico, se origina en las uvas empleadas para producir el vino [1]. Esta es por tanto una variable continua racional, pues el cero absoluto significa ausencia de ácidos fijos en el vino.
- **Volatile acidit.** Puede traducirse como *acidez volátil* y sus unidades están dadas en gramos de ácido acético ($CH_3 - COOH(C_2H_4O_2)$) por decímetro cúbico ($\frac{g[CH_3 - COOH(C_2H_4O_2)]}{dm^3}$). Estos ácidos son un componente de la acidez total del vino que se diferencian de los ácidos fijos porque son destilables al vapor. Una alta concentración de estos ácidos en un vino suele ser indicador de deterioro y produce un sabor semejante al del vinagre [2]. Así, se puede definir que la acidez volátiles es una variable continua racional.
- **Citric acid.** Esta variable puede ser traducida al castellano como *ácido cítrico* y se expresa en gramos por decímetro cúbico $\frac{g}{dm^3}$. Estos ácidos se diferencian del resto por ser ácidos débiles inorgánicos y que son frecuentemente empleados como preservativos naturales o para agregar un sabor agrio a la comida. Además, puede emplearse para eliminar o disminuir la cantidad de mohos y bacteria en los vinos [3]. Con esto, se toma al ácido cítrico como una variable continua racional.
- **Residual sugar.** Esta variable se interpreta en el español como *azúcar residual* y sus unidades están dadas en gramos por decímetro cúbico $\frac{g}{dm^3}$. Este componente del vino se asocia con la cantidad de azúcar que queda en el vino luego del proceso de fermentación. A partir de esta variable se pueden clasificar los vinos como *secos*, que tienen de cero a cuatro gramos de azúcar por litro; *semisecos*, que son aquellos vinos con una concentración de cuatro a doce gramos de azúcar por litro; vinos *semidulces*, que se caracterizan porque su contenido de azúcar va desde los ocho hasta los 45 gramos por litro y por último los vinos *dulces*, los cuales poseen más de 45 gramos de azúcar por litro [4]. Teniendo la anterior clasificación presente, se puede decir que los azúcares residuales son una variable continua racional.
- **Chlorides.** En español se entiende esta variable como *cloruros* y se mide en gramos de cloruro de sodio por decímetro cúbico ($\frac{g[NaCl]}{dm^3}$). Los cloruros son útiles para balancear la cantidad de ácidos y alcalinos [5]. Esta variable es, por tanto, continua racional.
- **Quality.** Traducida como *calidad*, es una variable discreta ordinal que clasifica los vinos en un puntaje de cero a diez, donde diez implica la mejor calidad posible y cero la peor calidad posible.

Aspectos iniciales para el modelo de regresión lineal

Ahora bien, el objetivo es plantear un **modelo de regresión lineal múltiple**, y atendiendo al contexto y según el propio objetivo de los investigadores con técnicas más avanzadas de *machine learning* (aprendizaje de máquina en castellano), se puede establecer que la variable de respuesta es la **calidad**, en tanto los productores de vino están interesados en conocer cuál será la calidad de los vinos que producen en sus viñedos a partir de las demás variables (concentraciones de ácidos fijos, volátiles y cítricos, azúcares residuales y cloruros en el vino) para poder tomar decisiones encaminadas en la obtención de mejores vinos que les permitan ser más competitivos y tener mejor reputación en el mercado; asimismo, esto interesa a los consumidores en tanto estarán informados respecto a qué vinos tienen mejor calidad y por tanto merecen más la pena ser comprados.

Teniendo este presente, es útil considerar en este análisis descriptivo la estructura de varianzas y covarianzas.

Punto 2. Análisis descriptivo.

La calidad es una variable numérica discreta que puede ser estudiada inicialmente mediante el siguiente esquema de resúmenes numéricos:

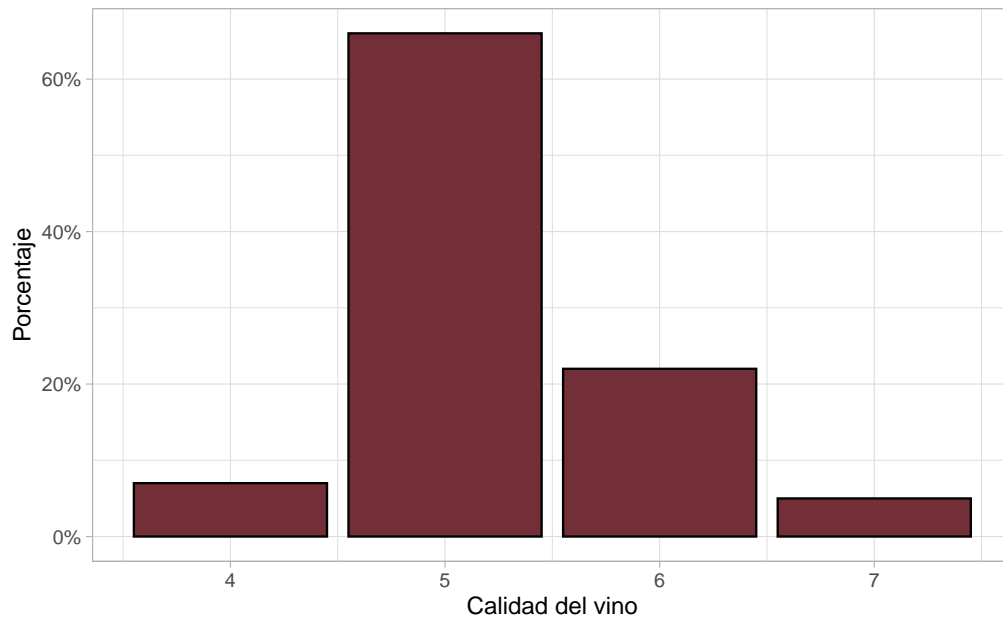
Parámetro	Valor
Media	5.25
Desviación estándar	0.66
Mínimo	4
Primer cuantil (Q1)	4
Mediana (Q2)	5
Tercer cuantil (Q3)	6
Máximo (Q4)	7
Rango intercuartílico	1
Coefficiente de variación	0.13
Coefficiente de asimetría	0.75
Curtosis	0.82

Entonces, se comienza mencionando que la calidad promedio de los vinos de la muestra de los investigadores es de 5.25, con una desviación estándar de 0.66. Por otro lado, se tiene que el vino de peor calidad tiene un puntaje de cuatro puntos, toda vez que el mejor ranqueado destaca con siete puntos de diez. Asimismo, se tiene que la mediana ocurre en los cinco puntos, al igual que el primer cuantil, lo que quiere decir que al menos el 50 % de los vinos de esta base de datos tiene una calidad puntuada entre los cinco y los siete puntos, mientras que los demás tienen cuatro puntos; asimismo, se cumple que el tercer cuantil ocurre a los seis puntos y, en consecuencia, el rango intercuartílico es de un punto únicamente, lo cual ya anticipa una concentración importante de valores al rededor de este rango.

Otras características de la distribución de esta variable es que el coeficiente de asimetría es de 0.75, lo cual da cuenta de una concentración importante de clasificaciones de calidad cercanas al mínimo, mientras que la curtosis es de 0.82 y, entonces, se tiene que hay una mayor cantidad de valores atípicos en comparación con una distribución normal.

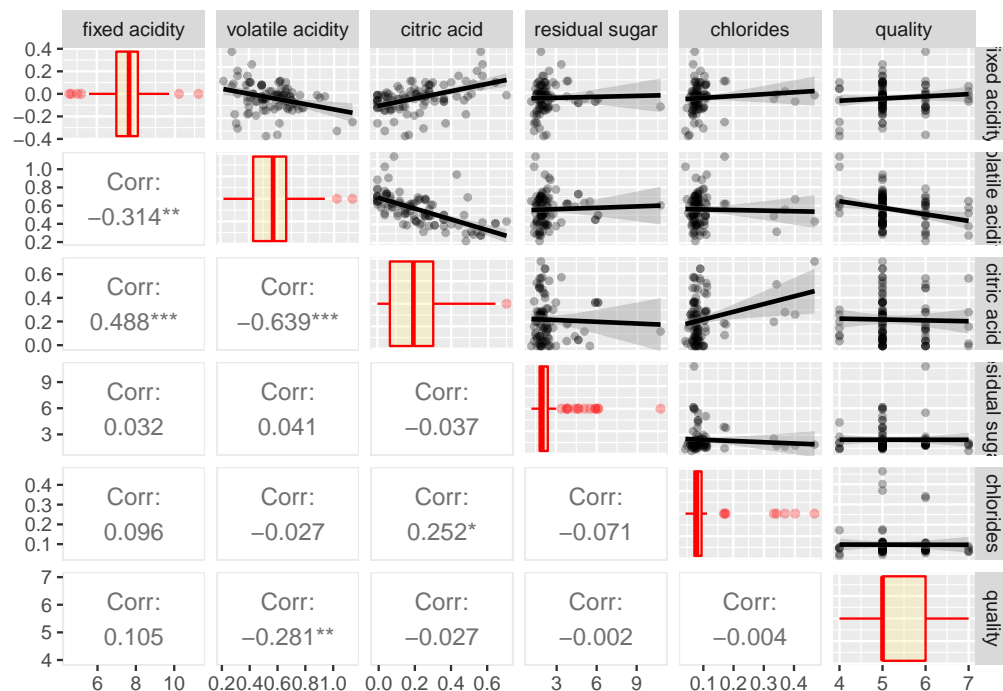
Ahora bien, para poder entender mejor esta variable vale la pena considerar el siguiente gráfico de barras:

Diagrama de barras para la calidad de los vinos
Calificación de la calidad de los vinos en una escala de uno a diez



Y como se puede observar, más del 60 % de los vinos incluidos en la base de datos que se está estudiante poseen una calidad de cinco puntos de diez, y la segunda clasificación de calidad más frecuente es la de seis puntos, con poco más del 20 % del total. Esto muestra que la mayoría de vinos de esta base de datos tienen clasificaciones de calidad regulares considerando que este parámetro puede tomar valores entre cero y diez.

Estructura de varianzas y covarianzas



Del gráfico anterior se observa pues que las dos variables que presentan la mayor relación lineal son los **ácidos volátiles** y los **ácidos cítricos**, puesto que su coeficiente de correlación es de -0.639 , lo cual indica que tienen una correlación lineal negativa moderada a fuerte. Después de esta, vale la pena destacar también a la **acidez fija** con la **acidez cítrica**, teniendo un coeficiente de correlación de 0.488 , lo que implica que este par de variables presentan una correlación lineal positiva moderada. A continuación, destacan la **acidez fija** con la **acidez volátil**, puesto que el coeficiente de correlación entre este par de variables es de -0.314 , lo que significa que tiene una correlación lineal negativa moderada a débil. Ya en tercer lugar se tiene a la **acidez volátil** con la **calidad**, teniendo un coeficiente de correlación lineal de -0.281 , lo que significa que se trata de una correlación lineal negativa moderada a débil. Es importante notar pues que de las cuatro correlaciones lineales más importantes que se evidencian, tres de ellas implican a la acidez volátil, siendo todas ellas correlaciones lineales negativas, y dos tienen en cuenta a la acidez fija y otros dos a la acidez cítrica.

Ahora bien, al ceñirse únicamente a la calidad, solo se destaca la correlación lineal negativa moderada a débil que se mencionó previamente entre esta variable y la concentración de ácidos volátiles, mientras que con las demás variables se tienen correlaciones lineales débiles, destacándose la que se tiene con las concentraciones de azúcares residuales y los cloruros, pues los coeficientes de correlación son de -0.002 y -0.004 respectivamente.

Punto 3. Modelo de regresión.

Para plantear el modelo de regresión lineal, se van a considerar las siguientes variables:

- Y_i . Calidad del i -ésimo vino analizado.
- X_{1i} . Concentración de ácidos fijos i -ésimo vino analizado en $\frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$.
- X_{2i} . Concentración de ácidos volátiles en el i -ésimo vino analizado en $\frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$.
- X_{3i} . Concentración de ácidos cítricos en el i -ésimo vino analizado en $\frac{g}{dm^3}$.
- X_{4i} . Concentración de azúcares residuales en el i -ésimo vino analizado en $\frac{g}{dm^3}$.
- X_{5i} . Concentración de cloruros en el i -ésimo vino analizado en $\frac{g[NaCl]}{dm^3}$.
- E_i . Error aleatorio de la regresión.

Notar que para cada una de las variables el índice i es tal que $i = 1, 2, \dots, n$, con $n = 100$, puesto que se está considerando una muestra de cien vinos. Con esto presente, el modelo de regresión lineal múltiple que se va a ajustar es el siguiente:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + E_i, \quad E_i \stackrel{iid}{\sim} Normal(0, \sigma^2), \quad i = 1, 2, \dots, 100$$

Y al realizar el ajuste del modelo ayuda de [R](#), se obtiene lo siguiente:

	<i>Dependent variable:</i>
	Calidad
Fija	0.096 (0.066)
Volatil	-2.088*** (0.495)
Citrico	-1.686*** (0.511)
Azucar	0.002 (0.045)
Cloruros	0.787 (0.941)
Constant	5.973*** (0.584)
Observations	100
R ²	0.175
Adjusted R ²	0.131
Residual Std. Error	0.612 (df = 94)
F Statistic	3.997*** (df = 5; 94)

Note: *p<0.1; **p<0.05; ***p<0.01

Es decir, el modelo ajustado está dado por:

$$\hat{Y}_i = 5.9729 + 0.0964X_{1i} - 2.0875X_{2i} - 1.6863X_{3i} + 0.0018X_{4i} + 0.7868X_{5i} \quad \langle 2 \rangle$$

Ahora bien, la tabla **ANOVA** para este modelo es la siguiente:

```
## Analysis of Variance Table
##
## Response: Calidad
##
##              Df Sum Sq Mean Sq F value    Pr(>F)
## F0(Fija, Volatil, Citrico, Azucar, Cloruros)  5   7.495  1.49904   3.9969 0.002482
## Residuals                                94  35.255  0.37505
##
## F0(Fija, Volatil, Citrico, Azucar, Cloruros) **
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Entonces, si plantean las siguientes hipótesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \iff$ el modelo **no** es significativo.

$H_1 : \exists j : \beta_j \neq 0, j = 1, 2, 3, 4, 5 \iff$ el modelo **es** significativo.

Y para este test, si se toma un nivel de significancia de $\alpha = 0.05$ y se considera la tabla ANOVA anterior, el valor p asociado a esta prueba de hipótesis es $V_p = 0.0002482 < 0.05 = \alpha$, por lo que se rechaza la

hipótesis nula, esto es, hay evidencia muestral suficiente para sugerir que el modelo de regresión lineal múltiple planteado en la ecuación (2) **es significativo**.

Finalmente, como se pudo observar en la tabla uno, se obtuvo un $R^2 = 0.1753$, lo que quiere decir que el 17.53 % de la variabilidad de la calificación de calidad de un vino está explicado por el modelo de regresión lineal múltiple, el cual incluye a las variables de concentraciones de acidez fija, acidez volátil acidez cítrica, azúcares residuales y cloruros en el vino. Como se puede observar, este es un valor muy bajo y por tanto se tiene un modelo que no logra explicar adecuadamente la variabilidad de la calidad del vino.

Punto 4. Coeficientes de regresión estandarizados.

A continuación se muestra una tabla que exhibe el valor de los coeficientes estandarizados, esto es, despojándolos del efecto que puedan tener las unidades de cada uno de ellos:

##	Estimaci.on	Límites.2.5..	Límites.97.5..	Vif	Coef.Std
## (Intercept)	5.972901620	4.81264537	7.13315787	0.000000	0.000000000
## Fija	0.096392752	-0.03521555	0.22800105	1.317655	0.156355792
## Volatil	-2.087518622	-3.07030163	-1.10473561	1.748553	-0.522353100
## Citrico	-1.686348009	-2.70000167	-0.67269435	2.188026	-0.457652554
## Azucar	0.001826432	-0.08834673	0.09199959	1.009790	0.003785243
## Cloruros	0.786834740	-1.08080972	2.65447920	1.109613	0.082532929

De la tabla anterior se puede extraer que $|\beta_2| > |\beta_3| > |\beta_1| > |\beta_5| > |\beta_4|$, lo que significa que es la concentración de ácidos cítricos la variable que tiene mayor efecto en la calidad de los vinos según el modelo de regresión lineal múltiple planteado en [2].

Punto 5. Significancia individual.

Queremos probar la significancia individual de cada uno de los parámetros del modelo (excepto intercepto) para ello usaremos la prueba t; los resultados son los siguientes:

Parámetro	Estimación	Std. Error	T_0	$P(t > T_0)$	Test asociado
β_1	0.096393	0.066284	1.454	0.14921	$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$
β_2	-2.087519	0.494974	-4.217	5.68e-05	$H_0: \beta_2 = 0$ vs $H_A: \beta_2 \neq 0$
β_3	-1.686348	0.510522	-3.303	0.00135	$H_0: \beta_3 = 0$ vs $H_A: \beta_3 \neq 0$
β_4	0.001826	0.045415	0.040	0.96801	$H_0: \beta_4 = 0$ vs $H_A: \beta_4 \neq 0$
β_5	0.786835	0.940631	0.836	0.40500	$H_0: \beta_5 = 0$ vs $H_A: \beta_5 \neq 0$

Usando el hecho de que si el valor P es menor al nivel de significancia que establecimos como $\alpha = 0.05$, el estadístico de prueba t cae en la región de rechazo decretamos como criterio de rechazo el valor P.

- **Significancia de β_1 :** No hay evidencia suficiente para rechazar la hipótesis nula, por lo tanto el ácido fijo **no es significativo** para explicar la calidad del vino dado que las otras covariables están en el modelo.
- **Significancia de β_2 :** Hay evidencia suficiente para rechazar la hipótesis nula, por lo tanto el ácido volátil **es significativo** para explicar la calidad del vino dado que las otras covariables están en el modelo.
- **Significancia de β_3 :** Hay evidencia suficiente para rechazar la hipótesis nula, por lo tanto la ácido cítrico **es significativo** para explicar la calidad del vino dado que las otras covariables están en el modelo.

- **Significancia de β_4 :** No hay evidencia suficiente para rechazar la hipótesis nula, por lo tanto la azúcar residual **no es** significativa para explicar la calidad del vino dado que las otras covariables están en el modelo.
- **Significancia de β_5 :** No hay evidencia suficiente para rechazar la hipótesis nula, por lo tanto los cloruros **no son** significativos para explicar la calidad del vino dado que las otras covariables están en el modelo.

Punto 6. Prueba significancia simultánea.

Como en el numeral anterior llegamos a la conclusión de que β_1 , β_4 y β_5 no eran significativas de manera individual para explicar la calidad del vino, queremos probar si de manera conjunta siguen sin ser significativas y con esta información podemos considerar postular un nuevo modelo que contenga menos parámetros, esto puede ser conveniente ya que preferimos modelos parsimoniosos.

- Modelo reducido: $Y_i = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + E_i$, $E_i \sim N(0, \sigma^2)$
- Modelo completo: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + E_i$, $E_i \sim N(0, \sigma^2)$
- $H_0 = \begin{cases} \beta_1 = 0 \\ \beta_4 = 0 \\ \beta_5 = 0 \end{cases}$ vs $H_A = \begin{cases} \beta_1 \neq 0 \text{ ó} \\ \beta_4 \neq 0 \text{ ó} \\ \beta_5 \neq 0 \end{cases}$

Fuente	DF errores	SC residuos	Df(SSR parcial)	SSR parcial	F_0	$\Pr(f_{3,94} > F_0)$
Modelo Reducido (MR)	97	36.287				
Modelo Completo (MF)	94	35.255	3	1.0326	0.9177	0.4355
SSR parcial = SSE(MR)-SSE(MF)						

- El estadístico de prueba lo construimos así:

$$F_0 = \frac{SSR \text{ parcial}}{MSE(MF)} = \frac{SSE(MR) - SSE(MF)/g.l[SSE(MR)] - g.l[SSE(MF)]}{SSE(MF)/g.l[SSE(MF)]}$$

$$F_0 = \frac{36.287 - 35.255/97 - 94}{35.255/94} = \frac{1.0326/3}{35.255/94} = 0.9177$$

Recordemos que la distribución del estadístico es $F_0 \sim f_{g.l[SSE(MR)]-g.l[SSE(MF)], n-k-1}$, que en nuestro caso equivale a $F_0 \sim f_{3,94}$

- Calcularemos el valor P, con la ayuda de R así:

```
pf(0.9177,3,94, lower.tail = F)
```

```
## [1] 0.4355086
```

El valor P es mayor que el nivel de significancia que fijamos como $\alpha = 0.05$, por tanto el valor de nuestro estadístico de prueba no cae en la región de rechazo; no hay evidencia suficiente para rechazar H_0 , por lo tanto podemos decir que las variables *acidez fija*, *azúcar residual* y *cloruros* no ayudan a explicar la calidad de los vinos, dado que en el modelo están las variables *acidez volátil* y *ácido cítrico*.

Punto 7. Análisis de errores secuenciales y parciales

Fuente	SS1	Df	F_0	$\Pr(f_{1,94} > F_0)$	Test asociado
Fija	$\text{SSR}(X_1)=0.474$	1	1.2632	0.2639	$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$
Volatil	$\text{SSR}(X_2 X_1)=2.919$	1	7.7829	0.006386	$H_0: \beta_2 = 0$ vs $H_A: \beta_2 \neq 0$
Citrico	$\text{SSR}(X_3 X_1, X_2)=3.840$	1	10.2384	0.001876	$H_0: \beta_3 = 0$ vs $H_A: \beta_3 \neq 0$
Azucar	$\text{SSR}(X_4 X_1, X_2, X_3)=0.000$	1	0.0003	0.987118	$H_0: \beta_4 = 0$ vs $H_A: \beta_4 \neq 0$
Cloruros	$\text{SSR}(X_5 X_1, X_2, X_3, X_4)=0.262$	1	0.6997	0.404997	$H_0: \beta_5 = 0$ vs $H_A: \beta_5 \neq 0$
Error	$\text{SSE}(X_1, X_2, X_3, X_4, X_5)=35.255$	94			

Empezando por las sumas de cuadrados secuenciales (tipo 1), tenemos la anterior tabla, como podemos observar los menores valores para las sumas de cuadrados de tipo I son:

1. $\text{SSR}(X_4|X_1, X_2, X_3) = 0.000$
2. $\text{SSR}(X_5|X_1, X_2, X_3, X_4) = 0.262$
3. $\text{SS1}_{X_1} = 0.474$

La variable con menor valor en la suma de cuadrados en este caso es Azúcar, lo que significa que al añadir la variable Azúcar dado que las covariables Acidez fija, Acidez Volátil y Ácido cítrico están en el modelo, esta no ayuda a reducir en mayor medida la suma de cuadrados del error lo que no es lo mas conveniente, ya que lo que buscamos es que los residuales de nuestro nuevo modelo sea cada vez más cercano a cero; lo que puede ser un indicio de que la azúcar residual no es significativa para explicar la calidad del vino dado que las otras covariables mencionadas anteriormente están en el modelo; de manera similar sucede con las covariables Fija y Cloruros.

Fuente	SS2	Df	F_0	$\Pr(f_{1,94} > F_0)$	Test asociado
Fija	$\text{SSR}(X_1 X_2, X_3, X_4, X_5)=0.793$	1	2.1148	0.149211	$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$
Volatil	$\text{SSR}(X_2 X_1, X_3, X_4, X_5)=6.671$	1	17.7867	5.685e-05	$H_0: \beta_2 = 0$ vs $H_A: \beta_2 \neq 0$
Citrico	$\text{SSR}(X_3 X_1, X_2, X_4, X_5)=4.092$	1	10.9110	0.001353	$H_0: \beta_3 = 0$ vs $H_A: \beta_3 \neq 0$
Azucar	$\text{SSR}(X_4 X_1, X_2, X_3, X_5)=0.001$	1	0.0016	0.968006	$H_0: \beta_4 = 0$ vs $H_A: \beta_4 \neq 0$
Cloruros	$\text{SSR}(X_5 X_1, X_2, X_3, X_4)=0.262$	1	0.6997	0.404997	$H_0: \beta_5 = 0$ vs $H_A: \beta_5 \neq 0$
Error	$\text{SSE}(X_1, X_2, X_3, X_4, X_5)=35.255$	94			

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. $\text{SS2}_{X_1} = 0.793$

Para la suma de cuadrados parciales el significado es el mismo lo que cambia esta vez son las covariables que ya están en el modelo antes de agregar la variable de interes, asi las conclusiones que podemos sacar usando la tabla Anova son (usando un nivel de significancia de $\alpha = 0.05$):

- Para la variable Fija no hay evidencia suficiente para rechazar la hipótesis nula, lo que implica que Acidez Fija no es significativa para explicar la calidad del vino dado que las covariables Acidez Volátil, Ácido cítrico, azúcar y cloruros estan en el modelo.
- Para la variable Azúcar no hay evidencia suficiente para rechazar la hipótesis nula, lo que implica que la azúcar residual no es significativa para explicar la calidad del vino dado que las covariables Acidez fija, Acidez Volátil, Ácido cítrico, y cloruros estan en el modelo.
- Para la variable Cloruros no hay evidencia suficiente para rechazar la hipótesis nula, lo que implica que los cloruros no son significativos para explicar la calidad del vino dado que las covariables Acidez fija, Acidez Volátil, Ácido cítrico, y azúcar están en el modelo.


```
## 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. $SS1_{X_4} = 0.000$ 2. $SS1_{X_5} = 0.262$ 3. $SS1_{X_1} = 0.474$ Nuestra tabla anova tambien nos dice que: $SSR(X_4|X_1, X_2, X_3) = 0.000$ $SSR(X_5|X_1, X_2, X_3, X_4) = 0.262$ $SSR(X_1) = 0.474$, 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
```

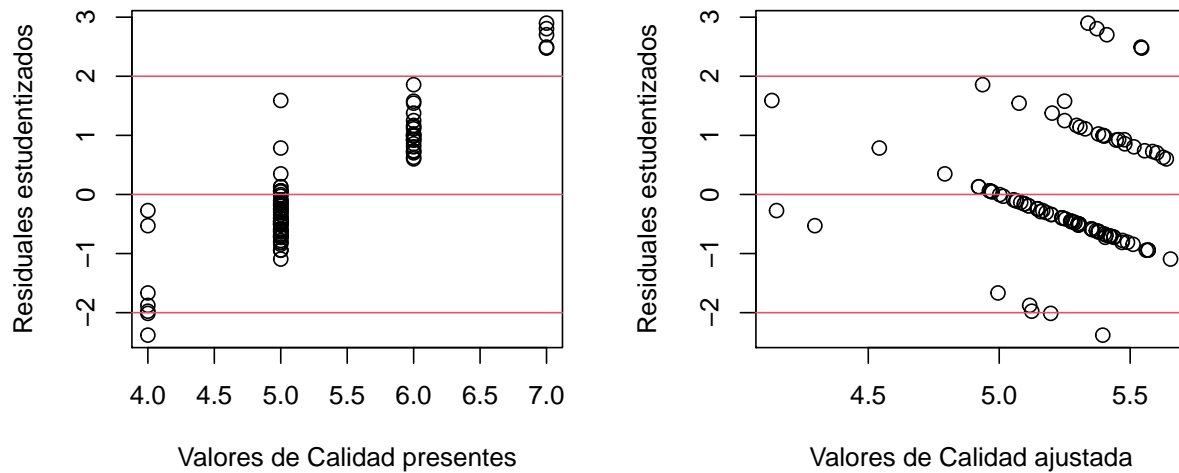
Como podemos observar en la tabla anterior los menores valores para las sumas de cuadrados de tipo II son:

1. $SS2_{X_4} = 0.001$
2. $SS2_{X_5} = 0.262$
3. $SS2_{X_1} = 0.793$

Cada valor nos dice el SSR de cada variable en el modelo completo dadas las demás (ej. $SSR(X_4|X_1, X_2, X_3, X_4) = 0.001$), 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.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + E_i, \text{ con } E \sim N(0, \sigma^2) \text{ } H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta \neq 0$$

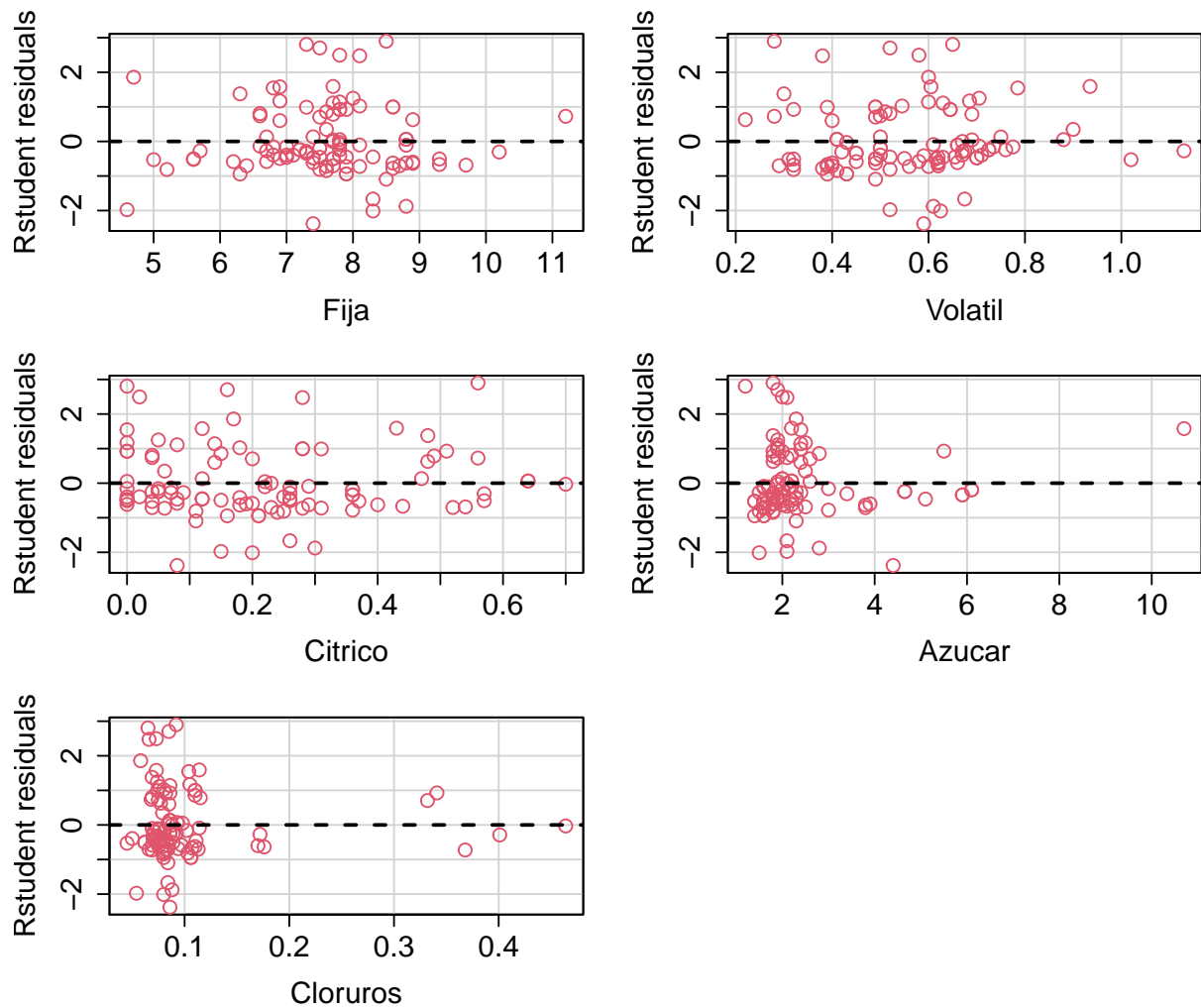
Punto 8. Gráficos de los residuales estudentizados vs. Valores ajustados y contra las variables de regresión utilizadas.



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 influenciadoras.

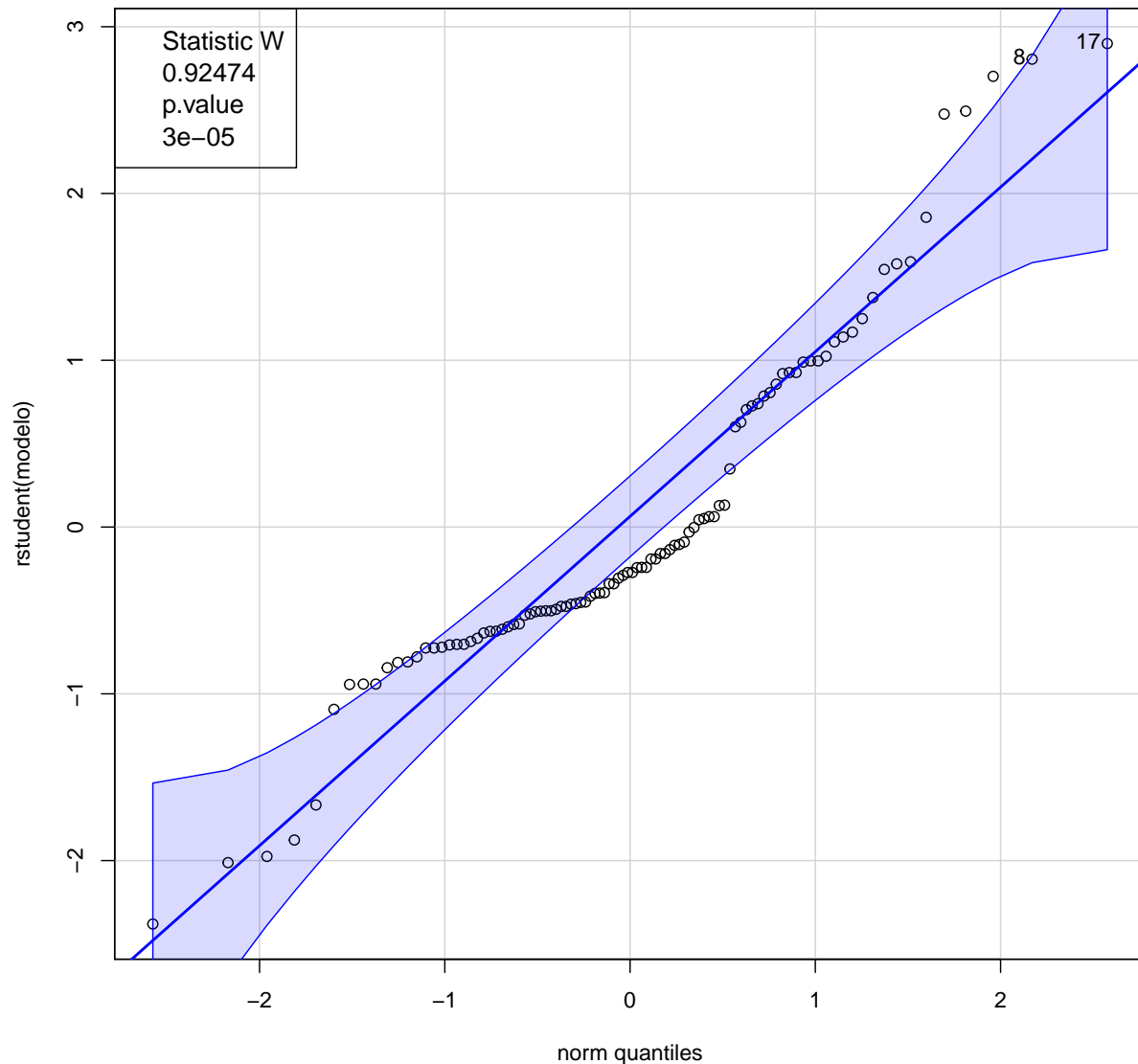
En la gráfica de valores de la calidad ajustada por el modelo vs. residuales estudentizados podemos ver que tiene un efecto similar, pero en este los residuales están un poco más centrados.



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 de los residuales.

Punto 9. Gráfica de probabilidad normal para los residuales estu-
dentizados. ¿Existen razones para dudar de la hipótesis de normal-
idad 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 10. Presencia de observaciones atípicas, de balanceo y/o influyentes.

Para este punto, vale la pena comenzar revisando la tabla que indica qué observaciones pueden ser influyentes de acuerdo a diferentes medidas, la cual puede ser consultada haciendo *[clic aquí](#)*. En esta se pueden

ver qué observaciones son catalogadas como tal según varios criterios revisando cuáles tienen una letra T.

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

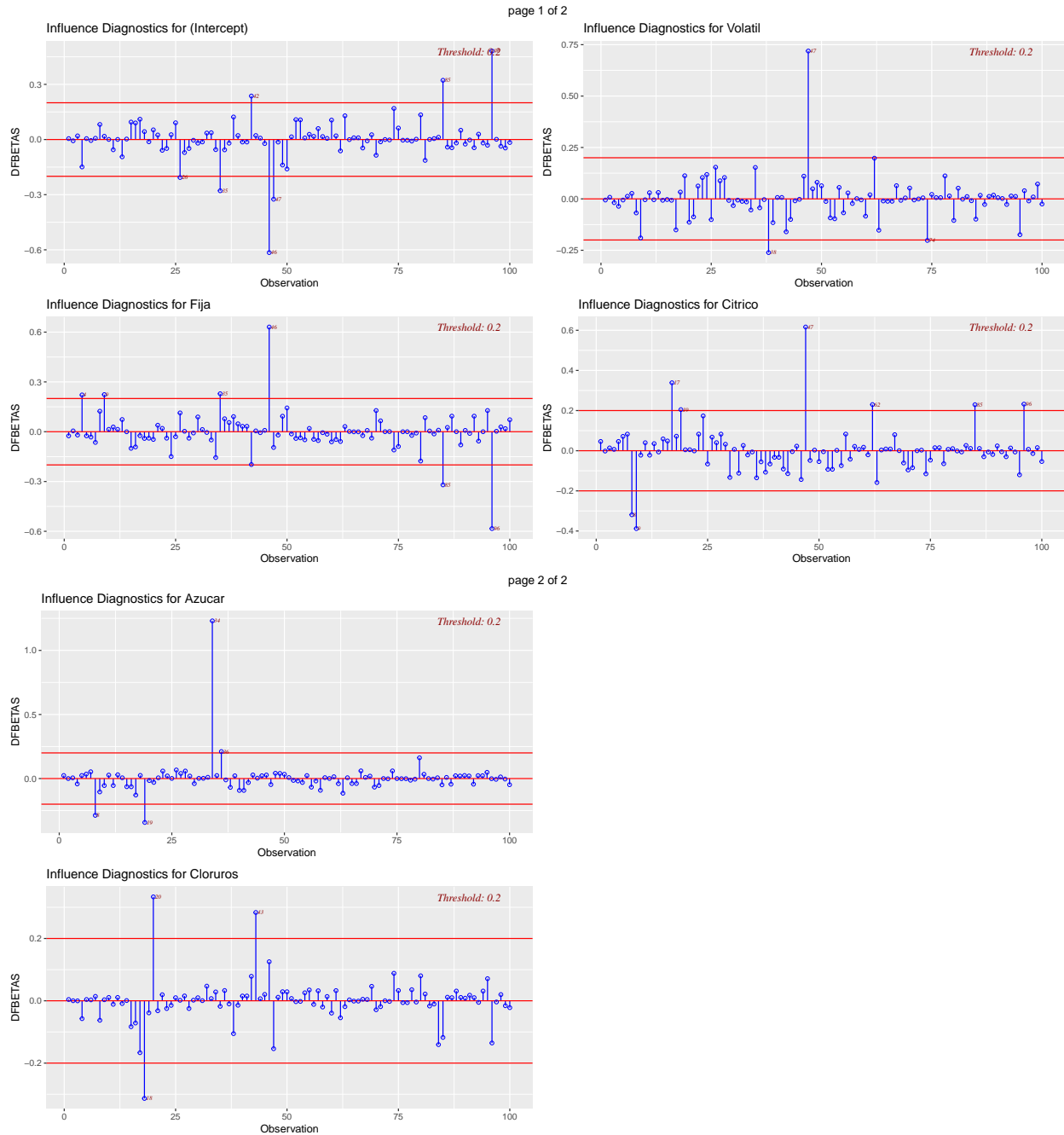
Según la medida DFBetas, los datos influenciados son: 34

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

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

Según la Distancia de Cook ningún dato es influenciado.

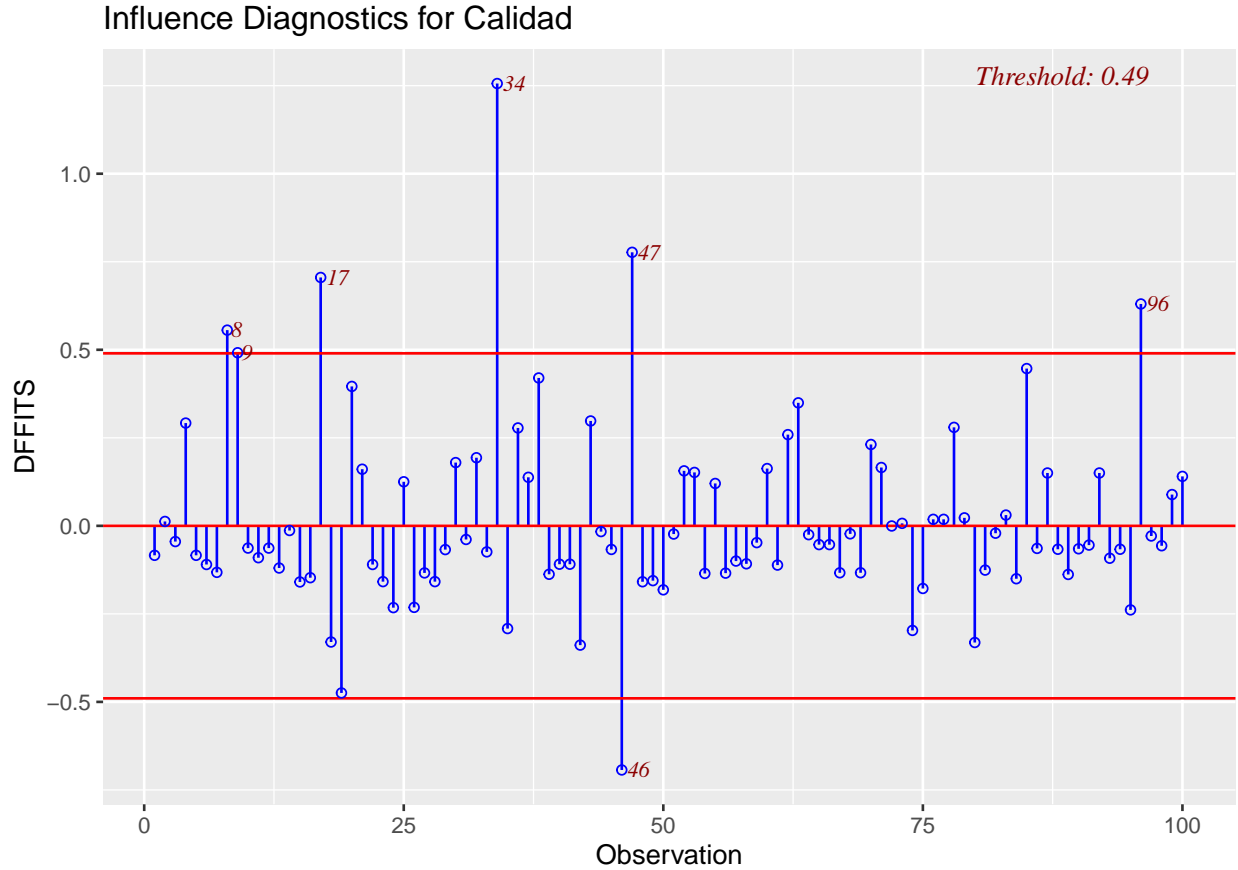
DFBETAS



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

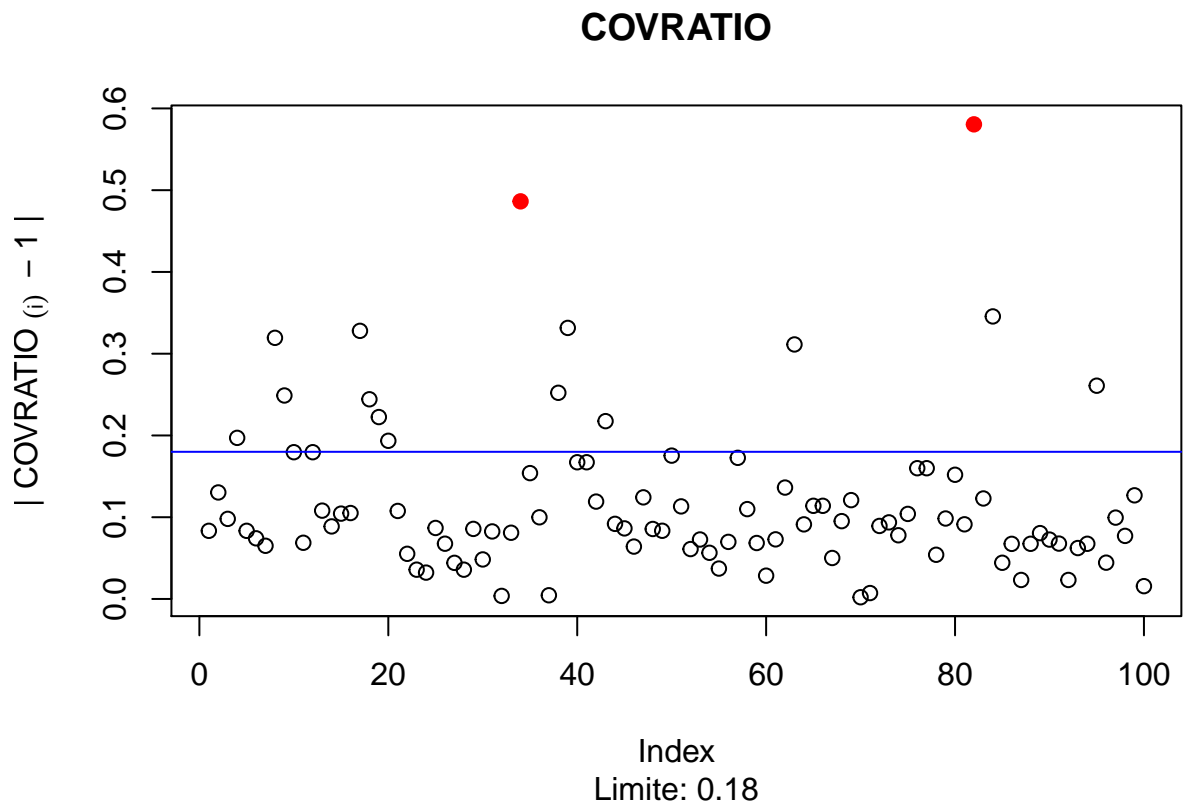
DFFITS

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

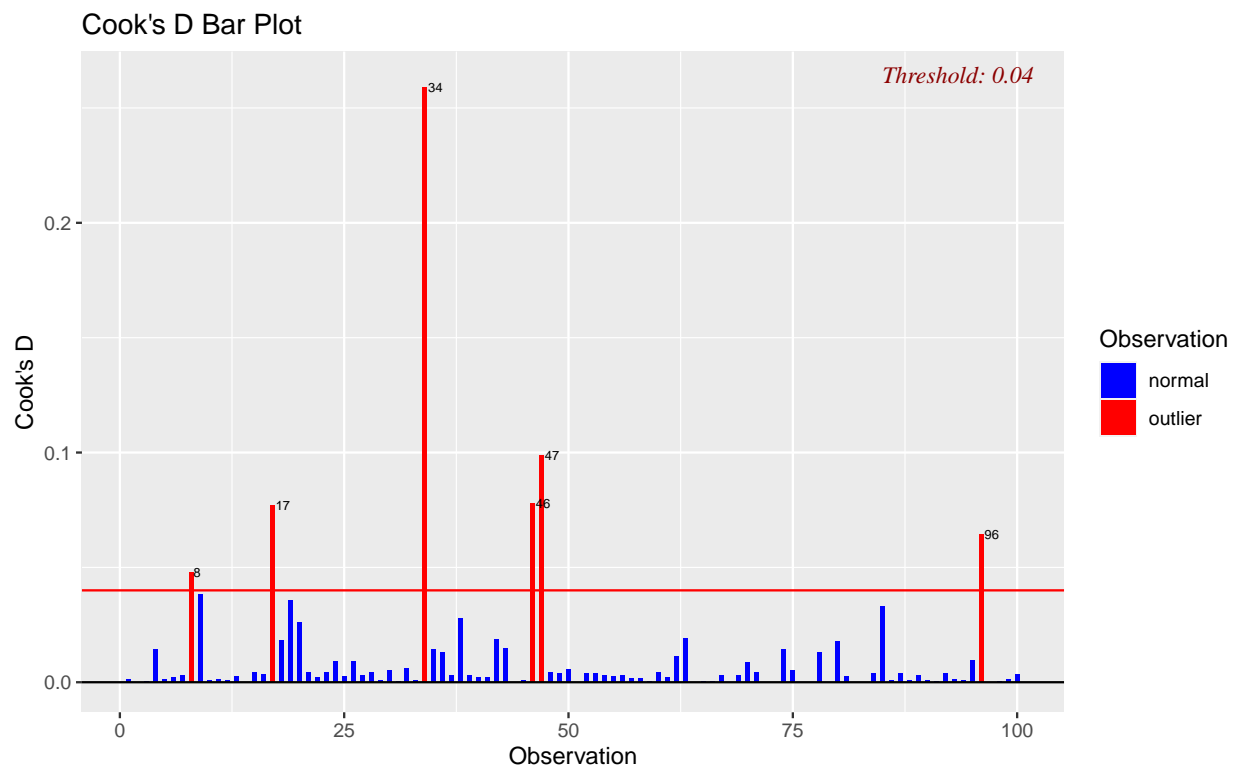


COVRATIO

Como podemos ver en la gráfica que hay varios datos que superan el límite fijado, recordemos que una observación es candidata a ser influyente si $|COVRATIO_i - 1| > 3(k+1)/n$; usamos R y encontramos los datos que cumplen la condición, a continuación los datos potencialmente influyentes y su $COVRATIO$, destacamos las observaciones 34 y 82, ya que, según las medidas DFFITS y DFBetas vistas anteriormente, el dato 34 es candidato a ser influyente y el dato 82 es el que tiene un valor mayor en la desigualdad:

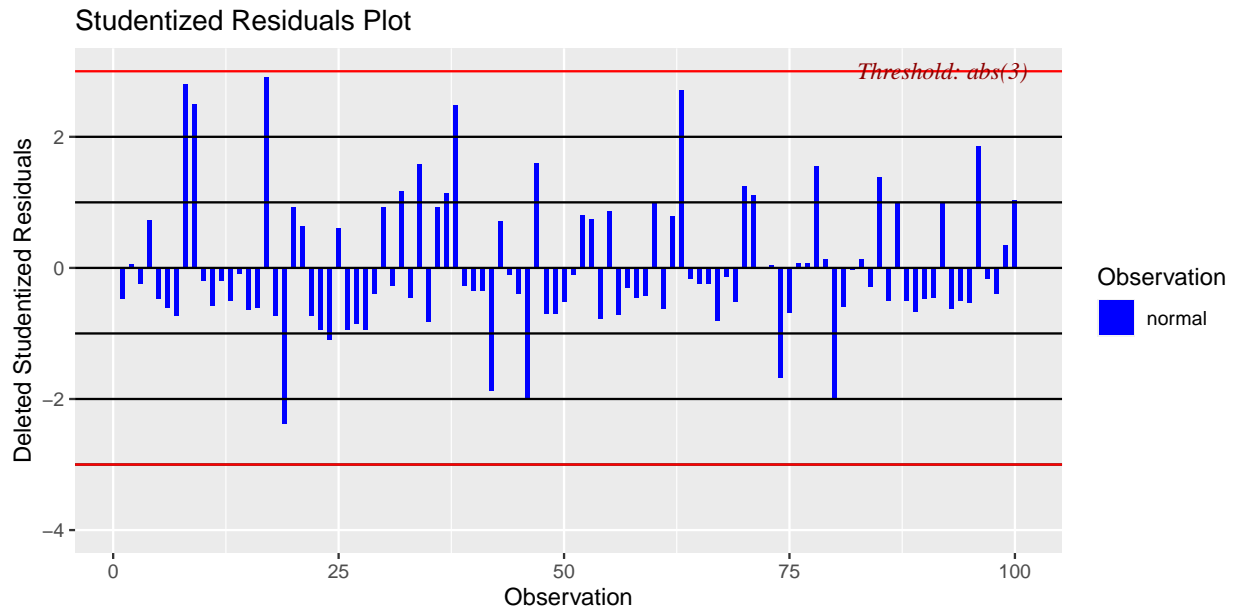


DISTANCIA DE COOK

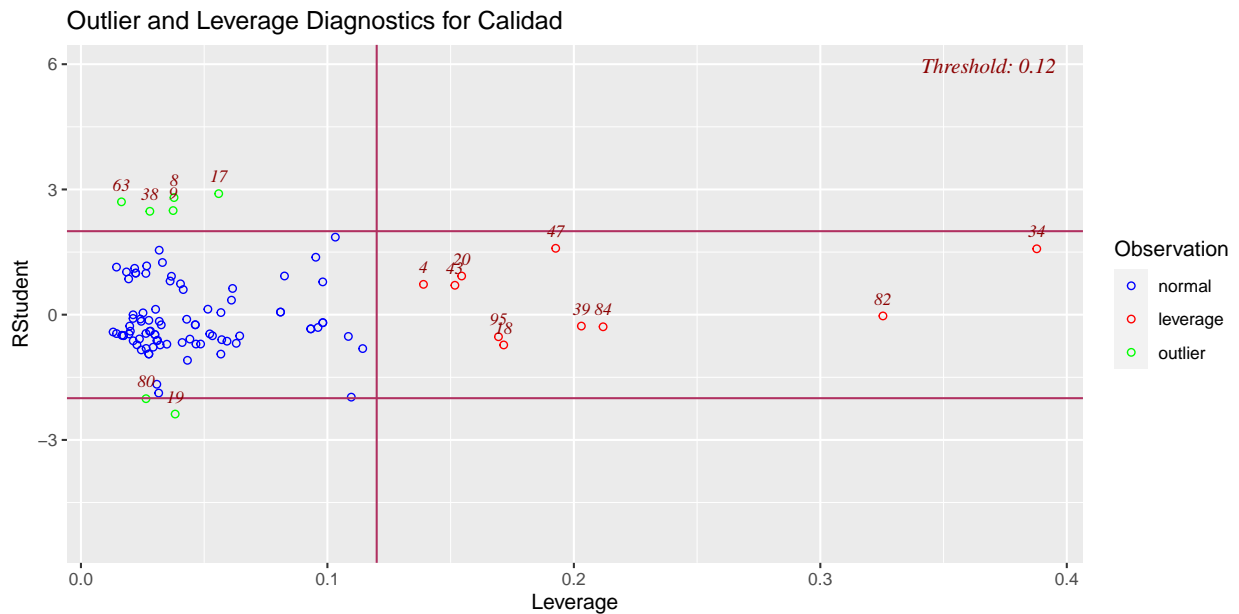


Validamos y encontramos que efectivamente ningún valor sobrepasa 1, pero podemos ver que la observación 34 está demasiado lejos de las demás, por lo que la tendremos en cuenta.

Residuales estudentizados (internamente estudentizados)



En este gráfico podemos ver que ninguna observación sobrepasa el límite 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 límite diferente, según hemos visto en clase el límite que fijamos es ± 3 , con el cual obtendríamos que ningún dato es una observación atípica, pero esta gráfica usa otro límite, que causa que sean varios los datos que podrían ser observaciones atípicas.

En la gráfica anterior también 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 límite utilizado $h_{ii} > 2(k+1)/n$, que equivale a $h_{ii} > 0.12$, como se puede observar en la gráfica.

Como conclusión podemos decir que las observaciones 82 y 34 son puntos de balanceo y además influenciados, de acuerdo a varios diagnósticos usados y explicados.

```
## Rows: 1599 Columns: 12

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

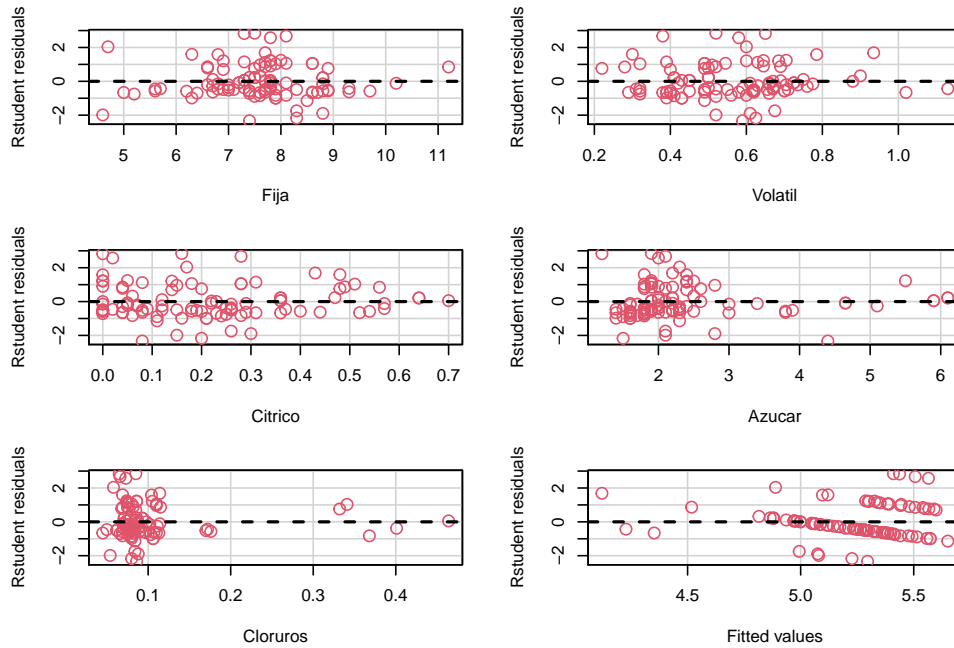
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

Punto 11.

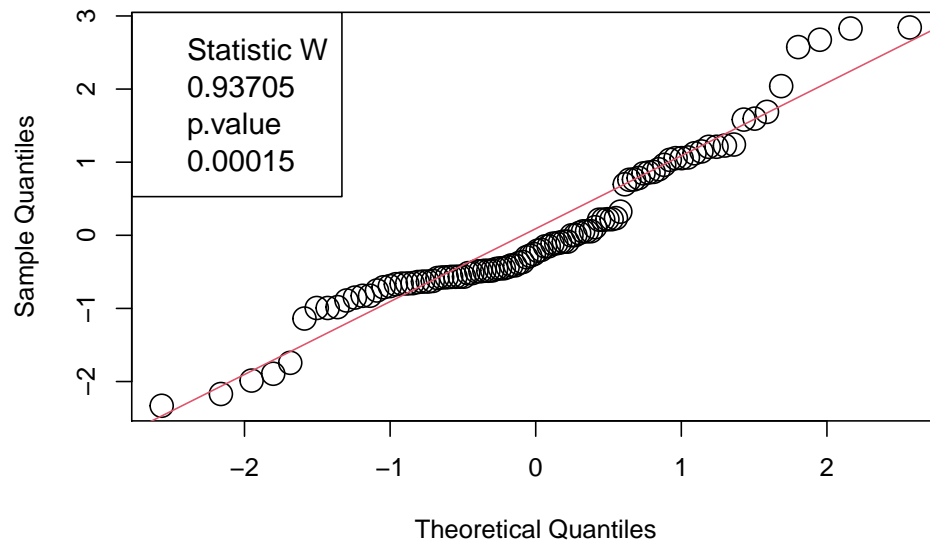
- *Ajuste el modelo de regresión sin las observaciones 17 y 34, suponga que se establece que hay un error de digitación con estas dos observaciones, presente sólo la tabla de parámetros ajustados resultante ¿Cambian notoriamente las estimaciones de los parámetros, sus errores estándar y/o la significancia? ¿Qué concluye al respecto? Evalúe el gráfico de normalidad para los residuales estudentizados para este ajuste ¿mejoró la normalidad? Concluya sobre los efectos de este par de observaciones.*

```
##
## Call:
## lm(formula = Calidad ~ ., data = datos2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2959 -0.3336 -0.1336  0.4274  1.5897
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.94188    0.55924  10.625 < 2e-16 ***
## Fija         0.10754    0.06366   1.689 0.094535 .
## Volatil     -1.99180    0.47394  -4.203 6.1e-05 ***
## Citrico     -1.83987    0.49129  -3.745 0.000314 ***
## Azucar      -0.04494    0.05512  -0.815 0.416997
## Cloruros     0.91020    0.90062   1.011 0.314839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5853 on 92 degrees of freedom
## Multiple R-squared:  0.1932, Adjusted R-squared:  0.1493
## F-statistic: 4.406 on 5 and 92 DF,  p-value: 0.001216
```

- En la anterior tabla se puede observar que, la estimación de los parámetros β_j no presentan cambios notorios, excepto para β_4 , pasa de 0.001826 (signo positivo) considerando las observaciones 17 y 34, a -0.04494 (signo negativo) sin considerar las observaciones 17 y 34. Ahora, para los errores estándar se puede observar que no cambian notoriamente sin considerar las observaciones 17 y 34, sin embargo, para la significancia, si se toma un nivel de significancia de 10%, $\alpha = 0.1$, se tiene que β_1 es significativo, pues $0.094535 < 0.1$, mientras que considerando las observaciones 17 y 34 esto no pasa, pues $0.14921 > 0.1$.



Normal Q-Q Plot



- En la anterior gráfica se puede observar que, sin considerar las observaciones 17 y 34, esto no mejora la normalidad para los residuales estudentizados, pues el valor p es muy pequeño < 0.05 , por lo que se rechaza la hipótesis nula H_0 : los residuales estudentizados se distribuyen como una normal.
- ¿Cuál sería el efecto de estas 2 observaciones?

El efecto de las observaciones 17 y 34 en el modelo es de tipo *influenciable*, puesto que cuando estas no se incluyen en el modelo, el valor del coeficiente β_4 , asociado a los cloruros cambia de signo.

Punto 12. Diagnóstico de multicolinealidad.

- Para el modelo con todas las variables y sin las observaciones 17 y 34, realice diagnósticos de multicolinealidad mediante XXXX.

Literal A. Matriz de correlación de las variables predictoras

```
##           Fija      Volatil      Citrico      Azucar      Cloruros
## Fija      1.00000000 -0.30395236  0.48008494  0.095188579  0.095194155
## Volatil  -0.30395236  1.00000000 -0.626393071  0.022231186 -0.028960549
## Citrico   0.48008494 -0.62639307  1.000000000  0.006279926  0.257846311
## Azucar    0.09518858  0.02223119  0.006279926  1.000000000 -0.059467342
## Cloruros  0.09519416 -0.02896055  0.257846311 -0.059467342  1.000000000
## Calidad   0.09399282 -0.25265651 -0.077737119 -0.080289809  0.004570133
##           Calidad
## Fija      0.093992822
## Volatil  -0.252656514
## Citrico  -0.077737119
## Azucar   -0.080289809
## Cloruros  0.004570133
## Calidad  1.000000000
```

- Matriz de correlaciones: Se detecta una asociación lineal alta entre las variables cítrico y volátil, con un valor de -0.626393071.

Literal B. VIFs

```
## Coeficientes estimados, sus I.C, Vifs y Coeficientes estimados estandarizados
```

```
##           Estimación Límites.2.5.. Límites.97.5..      Vif      Coef.Std
## (Intercept)  5.94188241      4.8311878      7.0525770  0.0000000  0.000000000
## Fija         0.10753755     -0.0188877      0.2339628  1.315018  0.18141877
## Volatil     -1.99179719     -2.9330746     -1.0505198  1.701883 -0.51343135
## Citrico     -1.83986936     -2.8156132     -0.8641255  2.128785 -0.51169020
## Azucar      -0.04493921     -0.1544098      0.0645314  1.016979 -0.07699702
## Cloruros     0.91020023     -0.8785036      2.6989040  1.111713  0.09978949
```

- Con los valores VIFs: no se observa valores superando la cota de 10. Por este método no se detecta multicolinealidad-

Literal C. Proporciones de varianza

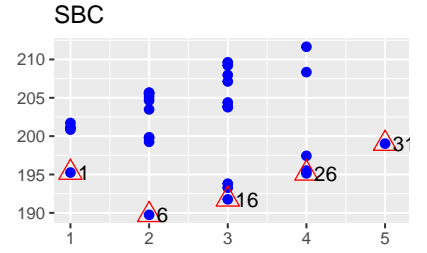
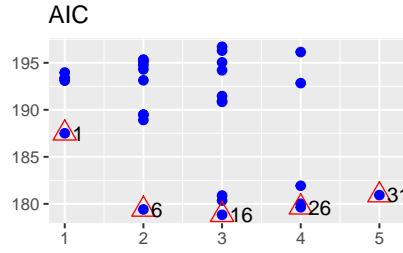
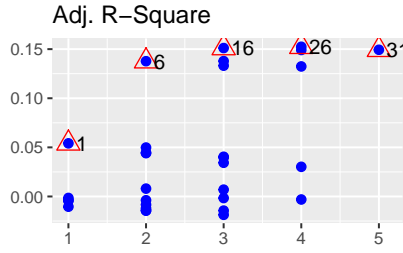
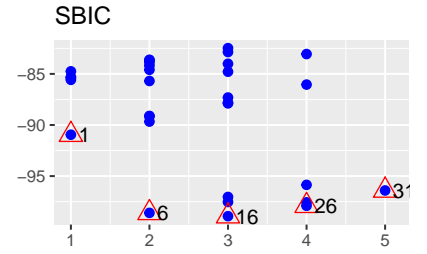
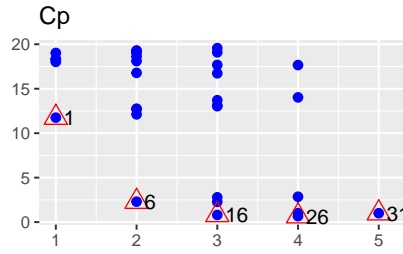
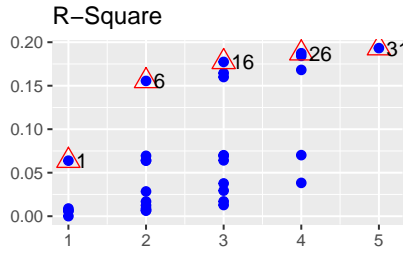
```
## Condition
## Index      Variance Decomposition Proportions
##           intercept Fija  Volatil Citrico Azucar Cloruros
## 1    1.000  0.000      0.001 0.002   0.005  0.005  0.008
## 2    3.503  0.001      0.000 0.022   0.288  0.027  0.040
## 3    4.409  0.000      0.001 0.001   0.087  0.076  0.789
## 4    6.431  0.007      0.008 0.051   0.016  0.868  0.146
## 5   13.796  0.046      0.184 0.767   0.599  0.023  0.016
## 6   27.141  0.946      0.807 0.158   0.005  0.001  0.001
```

- Con las proporciones de descomposición de varianza: se puede observar que, en la quinta fila, π_{52} y π_{53} superan 0.5, y no existe otra fila i donde 2 π_{ij} superen esta cota, luego, con estos índices se detecta que volátil y cítrico están involucradas en una relación de multicolinealidad.

Punto 13. Modelos de regresión con métodos de selección.

- Sin las observaciones 17 y 34, construya modelos de regresión utilizando los métodos de selección (muestre de cada método sólo la tabla de resumen de este y la tabla ANOVA y la de parámetros estimados del modelo finalmente resultante).

##	mindex	n	predictors	rsquare	adjr	cp
## 2	1	1	Volatil	0.06383531	0.054084	12.750
## 1	2	1	Fija	0.00883465	-0.001490	19.022
## 4	3	1	Azucar	0.00644645	-0.003903	19.294
## 3	4	1	Citrico	0.00604306	-0.004311	19.340
## 5	5	1	Cloruros	0.00002089	-0.010396	20.027
## 10	6	2	Volatil Citrico	0.15549564	0.137717	4.298
## 11	7	2	Volatil Azucar	0.06941412	0.049823	14.114
## 6	8	2	Fija Volatil	0.06416116	0.044459	14.713
## 12	9	2	Volatil Cloruros	0.06384287	0.044134	14.749
## 7	10	2	Fija Citrico	0.02845080	0.007997	18.785
## 8	11	2	Fija Azucar	0.01687068	-0.003827	20.105
## 13	12	2	Citrico Azucar	0.01241161	-0.008380	20.614
## 9	13	2	Fija Cloruros	0.00885399	-0.012012	21.020
## 14	14	2	Citrico Cloruros	0.00669208	-0.014220	21.266
## 15	15	2	Azucar Cloruros	0.00644650	-0.014470	21.294
## 16	16	3	Fija Volatil Citrico	0.17736041	0.151106	3.805
## 23	17	3	Volatil Citrico Cloruros	0.16452023	0.137856	5.269
## 22	18	3	Volatil Citrico Azucar	0.15996648	0.133157	5.788
## 17	19	3	Fija Volatil Azucar	0.07010037	0.040423	16.036
## 24	20	3	Volatil Azucar Cloruros	0.06946534	0.039767	16.108
## 18	21	3	Fija Volatil Cloruros	0.06418056	0.034314	16.711
## 19	22	3	Fija Citrico Azucar	0.03767912	0.006967	19.733
## 20	23	3	Fija Citrico Cloruros	0.02938369	-0.001593	20.679
## 21	24	3	Fija Azucar Cloruros	0.01698351	-0.014389	22.093
## 25	25	3	Citrico Azucar Cloruros	0.01283068	-0.018675	22.566
## 27	26	4	Fija Volatil Citrico Cloruros	0.18735973	0.152407	4.665
## 26	27	4	Fija Volatil Citrico Azucar	0.18423199	0.149145	5.021
## 30	28	4	Volatil Citrico Azucar Cloruros	0.16816091	0.132383	6.854
## 28	29	4	Fija Volatil Azucar Cloruros	0.07019572	0.030204	18.025
## 29	30	4	Fija Citrico Azucar Cloruros	0.03829514	-0.003069	21.662
## 31	31	5	Fija Volatil Citrico Azucar Cloruros	0.19318929	0.149341	6.000



Literal A. Selección según el R_{adj}^2

Según el R_{adj}^2 , los mejores modelos son el 6, 16, 26 y 31, y como estos 3 últimos no muestran un incremento significativo en este estadístico, con respecto al modelo 6, entonces aplicando el principio de parsimonia, se escogería el modelo 6:

$$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + E_i, \quad E_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

.

Literal B. Selección según el estadístico C_p

Teniendo en cuenta que con este estadístico se busca que el modelo con el menor valor $|C_p - p|$, los mejores candidatos son el modelo 6: $|C_p - p| = |4.298177 - 3| = 1.298177$, el modelo 16: $|C_p - p| = |3.804954 - 4| = 0.195046$, el modelo 26: $|C_p - p| = |4.664740 - 5| = 0.33526$ y el modelo 31: $|C_p - p| = |6 - 6| = 0$, pero de acuerdo con la ecuación

$$C_p = \frac{SSE_p}{MSE(X_1, X_2, \dots, X_k)} - (n - 2p)$$

, esto siempre ocurre con el modelo con todas las variables, por lo tanto, teniendo en cuenta que el modelo 16 tiene el valor más pequeño, entonces por este criterio se selecciona el modelo 16:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + E_i$$

, $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$.

Literal C. Stepwise

```
## Stepwise Selection Method
## -----
##
## Candidate Terms:
##
## 1. Fija
```

```

## 2. Volatil
## 3. Citrico
## 4. Azucar
## 5. Cloruros
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - Volatil added
##
##
##           Model Summary
## -----
## R                0.253      RMSE                0.617
## R-Squared        0.064      Coef. Var            11.813
## Adj. R-Squared   0.054      MSE                 0.381
## Pred R-Squared   0.028      MAE                 0.460
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##
##           ANOVA
## -----
##           Sum of
##           Squares      DF      Mean Square      F      Sig.
## -----
## Regression      2.493        1          2.493      6.546      0.0121
## Residual       36.568       96          0.381
## Total          39.061       97
## -----
##
##
##           Parameter Estimates
## -----
##           model      Beta      Std. Error      Std. Beta      t      Sig      lower      upper
## -----
## (Intercept)    5.776        0.224                25.732      0.000      5.331      6.222
## Volatil       -0.980        0.383        -0.253      -2.559      0.012     -1.741     -0.220
## -----
##
##
## Stepwise Selection: Step 2
##
## - Citrico added
##
##
##           Model Summary
## -----
## R                0.394      RMSE                0.589
## R-Squared        0.155      Coef. Var            11.279
## Adj. R-Squared   0.138      MSE                 0.347
## Pred R-Squared   0.105      MAE                 0.446
## -----

```

RMSE: Root Mean Square Error
 ## MSE: Mean Square Error
 ## MAE: Mean Absolute Error

##

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	6.074	2	3.037	8.746	3e-04
Residual	32.987	95	0.347		
Total	39.061	97			

##

##

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	6.608	0.336		19.654	0.000	5.941	7.276
Volatil	-1.924	0.469	-0.496	-4.100	0.000	-2.855	-0.992
Citrico	-1.397	0.435	-0.388	-3.211	0.002	-2.260	-0.533

##

##

##

##

Model Summary

R	0.394	RMSE	0.589
R-Squared	0.155	Coef. Var	11.279
Adj. R-Squared	0.138	MSE	0.347
Pred R-Squared	0.105	MAE	0.446

##

RMSE: Root Mean Square Error

MSE: Mean Square Error

MAE: Mean Absolute Error

##

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	6.074	2	3.037	8.746	3e-04
Residual	32.987	95	0.347		
Total	39.061	97			

##

##

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	6.608	0.336		19.654	0.000	5.941	7.276
Volatil	-1.924	0.469	-0.496	-4.100	0.000	-2.855	-0.992
Citrico	-1.397	0.435	-0.388	-3.211	0.002	-2.260	-0.533

##

```

##
##
##
## No more variables to be added/removed.
##
##
## Final Model Output
## -----
##
##                               Model Summary
## -----
## R                0.394          RMSE                0.589
## R-Squared        0.155          Coef. Var            11.279
## Adj. R-Squared   0.138          MSE                0.347
## Pred R-Squared   0.105          MAE                0.446
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##                               ANOVA
## -----
##                Sum of
##                Squares      DF      Mean Square      F      Sig.
## -----
## Regression        6.074         2          3.037      8.746      3e-04
## Residual         32.987        95          0.347
## Total            39.061        97
## -----
##
##                               Parameter Estimates
## -----
##      model      Beta      Std. Error      Std. Beta      t      Sig.      lower      upper
## -----
## (Intercept)    6.608         0.336          19.654      0.000      5.941      7.276
## Volatil       -1.924         0.469          -0.496     -4.100      0.000     -2.855     -0.992
## Citrico        -1.397         0.435          -0.388     -3.211      0.002     -2.260     -0.533
## -----
##
##                               Stepwise Selection Summary
## -----
##      Step      Variable      Added/      R-Square      Adj.      C(p)      AIC      RMSE
##      Step      Variable      Removed      R-Square      R-Square      C(p)      AIC      RMSE
## -----
##      1      Volatil      addition      0.064      0.054      12.7500      187.5034      0.6172
##      2      Citrico      addition      0.155      0.138      4.2980      179.4054      0.5893
## -----

```

Según el método *stepwise*, el modelo a usar es el modelo 6:

$$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + E_i$$

$$, E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2).$$

Literal D. Selección hacia adelante o forward

```
## Forward Selection Method
## -----
##
## Candidate Terms:
##
## 1. Fija
## 2. Volatil
## 3. Citrico
## 4. Azucar
## 5. Cloruros
##
## We are selecting variables based on p value...
##
## Forward Selection: Step 1
##
## - Volatil
##
##                               Model Summary
## -----
## R                0.253      RMSE                0.617
## R-Squared        0.064      Coef. Var            11.813
## Adj. R-Squared   0.054      MSE                 0.381
## Pred R-Squared   0.028      MAE                 0.460
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##                               ANOVA
## -----
##                Sum of      DF      Mean Square      F      Sig.
##                Squares
## -----
## Regression      2.493        1          2.493      6.546    0.0121
## Residual       36.568       96          0.381
## Total          39.061       97
## -----
##
##                               Parameter Estimates
## -----
##                model      Beta      Std. Error      Std. Beta      t      Sig      lower      upper
## -----
## (Intercept)     5.776        0.224                25.732     0.000      5.331      6.222
## Volatil        -0.980        0.383         -0.253     -2.559     0.012     -1.741     -0.220
## -----
##
##
## Forward Selection: Step 2
##
## - Citrico
```

```

##
##                               Model Summary
## -----
## R                               0.394          RMSE              0.589
## R-Squared                       0.155          Coef. Var        11.279
## Adj. R-Squared                   0.138          MSE              0.347
## Pred R-Squared                   0.105          MAE              0.446
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##                               ANOVA
## -----
##                               Sum of
##                               Squares      DF      Mean Square      F      Sig.
## -----
## Regression      6.074          2          3.037      8.746      3e-04
## Residual        32.987         95          0.347
## Total           39.061         97
## -----
##
##                               Parameter Estimates
## -----
##                               model      Beta      Std. Error      Std. Beta      t      Sig      lower      upper
## -----
## (Intercept)      6.608          0.336          19.654      0.000      5.941      7.276
## Volatil           -1.924          0.469          -0.496      -4.100      0.000      -2.855      -0.992
## Citrico           -1.397          0.435          -0.388      -3.211      0.002      -2.260      -0.533
## -----
##
##
##
## No more variables to be added.
##
## Variables Entered:
##
## + Volatil
## + Citrico
##
##
## Final Model Output
## -----
##
##                               Model Summary
## -----
## R                               0.394          RMSE              0.589
## R-Squared                       0.155          Coef. Var        11.279
## Adj. R-Squared                   0.138          MSE              0.347
## Pred R-Squared                   0.105          MAE              0.446
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error

```

```
##
##                               ANOVA
## -----
##           Sum of
##           Squares           DF      Mean Square      F      Sig.
## -----
## Regression      6.074           2           3.037      8.746      3e-04
## Residual       32.987          95           0.347
## Total          39.061          97
## -----
##
##                               Parameter Estimates
## -----
##           model      Beta      Std. Error      Std. Beta      t      Sig      lower      upper
## -----
## (Intercept)      6.608           0.336           19.654      0.000      5.941      7.276
## Volatil          -1.924           0.469           -0.496     -4.100      0.000     -2.855     -0.992
## Citrico          -1.397           0.435           -0.388     -3.211      0.002     -2.260     -0.533
## -----
##
##                               Selection Summary
## -----
##           Variable      Adj.
## Step   Entered      R-Square      R-Square      C(p)      AIC      RMSE
## -----
## 1      Volatil      0.0638      0.0541      12.7501      187.5034      0.6172
## 2      Citrico      0.1555      0.1377      4.2982      179.4054      0.5893
## -----
```

Según el método *forward*, nuevamente, el modelo seleccionado es el modelo seis.

Literal E. Selección hacia atrás o backward

```
## Backward Elimination Method
## -----
##
## Candidate Terms:
##
## 1 . Fija
## 2 . Volatil
## 3 . Citrico
## 4 . Azucar
## 5 . Cloruros
##
## We are eliminating variables based on p value...
##
## - Azucar
##
## Backward Elimination: Step 1
##
## Variable Azucar Removed
##
```

```

##                               Model Summary
## -----
## R                               0.433      RMSE                0.584
## R-Squared                       0.187      Coef. Var          11.182
## Adj. R-Squared                  0.152      MSE                 0.341
## Pred R-Squared                  0.104      MAE                 0.439
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##                               ANOVA
## -----
##                               Sum of
##                               Squares      DF      Mean Square      F      Sig.
## -----
## Regression      7.319          4          1.830      5.36      6e-04
## Residual       31.743         93          0.341
## Total          39.061         97
## -----
##
##                               Parameter Estimates
## -----
##                               model      Beta      Std. Error      Std. Beta      t      Sig      lower      upper
## -----
## (Intercept)    5.881          0.553              10.631      0.000      4.782      6.979
## Fija           0.102          0.063              1.617      0.109     -0.023      0.228
## Volatil       -2.009          0.473             -0.518     -4.252      0.000     -2.948     -1.071
## Citrico       -1.841          0.490             -0.512     -3.755      0.000     -2.815     -0.867
## Cloruros       0.960          0.897              0.105      1.070      0.288     -0.822      2.741
## -----
##
##
## - Cloruros
##
## Backward Elimination: Step 2
##
## Variable Cloruros Removed
##
##                               Model Summary
## -----
## R                               0.421      RMSE                0.585
## R-Squared                       0.177      Coef. Var          11.191
## Adj. R-Squared                  0.151      MSE                 0.342
## Pred R-Squared                  0.104      MAE                 0.445
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##                               ANOVA
## -----
##                               Sum of
##                               Squares      DF      Mean Square      F      Sig.

```

```

## -----
## Regression      6.928      3      2.309      6.755      4e-04
## Residual       32.133     94      0.342
## Total          39.061     97
## -----
##
##                               Parameter Estimates
## -----
##      model      Beta      Std. Error      Std. Beta      t      Sig      lower      upper
## -----
## (Intercept)    5.911      0.553              10.691    0.000      4.813      7.009
##      Fija      0.100      0.063      0.169      1.581    0.117     -0.026      0.225
##      Volatil  -1.920      0.466     -0.495     -4.125    0.000     -2.845     -0.996
##      Citrico  -1.685      0.469     -0.469     -3.597    0.001     -2.616     -0.755
## -----
##
##
## - Fija
##
## Backward Elimination: Step 3
##
## Variable Fija Removed
##
##                               Model Summary
## -----
## R              0.394      RMSE              0.589
## R-Squared      0.155      Coef. Var      11.279
## Adj. R-Squared 0.138      MSE              0.347
## Pred R-Squared 0.105      MAE              0.446
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##                               ANOVA
## -----
##      Sum of
##      Squares      DF      Mean Square      F      Sig.
## -----
## Regression      6.074      2      3.037      8.746      3e-04
## Residual       32.987     95      0.347
## Total          39.061     97
## -----
##
##                               Parameter Estimates
## -----
##      model      Beta      Std. Error      Std. Beta      t      Sig      lower      upper
## -----
## (Intercept)    6.608      0.336              19.654    0.000      5.941      7.276
##      Volatil  -1.924      0.469     -0.496     -4.100    0.000     -2.855     -0.992
##      Citrico  -1.397      0.435     -0.388     -3.211    0.002     -2.260     -0.533
## -----
##
##

```

```

##
## No more variables satisfy the condition of p value = 0.05
##
##
## Variables Removed:
##
## - Azucar
## - Cloruros
## - Fija
##
##
## Final Model Output
## -----
##
##                               Model Summary
## -----
## R                               0.394          RMSE                0.589
## R-Squared                       0.155          Coef. Var          11.279
## Adj. R-Squared                   0.138          MSE                0.347
## Pred R-Squared                   0.105          MAE                0.446
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##                               ANOVA
## -----
##                               Sum of
##                               Squares      DF      Mean Square      F      Sig.
## -----
## Regression      6.074          2          3.037      8.746      3e-04
## Residual        32.987         95          0.347
## Total           39.061         97
## -----
##
##                               Parameter Estimates
## -----
##                               model      Beta      Std. Error      Std. Beta      t      Sig      lower      upper
## -----
## (Intercept)      6.608          0.336          19.654      0.000      5.941      7.276
## Volatil          -1.924          0.469          -0.496      -4.100      0.000      -2.855      -0.992
## Citrico          -1.397          0.435          -0.388      -3.211      0.002      -2.260      -0.533
## -----
##
##
##                               Elimination Summary
## -----
##                               Variable
##                               Removed      R-Square      Adj.
##                               R-Square      C(p)      AIC      RMSE
## -----
## 1      Azucar      0.1874      0.1524      4.6647      179.6362      0.5842
## 2      Cloruros      0.1774      0.1511      3.8050      178.8347      0.5847
## 3      Fija      0.1555      0.1377      4.2982      179.4054      0.5893

```

Según el método backward, nuevamente, el modelo seleccionado es el modelo 6.

Punto 14. Selección final y justificación