# Regresión Múltiple

En esta lección se presenta un método para analizar una relación lineal que incluye más de dos variables. Nos enfocamos en tres elementos fundamentales: 1. la ecuación de regresión múltiple, 2. el valor de R2 ajustada y 3. el valor P.

**Ecuación de regresión múltiple**

Se usa la siguiente ecuación de regresión múltiple para describir relaciones lineales que incluyen más de dos variables.

Una **ecuación de regresión múltiple** expresa una relación lineal entre una variable de respuesta *y* y dos o más variables de predicción $(x_1, x_2, \cdots, x_k)$. La forma general de una ecuación de regresión múltiple es

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

**Notación**

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$ (forma general de la ecuación de regresión múltiple estimada).

n=tamaño de la muestra.

k=número de variables de predicción (las variables de predicción también se conocen como variables independientes o variables x).

$\hat{y}$ = valor predicho de *y* (se calcula por medio de la ecuación de regresión múltiple).

$x_1, x_2, \cdots, x_k$ = son las variables de predicción.

$\beta_0$= intercepto y, o el valor de *y* cuando todas las variables de predicción son 0 (este valor es un parámetro poblacional).

$b_0$ = estimado de $\beta_0$basado en los datos muestrales ($b0$ es un estadístico muestral).

$\beta_1, \beta_2, \cdots \beta_k$ son los coeficientes de las variables de predicción $x_1, x_2, \cdots, x_k$

$b_1, b_2, \cdots b_k$ son estimados muestrales de los coeficientes$\beta_1, \beta_2, \cdots \beta_k$.

# Regresión Múltiple

Para cualquier conjunto específico de valores de x, la ecuación de regresión está asociada con un error aleatorio que suele denotarse por $\varepsilon$, y suponemos que estos errores se distribuyen normalmente, con una media de 0 y una desviación estándar de $\sigma$, y que los errores aleatorios son independientes.

Si una ecuación de regresión múltiple se ajusta bien a los datos muestrales, se puede emplear para hacer predicciones.

## $R^2$ ajustada

$R^2$ denota el **coeficiente múltiple de determinación**, que es una medida de lo bien que se ajusta la ecuación de regresión múltiple a los datos muestrales. Un ajuste perfecto daría como resultado $R^2 = 1$, y un ajuste muy bueno daría por resultado un valor cercano a 1. Un ajuste muy deficiente se relaciona con un valor de $R^2$cercano a 0.

La $R^2$ más grande se obtiene por el simple hecho de incluir todas las variables disponibles, pero la mejor ecuación de regresión múltiple no necesariamente utiliza todas las variables disponibles. A causa de esta desventaja, la comparación de diferentes ecuaciones de regresión múltiple se logra mejor con el coeficiente ajustado de determinación, que es $R^2$ ajustada para el número de variables y el tamaño de la muestra.

### Definición

El **coeficiente ajustado de determinación** es el coeficiente múltiple de determinación $R^2$ modificado para justificar el número de variables y el tamaño de la muestra. Se calcula por medio de la fórmula:

$$R^2\ ajustada = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

Donde

n=tamaño muestral.

k=número de variables de predicción (x).

**Valor P**

El valor P es una medida de la significancia general de la ecuación de regresión múltiple.

**Lineamientos para el cálculo de la mejor ecuación de regresión múltiple**

1. Utilice el sentido común y consideraciones prácticas para incluir o excluir variables. Por ejemplo, podríamos excluir la variable de la estatura después de saber que esta variable es un estimado visual y no una medida exacta.

2. Considere el valor P. Seleccione una ecuación que tenga significancia general, tal como lo determina el valor P indicado en los resultados del programa de cómputo.

3. Considere ecuaciones con valores altos de $R^2$ ajustada y trate de incluir solo unas cuantas variables. En vez de incluir casi todas las variables disponibles, trate de incluir relativamente pocas variables de predicción (x). Utilice los siguientes lineamientos:

- Seleccione una ecuación que tenga un valor de $R^2$ ajustada con esta propiedad: si se incluye una variable independiente adicional, el valor de $R^2$ ajustada no se incrementa de manera sustancial.
- Para un número dado de variables de predicción (x), seleccione la ecuación con el valor más grande de la $R^2$ ajustada.
- Para eliminar las variables de predicción (x) que no tienen mucho efecto sobre la variable de respuesta (y), sería útil calcular el coeficiente de correlación

lineal r para cada par de variables en consideración. Si dos valores de predicción tienen un coeficiente de correlación lineal muy alto, no es necesario incluir a ambos, y debemos excluir la variable con el valor de r más bajo.

**Referencia:** Triola, M., (2013). Estadística. Decimoprimera edición. Pearson educación. México