

Análisis de Regresión

La ecuación de regresión expresa una relación entre x (llamada variable explicativa, variable de predicción o variable independiente) y \hat{y} (llamada variable de respuesta o variable dependiente). La ecuación típica de una línea recta $y = mx + b$ está expresada en la forma $\hat{y} = b_0 + b_1x$, donde b_0 es el intercepto y (o la intersección en y) y b_1 es la pendiente. La notación dada muestra que b_0 y b_1 son estadísticos muestrales utilizados para estimar los parámetros poblacionales β_0 y β_1 .

Requisitos

1. La muestra de datos apareados (x, y) es una muestra aleatoria de datos cuantitativos.
2. El examen visual del diagrama de dispersión indica que los puntos se aproximan al patrón de una línea recta.
3. Se debe eliminar cualquier valor extremo si se sabe que es un error. Es importante tomar en cuenta los efectos de cualquier valor extremo que no sea un error conocido.

Nota: Los requisitos 2 y 3 representan una verificación simplificada de los siguientes requisitos formales del análisis de regresión:

- Para cada valor fijo de x , los valores correspondientes de y tienen una distribución en forma de campana.
- Para los distintos valores fijos de x , las distribuciones de los valores correspondientes de y tienen la misma varianza.
- Para los distintos valores fijos de x , las distribuciones de los valores correspondientes de y tienen medias que se ubican en la misma línea recta.
- Los valores de y son independientes.

Análisis de Regresión

Los resultados no se ven muy afectados si la distribución no se aleja demasiado de la normalidad y si las varianzas no son demasiado diferentes.

Definiciones

Dado un conjunto de datos muestrales apareados, la ecuación de regresión

$$\hat{y} = b_0 + b_1x$$

describe algebraicamente la relación entre las dos variables. La gráfica de la ecuación de regresión se denomina recta de regresión (*o recta del mejor ajuste o recta de mínimos cuadrados*).

Notación para la ecuación de regresión

	Parámetro poblacional	Estadístico muestral
Intercepto y de la ecuación de regresión	β_0	b_0
Pendiente de la ecuación de regresión	β_1	b_1
Ecuación de la recta de regresión	$y = \beta_0 + \beta_1x$	$\hat{y} = b_0 + b_1x$

Análisis de Regresión

Cálculo de la pendiente b_1 y de b_0 (el intercepto y) en la ecuación de regresión

$$\hat{y} = b_0 + b_1x$$

Pendiente:

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Intercepto y :

$$b_0 = \bar{y} - b_1\bar{x}$$

Uso de la ecuación de regresión para hacer predicciones

Se debe utilizar la ecuación de regresión para hacer predicciones solo si existe una correlación lineal. En ausencia de una correlación lineal, no debemos usar la ecuación de regresión para proyectar o predecir; en vez de ello, nuestro mejor estimado de la segunda variable es simplemente su media muestral.

Al predecir un valor de y con base en algún valor dado de x . . .

1. Si no existe una correlación lineal, el mejor valor predicho de y es \bar{y} .
2. Si existe una correlación lineal, el mejor valor predicho de y se calcula sustituyendo el valor de x en la ecuación de regresión.

Análisis de Regresión

Lineamientos para el uso de la ecuación de regresión

1. Si no existe una correlación lineal, no utilice la ecuación de regresión para hacer predicciones.
2. Cuando utilice la ecuación de regresión para hacer predicciones, permanezca en el ámbito de los datos muestrales disponibles. Si usted calcula una ecuación de regresión que relaciona la estatura y el número de calzado de mujeres, es absurdo predecir el número de calzado de una mujer que mide 10 pies de estatura.
3. Una ecuación de regresión que está basada en datos antiguos no necesariamente es válida ahora. La ecuación de regresión que relaciona precios de automóviles usados con la antigüedad de los automóviles ya no es útil si está basada en datos de la década de 1990.
4. No haga predicciones acerca de una población distinta de la población de donde se obtuvieron los datos muestrales. Si reunimos datos muestrales de hombres y desarrollamos una ecuación de regresión que relaciona la edad con el uso del control remoto del televisor, los resultados no necesariamente se aplican a las mujeres. Si empleamos promedios estatales para desarrollar una ecuación de regresión que relaciona las calificaciones de matemáticas del SAT con las calificaciones verbales del SAT, los resultados no necesariamente se aplican a los individuos.

Referencia:

Triola, M., (2013). Estadística. Decimoprimer edición. Pearson educación. México