

ENRIC RUIZ MORILLAS

ESTADÍSTICA APLICADA A EXPERIMENTOS Y MEDICIONES

REGRESIÓN LINEAL



ESTADÍSTICA APLICADA A EXPERIMENTOS Y MEDICIONES

REGRESIÓN LINEAL

ENRIC RUIZ MORILLAS

ESTADÍSTICA APLICADA A EXPERIMENTOS Y MEDICIONES

REGRESIÓN LINEAL



Madrid • Buenos Aires • México • Bogotá

© Enric Ruiz Morillas, 2025

Edición ebook

Reservados todos los derechos.

«No está permitida la reproducción total o parcial de este libro, ni su tratamiento informático, ni la transmisión de ninguna forma o por cualquier medio, ya sea electrónico, mecánico por fotocopia, por registro u otros métodos, sin el permiso previo y por escrito de los titulares del Copyright.»

Ediciones Díaz de Santos

Internet: <http://www.editdiazdesantos.com>

E-mail: ediciones@editdiazdesantos.com

ISBN: 978-84-9052-564-7 (edición papel)

e-ISBN: 978-84-9052-565-4 (edición digital)

Diseño de cubierta y Fotocomposición: P55 Servicios Culturales

Per a l'Anna, la Andrea i la Sarai

Índice

1. MODELOS ESTADÍSTICOS	1
2. EL MODELO DE REGRESIÓN LINEAL SIMPLE	3
3. CORRELACIÓN.....	31
4. EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE.....	41
5. MULTICOLINEALIDAD.....	55
6. EL MODELO DE REGRESIÓN POLINÓMICO	91
7. LA MEDIDA INSTRUMENTAL	95
8. EL LÍMITE DE DETECCIÓN.....	99
Bibliografía	107

MODELOS ESTADÍSTICOS

El Análisis de Regresión es una parte de la Estadística que tiene como objetivo obtener modelos matemáticos que describan las *relaciones* existentes entre variables.

Las relaciones estadísticas no implican necesariamente relaciones causales, sin embargo, sí son un punto de partida para establecerlas.

Una vez se ha comprobado la existencia de una relación estadística entre variables, la modelización matemática permite realizar *predicciones*; hay que tener en cuenta que el modelo puede no ser válido fuera del rango normal de variación de las variables consideradas como explicativas.

En una primera aproximación al problema de la regresión, es útil representar como puntos en la denominada *gráfica de dispersión* los valores obtenidos experimentalmente de las variables cuya relación se quiere analizar y, a continuación, ajustar una curva a través de los puntos de manera que los mismos queden lo más cerca posible de la curva. En la mayor parte de los casos no se obtendrá un ajuste exacto, porque al menos una de las variables está afectada por condiciones no controladas y/o por errores de medida que introducen una variabilidad de carácter aleatorio.

Normalmente se quiere estudiar cómo una variable (llamada *variable respuesta* o *dependiente*) depende de una o varias variables (llamadas *variables regresoras* o *independientes*).

El caso más simple se tiene cuando hay solo *dos variables relacionadas de forma lineal* y una de ellas es controlada por el experimentador. El modelo de regresión en este caso es una línea recta que podemos denominar *recta de regresión*, de la forma:

$$y = \beta_0 + \beta_1 x$$

Sin embargo, debido a que los puntos representados en el diagrama de dispersión no se sitúan exactamente sobre una línea recta, esta ecuación tiene que ser modificada para tener en cuenta este hecho.

Se llama ε a la diferencia entre el valor observado de la variable respuesta y la línea recta. El error tiene en cuenta la falta de capacidad del modelo para ajustar los datos de manera exacta. De este modo, una manera más adecuada de modelizar la variable y frente a la variable x es mediante la siguiente expresión:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

La cual recibe el nombre de *modelo de regresión lineal simple*.

De forma general, la variable respuesta y puede estar relacionada con un número k de variables regresoras $x_1, x_2, \dots, x_k$. Si la relación es lineal, puede modelizarse la variable y frente a las variables $x_1, x_2, \dots, x_k$, mediante la siguiente expresión:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

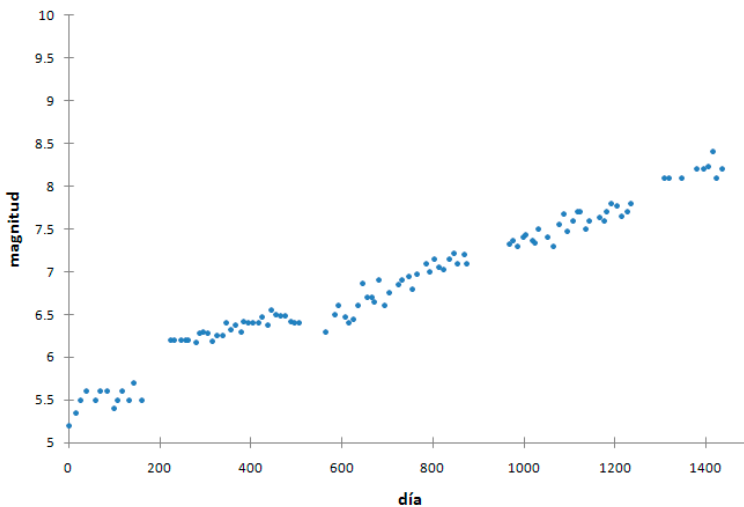
La cual recibe el nombre de *modelo de regresión lineal múltiple*.

Los parámetros $\beta_0, \beta_1, \dots, \beta_k$ son los *coeficientes de regresión* del modelo, y se estiman a partir de un conjunto de datos obtenidos mediante experimentación. Hay que tener en cuenta que las técnicas del análisis de regresión lineal para la estimación de los coeficientes de regresión podrán aplicarse *siempre que la relación sea lineal en los parámetros* $\beta_0, \beta_1, \dots, \beta_k$. Así, mediante el análisis de regresión lineal será posible establecer modelos de regresión, aunque no exista una relación lineal entre la variable respuesta y la variable regresora, como es el caso, por ejemplo, de una relación polinómica:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

EL MODELO DE REGRESIÓN LINEAL SIMPLE

Consideremos el experimento en el que a lo largo de una serie de días se observa la magnitud de una estrella. Si representamos los datos experimentales en la gráfica de dispersión se obtiene la siguiente representación:



Cada uno de los días en los que se ha realizado una observación de la magnitud de la estrella es un valor de una variable que denominaremos *variable regresora*, y que representaremos por la letra x . Los valores de esta variable están controlados por el experimentador. Las magnitudes observadas son los valores posibles y de una variable que denominaremos *variable respuesta*, y que representaremos por la letra Y . Los valores de esta variable están sometidos a condiciones no controladas por el

experimentador y están afectados por errores de medida, por lo que la variable Y es una variable aleatoria.

Mediante la gráfica de dispersión se puede observar que existe una dependencia de tipo lineal entre los valores y de la variable aleatoria Y , y los valores x fijados por el experimentador. Relacionaremos la *respuesta* y con el *regresor* x mediante la ecuación denominada *modelo de regresión lineal simple*:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

donde los *parámetros* β_0 y β_1 se llaman *coeficientes de regresión*. ε es la componente aleatoria de los valores posibles de la variable Y , y la denominaremos *error*.

Consideraremos que ε son los valores posibles de una variable aleatoria ε que sigue una ley de distribución de probabilidades *normal*, y que asumiremos que tiene de media cero y varianza σ^2 . En consecuencia, consideraremos que la variable aleatoria Y es una variable aleatoria *normal*.

Por tanto, por haber considerado que el error es una variable aleatoria $N(0, \sigma)$, se cumple que:

$$E(Y) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x$$

Lo cual nos indica que *el valor medio de los posibles valores y para una valor dado x* es una función lineal de x , y podemos escribir:

$$E(Y) = E(y|x) = E(y) = \beta_0 + \beta_1 x$$

que, de forma simplificada, puede leerse como que *el valor medio de y es una función lineal de x* .

Así, podemos ver que la *recta de regresión* proporciona el *valor medio de los posibles valores y* que pueden obtenerse mediante experimentación *para una valor dado x* .

Por otra parte, por haber considerado que el error es una variable aleatoria $N(0, \sigma)$, se cumple que:

$$\text{var}(Y) = \text{var}(y|x) = \text{var}(y) = \text{var}(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$$

lo que indica que la varianza de los posibles valores de y no depende del valor de x . Es decir, que la *dispersión* de los posibles valores y que pueden obtenerse mediante experimentación para una valor dado x , *es la misma para cada x* .

Los *coeficientes de regresión* β_0 y β_1 , y la *varianza del error* σ^2 son parámetros desconocidos y deben ser *estimados* a partir de un conjunto de datos obtenidos mediante experimentación.

Estimación de β_0 y β_1

Los coeficientes de regresión β_0 y β_1 son parámetros poblacionales que se estiman a partir de una muestra formada por los n pares de datos (x_1, y_1) , $(x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)$ obtenidos en el experimento dado.

Para ello se utiliza el *método de los mínimos cuadrados*. Se estima β_0 y β_1 , de forma que la suma de los cuadrados de los errores correspondientes a cada una de las observaciones experimentales ε_i sea mínima. Los errores ε_i se determinan por el conjunto de ecuaciones que se obtienen al sustituir los n pares de datos del experimento en el modelo de regresión lineal simple:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \dots, n$$

Por tanto, la suma de los cuadrados de los errores ε_i es:

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Y las condiciones que tienen que cumplirse para que $S(\beta_0, \beta_1)$ sea mínima son:

$$\left. \frac{\partial S}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left. \frac{\partial S}{\partial \beta_1} \right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son las *estimaciones puntuales* de β_0 y β_1 y respectivamente.

A la función $S(\beta_0, \beta_1)$ se la denomina *función de mínimos cuadrados*.

Estas dos ecuaciones simplificadas quedan de la forma:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$