



Centro Internacional de Agricultura Tropical
Desde 1967 *Ciencia para cultivar el cambio*

Introducción a regresión lineal multiple y análisis de varianza en R

Marzo 2019

Autores

Hugo Andrés Dorado

Juan Camilo Rivera

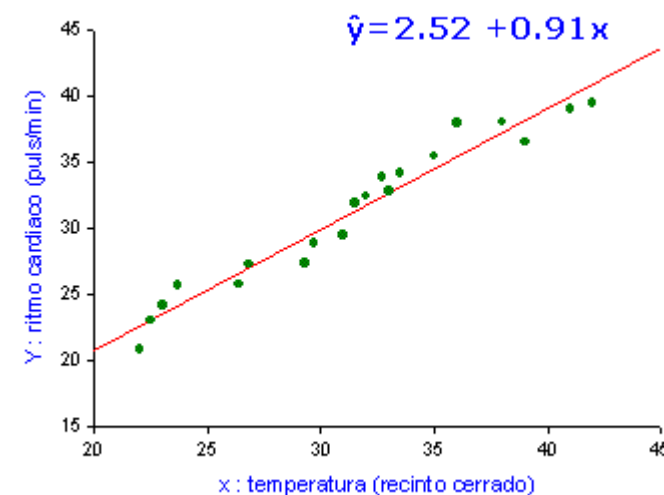
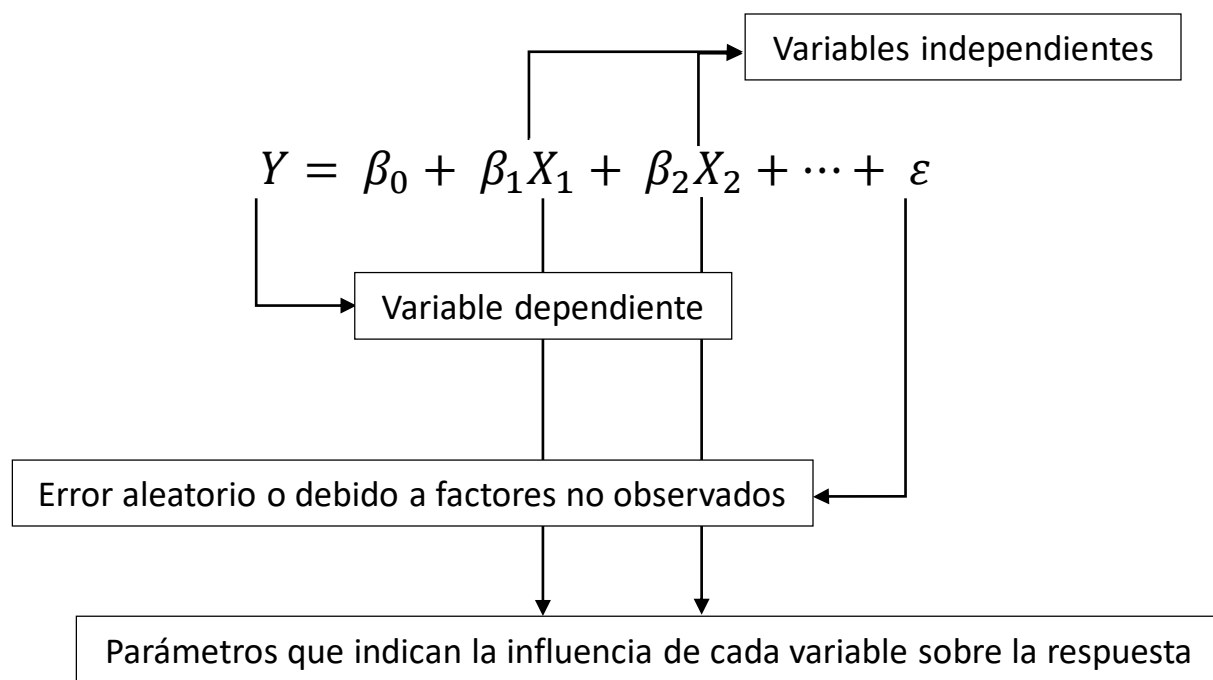
h.a.dorado@cgiar.org , j.c.rivera@cgiar.org



El CIAT es un Centro de Investigación de CGIAR

Regresión lineal multiple

Es un modelo matemático que busca ajustar una ecuación lineal que maximice las relaciones entre una variable dependiente 'Y' y un conjunto de variables independientes (X1,X2,...,Xn) y un término de error.



Fuente: e-stadistica.bio.ucm.es

Principales supuestos

- Relaciones lineales entre variables
- Las mediciones deben ser independientes
- Los errores deben tener varianza constante
- Los errores deben seguir una distribución normal

Ejemplo regresión lineal multiple

El **ozono** (O_3) es una sustancia cuya molécula está compuesta por tres átomos de oxígeno, Actúa en la atmósfera como depurador del aire y sobre todo como filtro de los rayos ultravioletas procedentes del Sol. Sin ese filtro la existencia de vida en la Tierra sería completamente imposible.

Fuente: <https://es.wikipedia.org/wiki/Ozono>

Conjunto de datos de prueba

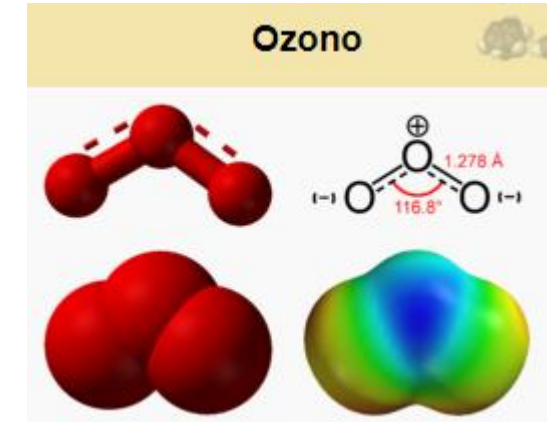
Mediciones diarias de la calidad del aire en Nueva York, de mayo a septiembre de 1973.

Conjunto de datos 154 observaciones sobre 6 variables.

Y = Ozone Ozono numérico (ppb)
X1 = Solar.R Radiación solar (lang)
X2 = Wind Viento (mph)
X3 = Temp Temperatura (grados F)

Propósito

Construir un modelo de regression lineal multiple para predecir el ozono en función de variables climáticas.



Fuente: <https://es.wikipedia.org/wiki/Ozono>

Ejemplo regresión lineal multiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

```
lm.calidadAire <- lm(Ozone~.,data=calidadAire)
summary(lm.calidadAire)
```

Coefficients:

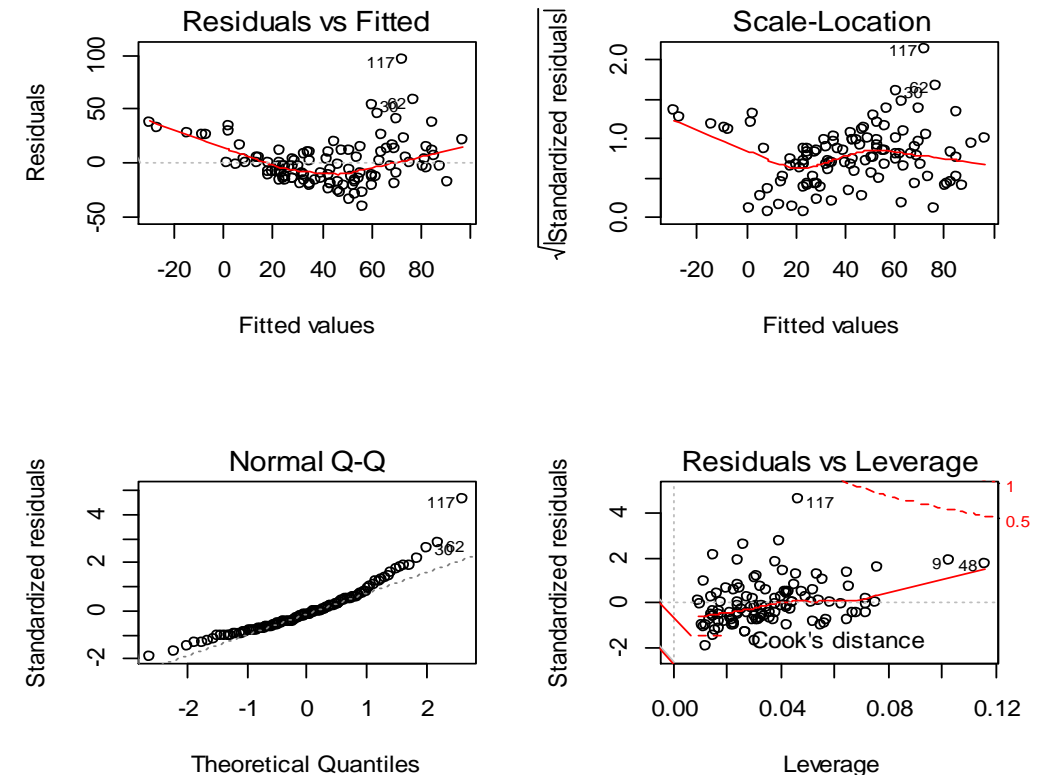
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-64.34208	23.05472	-2.791	0.00623 **
Solar.R	0.05982	0.02319	2.580	0.01124 *
wind	-3.33359	0.65441	-5.094	1.52e-06 ***
Temp	1.65209	0.25353	6.516	2.42e-09 ***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.18 on 107 degrees of freedom
Multiple R-squared: 0.6059, Adjusted R-squared: 0.5948
F-statistic: 54.83 on 3 and 107 DF, p-value: < 2.2e-16

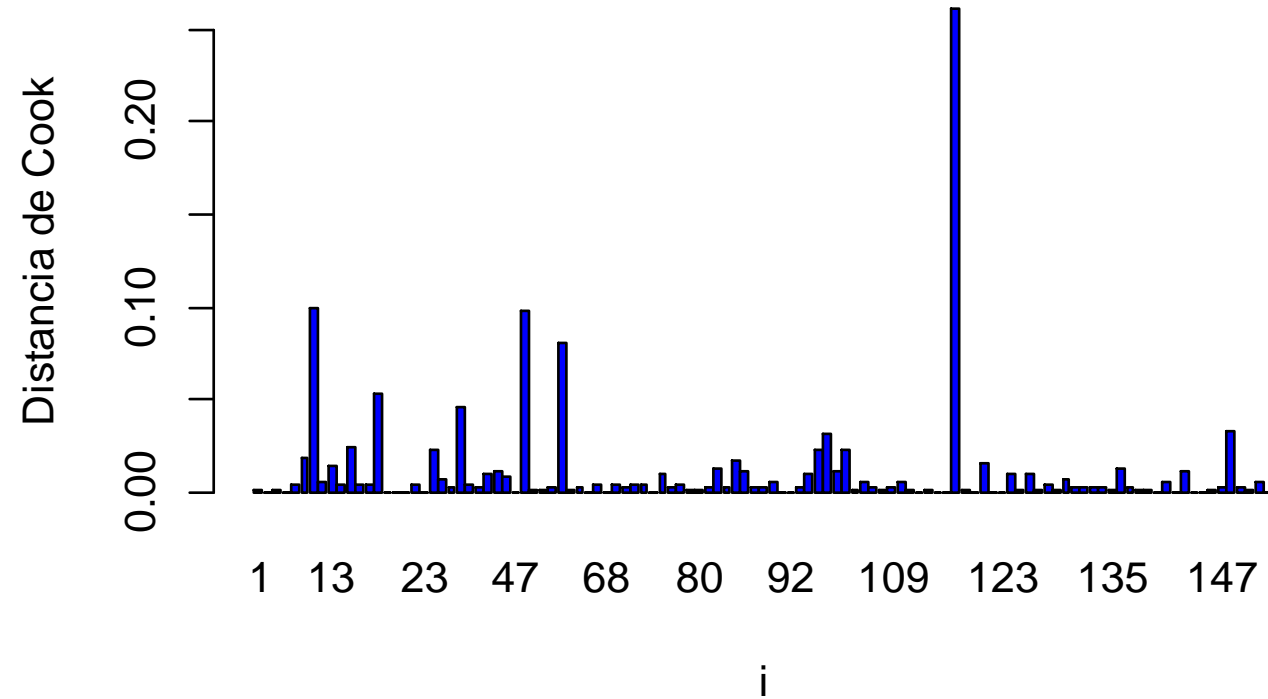
$$Y = -64 + 0.06_1 X_1 - 3.33 X_2 + 1.66 X_3 + \varepsilon$$

```
layout(matrix(1:4,2,2))
plot(lm.calidadAire)
```



Detección de outliers con distancia de cook

```
cook = cooks.distance(lm.calidadAire)  
barplot(cook,col='blue',xlab='i',ylab='Distancia de Cook')
```



Análisis de varianza

El objetivo principal de muchos experimentos consiste en determinar el efecto que sobre alguna variable dependiente **Y** tienen distintos niveles de algún factor **X** (variable independiente y discreta).

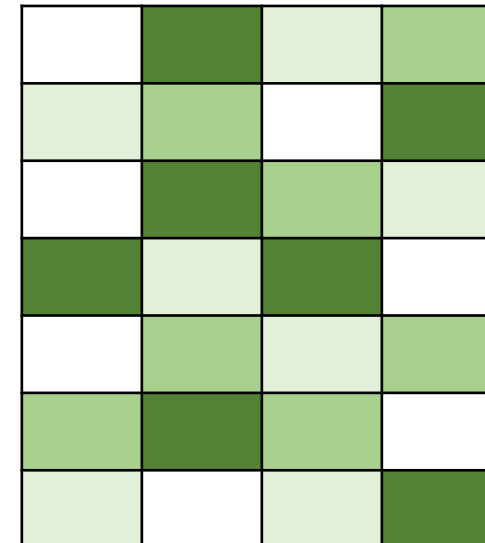
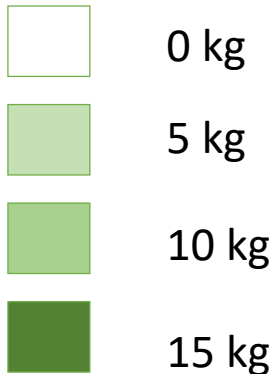
El efecto se evalúa a través de la comparación de las medias de cada nivel de la variable discreta X.

Fuente: <https://www.uoc.edu/in3/emath/docs/ANOVA.pdf>

$$Y = \mu + \tau + \varepsilon$$

Hay algún efecto sobre el rendimiento (Y) de acuerdo a la cantidad de nitrógeno (X) aplicada en mi finca?, donde se presentan las diferencias?

Cantidad aplicada de nitrógeno



Ejemplo de análisis de varianza

El conjunto de datos **Iris flor** es un [conjunto de datos multivariante](#) introducido por [Ronald Fisher](#) 1936, coleccionó la data usada para cuantificar la variación [morfológica](#) del [Iris](#) con las flores de tres especies relacionadas

Fuente: https://es.wikipedia.org/wiki/Iris_flor_conjunto_de_datos

Iris es un marco de datos con 150 casos (filas) y 5 variables (columnas)

Sepal.Length: Largo de sépalos

Sepal.Width: Ancho de sépalos

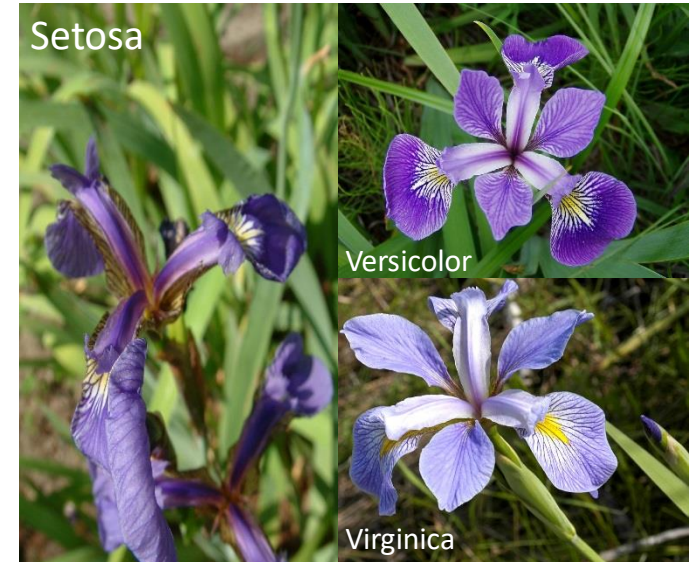
Petal.Length: Largo de pétalos

Petal.Width: Ancho de pétalos

Species: Especie (Setosa, Versicolor y Virginica)

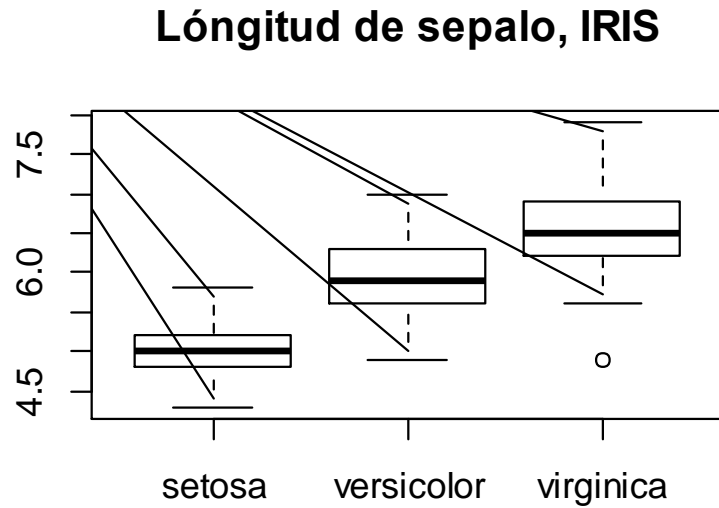
Propósito

Evaluar si hay un efecto entre las dimensiones del sépalos y la especie y en caso de encontrarlo identificar entre que cual de ellas se presenta dicha diferencia.

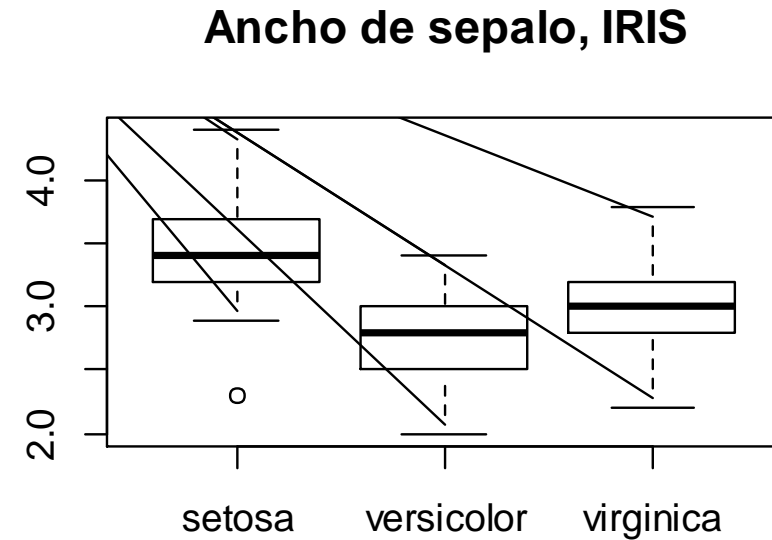


Ejemplo de análisis de varianza

```
boxplot(Sepal.Length~Species,data=iris,main="Lóngitud de sepalo, IRIS")
```



```
boxplot(Sepal.Width~Species,data = iris,main="Ancho de sepalo, IRIS")
```

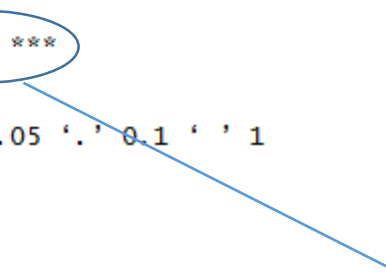


Ejemplo de análisis de varianza

```
> aov.irisLength <- aov(Sepal.Length~Species,data=iris)
> summary(aov.irisLength)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	63.21	31.606	119.3	<2e-16 ***
Residuals	147	38.96	0.265		

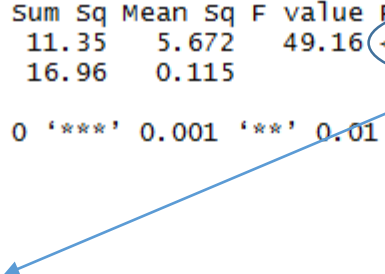
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



```
> aov.irisSepal <- aov(Sepal.width~Species,data=iris)
> summary(aov.irisSepal)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	11.35	5.672	49.16	<2e-16 ***
Residuals	147	16.96	0.115		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Significancia

Prueba de TukeyHSD

```
> TukeyHSD(aov.irisLength)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = Sepal.Length ~ Species, data = iris)
```

\$Species		diff	lwr	upr	p adj
versicolor-setosa		0.930	0.6862273	1.1737727	0
virginica-setosa		1.582	1.3382273	1.8257727	0
virginica-versicolor		0.652	0.4082273	0.8957727	0

```
> TukeyHSD(aov.irisSepal)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = Sepal.width ~ Species, data = iris)
```

\$Species		diff	lwr	upr	p adj
versicolor-setosa		-0.658	-0.81885528	-0.4971447	0.0000000
virginica-setosa		-0.454	-0.61485528	-0.2931447	0.0000000
virginica-versicolor		0.204	0.04314472	0.3648553	0.0087802

Mas información

Regresión lineal multiple.

- <http://r-statistics.co/Linear-Regression.html>
- <https://www.r-bloggers.com/simple-linear-regression-2/>
- <https://datascienceplus.com/how-to-apply-linear-regression-in-r/>

Análisis de varianza y diseños experimentales.

- <http://www.r-tutor.com/elementary-statistics/analysis-variance>
- <https://cran.r-project.org/web/packages/agricolae/vignettes/tutorial.pdf>
- <https://www.jstatsoft.org/article/view/v043b05/v43b05.pdf>

¡Gracias!



NOS ENORGULLECE
HABER CELEBRADO 50 AÑOS
DE INVESTIGACIÓN AGRÍCOLA
PARA EL DESARROLLO

Centro Internacional de Agricultura Tropical - CIAT

Sede Principal y Oficina Regional
para Suramérica y el Caribe

+57 2 445 0000

Km 17 Recta Cali-Palmira
A.A. 6713, Cali, Colombia

✉ ciat@cgiar.org

🌐 ciat.cgiar.org



El CIAT es un Centro de Investigación de CGIAR