Multiple Linear Regression

Tech Lead Data Science

Master en Data Science
2022-2023

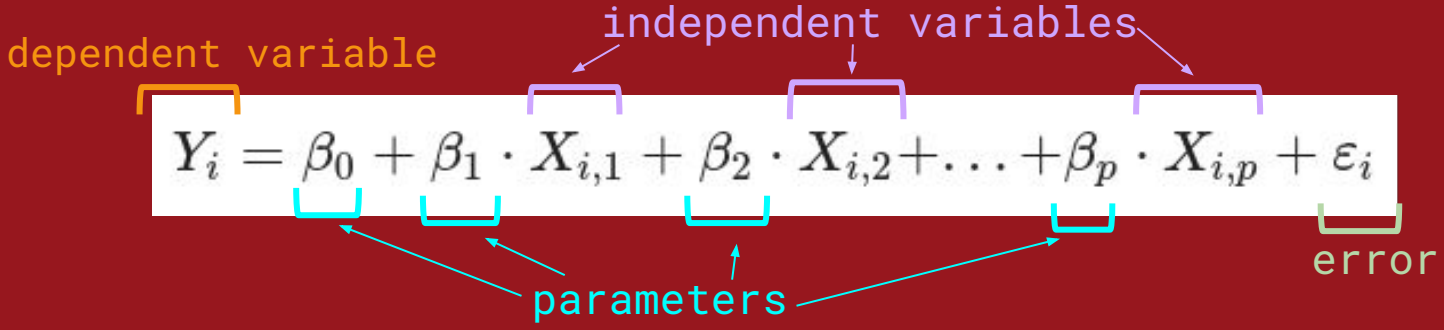
Assembler
Institute of Technology

Index

Introduction to MLR

Multiple linear regression

- It's used to predict a continuous numerical variable Y (dependent variable) from a set of covariables X (independent variables).
- The model assumes that the **variable Y** is a linear function of the **covariables X** and an **error**.
- The maths formula for one observation *i* is:



Multiple linear regression

X : variables independientes, predictoras o exógenas

Y : variable dependiente, a predecir o endógena

$\hat{Y}$: predicción, valor predicho

β_p : parámetro

$\hat{\beta}_p$: estimador

$\hat{\beta}_p$ = número : estimación, coeficiente estimado

Assumptions

- The means of the errors are zero, $E(\varepsilon_i) = 0$
- **Homoscedasticity:** The ε_i have the same variance (σ^2), being $Var(\varepsilon_i) = \sigma^2$
- ε_i are normally distributed
- ε_i are independent amongst them and they are not correlated to X_i

GAUSS-MARKOV ASSUMPTIONS

Assumptions

- The means of the errors are zero, $E(\varepsilon_i) = 0$
- **Homoscedasticity:** The ε_i have the same variance (σ^2), being $Var(\varepsilon_i) = \sigma^2$
- ε_i are normally distributed
- ε_i are independent amongst them and they are not correlated to X_i

GAUSS-MARKOV ASSUMPTIONS

$$\epsilon_i \sim N(0, \sigma^2) \text{ para todo } 1 < i < n, \text{ independientes entre sí}$$

Assumptions

- The means of the errors are zero, $E(\varepsilon_i) = 0$
- **Homoscedasticity:** The ε_i have the same variance (σ^2), being $Var(\varepsilon_i) = \sigma^2$
- ε_i are normally distributed
- ε_i are independent amongst them and they are not correlated to X_i

GAUSS-MARKOV ASSUMPTIONS

$$\epsilon_i \sim N(0, \sigma^2) \text{ para todo } 1 < i < n, \text{ independientes entre sí}$$

Reformulation

ASSUMPTION: $\epsilon_i \sim N(0, \sigma^2)$ para todo $1 < i < n$, independientes entre sí

ORIGINAL FORMULATION OF THE MODEL:

$$Y_i = \beta_0 + \beta_1 \cdot X_{i,1} + \beta_2 \cdot X_{i,2} + \ldots + \beta_p \cdot X_{i,p} + \varepsilon_i$$

Reformulation

ASSUMPTION: $\epsilon_i \sim N(0, \sigma^2)$ para todo $1 < i < n$, independientes entre sí

ORIGINAL FORMULATION OF THE MODEL:

$$Y_i = \beta_0 + \beta_1 \cdot X_{i,1} + \beta_2 \cdot X_{i,2} + \ldots + \beta_p \cdot X_{i,p} + \varepsilon_i$$

$$E(Y|X_1, X_2, \ldots, X_p) = E(\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p + \varepsilon|X_1, X_2, \ldots, X_p)$$

Reformulation

ASSUMPTION: $\epsilon_i \sim N(0, \sigma^2)$ para todo $1 < i < n$, independientes entre sí

ORIGINAL FORMULATION OF THE MODEL:

$$Y_i = \beta_0 + \beta_1 \cdot X_{i,1} + \beta_2 \cdot X_{i,2} + \ldots + \beta_p \cdot X_{i,p} + \varepsilon_i$$

$$E(Y|X_1, X_2, \ldots, X_p) = E(\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p + \varepsilon|X_1, X_2, \ldots, X_p)$$

$$E(Y|X_1, X_2, \ldots, X_p) = E(\beta_0|.) + E(\beta_1 \cdot X_1|.) + E(\beta_2 \cdot X_2|.) + \ldots + E(\beta_p \cdot X_p|.) + E(\varepsilon|.)$$

Reformulation

ASSUMPTION: $\epsilon_i \sim N(0, \sigma^2)$ para todo $1 < i < n$, independientes entre sí

ORIGINAL FORMULATION OF THE MODEL:

$$Y_i = \beta_0 + \beta_1 \cdot X_{i,1} + \beta_2 \cdot X_{i,2} + \ldots + \beta_p \cdot X_{i,p} + \varepsilon_i$$

$$E(Y|X_1, X_2, \ldots, X_p) = E(\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p + \varepsilon|X_1, X_2, \ldots, X_p)$$

$$E(Y|X_1, X_2, \ldots, X_p) = E(\beta_0|.) + E(\beta_1 \cdot X_1|.) + E(\beta_2 \cdot X_2|.) + \ldots + E(\beta_p \cdot X_p|.) + E(\varepsilon|.)$$

$E(\beta_j . X_j) = \beta_j . X_j$

$E(\varepsilon|X_1, X_2, \ldots, X_p)$

$$E(Y|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p + E(\varepsilon)$$

Reformulation

ASSUMPTION: $\epsilon_i \sim N(0, \sigma^2)$ para todo $1 < i < n$, independientes entre sí

ORIGINAL FORMULATION OF THE MODEL:

$$Y_i = \beta_0 + \beta_1 \cdot X_{i,1} + \beta_2 \cdot X_{i,2} + \ldots + \beta_p \cdot X_{i,p} + \varepsilon_i$$

$$E(Y|X_1, X_2, \ldots, X_p) = E(\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p + \varepsilon | X_1, X_2, \ldots, X_p)$$

$$E(Y|X_1, X_2, \ldots, X_p) = E(\beta_0|.) + E(\beta_1 \cdot X_1|.) + E(\beta_2 \cdot X_2|.) + \ldots + E(\beta_p \cdot X_p|.) + E(\varepsilon|.)$$

$E(\beta_j . X_j) = \beta_j . X_j$

$E(\varepsilon | X_1, X_2, \ldots, X_p)$

$$E(Y|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p + E(\varepsilon)$$

$E(\varepsilon) = 0$

$$E(Y|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p$$

Parameter estimation and interpretation

PARAMETER ESTIMATION

- The parameter estimation can be done in two different ways:
  - **Optimization method:** gradient descend respect of the Mean Squared Error (MSE)
  - **Estimation of Ordinary Least Squares (OLS):** Find a solution to the normal equations of the model.
- As the solution is unique, with both methods we should arrive to the same solution.
- The estimations by OLS give additional information about the parameters estimation.

PARAMETER ESTIMATION

scikit learn

General ML library

It has the tools to implement and evaluate a multiple linear regression

It’s not focused in general statistics

statsmodels

Statistics model library

It has the tools to implement a multiple linear regression

It’s not focused in ML

PARAMETER ESTIMATION

After the estimation process, we'll have the numerical values of the parameters: they are the estimated coeficients

$$E(Y|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p$$

estimation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot X_2 + \ldots + \hat{\beta}_p \cdot X_p$$

PARAMETER ESTIMATION

$\hat{\beta}_0$ **Ordenada al origen / y interception**

Es el valor esperado o promedio de y cuando todas las variables predictoras son iguales a cero. No siempre tiene una interpretación válida en el contexto del problema.

$\hat{\beta}_j$ **Pendiente / gradient**

El valor esperado o promedio de y cambiará en $\hat{\beta}_j$ unidades cuando X_j aumenta en una unidad, dadas el resto de las variables (X) constantes

PARAMETER ESTIMATION

Supongamos que definimos un modelo para estimar el precio de propiedades en base a la superficie total (medida en metros cuadrados) y la cantidad de baños. Su especificación es:

$$precio = \beta_0 + \beta_1 \cdot superficieTotal + \beta_2 \cdot baños$$

Luego del proceso de estimación obtenemos:

$$\hat{precio} = -107213 + 2069 \cdot superficieTotal + 113359 \cdot baños$$

Assembler
Institute of Technology

PARAMETER ESTIMATION

$$precio = \beta_0 + \beta_1 \cdot superficieTotal + \beta_2 \cdot baños$$

$$\hat{precio} = -107213 + 2069 \cdot superficieTotal + 113359 \cdot baños$$

PARAMETER ESTIMATION

$$precio = \beta_0 + \beta_1 \cdot superficieTotal + \beta_2 \cdot baños$$

$$\hat{precio} = -107213 + 2069 \cdot superficieTotal + 113359 \cdot baños$$

$\hat{\beta}_0 = -107213$

El valor esperado/promedio/predicho de una propiedad sin superficie ni baños es de -107213 dólares

$\hat{\beta}_1 = 2069$

El valor esperado/promedio/predicho de una propiedad aumenta en 2069 dólares frente a un aumento de 1 metro cuadrado de la superficie total dada la cantidad de baños

$\hat{\beta}_2 = 113359$

El valor esperado/promedio/predicho de una propiedad aumenta en 113359 dólares frente a un aumento de 1 baño dada la superficie total

Assembler
Institute of Technology

Categorical variables

Variables categóricas

Las variables categóricas son aquellas que tienen una cantidad finita de valores posibles (categorías). Pueden ser:

Binarias

Tienen dos categorías

Por ejemplo: si una propiedad es un departamento o no

Multiclase

Tienen tres o más categorías

Por ejemplo: si una propiedad es un departamento, condominio o casa

Variables binarias

Una variable binaria indica si una observación presenta un determinado atributo o no:

$$X_{binaria} = \begin{cases} 0 & \text{si la observación no presenta el atributo} \\ 1 & \text{si la observación presenta el atributo} \end{cases}$$

Observemos cómo incluimos una variable binaria en un modelo de regresión múltiple y cómo se interpreta su coeficiente

Regresión con variables binarias

Definimos un modelo con una variable continua y una binaria:

$$E(Y|X) = \beta_0 + \beta_1 \cdot X_1 + \beta_{binaria} \cdot X_{binaria}$$

Las dos situaciones posibles respecto a la variable son:

Si $X_{binaria} = 0$:

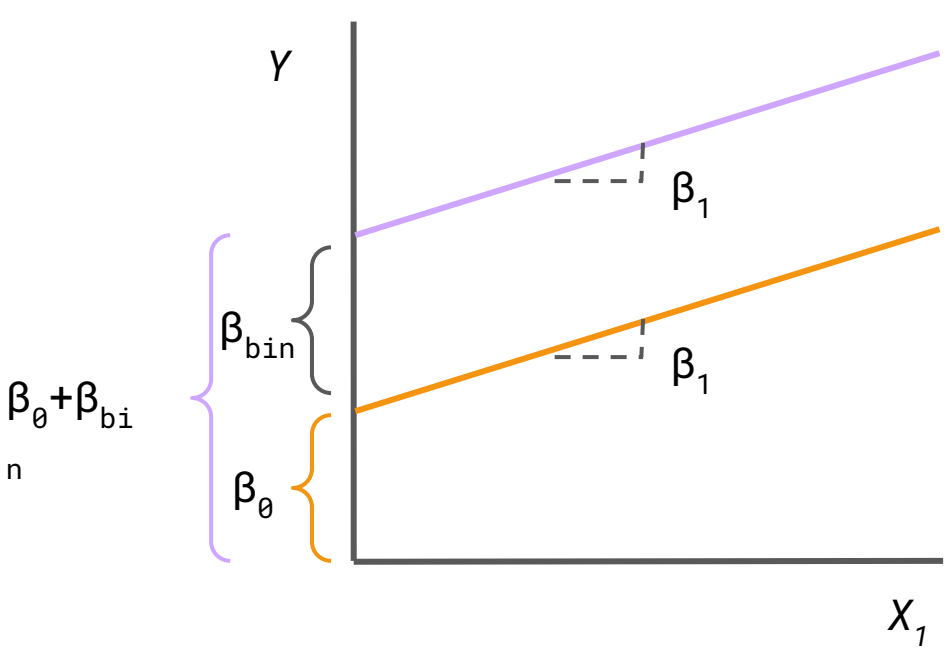
$$E(Y|X) = \beta_0 + \beta_1 \cdot X_1$$

Si $X_{binaria} = 1$:

$$E(Y|X) = (\beta_0 + \beta_{binaria}) + \beta_1 \cdot X_1$$

Assembler
Institute of Technology

Regresión con variables binarias



$$E(Y|X) = (\beta_0 + \beta_{binaria}) + \beta_1 \cdot X_1$$

$$E(Y|X) = \beta_0 + \beta_1 \cdot X_1$$

Si la observación presenta la variable binaria se produce un **desplazamiento en la ordenada al origen**

Interpretación de los coeficientes: ejemplo

Supongamos que definimos un modelo para estimar el precio de propiedades en base a la superficie total (medida en metros cuadrados) y si tiene pileta o no

$$precio = \beta_0 + \beta_1 \cdot superficieTotal + \beta_2 \cdot tienePile$$

La variable binaria es:

$$X_2 = \begin{cases} 0 & \text{si la propiedad no tiene pileta} \\ 1 & \text{si la propiedad tiene pileta} \end{cases}$$

Luego del proceso de estimación obtenemos:

$$\hat{precio} = -80000 + 2500 \cdot superficieTotal + 100000 \cdot tienePile$$

Assembler
Institute of Technology

Interpretación de los coeficientes: ejemplo

$$\hat{\beta}_0 = -80000$$

El valor esperado de una propiedad sin superficie y **sin pileta** es **-80000 dólares**

$$\hat{\beta}_1 = 2500$$

El valor esperado de una propiedad **aumenta en 2500 dólares** frente a un aumento de 1 metro cuadrado de la superficie total tanto si tiene pileta como si no

$$\hat{\beta}_2 = 100000$$

El valor esperado de una propiedad **aumenta en 100000 dólares** si tiene pileta dada la superficie total

Assembler
Institute of Technology

Interacción variable binaria con numérica

Es posible definir una interacción entre una variable binaria con una numérica creando un nuevo término. El modelo es:

$$E(Y|X) = \beta_0 + \beta_1 \cdot X_1 + \beta_{binaria} \cdot X_{binaria} + \beta_{interaccion} \cdot \underbrace{(X_{binaria} \cdot X_1)}_{X_{interaccion}}$$

Si $X_{binaria} = 0$:

$$E(Y|X) = \beta_0 + \beta_1 \cdot X_1$$

Si $X_{binaria} = 1$:

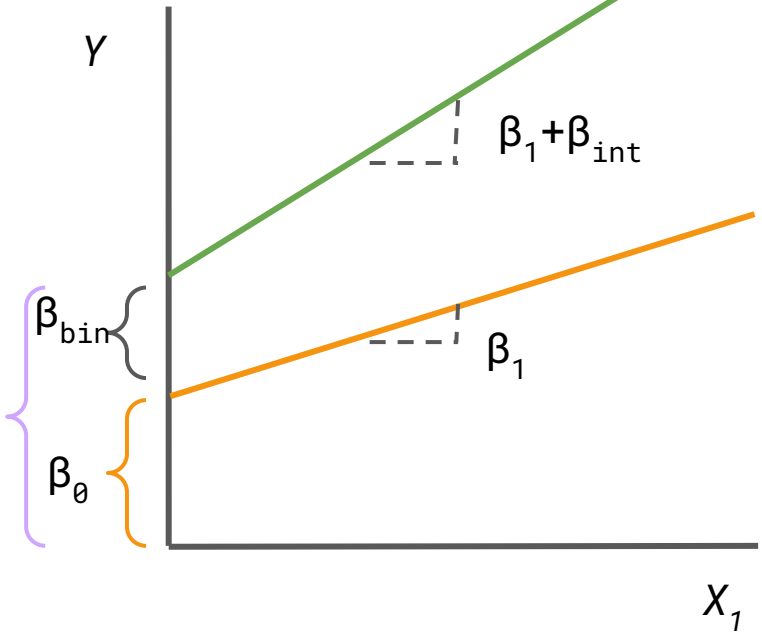
$$E(Y|X) = (\beta_0 + \beta_{binaria}) + (\beta_1 + \beta_{interaccion}) \cdot X_1$$

Assembler
Institute of Technology

Interacción variable binaria con numérica

$$E(Y|X) = (\beta_0 + \beta_{binaria}) + (\beta_1 + \beta_{interaccion}) \cdot X_1$$

$$E(Y|X) = \beta_0 + \beta_1 \cdot X_1$$



Si la observación presenta la variable binaria se produce un **desplazamiento en la ordenada al origen** y un **cambio en la pendiente**

Assembler
Institute of Technology

Trampa dummy

Al incluir una variable binaria (dos clases) en el modelo lo hicimos con una única variable indicadora. ¿Qué sucede si definimos tantas variables como clases?

$$X_{binaria1} = \begin{cases} 0 & \text{si la observación NO presenta el atributo} \\ 1 & \text{si la observación SI presenta el atributo} \end{cases}$$

$$X_{binaria2} = \begin{cases} 0 & \text{si la observación SI presenta el atributo} \\ 1 & \text{si la observación NO presenta el atributo} \end{cases}$$

El modelo quedaría definido:

$$E(Y|X) = \beta_0 + \beta_1 \cdot X_1 + \beta_{binaria1} \cdot X_{binaria1} + \beta_{binaria2} \cdot X_{binaria2}$$

Trampa dummy

Esta especificación del modelo es incorrecta ya que ocasiona haya **multicolinealidad perfecta** y no sea posible estimar los parámetros.

Solución: para las variables categóricas con K clases se deben crear k-1 variables indicadoras/binarias

Ejemplos

- Para una variable binaria (k=2) se crea una variable dummy/binaria
- Para una variable con 4 clases (k=4) se crean 3 variables dummies/binarias

Variables con 3 o más categorías

Una variable categórica con k categorías deberá ser codificada como (k-1) variables binarias.

La clase que no es codificada en una variable queda contenida en el valor del parametro β_0.

Ejemplo: la variable tipo de propiedad tiene las categorías: departamento, casa y propiedad horizontal (PH). Entonces generamos 2 variables binarias:

$$X_{depto} = \begin{cases} 0 & \text{si la observación NO es un departamento} \\ 1 & \text{si la observación es un departamento} \end{cases}$$

$$X_{casa} = \begin{cases} 0 & \text{si la propiedad NO es una casa} \\ 1 & \text{si la observación es una casa} \end{cases}$$

Assembler
Institute of Technology

Variables con 3 o más categorías

El modelo queda especificado:

$$E(Y|X) = \beta_0 + \beta_1 \cdot superficieTotal + \beta_2 \cdot X_{casa} + \beta_3 \cdot X_{depto}$$

Si la propiedad es una casa, $X_{casa} = 1$ y $X_{depto} = 0$

$$E(Y|X) = (\beta_0 + \beta_2) + \beta_1 \cdot superficieTotal$$

Si la propiedad es un departamento, $X_{depto} = 1$

$$E(Y|X) = (\beta_0 + \beta_3) + \beta_1 \cdot superficieTotal$$

Si la propiedad es un PH, $X_{casa} = 0$ y $X_{depto} = 0$

$$E(Y|X) = \beta_0 + \beta_1 \cdot superficieTotal$$

Assembler
Institute of Technology

Evaluación del modelo

La evaluación del modelo consiste en la realización de tests estadísticos y cálculo de métricas sobre la performance del modelo.

Los **tests estadísticos** de significatividad individual y global nos permiten evaluar si uno o todos los coeficientes estimados son estadísticamente distintos de cero

Métricas como el coeficiente de determinación (R cuadrado) y Error Cuadrático Medio (ECM) nos permiten evaluar la **capacidad explicativa o predictiva** del modelo

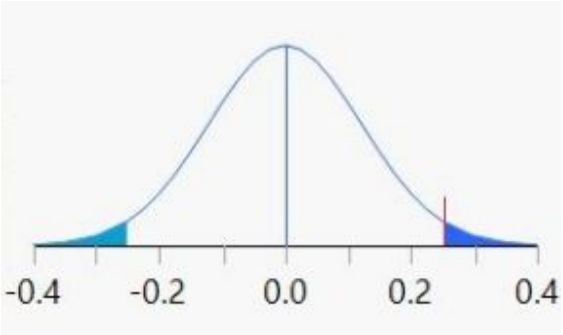
Test significatividad individual

La pregunta que se desea responder es: ¿La variable x_j tiene una relación lineal con la variable Y?

Esto equivale a preguntarse si el parámetro β_j es igual a cero o no. Las hipótesis quedan planteadas:

$H_0: \beta_j = 0$

$H_A: \beta_j \neq 0$



Para rechazar la hipótesis nula se suele requerir un **p-valor** del test menor o igual **0.05**

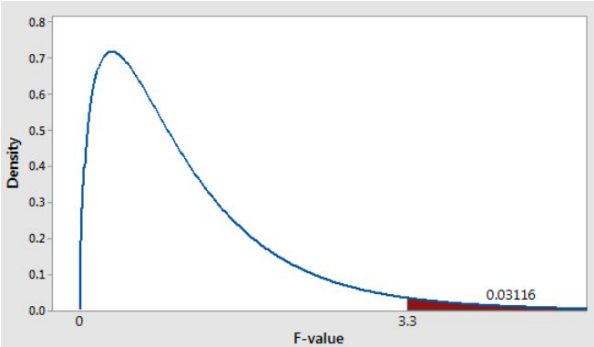
Assembler
Institute of Technology

Test significatividad global

La pregunta que se desea responder es: ¿Alguna de las variables x_j tiene una relación lineal con la variable Y?

Esto equivale a preguntarse si todos los parámetros β_j son iguales a cero o no. Las hipótesis quedan planteadas:

H_0: Todos los $\beta_j = 0$ $\quad$ H_A: Algún $\beta_j \neq 0$



Para rechazar la hipótesis nula se suele requerir un **p-valor** del test menor o igual **0.05**

Assembler
Institute of Technology

Coeficiente R-cuadrado

Es una métrica que permite medir qué porcentaje de la variabilidad de la variable Y es explicada por el modelo. Para calcularlo son necesarios:

$$SumaCuadradosTotal = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Discrepancia entre la observación y la predicción base (promedio)

$$SumaCuadradosResiduos = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Discrepancia entre la observación y predicción

$$SumaCuadradosModelo = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Discrepancia entre la predicción y predicción base

Coeficiente R-cuadrado

SCT = SCM + SCR

$$R^2 = \frac{SumaCuadradosModelo}{SumaCuadradosTotal} = 1 - \frac{SumaCuadradosResiduos}{SumaCuadradosTotal}$$

Ratio variabilidad explicada *Ratio variabilidad no explicada*

Propiedades

- $0 \leq R^2 \leq 1$
- No depende de las unidades de medición de la variable
- A medida que la capacidad predictiva del modelo aumenta, el coeficiente crece (hay excepciones!)

Assembler
Institute of Technology

Coeficiente R-cuadrado ajustado

Una **propiedad no deseable** del R^2 es que aumenta frente a la inclusión de variables en el modelo aunque las mismas tengan muy poco poder predictivo. La alternativa es utilizar el R^2 ajustado.

"penalización"

$$R^2_A = 1 - \frac{SumaCuadradosResiduos}{SumaCuadradosTotal} \cdot \frac{NumObservaciones - 1}{NumObservaciones - NumVariables}$$

ratio variabilidad no explicada

Si el número de variables aumenta pero el ratio variabilidad no explicada disminuye muy poco, el R^2 ajustado va a caer

Assembler
Institute of Technology

Error de predicción

Error Cuadrático Medio (Mean Squared Error)

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

Raíz Error Cuadrático Medio (Root Mean Squared Error)

$$RMSE = \sqrt{MSE}$$

Error Absoluto Medio (Mean Absolute Error)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$