Análisis inferencial (Regresión lineal simple), sobre la intensidad de color en el vino (variable dependiente) y la concentración de fenoles no flavonoides (variable independiente).

Anthony J. Servitá R.[1]

Resumen

El siguiente análisis se realizó sobre dos variables obtenidas desde la tabla de datos de las características físico-químicas de los vinos durante su proceso de destilación. En el análisis inferencial se observa que existe una relación lineal entre ambas variables lo que podemos inferir que la variable Y puede ser descrita por la variable X. Así mismo, también se ha encontrado una correlación alta superior a cero entre ambas variables lo que da como conclusión de que puede existir regresión línea las variables de la concentración, fenoles no flavonoides y el color de los vinos. Por lo que, un aumento de las concentraciones de fenoles no flavonoides en el vino, producen un aumento en la intensidad de color de los vinos. En los resultados obtenidos de la ecuación de la recta estimada a través del programa JASP encontramos un R^2 no muy elevado; lo que significa que la recta estimada no describe con exactitud la variación de Y.

Introducción

La regresión Lineal simple es un método estadístico, usado para realizar análisis sobre dos variables cuantitativas en estudio. La misma pretende analizar mediante la ecuación de la recta como es el comportamiento de una variable con respecto a la otra y ve si una es dependiente de la otra. Por ello se intenta encontrar la mejor recta que se adapte a los datos en estudio. Esta Recta es solo una estimación de la recta paramétrica. La recta estimada es llamada $\hat{y} = \hat{\alpha} + \hat{\beta}x$ y la recta paramétrica se denomina como $\mu_{y/x} = \alpha + \beta x$, donde en ambas partes **y** es la variable dependiente, α es el punto de corte con el eje Y, β es la pendiente de dicha recta, **x** es la variable independiente.

En la evaluación de la mejor recta estimada, debe usarse el análisis de varianza (ANDEVA) el cual es indicativo, de que si se obtiene un valor-p menor al $\alpha = 0.05$ entonces, la ecuación de la recta establecida puede utilizarse para predecir los valores de la Y. Además, se intenta calcular la bondad de ajuste o el R^2, el cual expresa valores que oscilan entre 0 y 1, y para un valor igual 1 revela que toda la variación de la variable dependiente es explicada a través de la regresión lineal.

La realización de este estudio, se hizo mediante dos variables presentes en las características de los vinos, éstas son: Variable independiente "fenoles no flavonoides" y variable dependiente "intensidad de color". El objetivo principal es saber si puede haber regresión lineal entre ambas variables, y encontrar la recta que mejor se ajuste a eso datos. Los datos usados en este análisis son de procedencia de una base de datos online UCI machine learning; para realizar medidas en las concentraciones de "fenoles no flavonoides" y la "intensidad de color" es necesario el uso de instrumentos y conocimientos químicos.

Materiales y métodos

Los análisis desarrollados en esta investigación fueron realizados con el software opensource JASP versión 0.13.1, cuyos algoritmos están basados sobre la sintaxis del lenguaje de programación R. El programa fue instalado en el sistema operativo Ubuntu Linux versión 20.04

Base de datos:

Los datos fueron colectados desde la base de datos del repositorio UCI machine learning cuyos "datasets" son publicados por medio de investigaciones que han sido corroborados y que por lo tanto son datos factibles sobre una buena practica real.

El datasets colectado tiene por nombre "*wine.data*", y es una data-set que contiene las concentraciones de varios componentes encontrados en 3 tipos de vinos. Las variables que aquí encontraremos son:

1. Alcohol: El cual contiene datos del tipo entero, el cual describe el tipo de vino mediante 1, 2 ó 3.
2. Ácido málico: Contiene las concentraciones de ácido málico que es posible encontrar en cada uno de estos 3 vinos. Estas concentraciones vienen en unidades de mg/mL.
3. Cenizas: El cual mide las concentraciones de Cenizas que se hayan luego de la fundición de la pulpa a altas temperatura.
4. Alcalinidad de las cenizas: Contiene los valores alcalinos de las cenizas para los tres tipos de vino.
5. Magnesio: Concentraciones de magnesio en los tres tipos de vino.
6. Fenoles totales: La cantidad bruta de fenol en los 3 tipo de vino.
7. Flavonoides: Concentraciones de flavonoides.
8. Pro-antocianinas: Concentraciones de Proantocianinas en los 3 tipos de vinos.
9. Intensidad de color: Es la dureza del color de los 3 tipos de vino.
10. Tono: Grado de tonalidad que toman los 3 tipos de vino.
11. Prolina: Concentraciones de Prolina en los mismos 3 tipos de vino.

Estas 11 variables pertenecientes al datasets "*wine.data",* no serán foco de este estudio *perse*, ya que la regresión lineal simple aborda solo una variable predictora y una variable a predecir. Es decir, de tan solo 2 variables. Si quisiéramos realizar, un análisis inferencial en todas estas variables entonces se podría hacer uso de la regresión lineal múltiple, la cual difiere de la simple por que la variable dependiente toma valores de varias variables independientes o, de entrada.

Análisis de los datos

En este estudio, se tomarán dos variables a analizar, denominadas como: Fenoles no flavonoides y la intensidad de color. Con estas dos variables se realizará la regresión lineal. Para esto, lo primero será hacer una visualización de los datos con un scatterplot (gráfico de dispersión), de esta manera se entenderá como es el comportamiento de las variables con respecto a la otra.

Gráfico 1.1: Gráfico de dispersión que muestra la correlación lineal entre ambas variables, así como su valor r *Pearson* y su valor ρ (rho) de *Spearman*.

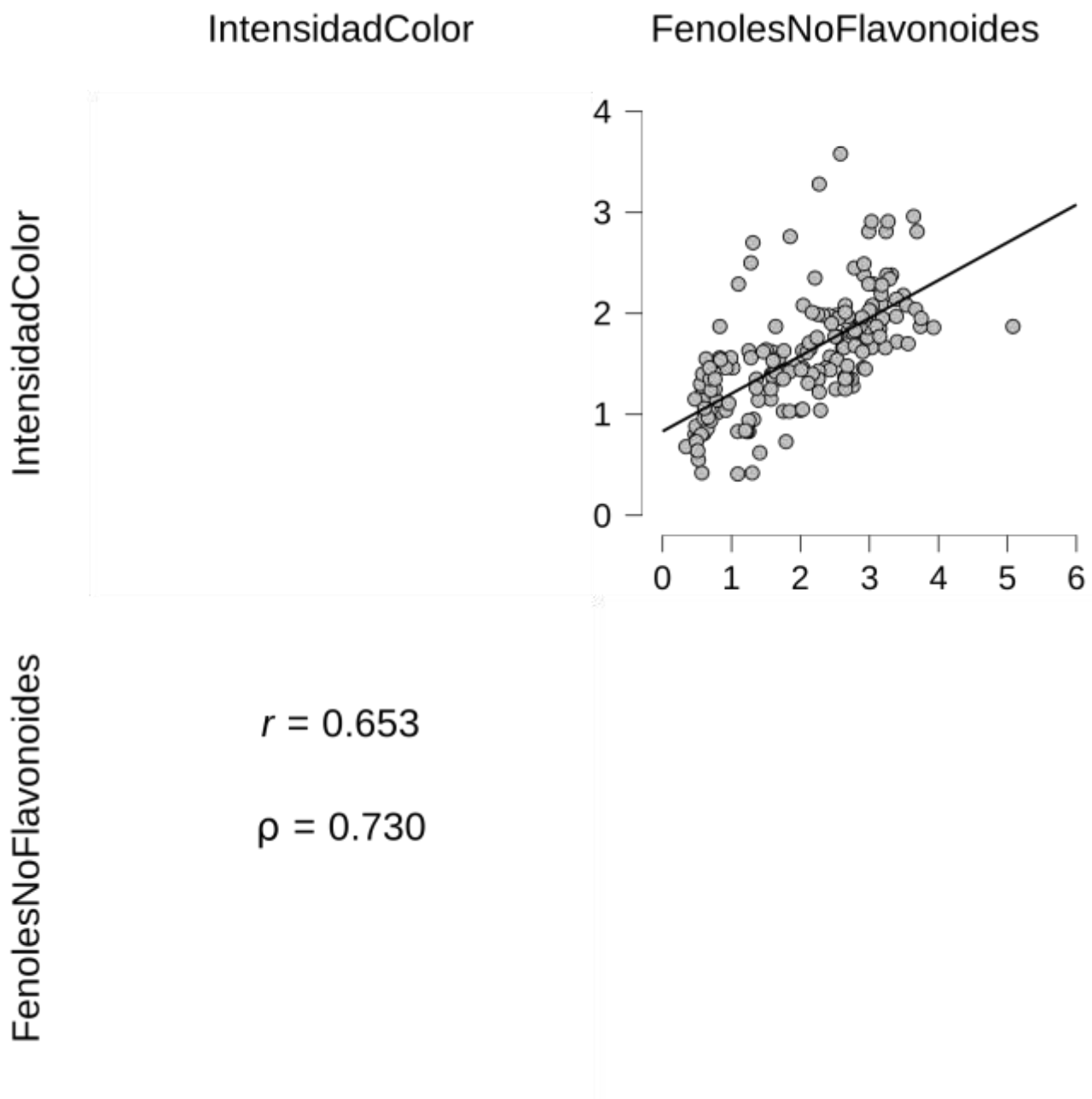


Tabla 1: Prueba de hipótesis para una correlación significativamente de cero.

Correlation

Spearman's Correlations

Variable		IntensidadColor	FenolesNoFlavonoides
1. IntensidadColor	Spearman's rho	—	
	p-value	—	
2. FenolesNoFlavonoides	Spearman's rho	0.730***	—
	p-value	< .001	—

* p < .05, ** p < .01, *** p < .001

Tabla 1.2: Prueba de normalidad de los datos multivariante de shapiro wilk.

Assumption checks

Shapiro-Wilk Test for Multivariate Normality

Shapiro-Wilk	p
0.936	< .001

El análisis previamente realizado, arrojo uno gráfico de dispersión que nos indica la relación lineal que existe entre ambas variables. El valor r = 0,653 encontrado en este gráfico, así como en la *tabla 1*, indica una correlación de *Pearson* positiva.

En dicha tabla, también podemos observar el valor-p que arroja dicha correlación el cual es menor al valor $\alpha = 0.05$, este valor es indicativo sobre la prueba de hipótesis, en el que el valor de correlación sea significativamente distinto de cero. Dicha hipótesis se establece como sigue:

$H0: \rho = 0$ (No existe correlación)
$H1: \rho \neq 0$ (Existe correlación)

Para el este caso, el valor-p calculado es menor a 0.05 y por lo tanto no existe evidencia significativa en estos datos para aceptar la hipótesis nula de que no existe correlación entre ambas variables. Por consiguiente, hay relación entre ambas variables.

La prueba de normalidad de los datos, también da un valor-p menor al $\alpha = 0.05$, lo que concluye en la no normalidad de los datos en ambas variables; por lo tanto, el coeficiente de correlación más significativo, es el calculado en la *tabla 1.2* ρ de *Spearman* de 0,730.

Resultados

Tabla 2.0. Estadísticos descriptivos sobre las variables Fenoles no flavonoides y la intensidad de color en el vino.

Descriptive Statistics

	FenolesNoFlavonoides	IntensidadColor
Valid	178	178
Missing	0	0
Mean	2.029	1.591
Std. Deviation	0.999	0.572
Shapiro-Wilk	0.955	0.981
P-value of Shapiro-Wilk	< .001	0.014
Minimum	0.340	0.410
Maximum	5.080	3.580

Como puede apreciarse en la *tabla 2.0*, en ambas variables el valor-p arrojado por la prueba de shapiro wilk aplicado independientemente a ambas variables nos muestra que no existe normalidad en los datos de las mismas. Por lo tanto, se tomará en cuenta el coeficiente de correlación de *spearman* como el valor significativo. No obstante, el valor r de *Pearson* calculado coincide un poco con el valor de *Spearman* los cuales están representados en el *gráfico 1;* de manera que, ambos están indicando un valor de correlación positiva.

Regresión Lineal

Tabla 2.1: Cómputos calculados sobre el análisis de regresión lineal.

Linear Regression

Model Summary - IntensidadColor

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	0.572
H_1	0.653	0.426	0.423	0.435

La *tabla 2.1* de regresión lineal contiene los valores de R^2 ajustado y RMSE que mejor se ajustan a la recta de regresión de los datos.

Tabla 2.2: Cómputos de la suma de cuadrados.

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	24.702	1	24.702	130.624	< .001
	Residual	33.283	176	0.189		
	Total	57.984	177			

Note. The intercept model is omitted, as no meaningful information can be shown.

La *tabla 2.2***,** contiene el análisis de varianzas entre las dos variables, el cual arroja un valor-p menor al 0,05 para la regresión.

Tabla 2.3: Cálculo de la estimación de la recta ajustada a la regresión.

Coefficients

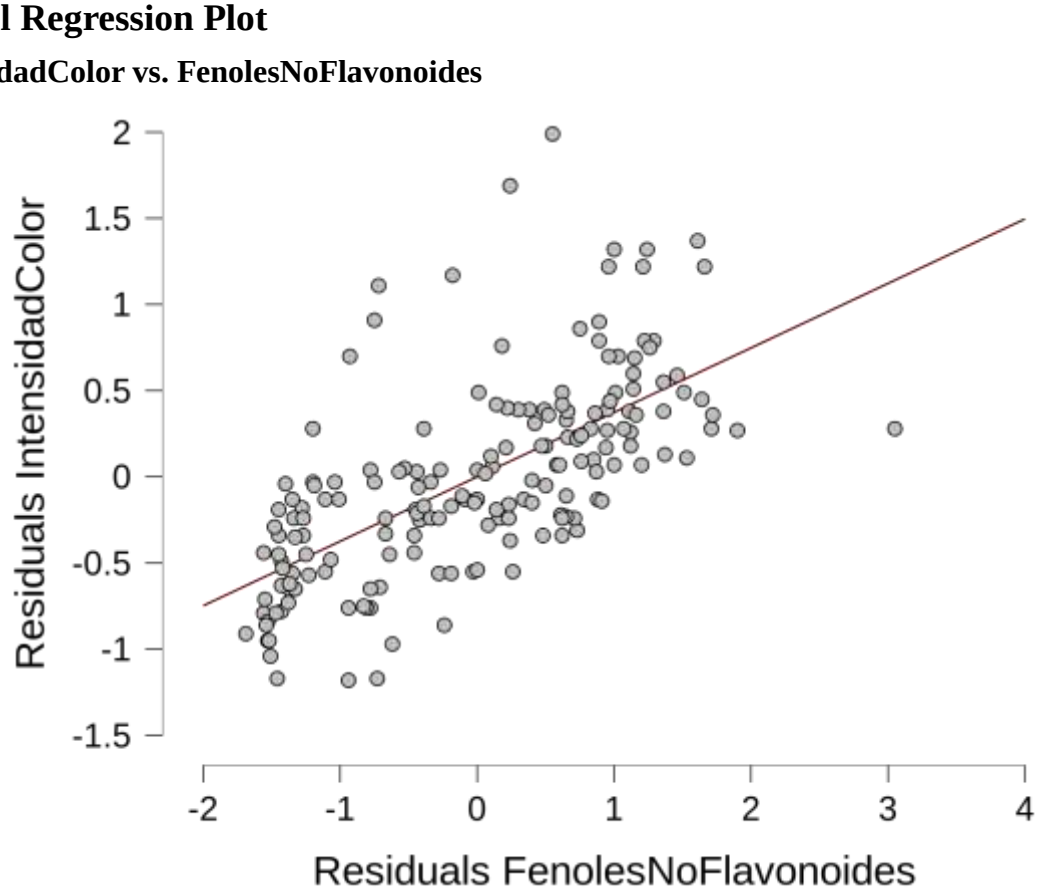
Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	1.591	0.043		37.084	< .001
H_1	(Intercept)	0.832	0.074		11.247	< .001
	FenolesNoFlavonoides	0.374	0.033	0.653	11.429	< .001

Los resultados presentados en la *tabla 2.3,* muestran que la recta que mejor se ajusta a nuestros datos viene dada por:

$$\hat{y} = 0{,}832 + 0{,}374x;$$

donde, ŷ es la variable dependiente intensidad de color y x en la variable independiente fenoles no flavonoides. Así mismo, la recta puede verse representada en el *gráfico 1.2.*

Gráfico 1.2: grafico de dispersión que representa la dependencia de la variable intensidad de color con respecto a las variables fenoles no flavonoides, en el mismo se representa el modelo de la recta ajustada a los datos.



Conclusiones

Los resultados obtenidos por medio del análisis de regresión lineal simple realizado para las variables Fenoles no flavonoides (variable independiente) e intensidad de color (variable dependiente). Satisfacen la ecuación de la recta estimada $\hat{y} = 0{,}832 + 0{,}374x$, y que gracias al valor-p $< \alpha = 0.05$ calculado en la *tabla 2.2*, aceptamos con un 95% de confianza que existe regresión lineal entre las concentraciones de fenoles no flavonoides y la intensidad de color en el vino.

Así mismo, La *tabla 2.1* muestra un valor de R^2 mayor a cero lo que indica que el 42.6% de la variación total de la variable intensidad de color se encuentra explicada por la regresión.

Este modelo de la ecuación de la recta, nos servirá para inferir sobre la cantidad (o intensidad) de color que poseen los 3 tipos de variedades de vino encontrados en estos datos, según sus concentraciones de Fenoles No Flavonoides, el cual sigue un modelo lineal; es decir: "A mayores concentraciones de Fenoles no Flavonoides generara mayor es la intensidad de color en el vino.

La recta estimada, $\hat{y} = \hat{\alpha} + \hat{\beta}x$ es solo una aproximación a la recta paramétrica $\mu_{y/x} = \alpha + \beta x$ la cual contiene la verdadera los valores que convergen entre la variable fenoles no flavonoides e intensidad de color.

La manera en que debe usarse esta ecuación es, que para cada dato nuevo de la variable independiente fenoles no flavonoides introducido en la ecuación obtenida $\hat{y} = 0{,}832 + 0{,}374x$ para x = concentración de fenoles no flavonoides, generara un nuevo dato $\hat{y}$ = intensidad de color. Así, es posible inferir sobre qué tan intenso será el color del vino y ciertas concentraciones de fenoles no flavonoides.

Bibliografía

MERAYO PATRICIA. Cómo evaluar si la correlación es significativa: pruebas de hipótesis para la correlación. Data science. https://www.maximaformacion.es/blog-dat/como-evaluar-si-la-correlacion-es-significativa-pruebas-de-hipotesis-para-la-correlacion/

SAMUEL E. SEGNINI F. Fundamentos de Bioestadística. Universidad de los andes Facultad de ciencias. Mérida-Venezuela 2004. Página 198-226.

JASP Team (2020). JASP (Versión 0.13.1) [Computer software].

UCI machine learning https://archive.ics.uci.edu/ml/datasets/wine