

Taller RLM 1

Sofía Cuartas García

Simón Cuartas Rendón
Deivid Zhang Figueroa

Julián Úsuga

Enero de 2022

Punto uno. 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 dos. Análisis descriptivo.

Ácido fijo.

A continuación se muestran los resúmenes numéricos para la variable *fixed acidity* (acidez fija):

```
## Descriptive Statistics
## datos1$'fixed acidity'
## N: 100
##
##              fixed acidity
## -----
##      Mean              7.58
##      Std.Dev            1.07
##      Min                4.60
##      Q1                 7.00
##      Median             7.65
##      Q3                 8.10
##      Max               11.20
##      MAD                0.89
##      IQR                1.10
##      CV                 0.14
##      Skewness           -0.08
##      SE.Skewness        0.24
##      Kurtosis           1.42
##      N.Valid            100.00
##      Pct.Valid          100.00
```

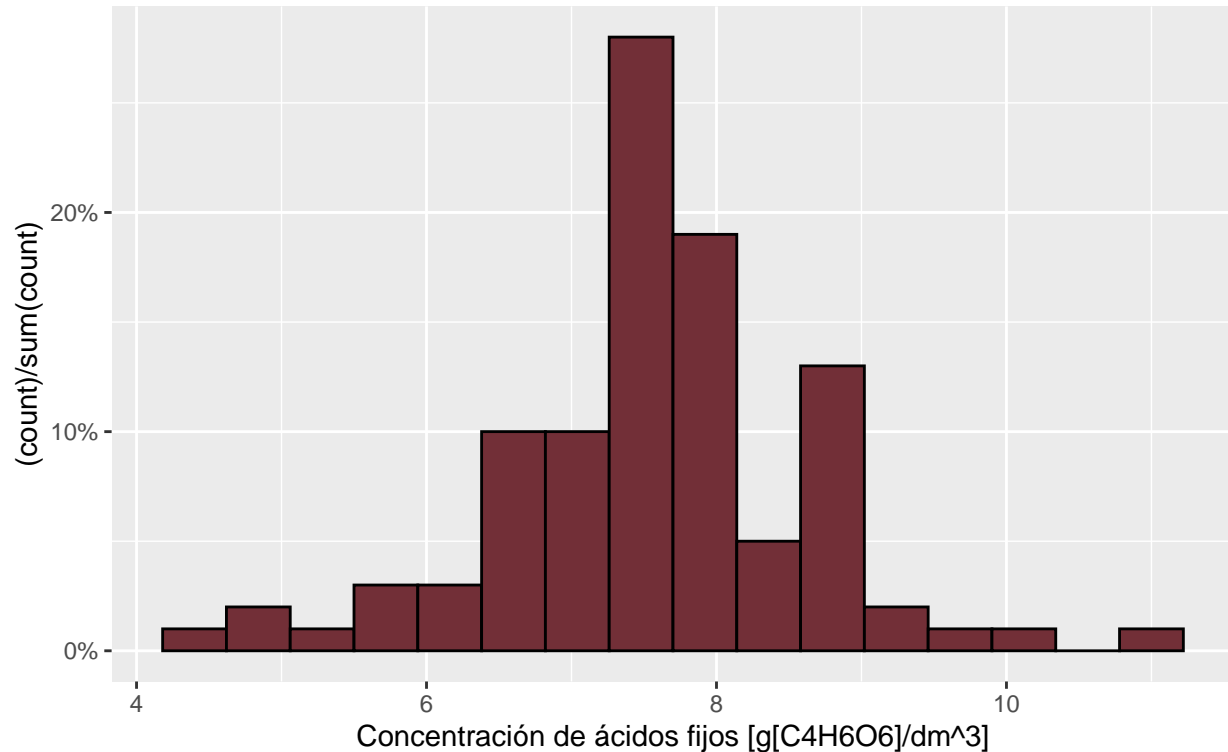
Y como se puede ver, la acidez fija promedi es de $7.58 \text{ g}[C_4H_6O_6]/dm^3$, con una desviación estándar de $1.07 \text{ g}[C_4H_6O_6]/dm^3$. Vale la pena destacar que el vino registrado con mayor concentración de ácidos fijos tenía $11.2 \text{ g}[C_4H_6O_6]/dm^3$, toda vez que el que menos registró $4.6 \text{ g}[C_4H_6O_6]/dm^3$. Por otro lado, se observa que la mediana se da en los $7.65 \text{ g}[C_4H_6O_6]/dm^3$, siendo un valor más alto que el promedio y lo que implica que al menos la mitad de los cien vinos observados tenían una concentración de ácidos fijos mayor o igual a dicha cantidad, mientras que el resto tenía una concentración inferior. Respecto a los demás cuantiles, vale la pena destacar que el primer cuantil está en los $7.00 \text{ g}[C_4H_6O_6]/dm^3$ y el tercer cuantil en los $8.10 \text{ g}[C_4H_6O_6]/dm^3$, lo que se traduce en un rango intercuartílico de $1.10 \text{ g}[C_4H_6O_6]/dm^3$.

Ahora bien, también se puede destacar que el coeficiente de asimetría es de -0.08 , lo que se traduce en que la distribución de la concentración de ácidos fijos medidos en $\text{g}[C_4H_6O_6]/dm^3$ presentan un pequeño sesgo

negativo, mientras que su curtosis es de 1.42, lo cual nos anticipa que la distribución de esta variable es platicúrtica, lo que se traduce en una baja concentración de valores entorno a su media. A continuación se observa un histograma para esta variable:

Histograma para la concentración de ácidos fijos

Concentración de ácidos fijos en los vinos de la base de datos



NULL

Y como se observa, hay una distribución aproximadamente simétrica, con una concentración de valores mayor hacia el mínimo respecto al máximo, lo cual confirma el valor obtenido por el coeficiente de asimetría, y también se evidencia que esta distribución puede tener colas más pesadas si se compara con el histograma típico de una distribución normal.

Ácido volátil

Una tabla con resúmenes numéricos para esta variable es la que se muestra enseguida:

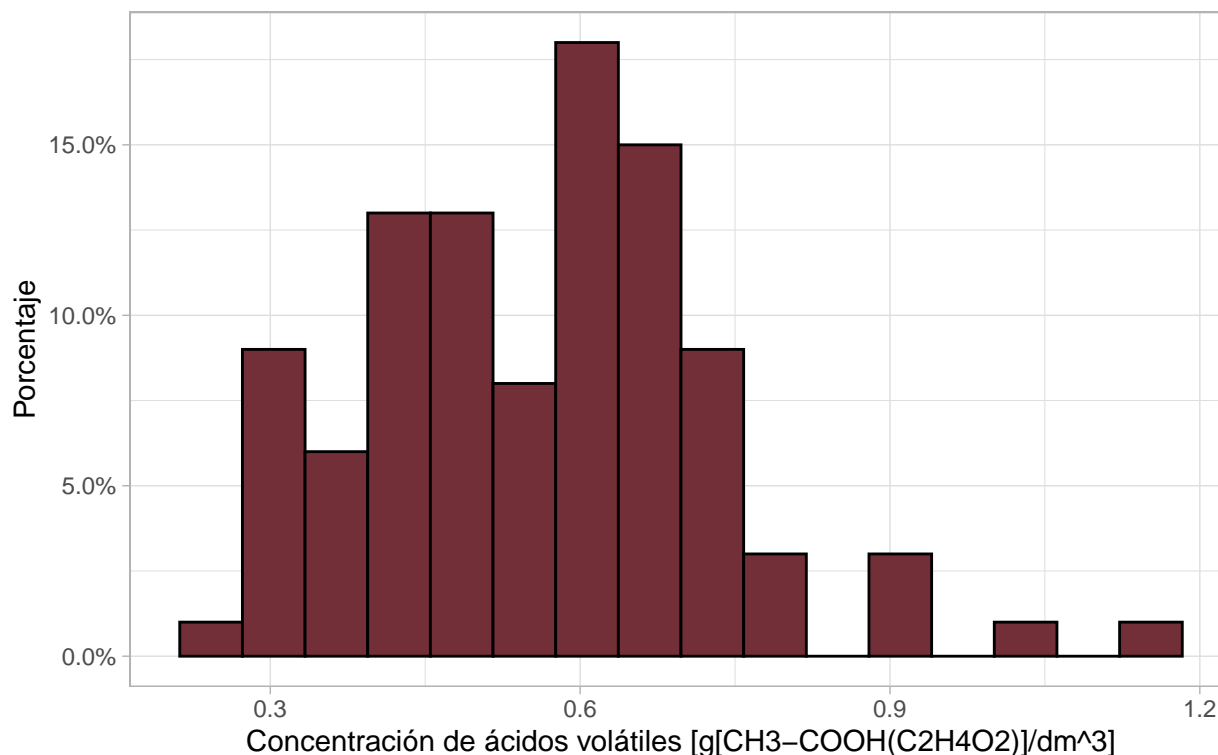
```
## Descriptive Statistics
## datos1$'volatile acidity'
## N: 100
##
##              volatile acidity
## -----
##           Mean              0.56
##          Std.Dev            0.16
##           Min               0.22
##           Q1                0.43
```

##	Median	0.57
##	Q3	0.66
##	Max	1.13
##	MAD	0.16
##	IQR	0.23
##	CV	0.29
##	Skewness	0.51
##	SE.Skewness	0.24
##	Kurtosis	0.71
##	N.Valid	100.00
##	Pct.Valid	100.00

Y como se observa, la concentración media de ácidos volátiles en los vinos registrados en la base de datos estudiada por los investigadores es de $0.56 \frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$, toda vez que su desviación estándar es de $0.16 \frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$. Se tiene, además, que la concentración mínima de ácidos volátiles registrada es de $0.33 \frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$ y la máxima es de $1.13 \frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$. También es importante destacar que la concentración mediana ácidos volátiles se da en los $0.57 \frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$, lo que significa que al menos la mitad de los vinos analizados y cuyos resultados están en la base de datos dada presentan una concentración de ácidos volátiles mayor o igual a $0.57 \frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$, mientras que los demás vinos tiene una concentración menor a dicha cantidad. Asimismo, se registra que el primer y el tercer cuantil ocurren a los $0.43 \frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$ y $0.63 \frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$ respectiva, lo cual se traduce en un rango intercuartílico de $0.23 \frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$.

Adicionalmente, se debe destacar que el coeficiente de asimetría es 0.51, lo cual da cuenta de una mayor asimetría que los ácidos fijos y, a diferencia de estos, la asimetría es positiva, lo cual resulta en una mayor acumulación de valores hacia el valor mínimo que hacia el máximo, y también se destaca una curtosis de 0.71, lo que implica que la distribución de la acidez volátil para esta muestra es de platicúrtica, que implica que hay una mayor concentración de valores en las colas de esta distribución si se compara con una distribución normal. Para tener mejor representación de esta información descriptiva, vale la pena observar el siguiente histograma:

Histograma para la concentración de ácidos volátiles
Concentración de ácidos volátiles en los vinos de la base de datos



De esta gráfica se puede constatar pues la concentración de valores hacia la izquierda de la media, lo cual fue anticipado por el alto valor positivo de la curtosis, así como la presencia de colas pesadas que refería la curtosis obtenida.

Ácido cítrico

Para el ácido cítrico vale la pena comenzar observando la siguiente tabla con resúmenes estadísticos asociados a esta variable:

```
## Descriptive Statistics
## datos1$'citric acid'
## N: 100
##
##          citric acid
## -----
##          Mean      0.22
##          Std.Dev   0.18
##          Min       0.00
##          Q1        0.06
##          Median    0.20
##          Q3        0.30
##          Max       0.70
##          MAD       0.18
##          IQR       0.23
##          CV        0.82
```

##	Skewness	0.73
##	SE.Skewness	0.24
##	Kurtosis	-0.30
##	N.Valid	100.00
##	Pct.Valid	100.00

Y como se observa, la concentración promedio de ácidos cítricos en los vinos observados en la base de datos registrada por los investigadores es de $0.22 \frac{g}{dm^3}$, y su desviación estándar es de

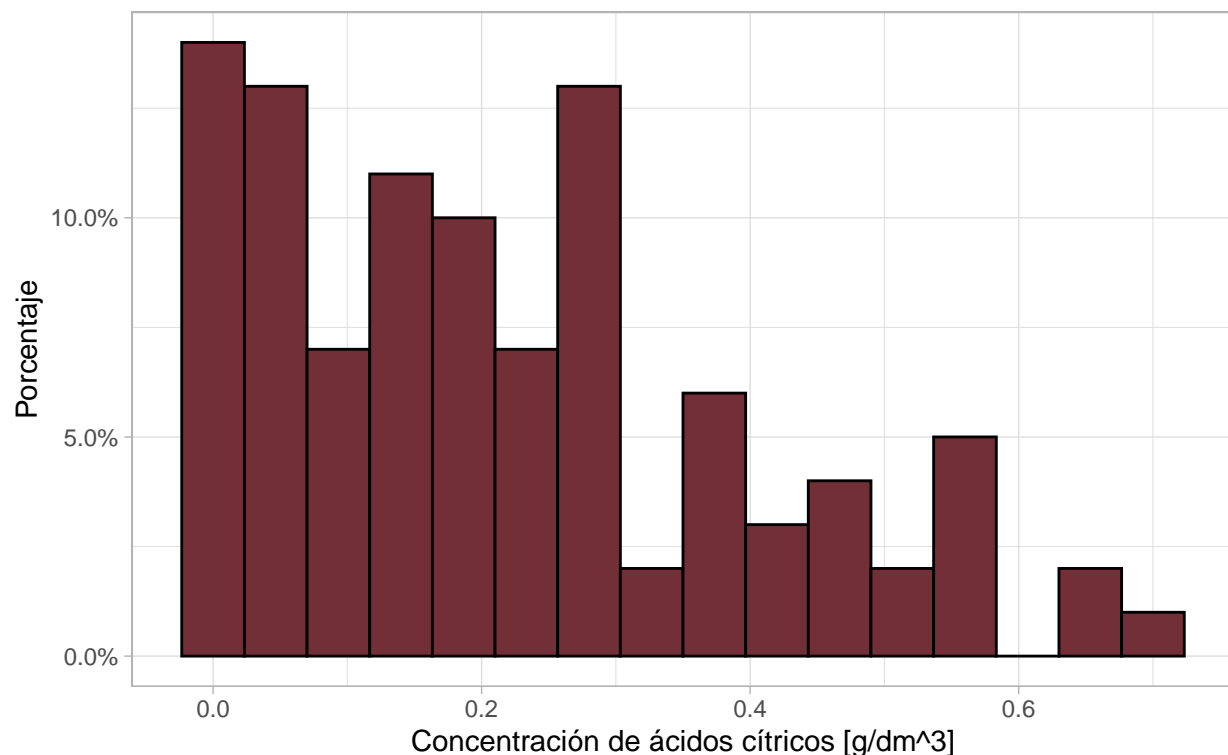
$$0.18 \frac{g}{dm^3}$$

. Además, se debe notar que la concentración mínima de ácidos cítricos registrada es de $0.00 \frac{g}{dm^3}$ y la máxima es de $0.70 \frac{g}{dm^3}$. También es destacable que la concentración mediana de ácidos cítricos se da en los $0.20 \frac{g}{dm^3}$, lo que significa que al menos la mitad de los vinos analizados y cuyos resultados están en la base de datos dada presentan una concentración de ácidos cítricos mayor o igual a $0.20 \frac{g}{dm^3}$, mientras que los demás vinos tienen una concentración de ácidos cítricos menor a dicha cantidad. Por otro lado, se registra que el primer y el tercer cuantil ocurren a los $0.06 \frac{g}{dm^3}$ y $0.30 \frac{g}{dm^3}$ respectivamente, lo cual se traduce en un rango intercuartílico de $0.23 \frac{g}{dm^3}$.

Adicionalmente, se debe destacar que el coeficiente de asimetría es 0.73, lo cual da cuenta de una mayor asimetría que los ácidos fijos y volátiles, y como la asimetría es positiva, la distribución muestral de esta variable resulta en una mayor acumulación de valores hacia el valor mínimo que hacia el máximo, y también se destaca una curtosis de -0.3. A continuación se observa un histograma que ilustra la distribución de esta variable.

Histograma para la concentración de ácidos cítricos

Concentración de ácidos cítricos en los vinos de la base de datos



Y así, lo primero que se destaca es la alta concentración de valores hacia el valor mínimo, lo cual fue anticipado por el alto coeficiente de asimetría positivo obtenido, así como lo pesadas que son las colas, lo cual también se conocía de antemano con el valor que se obtuvo para la curtosis.

Azúcar residual

Para el azúcar residual es útil iniciar con la siguiente tabla que lista varios resúmenes estadísticos asociados a esta variable:

```
## Descriptive Statistics
## datos1$'residual sugar'
## N: 100
##
##              residual sugar
## -----
##              Mean          2.46
##              Std.Dev       1.36
##              Min           1.20
##              Q1            1.80
##              Median        2.00
##              Q3            2.40
##              Max           10.70
##              MAD           0.44
##              IQR           0.60
##              CV            0.55
##              Skewness       3.17
##              SE.Skewness    0.24
##              Kurtosis       12.91
##              N.Valid        100.00
##              Pct.Valid      100.00
```

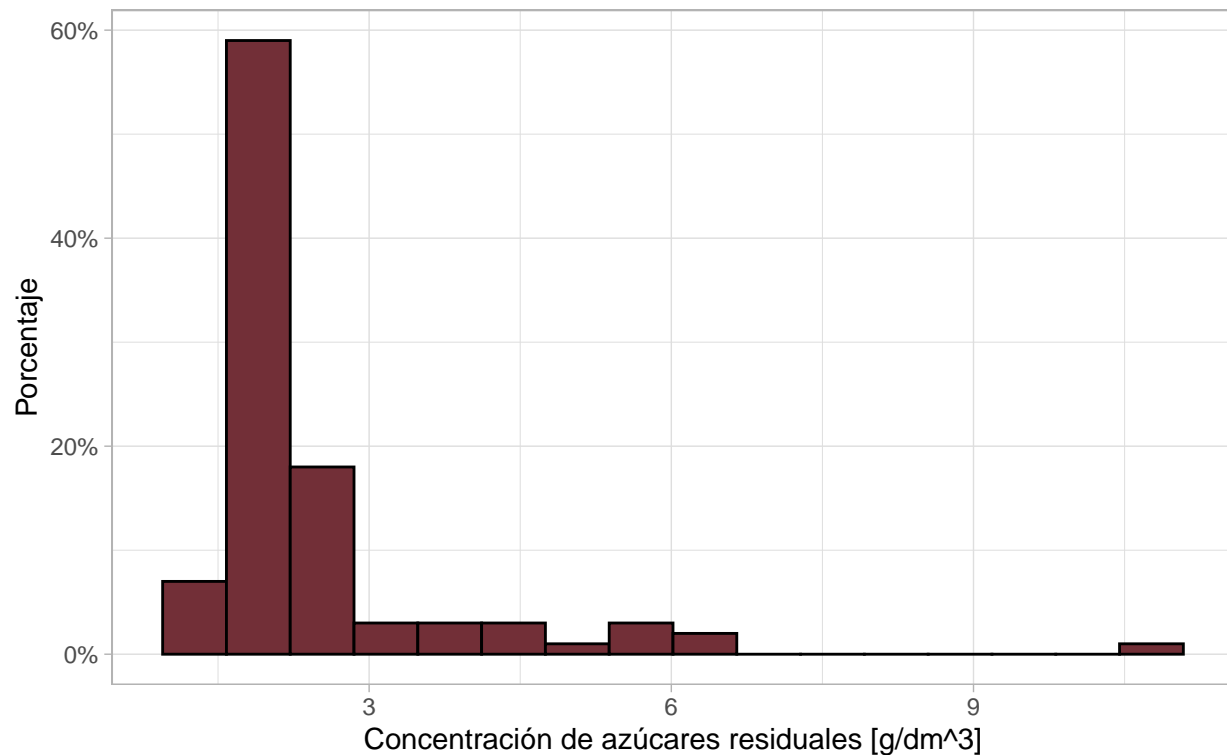
Y como se ve, la concentración promedio de azúcares residuales en los vinos observados en la base de datos registrada por los investigadores es de $2.46 \frac{g}{dm^3}$, y su desviación estándar es de

$$1.36 \frac{g}{dm^3}$$

. Además, se debe notar que la concentración mínima de azúcares residuales registrada es de $1.220 \frac{g}{dm^3}$ y la máxima es de $10.70 \frac{g}{dm^3}$, lo que implica que todos los vinos de esta muestra son *secos* y *semisecos*. También es destacable que la concentración mediana azúcares residuales se da en los $2.00 \frac{g}{dm^3}$, lo que significa que al menos la mitad de los vinos analizados y cuyos resultados están en la base de datos dada presentan una concentración de azúcares residuales mayor o igual a $2.00 \frac{g}{dm^3}$, mientras que los demás vinos tiene una concentración de azúcares residuales menor a dicha cantidad. Por otro lado, se registra que el primer y el tercer cuantil ocurren a los $1.80 \frac{g}{dm^3}$ y $2.40 \frac{g}{dm^3}$ respectivamente, lo cual se traduce en un rango intercuartílico de $0.60 \frac{g}{dm^3}$.

Adicionalmente, se debe destacar que el coeficiente de asimetría es 3.17, lo cual da cuenta de una fuerte asimetría, y dado que esta es positiva, la distribución muestral de esta variable resulta en una mayor acumulación de valores hacia el valor mínimo que hacia el máximo, y también se destaca una curtosis de 12.91, lo que implica que la distribución es leptocúrtica, lo que significa que hay una gran concentración de datos entorno a la media. A continuación se observa un histograma que ilustra la distribución de esta variable.

Histograma para la concentración de azúcares residuales
Concentración de azúcares residuales en los vinos de la base de datos



De esta es claramente destacable que hay una enorme concentración de valores hacia el mínimo, pues hay más de un 80 % de datos con concentraciones inferiores a los $3.00 \frac{g}{dm^3}$, lo que significa que más de cuatro quintas partes de los vinos observados pueden ser clasificados como secos.

Cloruros

A continuación se pueden observar varios resúmenes numéricos para los cloruros:

```
## Descriptive Statistics
## datos1$chlorides
## N: 100
##
##          chlorides
## -----
##      Mean      0.10
## Std.Dev      0.07
##      Min      0.04
##      Q1       0.07
##      Median    0.08
##      Q3       0.10
##      Max      0.46
##      MAD      0.01
##      IQR      0.02
##      CV       0.69
##      Skewness  3.67
```


##	SE.Skewness	0.24
##	Kurtosis	13.42
##	N.Valid	100.00
##	Pct.Valid	100.00

En la tabla anterior es posible observar que la concentración promedio de cloruros en los vinos analizados en la base de datos registrada por los investigadores es de $0.10 \frac{g[NaCl]}{dm^3}$, y su desviación estándar es de

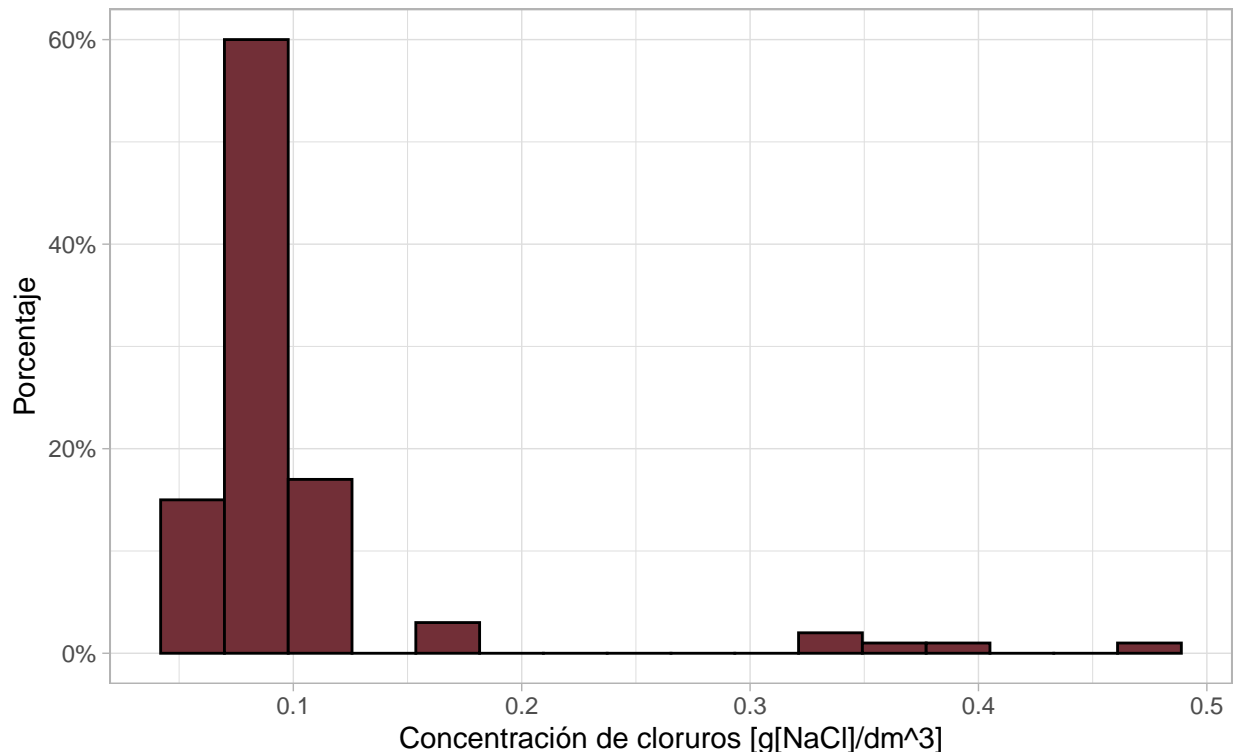
$$0.07 \frac{g[NaCl]}{dm^3}$$

. Además, se debe notar que la concentración mínima de cloruros registrada es de $0.04 \frac{g[NaCl]}{dm^3}$ y la máxima es de $0.46 \frac{g[NaCl]}{dm^3}$. También es destacable que la concentración mediana de cloruros se da en los $0.08 \frac{g[NaCl]}{dm^3}$, lo que significa que al menos la mitad de los vinos analizados y cuyos resultados están en la base de datos dada presentan una concentración de cloruros mayor o igual a $0.08 \frac{g[NaCl]}{dm^3}$, mientras que los demás vinos tienen una concentración de cloruros menor a dicha cantidad. Por otro lado, se registra que el primer y el tercer cuantil ocurren a los $0.07 \frac{g[NaCl]}{dm^3}$ y $0.10 \frac{g[NaCl]}{dm^3}$ respectivamente, lo cual se traduce en un rango intercuartílico de $0.02 \frac{g[NaCl]}{dm^3}$, lo cual da cuenta de una alta concentración de valores entorno a la mediana y la media, de la cual se debe notar que tanto el segundo cuantil (la mediana) como el tercero tienen el mismo valor.

Adicionalmente, se debe destacar que el coeficiente de asimetría es 3.67, lo cual da cuenta de una fuerte asimetría, y dado que esta es positiva, la distribución muestral de esta variable resulta en una mayor acumulación de valores hacia el valor mínimo que hacia el máximo, confirmando lo mencionado anteriormente respecto al rango intercuartílico; además, se destaca una kurtosis de 13.42, lo que implica que la distribución es leptocúrtica, lo que significa que hay una gran concentración de datos entorno a la media. A continuación se observa un histograma que ilustra la distribución de esta variable.

Histograma para la concentración de cloruros

Concentración de cloruros en los vinos de la base de datos



De la anterior gráfica se puede extraer entonces lo que ya se conoció con los resúmenes numéricos y es una muy alta concentración de valores cerca del valor mínimo de concentración de cloruros. De hecho, es destacable que aproximadamente el 90 % de los vinos analizados tienen una concentración de cloruros inferior o igual a $0.15 \frac{g[NaCl]}{dm^3}$.

Calidad

La calidad, al ser otra variable numérica, puede ser estudiada inicialmente mediante el siguiente esquema de resúmenes numéricos:

```
## Descriptive Statistics
## datos1$quality
## N: 100
##
##               quality
## -----
##           Mean      5.25
##         Std.Dev    0.66
##           Min      4.00
##           Q1       5.00
##          Median     5.00
##           Q3       6.00
##           Max      7.00
##           MAD      0.00
##           IQR      1.00
##           CV       0.13
##         Skewness    0.75
##       SE.Skewness    0.24
##         Kurtosis    0.82
##          N.Valid   100.00
##         Pct.Valid   100.00
```

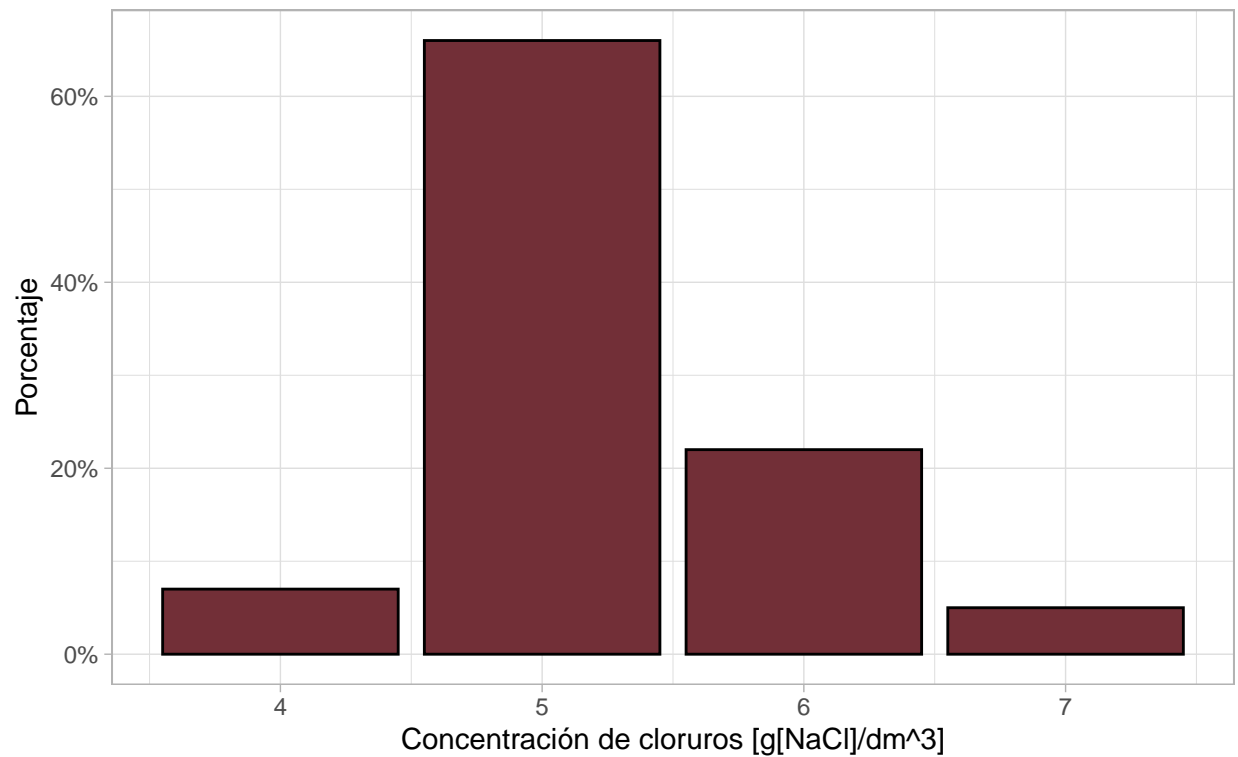
Entonces, se comienza mencionando que la calidad promedio de los vinos de la muestra de los investigadores es de 5.25, con una desviación estándar de 0.66. Por otro lado, se tiene que el vino de peor calidad tiene un puntaje de cuatro puntos, toda vez que el mejor ranqueado destaca con siete puntos de 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:

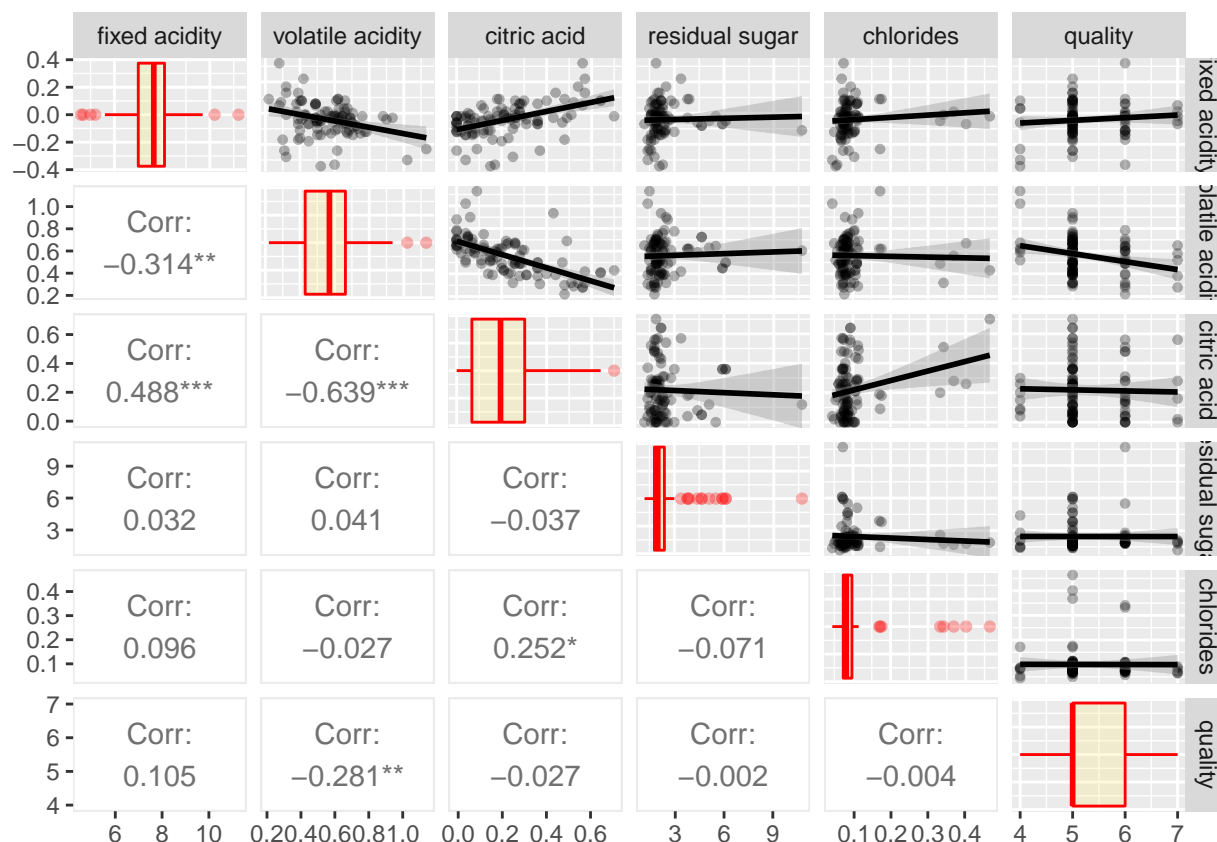
Histograma para la concentración de cloruros

Concentración de cloruros en los vinos de la base de datos



Y como se puede observar, más del 60 % de los vinos incluidos en la base de datos que se está estudiando 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.

A continuación se van a realizar los gráficos de dispersión entre el puntaje de calidad y las los ácidos volátiles y cítricos y entre el puntaje de calidad y ácidos fijos y los cítricos, pues son los que obtuvieron mayores correlaciones lineales.

PENDIENTE.

Punto tres. 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 XXXX.
- X_{2i} . Concentración de ácidos volátiles en el i -ésimo vino analizado en XXXX.
- X_{3i} . Concentración de ácidos cítricos en el i -ésimo vino analizado en XXXX.
- X_{4i} . Concentración de azúcares residuales en el i -ésimo vino analizado en XXXX.
- X_{5i} . Concentración de cloruros en el i -ésimo vino analizado en XXXX.
- 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:

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

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

```
##
## 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
```

Es decir, el modelo ajustado está dado por:

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

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:

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

$H_1 : \exists j : \beta_j \neq 0, j = 1, 2, 3, 4 \iff$ 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 cuatro. 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:

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

##	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].