Herramientas de Estadística (No)Paramétrica

(Parte 1)

Juan C. Correa

Material de uso exclusivo para
INGENIO PANTALEON, S.A.
Diagonal 6, 10-31, Zona 10

Ciudad de Guatemala

¿Por qué hablar de Estadística Paramétrica?

Aparte de describir, la estadística tiene el propósito de estimar la relación entre variables y para ello usa su recurso más popular conocido como Modelo Lineal.

Global Validation of Linear Model Assumptions

Edsel A. PEÑA and Elizabeth H. SLATE

An easy-to-implement global procedure for testing the four assumptions of the linear model is proposed. The test can be viewed as a Neyman smooth test and relies only on the standardized residual vector. If the global procedure indicates a violation of at least one of the assumptions, then the components of the global test statistic can be used to gain insight into which assumptions have been violated. The procedure can also be used in conjunction with associated deletion statistics to detect unusual observations. Simulation results are presented indicating the sensitivity of the procedure in detecting model violations under a variety of situations, and its performance is compared with three potential competitors, including a procedure based on the Box–Cox power transformation. The procedure is demonstrated by applying it to a new car mileage dataset and a water salinity dataset that has been used earlier to illustrate model diagnostics.

KEY WORDS: Box–Cox transformation; Deletion statistics; Model diagnostics and validation; Neyman smooth test; Outlier detection; Score test.

No obstante, para que la estimación de relaciones entre variables sea adecuada, se debe verificar la validez de los supuestos en los que se apoya el modelo lineal. Esto en la práctica, sin embargo, se omite e ignora frecuentemente.

Supuestos del Modelo Lineal

El modelo lineal establece que la relación entre una variable dependiente Y y un conjunto de variables independientes X está dada así

$$Y = X\beta + \sigma\epsilon \tag{1}$$

Siempre y cuando se verifique que se cumplen estos cuatro supuestos.

- A1 Linealidad: $E\{Y_i|X\} = x_i\beta$
- A2 Homoscedasticidad: $\text{var}\{Y_i|X\} = \sigma^2,\ i = 1, 2, ..., n$
- A3 Independencia: $\text{cov}\{Y_i, Y_j|X\} = 0(i \neq j)$
- A4 Normalidad: $(Y_1, Y_2, ..., Y_n|X) = \mathfrak{N}_n(\mu, \sigma\}$

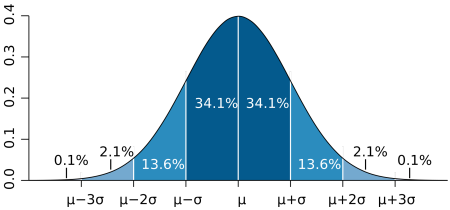
Supuestos del Modelo Lineal

$$
\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{2,1} & . & . & . & x_{k,1} \\ x_{1,2} & x_{2,2} & . & . & . & x_{k,2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{1,n} & x_{2,n} & . & . & . & x_{k,n} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix} \tag{2}
$$

En esta ecuación, se asume que para cada valor de $Y_i, i = 1, 2, ..., n$, la distribución de las variables independientes X es de tipo normal multivariada y también ϵ_i tiene una distribución normal. Afirmar que una variable tiene una distribución normal, significa que las siguientes propiedades son rigurosamente ciertas y comprobables...

Propiedades de Distribución Normal

1. La forma de la distribución es perfectamente simétrica
2. $\bar{X} = Md = Mod$
3. $Z_x \sim N(0,1)$
4. $As = Ku = 0$



Si alguna de estas propiedades no es comprobable, no podemos afirmar, en sentido estricto, que se dispone de una variable con distribución normal.

Estadística Paramétrica: Una Definición

La estadística paramétrica es una rama de la estadística inferencial que abarca a los procedimientos que, apoyándose de distribuciones con propiedades conocidas, nos permiten estimar relaciones entre variables a partir de un número finito y conocido de parámetros.

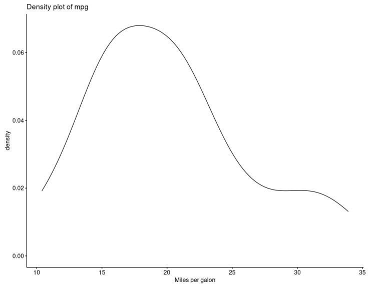
Para la inferencia paramétrica, se requiere como mínimo una variable con nivel de medición intervalo. Las relaciones entre variables nominales u ordinales no pueden ser estimables empleando técnicas de estadística paramétrica, sino a través de técnicas no paramétricas o semi-paramétricas, cuyos supuestos son más flexibles o relajados a los supuestos del modelo lineal.

Estadística No Paramétrica: Una Definición

La estadística no paramétrica es una rama de la estadística inferencial que abarca procedimientos que permiten estimar relaciones entre variables sin que se conozca, a priori, las propiedades de la distribución y el número de parámetros requeridos para tal propósito.

Técnica Gráfica (ggdensity)

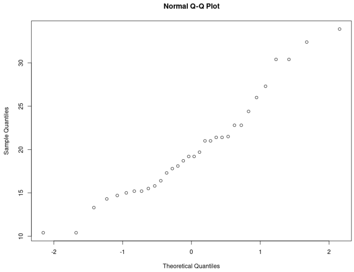
```
library("ggpubr")
ggdensity(mtcars$mpg,
          main = "Density plot of mpg",
          xlab = "Miles per galon")
```



Técnicas para Evaluar la Normalidad de una distribución

Técnica Gráfica (qqnorm)

```
library("ggpubr")
qqnorm(mtcars$mpg)
```



Aquí la nube de puntos debe formar una línea recta. De lo contrario, se debe interpretar que la distribución no es normal.

Técnicas para Evaluar la Normalidad de una distribución

Técnica de Cálculo (psych::describe)

```
library("psych")
describe(mtcars$mpg)
```

vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
1	32	20.09	6.03	19.2	19.7	5.41	10.4	33.9	23.5	$\boxed{0.61}$	$\boxed{-0.37}$	1.07

Asimetría ≠ 0

Kurtosis ≠ 0

También podría usar e1071::skewness(mtcars$mpg)

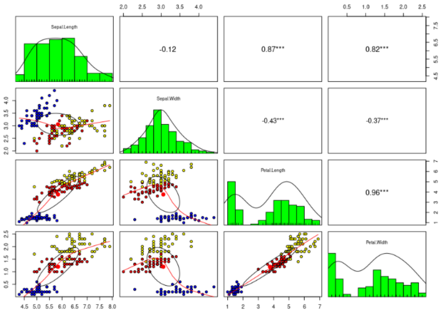
Relaciones Bivariadas

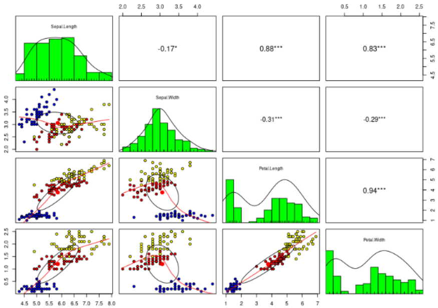
El concepto de correlación es posiblemente uno de los términos mejor conocidos para representar la asociación entre dos variables. En R, podemos calcular las siguientes formas de correlación:

1. Correlación de Pearson (r) (Técnica Paramétrica para variables de intervalo)
2. Correlación de Spearman (ρ) (Técnica No Paramétrica para una variable ordinal con otro ordinal, discretas o continuas)
3. Correlación de Kendall (τ) (Técnica No Paramétrica para una variable ordinal con otra ordinal discreta).
4. Correlación Punto-Biserial (r_{pb}) (Técnica paramétrica para una variable continua y otra variable dicotómica).

```
pairs.panels(iris[-5], method = "pearson",
        bg=c("blue","red","yellow")[iris$Species],
        hist.col = "green", pch=21, stars = TRUE)
```

```
pairs.panels(iris[-5], method = "spearman",
        bg=c("blue","red","yellow")[iris$Species],
        hist.col = "green", pch=21, stars = TRUE)
```





¿Es relevante entonces saber escoger el método de cálculo?