TECNOLÓGICO NACIONAL DE MÉXICO

Ciencia de DATOS

Regresión Lineal Simple y Múltiple

PASOS PARA UN ANÁLISIS DE REGRESIÓN

Un análisis de Regresión Lineal tiene lugar cuando deseamos explicar a qué se debe el comportamiento de una **variable de interés** (y) y suponemos una posible relación con otras variables que pueden estar influyendo sobre ella ($x_1, x_2, \dots, x_k$). De esta manera, el análisis de regresión consiste principalmente en ir respondiendo preguntas con respecto a esta posible relación, las cuales son contestadas en cada caso con una prueba o método estadístico.

De manera general, podemos listar las siguientes preguntas a responder con sus correspondientes herramientas para darles respuesta:

¿Es posible suponer una **relación lineal** entre las variables?

Gráfico de Dispersión.

↓

¿Cuál es el **modelo de regresión** que mejor ajusta la relación entre las variables?

Método de mínimos cuadrados para la obtención de parámetros del modelo.

↓

¿Existe una **relación significativa** entre las variables, con un nivel de confianza específico?

Prueba de significancia (Análisis de Varianza, prueba t).

↓

¿Qué porcentaje de variabilidad de y, se puede explicar a partir del modelo obtenido? ¿Es satisfactorio?

Coeficiente de Determinación R^2

↓

¿Qué tan fuerte es la relación lineal entre las variables ?

Coeficiente de Correlación R

↓

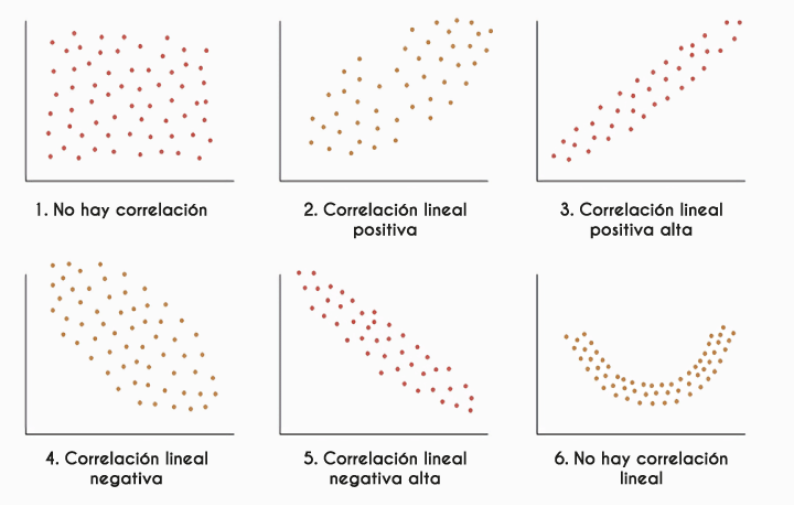
¿El modelo cumple los supuestos estadísticos?

Verificación de supuestos (normalidad, independencia, aleatoriedad)

- Cuando todas las preguntas anteriores han sido contestadas y se ha obtenido un modelo que ajuste satisfactoriamente el comportamiento de y en función de la o las variables $x's$, entonces pueden estimarse los valores de y para diferentes valores de x, de manera puntual y por intervalos. A continuación, se describe cada paso descrito en el esquema anterior.

1. Gráfico de Dispersión

Esta imagen muestra cómo interpretar la relación entre las variables de acuerdo a su apariencia en el gráfico de dispersión. En el gráfico 6 se representa solo un posible caso de relación **no lineal**, pero pueden existir otros patrones que indiquen una posible relación exponencial, logarítmica, polinomial, entre otros.



2. Modelo de regresión

A través del método de mínimos cuadrados se obtienen los **coeficientes de regresión** β_o, $\beta_1,\ \beta_2, \dots,\ \beta_k$ que definen el modelo de regresión.

$y = \beta_0 + \beta_1 x$ *en el caso de Regresión Lineal Simple*

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ *en el caso de Regresión Lineal Múltiple*

Donde k es el número de **variables regresoras** (x) que incluye el modelo.

3. Prueba de Significancia

Con un análisis de varianza o una prueba **t**, se prueban las hipótesis:

Prueba individual

$H_0: \beta_i = 0$ *No existe una relación significativa entre las variables* $\boldsymbol{y}$ *y* $\boldsymbol{x_i}$

$H_1: \beta_i \neq 0$ *Existe una relación significativa entre las variables* $\boldsymbol{y}$ *y* $\boldsymbol{x_i}$

Prueba general para Regresión Lineal

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ *Ninguna* $\boldsymbol{x}$ *tiene una relación significativa con* $\boldsymbol{y}$.

$H_1: \beta_i \neq 0$ *para al menos unai con* $i = 1, 2, \dots, k$

Al menos una $\boldsymbol{x}$ *tiene una relación significativa con* $\boldsymbol{y}$.

Rechazar Ho también implica que la variable x tiene un efecto significativo sobre el comportamiento de la variable y.

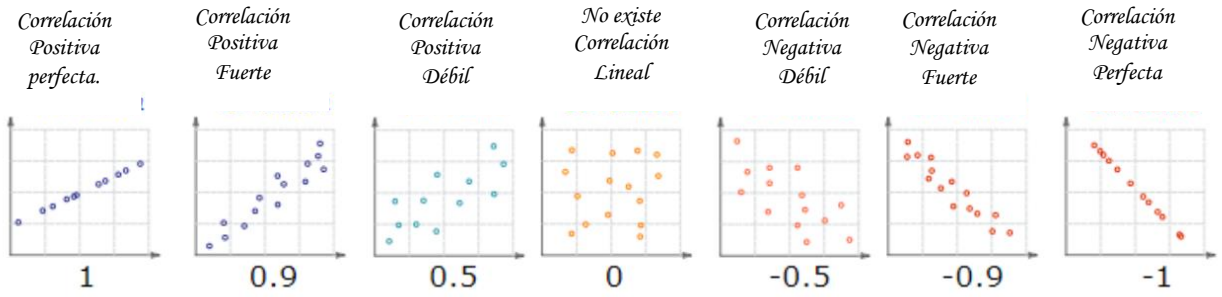
4. Coeficiente de Determinación

El coeficiente de Determinación mide la **variabilidad de y que puede ser explicada a partir de x**; puede tomar valores de 0 a 1, es decir, $0 \leq R^2 \leq 1$.
Un valor de **0** indica que la variable regresora (o las variables regresoras) no aportan información para explicar el comportamiento de la variable dependiente, mientras que un valor del **1**, implicaría que el **100%** de la variabilidad de y, puede ser explicada a partir de las variables x .

5. Coeficiente de Correlación

El coeficiente de correlación mide **qué tan fuerte es la relación lineal entre las variables**; puede tomar valores de -1 a 1, es decir, $-1 \leq R \leq 1$.
Un valor de **-1** indica que la existe una relación lineal fuerte, pero **negativa** (tendencia descendente –> si x aumenta, y disminuye), un valor de **1** indica que existe una relación lineal fuerte, **positiva** (tendencia ascendente –> si x aumenta, y también aumenta), valores de **0** indican una relación lineal **nula**.

En esta imagen se pueden ver algunos valores de R y lo que podría interpretarse con respecto a la relación entre las variables.



6. Verificación de Supuestos Estadísticos

Los métodos de prueba de Significancia y estimación de valores de y en función de x, están basados en los siguientes supuestos estadísticos:

a) Los residuos e_j, siguen una distribución normal $N(0, \sigma_e^2)$.
b) Los residuos e_j son independientes con respecto a x.
c) Los residuos e_j son independientes con respecto a los valores estimados $\hat{y}_j$.
d) Los residuos e_j son independientes con respecto al orden en que se obtuvieron las **n** observaciones.

Estos supuestos se prueban a partir de un **análisis de residuales** y en caso de que se cumplan **todos y cada uno de ellos**, se concluye que el modelo de regresión es adecuado para predecir a y en función de x, además que los resultados obtenidos en las pruebas de significancia son confiables.
Una vez comprobado lo anterior, pueden hacerse predicciones para la variable de interés a partir de la información disponible de las variables regresoras.