Aprendizaje Automático

Regresión Lineal Simple, Lineal Múltiple y Logística

2023-2Q

Introducción

Introducción

Los métodos de regresión estudian la construcción de modelos para explicar o representar la **dependencia** entre una **variable respuesta o dependiente Y** y la(s) **variable(s) explicativa(s) o independiente(s), X.**

$$Y = f(X) + e$$

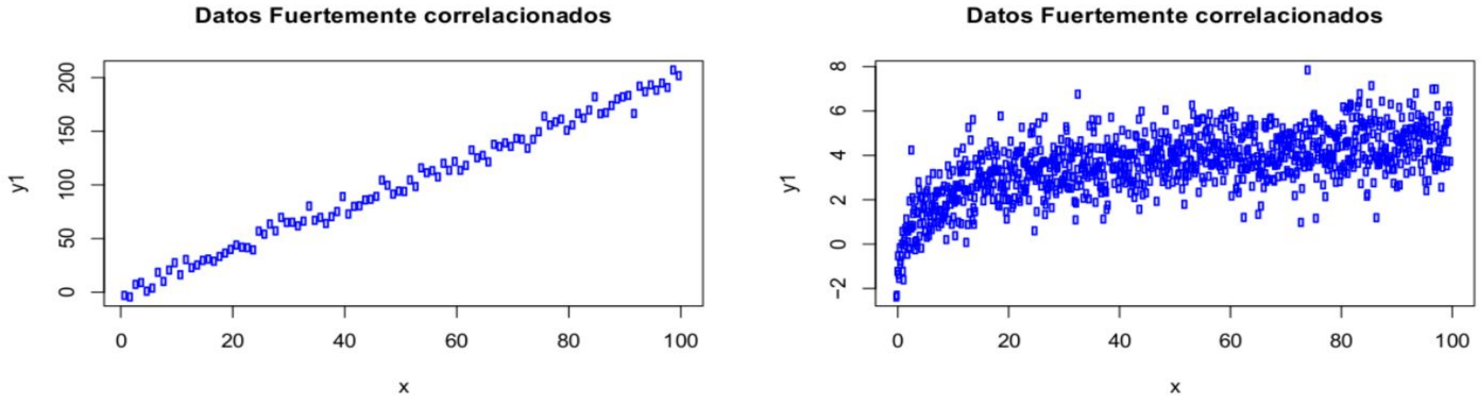
Introducción

Los métodos de regresión se utilizan para realizar inferencias cuando la variable respuesta no es categórica sino cuantitativa.

Clasificación vs. Regresión

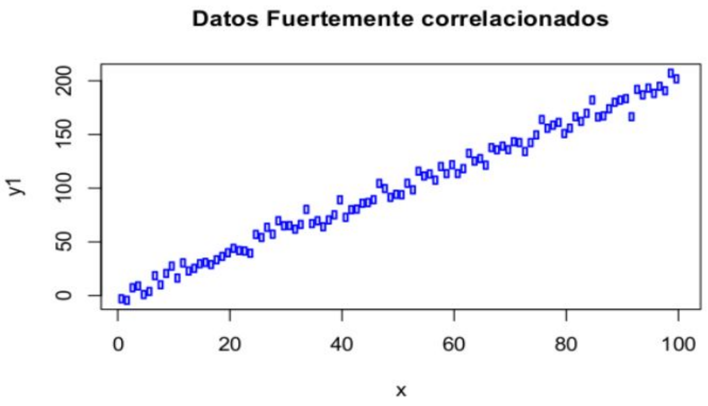
- Respuesta **cuantitativa** → Regresión
- Respuesta **cualitativa** → Clasificación

Introducción

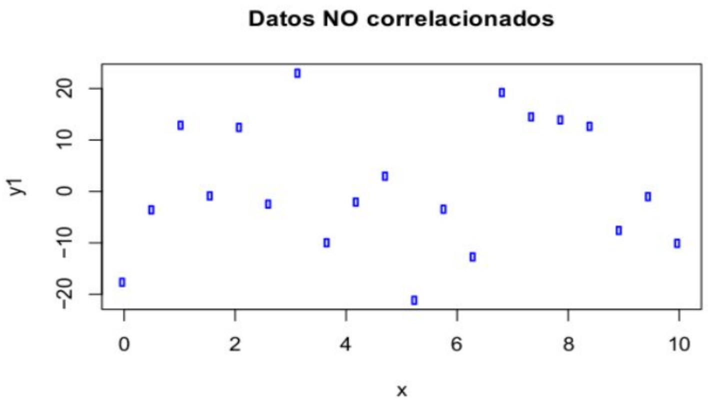


(a) Correlación Lineal (b) Correlación No Lineal

Introducción



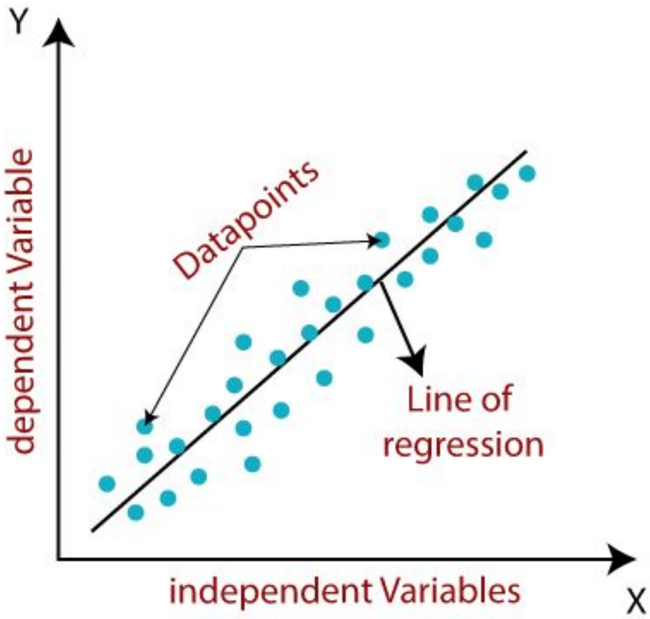
(c) Correlación Lineal



(d) Sin Correlación

Regresión Lineal

El modelo de regresión lineal, es aquel en el que la **dependencia** es de tipo **lineal**.



Regresión Lineal

Variable de interés o variable dependiente

- Se denota *Y*
- 1 variable aleatoria
- $\{y_1, \ldots, y_n\}$ son n observaciones de *Y*

Covariables o variables independientes:

- Se denota $X_1, \ldots, X_p$
- $\{x^1_1, \ldots, x^1_n\}$ son n observaciones de X_1

Nos preguntamos

- ¿Es significativo el efecto que una variable X causa sobre otra Y?
- ¿Es significativa la dependencia lineal entre esas dos variables?
- De ser así, utilizaremos el modelo de regresión lineal simple para explicar y predecir la variable dependiente Y a partir de valores observados en la independiente X.

Regresión Lineal

La idea es encontrar los coeficientes β_j tal que

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \beta_0 + e_i, \;\; i = 1, \ldots, n$$

y que el conjunto de errores $\{e_i, i = 1, \ldots n\}$ sea pequeño en algún sentido. Además consideramos $e_i \sim N(0, \sigma_2)$.

Regresión Lineal

Los métodos de regresión lineal difieren en minimizar estos errores o residuos.

Cuando solamente consideramos hallar los coeficientes β_0 y β_1, la llamamos **regresión lineal simple**, sino la llamamos **regresión lineal múltiple.**

Regresión Lineal

Regresión Lineal Simple

Dado un conjunto de pares de datos (x_i, y_i), $i = 1, \ldots, n$, se han desarrollado diversos métodos para ajustar una recta de la forma

$$Y = \beta_0 + \beta_1 X + E$$

al diagrama de dispersión de los datos.

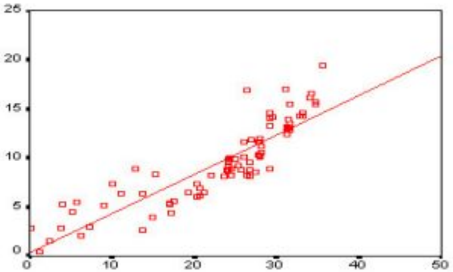


Figura: Ejemplo de Diagrama de dispersión y recta de ajuste.

Regresión Lineal Simple

Utilizamos el conjunto de entrenamiento (x_i, y_i), $i = 1, \ldots, n$, para estimar $\hat{\beta}_0$ y $\hat{\beta}_1$.

En un modelo de regresión lineal realizamos las siguientes suposiciones.

Regresión Lineal Simple

Suposiciones

La esperanza de Y |X

$E(Y\,|x_i) = \beta_0 + \beta_1 x_i$, $i = 1, \ldots, n$, o, equivalentemente, $E(e) = 0$, $i = 1, \ldots, n$.

Homocedasticidad (Para la varianza de Y |X)

La varianza es constante, $Var(Y|x_i) = \sigma_2$, $i = 1, \ldots, n$, o, equivalentemente, $Var(e) = \sigma_2$, $i = 1, \ldots, n$.
Entonces σ_2 es otro parámetro que deseamos estimar.

Regresión Lineal Simple

Los datos tienen distribución Gaussiana

$Y\,|x_i \sim N\,(\beta_0 + \beta_1 x_i, \sigma_2), i = 1, \ldots, n$

o, equivalentemente, $ei \sim N\,(0, \sigma_2), i = 1, \ldots, n.$

Las observaciones Y_i son independientes.

Estimación de los parámetros del modelo

Métodos de estimación

En el modelo de regresión lineal simple hay **tres parámetros** que se deben **estimar**:

- Los coeficientes de la recta de regresión, β_0 y β_1;
- La varianza de la distribución normal, σ^2.

El cálculo de estimadores para estos parámetros puede hacerse por diferentes métodos, los más utilizados son el **método de máxima verosimilitud** y el **método de mínimos cuadrados**.

Método de Cuadrados Mínimos

Recta de Ajuste por Cuadrados Mínimos

Dado un conjunto de pares (xi, yi), i = 1, . . . , n, buscamos la pendiente β1 y la ordenada al origen β0, tal que la recta y = β1x + β0 **minimice la suma de los residuos al cuadrado:**

Los residuos están dados por:

$$e_i = y_i - \beta_1 x_i - \beta_0, \;\; i = 1, \ldots, n$$

La pendiente y la ordenada al origen se **obtienen minimizando la suma de cuadrados de los residuos.**

Recta de Ajuste por Cuadrados Mínimos

La idea es

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

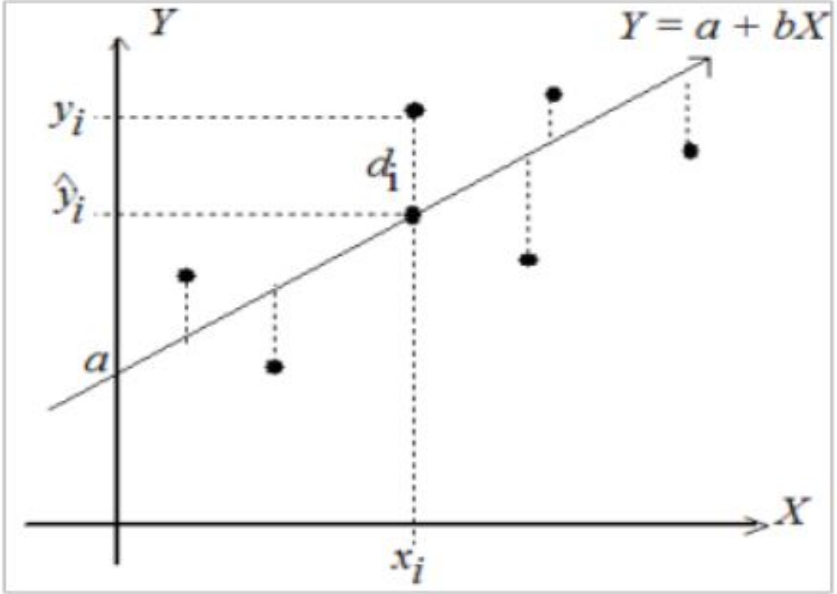
Derivando con respecto a β0 y a β1 e igualando a cero, obtenemos

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Recta de Ajuste por Cuadrados Mínimos

Lo que estamos haciendo es minimizar el error vertical



Errores estándar

El error estándar es el promedio de la diferencia entre el verdadero valor y estimado.

El error estándar de βˆ0

$$SE(\hat{\beta}_0)^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

El error estándar de βˆ1

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Residual Standard Error (RSE)

σ2 = Var(e)

No se conoce pero podemos estimarla.

Estimación de σ2:

$$RSE = \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Validación del Modelo

¿Cómo saber si el Modelo es Válido?

Test de Hipótesis

- Test t-student

Gráficos

- Gráfico de la dispersión
- Gráfico de la recta

¿Cómo saber si el Modelo es Válido?

Test Numérico

- Coeficiente de Determinación R2

Estimación de σ

$$RSE = \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Test t-Student

¿Es el efecto de *X* significativo sobre *Y*?

Student-test

Se Testea el efecto de β_1

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Test t-Student

Analizamos el estadístico

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

que, bajo la hipótesis nula, tiene distribución t- Student con n – 2 grados de libertad.

Analizamos el p-valor:

p – valor < 0,05 efecto significativo de β1 (rechazamos H0)

p – valor > 0,05 no hay efecto significativo de β1 (aceptamos H0)

¿Cómo saber si el Modelo es Válido?

Coeficiente de Determinación

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Además $0 \leq R2 \leq 1$.

Valores cercanos a 1 indica un mejor ajuste.

Métricas para evaluar el error

Mean Squared Error - Error Cuadrático Medio (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Mean Absolute Error - Error Absoluto Medio (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i)|$$

Donde *yi* es la observación verdadera de la variable explicada y *yˆi* es la estimación.

Diagnóstico del Modelo

- Los valores ajustados: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- Los residuos $e_i = \hat{y}_i - y_i$.
- Es útil estandarizar los residuos $es_i = \frac{e_i}{\hat{\sigma}}$.

Diagnóstico del Modelo

La hipótesis de normalidad

- QQ plot de los residuos.
- Un gráfico Cuantil-Cuantil permite observar qué tan cerca está la distribución de un conjunto de datos a alguna distribución ideal ó comparar la distribución de dos conjuntos de datos.
- Gráfico de las distribuciones empírica y teórica.
- Un test de Normalidad.

Variables predictoras categóricas

Ejemplo: una de las variables es el género: M, F.
Le asignamos:

$$x = \begin{cases} 0 & \text{si } x \text{ es } M \\ 1 & \text{si } x \text{ es } F \end{cases}$$

El modelo es

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x = \begin{cases} \beta_0 & \text{si } x \text{ es } M \\ \beta_0 + \beta_1 & \text{si } x \text{ es } F \end{cases}$$

Interpretación

- β_0 es el promedio de y para el género masculino.
- $\beta_0 + \beta_1$ es el promedio de y para el género femenino.
- β_1 es la diferencia entre femenino y masculino.

Resumen

Condiciones para la regresión lineal

- Linealidad: La relación entre ambas variables debe ser lineal.
- Distribución Normal de los residuos con media 0: Esto se puede comprobar con un histograma, con la distribución de cuantiles o con un test de hipótesis de normalidad.

Resumen

Independencia

- Independencia: Las observaciones deben ser independientes unas de otras. Puede detectarse estudiando si los residuos siguen un patrón o tendencia, utilizando un gráfico scatterplot.
- Dado que las condiciones se verifican a partir de los residuos, primero se genera el modelo y después se valida.

Resumen

Predicción

Una vez generado un modelo que se pueda considerar válido, es posible predecir el valor de la variable dependiente Y para nuevos valores de la variable predictora X.

Resumen

Limitaciones

- Limitarse al rango de valores dentro del que se encuentran las observaciones con las que se ha generado el modelo.
- Solo en esta región se tiene certeza de que se cumplen las condiciones para que el modelo sea válido.
- Para calcular las predicciones se emplea la ecuación generada por regresión.

Regresión Lineal Múltiple

Introducción

Regresión lineal múltiple

Representa una extensión de la regresión lineal simple en la que podemos incluir más de una variable independiente a la vez.

Introducción

Más de una variable predictora

Una opción sería ajustar un modelo de regresión a cada uno por separado.

pero…

cada ecuación de regresión estaría ignorando las demás a la hora de estimar los coeficientes de regresión.

Introducción

Variables Correlacionadas

Podría llevar a estimaciones erróneas haciendo el ajuste por separado. Por tanto, una ventaja de la regresión lineal múltiple es que **evalúa el efecto de cada variable en presencia del resto.**

Ejercicio Cervezas

Un distribuidor de cervezas está analizando el sistema de entregas de su producto. Está interesado en predecir el **tiempo** sugerido en repartir las botellas.

El ingeniero industrial a cargo del estudio ha sugerido que los factores que influyen sobre el tiempo de entrega son

- el **número de cajas** de cervezas.
- la **máxima distancia** que debe viajar el entregador de cajas.

Realizar un modelo que permita estimar el **tiempo** que el repartido necesita para repartir **29 cajas** a una **distancia máxima de 26 cuadras.**

El Modelo

Modelo de regresión Múltiple

La idea es encontrar los coeficientes βj tal que

$$Y_i = \sum_{j=1}^{p} \beta_j X_{ji} + \beta_0 + \epsilon_i,$$

- Con:
  - β0 ordenada al origen
  - βj, j = 1...p, el efecto de la covariable Xi
  - e el error
- Hipótesis:
  - ei (variable aleatoria), i ~ N(0, σ2)
  - ∀i 6= k, i, k son independientes.

Modelo de regresión Múltiple

En este caso

Por ejemplo, β_1 es la influencia que produce X_1 sobre Y, siendo que las otras variables están fijas.

En Forma Matricial

$$Y = X * \beta + \epsilon$$

donde

- $$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$
- $\beta = (\beta_0, \dots, \beta_p)^t$

Forma matricial

$$Y = X * \beta + \epsilon$$

Objetivo

Estimación de β, Var(β) y σ2

Notación

Los estimadores se denotan βˆ, Var(βˆ) y ˆσ2.

Los valores ajustados se denotan ˆy.

Los residuos se denotan ˆ = y − yˆ.

Solución Cuadrados Mínimos

Criterio de Cuadrados Mínimos

$$Y = X * \beta + \epsilon$$

Minimizar el RSS

$$minRSS = min \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Solución

El modelo

$$Y = X * \beta + \epsilon$$

Estimadores

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X^t X)^{-1}$$

$$\hat{\sigma}^2 = \frac{1}{n-p-1} RSS$$

Adj R2 (Coeficiente de determinación adjunto)

El R2 depende mucho de la cantidad variables que tenga el modelo. Un modelo con menor cantidad de variables tendrá siempre menor valor de R2 que uno con mayor cantidad.

Adj R^2

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - q}\right)$$

donde ***n* es el número de observaciones** y ***q* el número de coeficientes en el modelo**. Tiene en cuenta también la cantidad de coeficientes involucrados.

Nos preguntamos

1. ¿Es al menos, uno de los coeficientes significativamente diferente de 0?
2. ¿Es el efecto de X sobre Y significativo?
3. ¿Ajusta bien el modelo?

Test de Hipótesis para la significancia de los coeficientes

Test de Fisher

- H_0 : $\forall i, \beta_i = 0$
- H_1 : $\exists i, \beta_i != 0$

Analizamos el p – valor:

Si el p – valor < 0,05 rechazamos H_0 y entonces el modelo es válido.

Test de Fisher

El F-estadístico

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2, \; RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Test de Fisher

si

$$F > 1$$

hay evidencia para rechazar la hipótesis nula.

El estadístico F tiene distribución de Fisher, entonces analizamos el p – valor:

Si el p – valor < 0,05 hay evidencia para rechazar H0.

Selección de variables

1. Forward Selection: Se realizan todos los modelos lineales simples y se elige el que tiene menor RSS, y así sucesivamente agregando variables.
2. Backward Selection: Se realiza el modelo de regresión lineal múltiple con todas las variables, y se eliminan las que tienen el t estadístico cercano a 1, o el p-valor mayor que un umbral.
3. Mixed Selection: Se construye el modelo lineal agregando variables de a una, pero en el momento que el p-valor de una variable es mayor que un umbral, entonces esa variable se elimina del modelo.

Datos para evaluar un modelo

Estadístico	Criterio
R-Squared	Cuanto más grande mejor
Adj R-Squared	Cuanto más grande mejor
RSS	Lo más cercano a cero

Ejercicio: Análisis de Ventas en función de la publicidad

Utilizar el archivo *Advertising.csv* para analizar la influencia de la publicidad en las ventas de una empresa.

- Analizar la correlación entre las variables.
- Realizar los modelos de regresión lineal simple.
- Realizar el modelo de regresión lineal múltiple.
- Realizar el diagnóstico del modelo.
- En todos los casos analizar la influencia de cada variable en las ventas.
- En todos los casos anteriores dividir el conjunto total en un conjunto de entrenamiento y otro de prueba y calcular las matrices de confusión, accuracy, precision y recall.

¡Fin! ¿Alguna pregunta?