Taller RLM 1

Sofía Cuartas García Simón Cuartas Rendón Julián Úsuga
Deivid Zhang Figueroa

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 ($CH_3 - COOHC_2H_4O_2$)) por decímetro cúbrico ($\frac{g[CH_3-COOH(C_2H_4O_2)]}{dm^3}$). 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 vacteria 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 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 las 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.

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

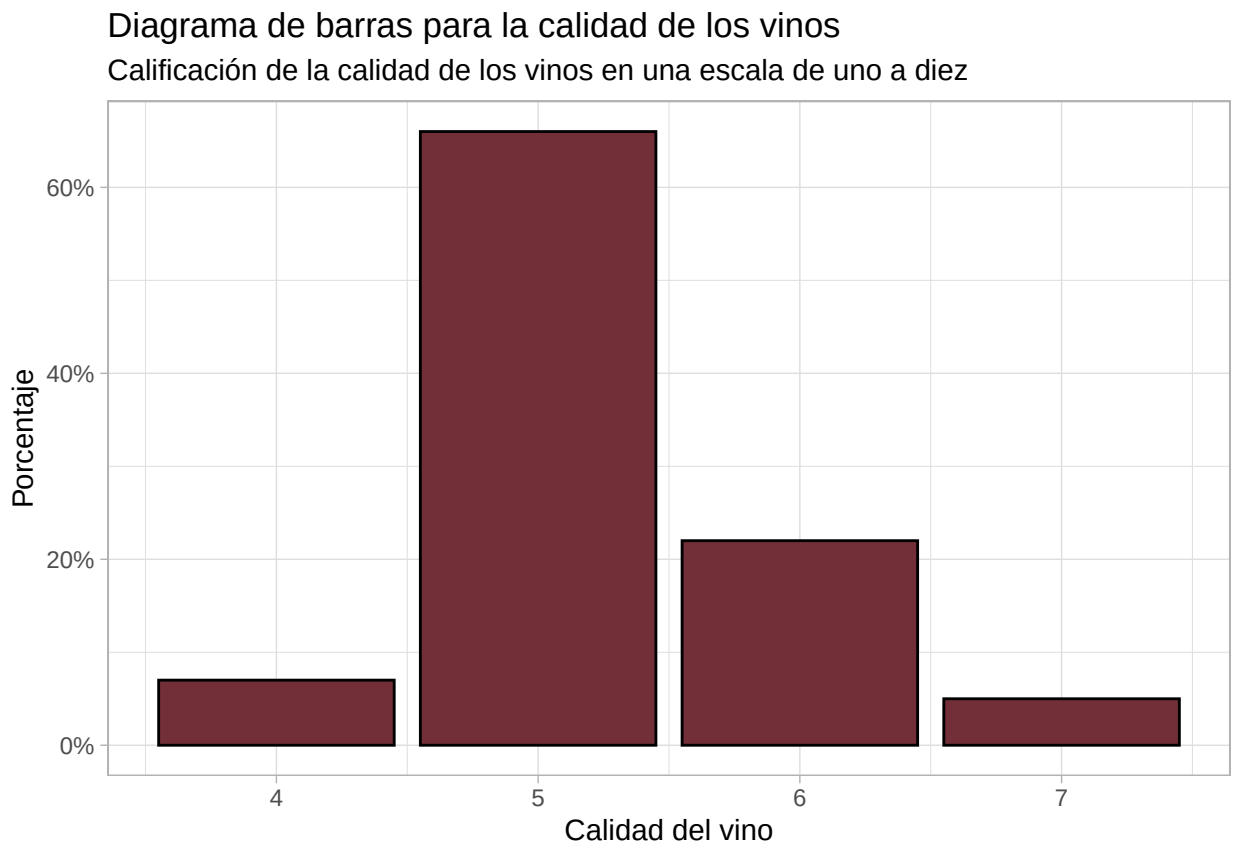
```
## Descriptive Statistics
## datos1$quality
## N: 100
##
##                      quality
## ----------------- ---------
##              Mean      5.25
##           Std.Dev      0.66
##               Min      4.00
##                Q1      5.00
##            Median      5.00
##                Q3      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 intercuartídico 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

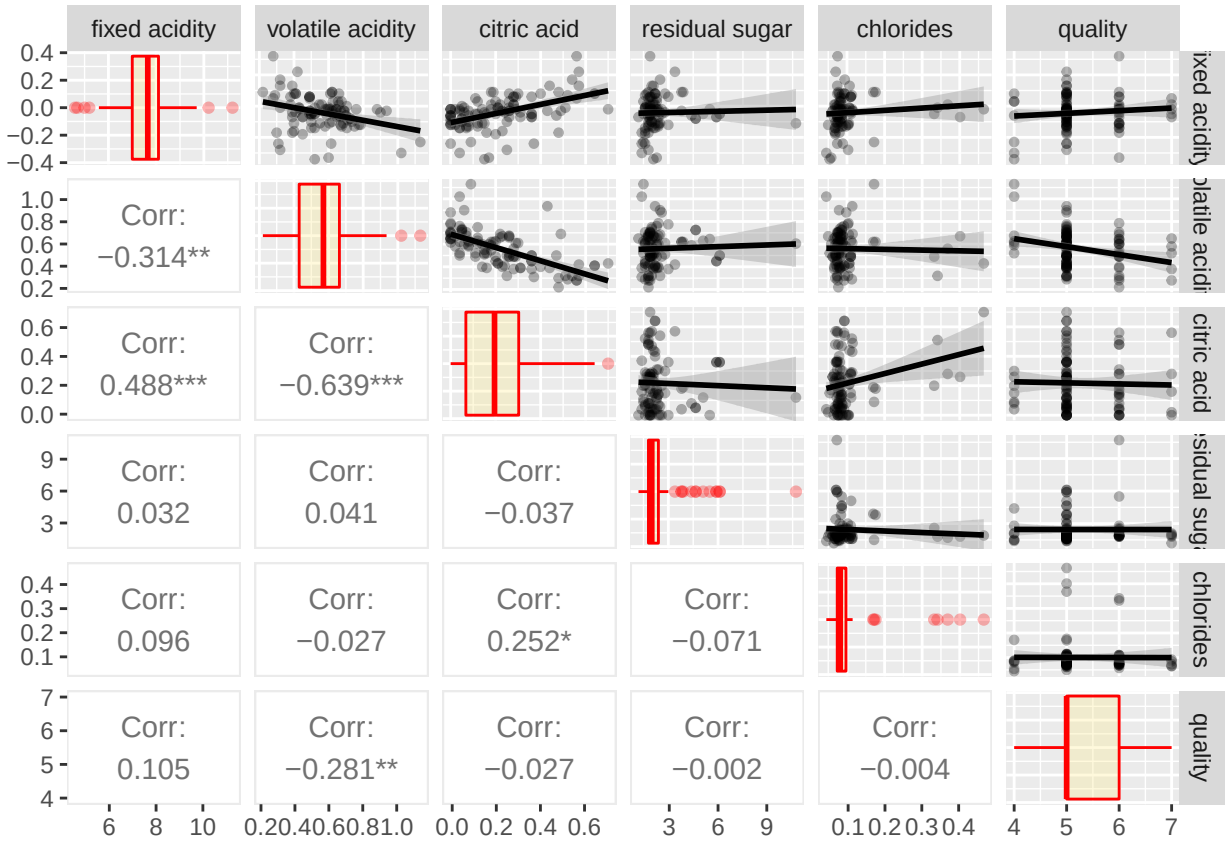
la curtosis es de 0.82 y, entonces, se tiene que hay una mayor cantidad de valores atípicos en comparación con una distribución normal.

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



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 ugar 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 $i = 1,\ 2,\ ...,\ n$, 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:

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

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:

$$\widehat{Y_i} = 5.9729 + 0.0964 X_{1i} - 2.0875 X_{2i} - 1.6863 X_{3i} + 0.0018 X_{4i} + 0.7868 X_{5i} \qquad \langle 2 \rangle$$

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)
## FO(Fija, Volatil, Citrico, Azucar, Cloruros)  5  7.495 1.49904  3.9969 0.002482
## Residuals                                    94 35.255 0.37505
##
## FO(Fija, Volatil, Citrico, Azucar, Cloruros) **
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Entonces, si plantean las siguientes hipótesis:

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

$H_1: \ \exists \ j : \beta_j \neq 0, \ j = 1, 2, 3, 4, 5 \quad \iff$ *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 $\langle 2 \rangle$ **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].

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:

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.

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 $H_A : \beta_5 \neq 0$

- **Significancia de β_1:** No hay evidencia suficiente para rechazara la hipótesis nula, por lo tanto el ácido fijo **no es** signifcativo para explicar la calidad del vino dado que las otras covariables están en el modelo.
- **Significancia de β_2:** Hay evidencia suficiente para rechazara la hipótesis nula, por lo tanto el ácido volátil **es** signifcativo para explicar la calidad del vino dado que las otras covariables están en el modelo.
- **Significancia de β_3:** Hay evidencia suficiente para rechazara la hipótesis nula, por lo tanto la ácido cítrico **es** signifcativo para explicar la calidad del vino dado que las otras covariables están en el modelo.
- **Significancia de β_4:** No hay evidencia suficiente para rechazara la hipótesis nula, por lo tanto la azúcar residual **no es** signifcativa para explicar la calidad del vino dado que las otras covariables están en el modelo.
- **Significancia de β_5:** No hay evidencia suficiente para rechazara la hipótesis nula, por lo tanto los cloruros **no son** signifcativos 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 signiticativas 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 párametros, 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:

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

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

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 estan las variables *acidez volátil* y *ácido cítrico*.