#AprendeDesdeCasa
#AprendeConLosPioneros

Online | UPC Perú

PROGRAMA DE ESPECIALIZACIÓN ANALÍTICA

ADVANCED DATA SCIENCE

Métodos Supervisados

SESIÓN 1

Docente: Brian Alarcon

PROGRAMA DE ESPECIALIZACIÓN ANALÍTICA

ADVANCED DATA SCIENCE

Online | DIHC Perú

PERFIL PROFESIONAL DOCENTE

- **NOMBRE:** JORGE BRIAN ALARCON FLORES
- **ESTUDIOS:**

LICENCIADO EN ESTADISTICA

MAESTRIA EN INFORMATICA
CIENCIAS DE LA COMPUTACION

- **EXPERIENCIA LABORAL:**

DATA SCIENCE MANAGER

Brian Alarcon Flores
brian.alarcon@pucp.edu.pe

PROGRAMA DE ESPECIALIZACIÓN ANALÍTICA

ADVANCED DATA SCIENCE

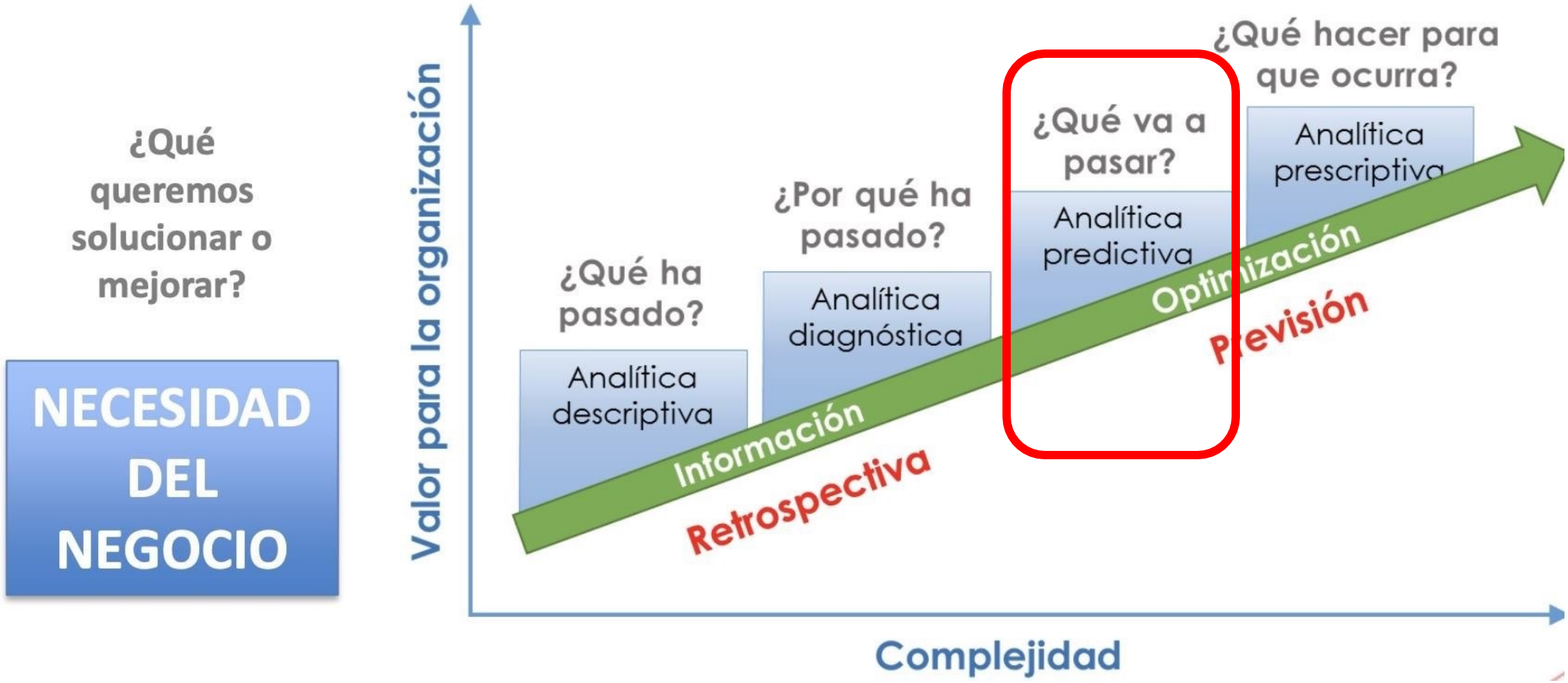
Online | DiHC Perú

ROADMAP DE TRABAJO

Sesion 1 7:30 -10:30 pm	Sesion 2 7:30 -10:30 pm	Sesion 3 7:30 -10:30 pm	Sesion 4 7:30 -10:30 pm	Sesion 5 7:30 -10:30 pm	Sesion 6 7:30 -10:30 pm
- Introducción a los Métodos Supervisados - Modelos Estadísticos de Regresión Lineal Simple y Múltiple	• Modelos de Regresión con Regularización • Modelos Computacionales de Regresión (No Lineales) • Evaluación de Modelos de Regresión	• Modelos Estadísticos de Clasificación - Modelos Computacionales de Clasificación I - Balanceo y Partición de datos	- Modelos Computacionales de Clasificación II - Optimización de Hiperparámetros - Evaluación de Modelos de Clasificación I	• Modelos Computacionales de Clasificación III (Stacking de Modelos) • Evaluación de Modelos de Clasificación II	- Introducción a las Redes Neuronales - Modelos de Redes Neuronales para Métodos Supervisados

- Evaluación de participación será contínua.
- Trabajo final grupal

¿Qué es lo que veremos hoy?



Machine Learning

01 Supervisado

- ¿Qué es?
  - Técnica para deducir una función a partir de datos de entrenamiento, previamente etiquetados y clasificados.
- Casos
  - Clasificación
  - Regresión (Lineal, Logística)

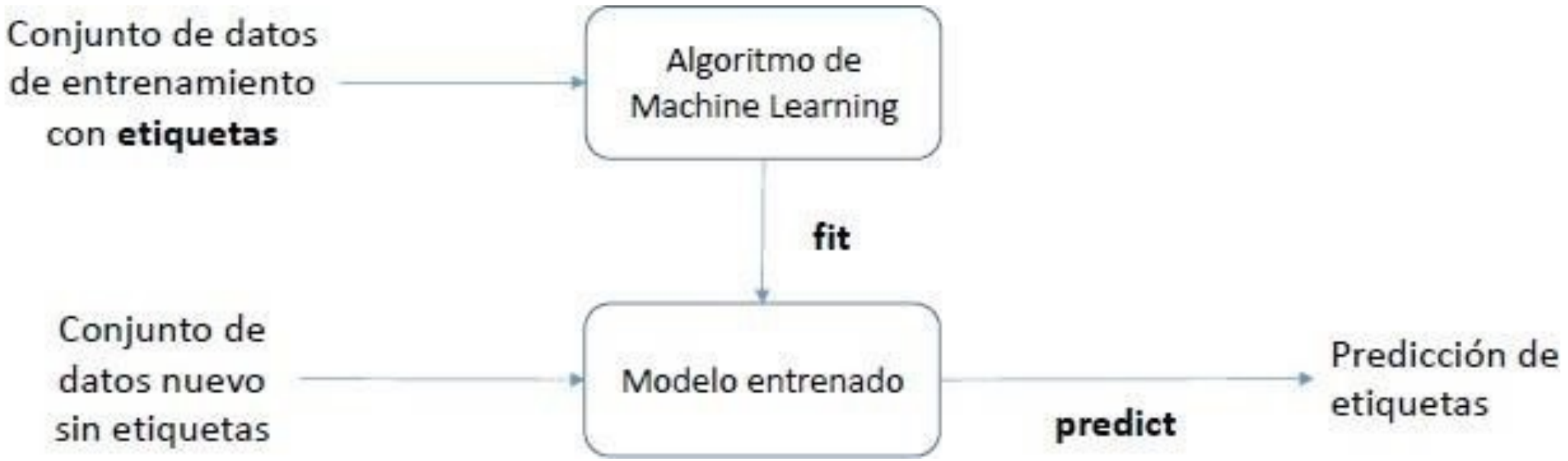
02 No Supervisado

- ¿Qué es?
  - Busca encontrar patrones en conjuntos de datos que no presentan etiquetado previo.
- Casos
  - Clustering
  - Reglas de asociación
  - Segmentación
  - Reducción de dimensionalidad

03 Refuerzo

- ¿Qué es?
  - Modelo acción-recompensa, que busca que el algoritmo se ajuste a la mejor “recompensa” dada por el ambiente.
- Casos
  - Procesos de decisión
  - Q-Learning

Entendiendo los Métodos Supervisados



Introduciéndonos a la Regresión

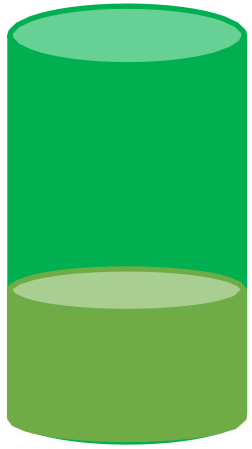
- Dada una colección de registros (Conjunto de Entrenamiento), cada registro contiene un conjunto de variables (***atributos, drivers***) denominado X, con una variable adicional que es un número denominada y (***Target, variable objetivo***).

- El objetivo de la ***regresión*** es encontrar un modelo (una función) para estimar o pronosticar el **valor numérico** que toma cada registro, ésta estimación debe hacerse con la mayor precisión posible.

- Un conjunto de prueba **(tabla de testing)** se utiliza para determinar **la precisión del modelo**. Por lo general, el conjunto de datos dado se divide en dos conjuntos al azar de el de **entrenamiento y el de prueba.**

Introduciéndonos a la Regresión

División de los datos: Datos Históricos



Entrenamiento

36,538 registros
(66.7%)

Validación

18,270 registros
(33.3%)

54,808 datos

Objetivo:
Ingresos

- **Ingreso de cada persona**

12 Variables Independientes

- 5 vars. categóricas
- 5 vars. numéricas
- 2 vars. binarias

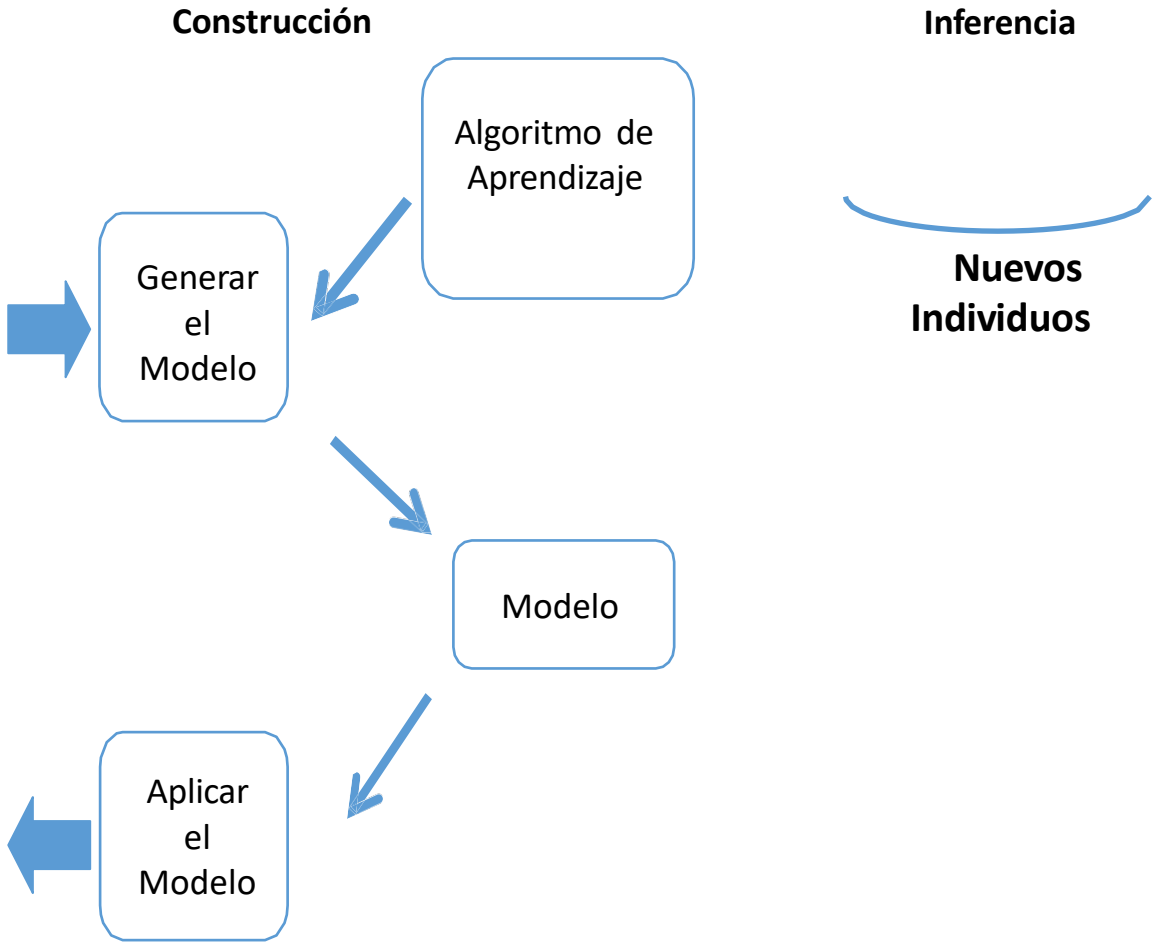
Introduciéndonos a la Regresión

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES
1	SI	SOLTERO	S/ 1,000
2	SI	CASADO	S/ 5,000
3	NO	CASADO	S/ 3,500
4	SI	VIUDO	S/ 4,500
5	NO	SOLTERO	S/ 2,000
6	NO	SOLTERO	S/ 1,500

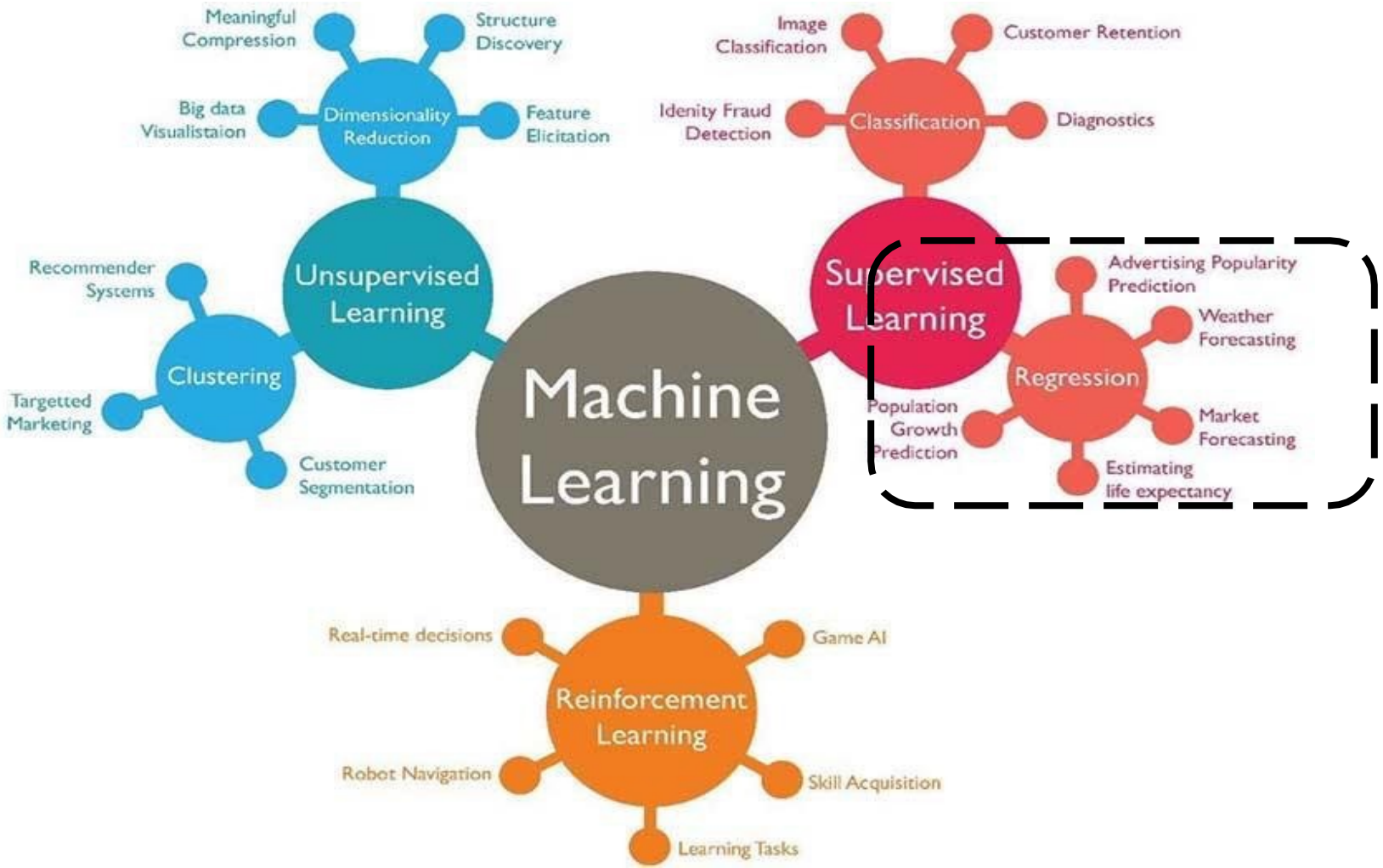
Tabla de Aprendizaje

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES
7	SI	SOLTERO	S/ 4,000
8	SI	CASADO	S/ 5,500
9	NO	CASADO	S/ 6,500

Tabla de Testing



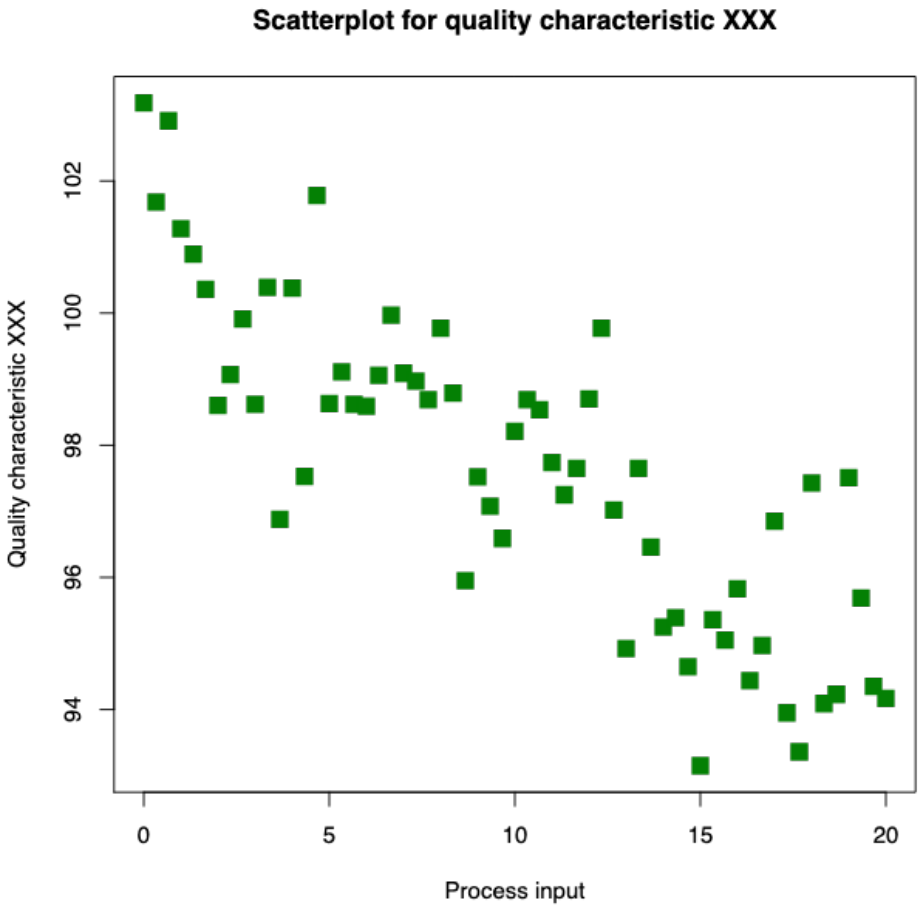
Introduciéndonos a la Regresión



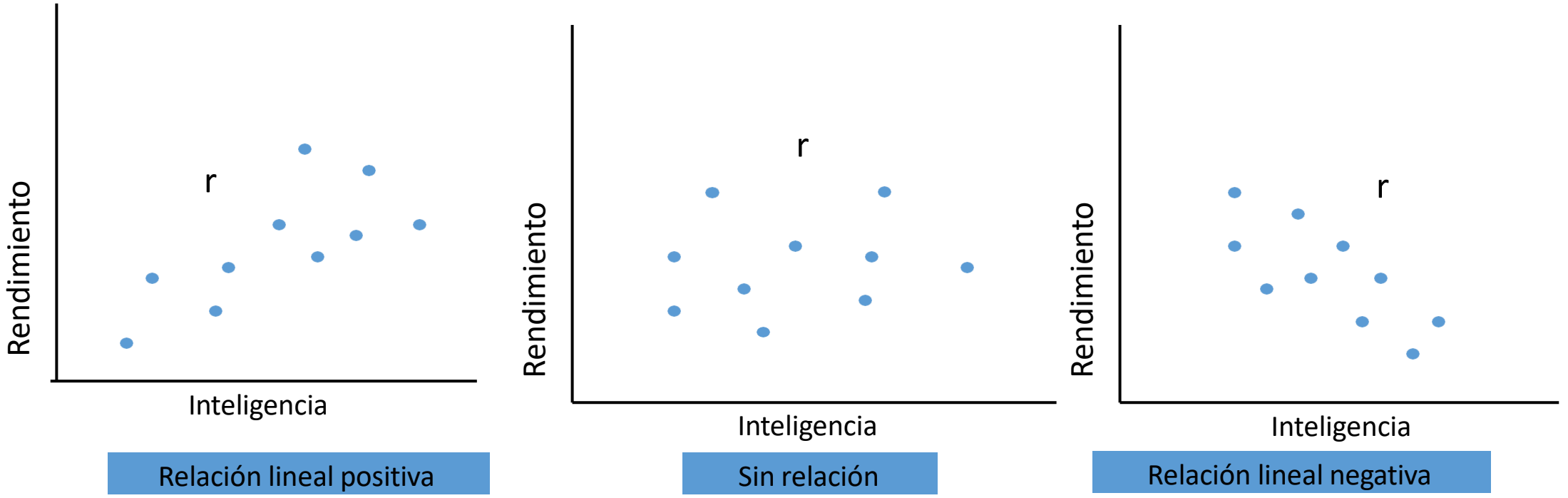
Introduciéndonos a la Regresión

- El diagrama de dispersión es una representación gráfica de la relación existente entre dos variables, nos permite visualizar:

  - Existe relación entre variables.
  - La relación es lineal o de otro tipo.
  - Intensidad de la relación (por la estrechez de la nube de puntos).
  - Valores anómalos (Outliers) distorsionan la relación.
  - La dispersión de los datos es o no uniforme (homocedasticidad vs. heterocedasticidad).

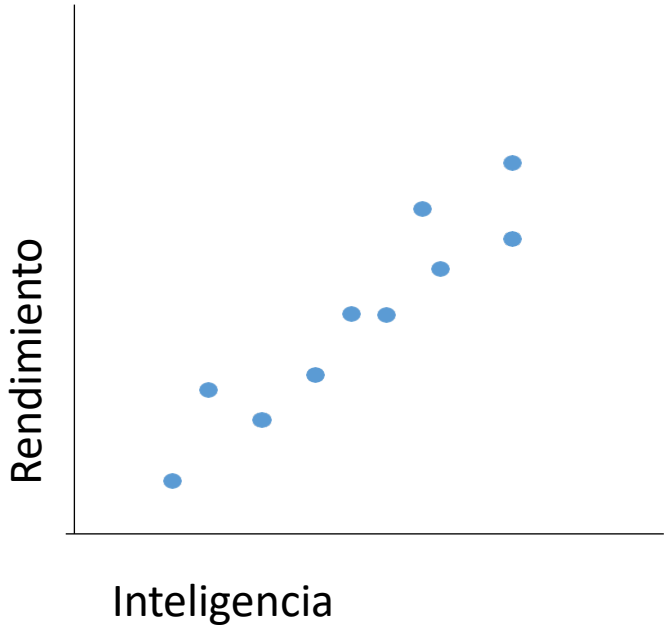


Introduciéndonos a la Regresión

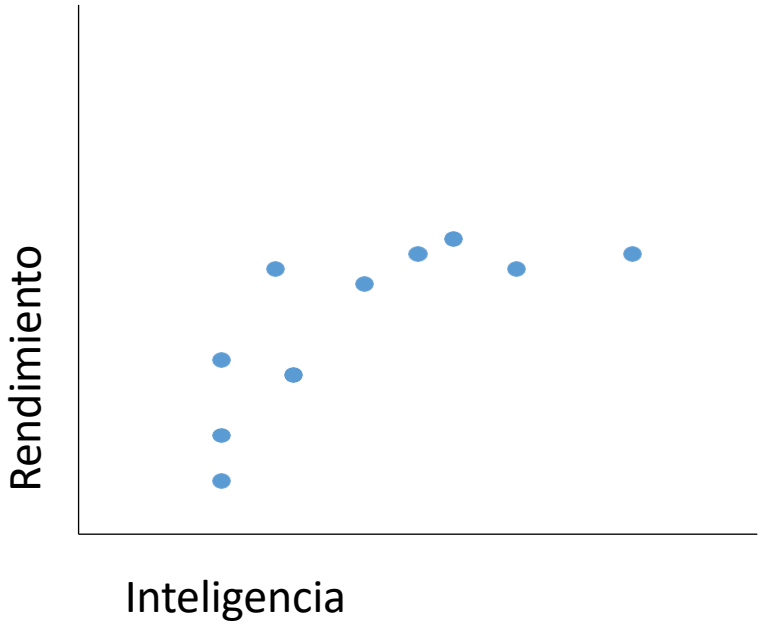


Nota: El coeficiente de **Correlación de Pearson** mide relación LINEAL.

Introduciéndonos a la Regresión



Relación lineal



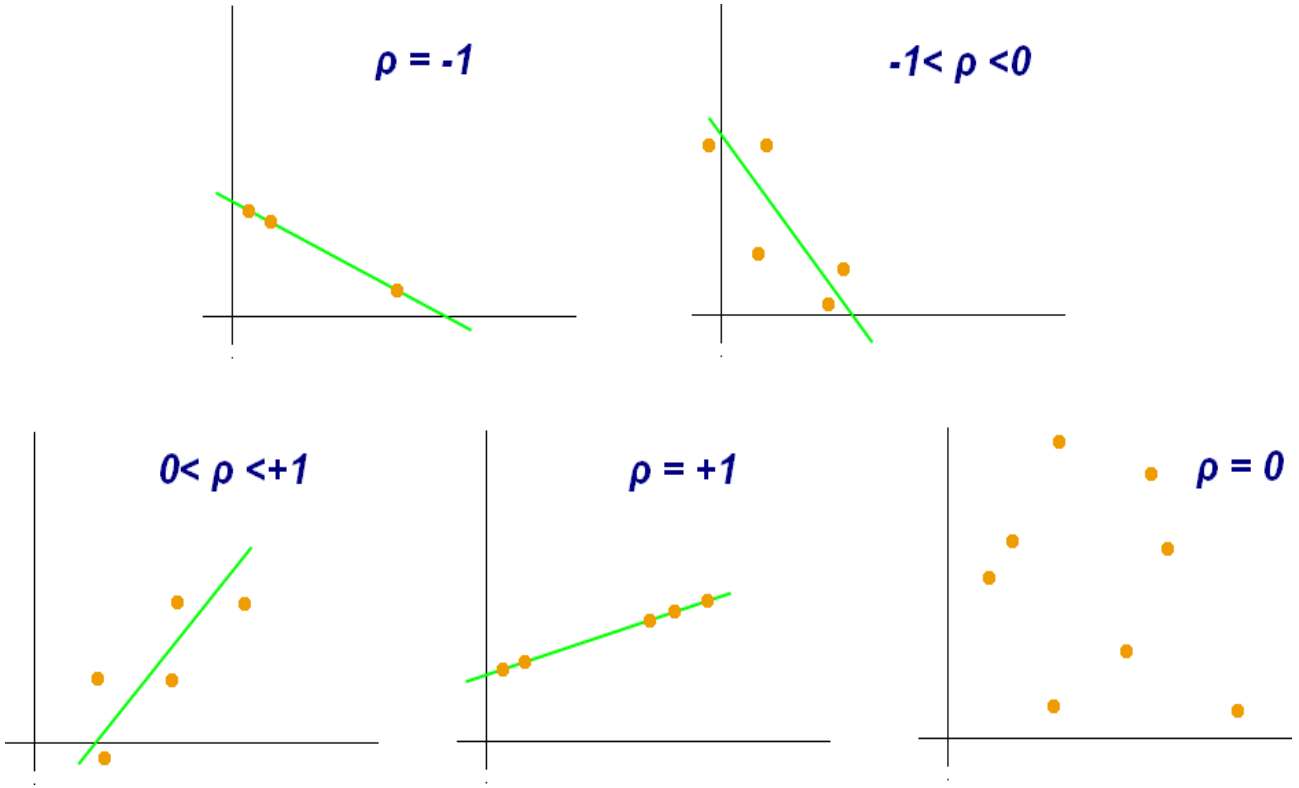
Relación no lineal

Nota: El coeficiente de **correlación de Pearson** mide relación **LINEAL**.

Introduciéndonos a la Regresión

r = 1	correlación perfecta
0,8 < r < 1	correlación muy alta
0,6 < r < 0,8	correlación alta
04 < r < 0,6	correlación moderada
0,2 < r < 0,4	correlación baja
0 < r < 0,2	correlación muy baja
r = 0	correlación nula

COEFICIENTE DE CORRELACIÓN DE PEARSON



Regresión Lineal Simple

- Determinar la ecuación de regresión sirve para:

  Describir de manera concisa la relación entre variables.

  Predecir los valores de una variable en función de la otra.

- Veremos EXCLUSIVAMENTE relaciones lineales.
- La regresión lineal simple estudia la relación entre sólo dos variables (el caso de relación más sencillo posible).

Regresión Lineal Simple

DENOMINACIÓN DE LAS VARIABLES	
X	Y
Predictora, regresor	Criterio
Explicativa	Explicada
Predeterminada	Respuesta
Independiente	Dependiente
Exógena	Endógena
(Explica la variabilidad de otra variable)	(Su variabilidad es explicada por otra variable)

Regresión Lineal Simple

1° **Información histórica** con la que contamos.

2° Hallamos los **valores de los coeficientes** y construimos el **modelo de regresión.**

$$Y = \alpha + \beta$$

$$\hat{Y} = \alpha + \beta X$$

3° Con el modelo de regresión **predecimos valores** y los **comparamos contra los que ya teníamos.**

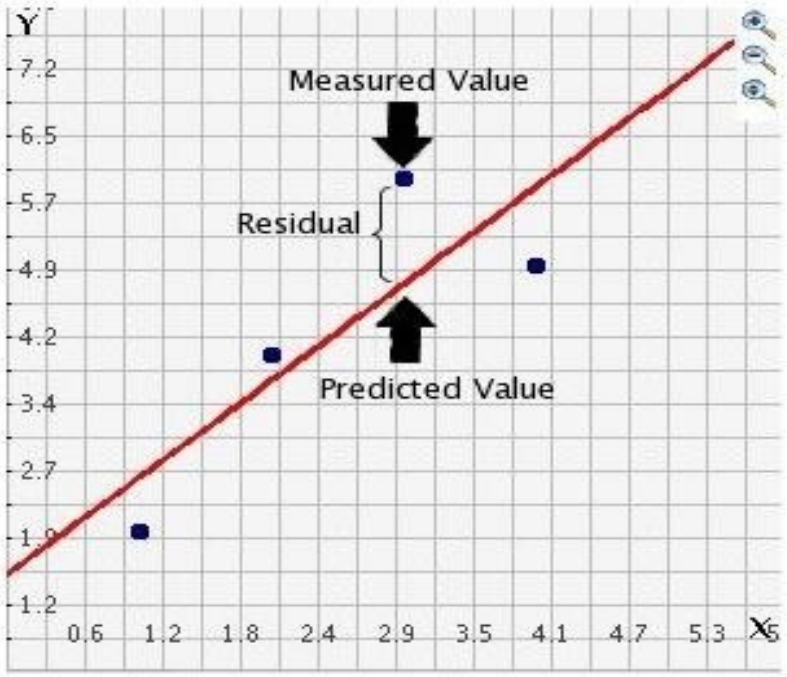
$$r = Y - \hat{Y}$$

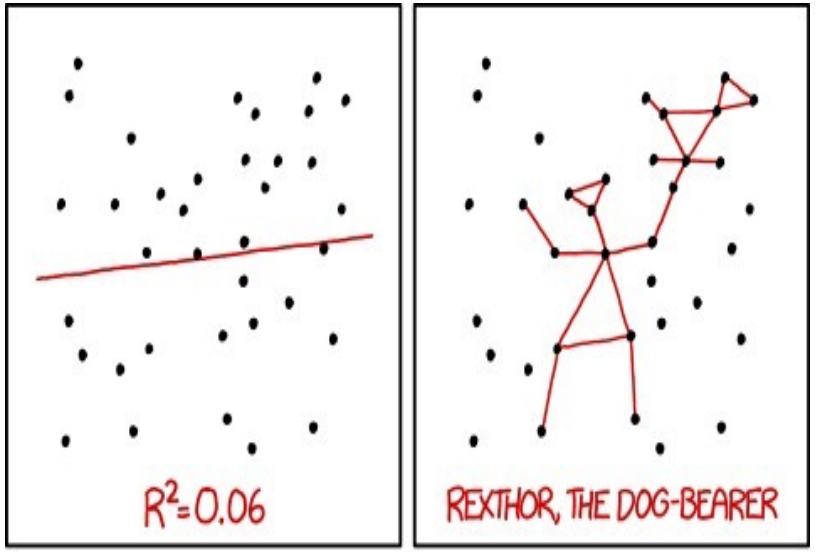
4° El resultado nos sirve para **validar el modelo construido.** → r

- Puede denominarse:
  - Error aleatorio.
  - Perturbación aleatoria.
- Se debe fundamentalmente a:
  - Medición incorrecta de la variable.
  - Influencia de otras variables no incluidas en el modelo.
  - Variabilidad inherente a la conducta humana.

Regresión Lineal Simple

QUE SON LOS ERRORES O RESIDUALES DE UN MODELO DE REGRESION ?





I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Regresión Lineal Simple

Interpretación del modelo de regresión lineal

En el modelo teórico de regresión lineal

$$Y = a + bX + e$$

distinguimos los siguientes elementos:

- e → error de estimación o puntuaciones residuales, parte aleatoria; aquello no explicado por el modelo.

$$\hat{Y} = a + bX$$

- $\hat{Y}$ → puntuación estimada: valor promedio previsto para todos los sujetos que han obtenido en la variable X un valor de Xi.
- b → pendiente de la recta: cambio en Y por cada unidad de cambio en X.
- a → ordenada en el origen: valor medio de Y cuando X=0.

Regresión Lineal Simple

$$\hat{Y} = 600 + 300\,X$$

Supongamos que tenemos la ecuación de regresión, donde X es el número de años de experiencia profesional, e Y es el sueldo mensual.

- b=300 → Cambio en Y por cada unidad de cambio en X. Por cada año de experiencia laboral, el sueldo mensual aumenta 300 €.
- a=600 → Valor medio de Y cuando X=0. Sueldo medio de aquellas personas sin experiencia laboral.

Una persona con 3 años de experiencia laboral, ¿qué sueldo mensual tendrá? Interpreta el resultado.

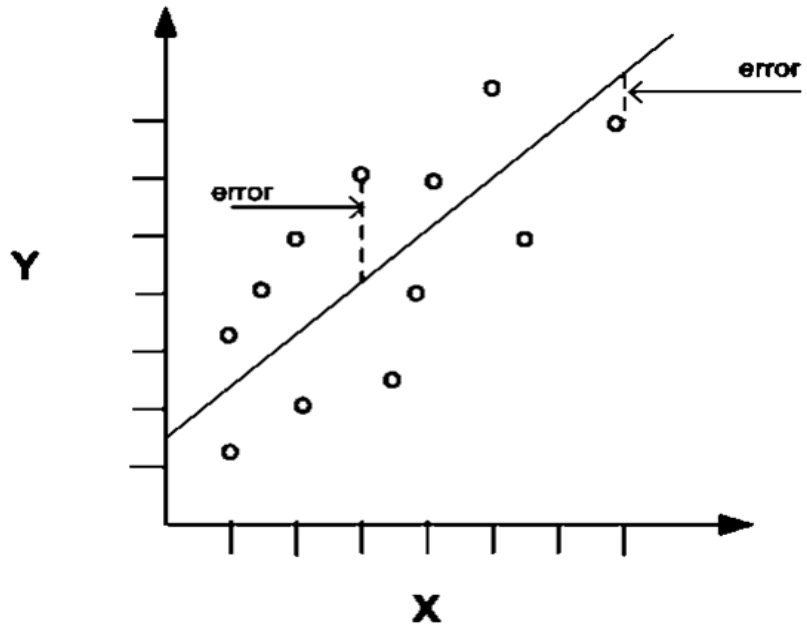
- $$X = 3 \Rightarrow \hat{Y} = 600 + 300 * 3 = 1500$$

→ Valor promedio previsto para todos los sujetos que han obtenido en la variable X un valor de Xi. Las personas con 3 años de experiencia tienen un sueldo promedio de 1500 €.

Regresión Lineal Simple

EVALUACIÓN DE MODELOS DE REGRESIÓN

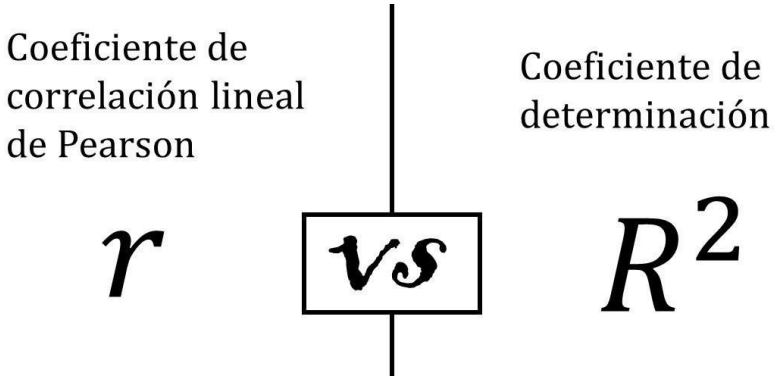
Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$



Regresión Lineal Simple

El coeficiente de determinación, se define como la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también llamado R cuadrado, refleja la bondad del ajuste de un modelo a la variable que pretender explicar.

Es importante saber que el resultado del coeficiente de determinación oscila entre 0 y 1. Cuanto más cerca de 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que estamos intentando explicar. De forma inversa, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, menos fiable será.



Regresión Lineal Múltiple

Regresión lineal múltiple

- Se ha visto el tema del análisis de regresión simple:

$$\text{Precio de la casa} = \beta_0 + \beta_1(\text{Área de la casa}) + \varepsilon$$

- Pero en general, una variable dependiente depende de más de una variable independiente:

- Precio de la casa puede depender de:
  - Área
  - Antigüedad
  - Número de baños
  - Área del garaje, etc.

Regresión Lineal Múltiple

- Para tratar este tipo de problemas se requiere expandir el análisis de regresión:

Regresión Lineal Simple

Regresión Lineal Múltiple

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots + \beta_p x_p + \varepsilon$$

Regresión Lineal Múltiple

Multicolinealidad

La multicolinealidad ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo. La consecuencia es no poder identificar de forma precisa el efecto individual que tiene cada predictor sobre la variable respuesta, lo que se traduce en un incremento de la varianza de los coeficientes de regresión estimados hasta el punto de que resulta imposible establecer su significancia estadística.

Como reconocerlo:

- Si el R2 es alto pero los coeficientes son no significativos, es un indicio.
- Si existe correlación entre las variables, si en su mayoría son mayores a 0.5, es un indicio.

Factor de Inflación de la Varianza (**VIF**),

$$VIF_{\hat{\beta}_j} = \frac{1}{1 - R^2}$$

VIF = 1, ausencia total
1<VIF<5, moderado
5 <VIF< 10, muy afectado

Regresión Lineal Múltiple

Algunas de las posibles soluciones al problema de multicolinealidad son las siguientes:

- mejora del diseño muestral estrayendo la información máxima de la variables observadas.
- eliminación de las variables que se sospechan son causantes de la multicolinealidad.
- en caso de disponer de pocas observaciones, aumentar el tamaño de la muestra.
- utilizar la relación extramuestral que permita realizar relaciones entre los parámetros (información a priori) que permita estimar el modelo por mínimos cuadrados restringidos.

Regresión Lineal Múltiple

Variables cualitativas

- Muchas veces en el modelo de regresión aparecen factores cualitativos (sexo, raza, estado civil,….). En estos casos la información relevante se puede representar con la ayuda de variables ficticias.
- Las variables ficticias son variables binarias que toman valor 0,1.
- Al definir una variable ficticia debemos decidir a qué acontecimiento se le asigna el valor 1, y a cuál el 0.

Ejemplo:
Supóngase que se pretende transformar la variable **"medios de transporte más comunes"** de tres categorías: 1=autobús, 2=tren y 3=avión.
La conversión podría efectuarse por medio de dos variables dicotómicas, F1 , F2 y F3. Los valores que éstas tomarían para representar cada categoría serían los siguientes:

Categoría	F1	F2	F3
Autobús	1	0	0
Tren	0	1	0
Avión	0	0	1

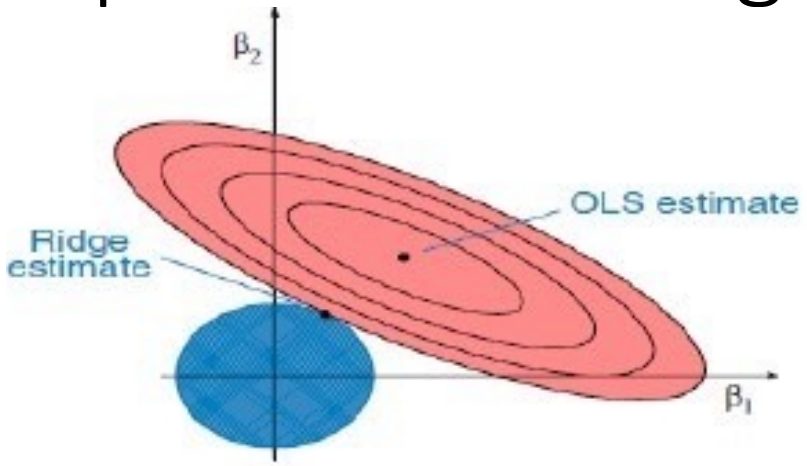
Problemas con la Regresión

La **regresión lineal** trata de modelar la relación entre una variable continua y una o más variables independientes mediante el ajuste de una ecuación lineal. Tres de las limitaciones que aparecen en la práctica al tratar de emplear este tipo de modelos (ajustados por mínimos cuadrados ordinarios) son:

- Se ven perjudicados por la incorporación de predictores correlacionados.
- No realizan selección de predictores, todos los predictores se incorporan en el modelo aunque no aporten información relevante.
- No pueden ajustarse cuando el número de predictores es superior al número de observaciones.

Una forma de atenuar el impacto de estos problemas es utilizar estrategias de regularización como ***ridge***, ***Lasso*** o ***Elastic Net***, que fuerzan a que los coeficientes del modelo tiendan a cero, minimizando así el riesgo de ***overfitting***, reduciendo varianza, atenuado el efecto de la correlación entre predictores y reduciendo la influencia en el modelo de los predictores menos relevantes.

Modelo de regresión penalizada: Ridge



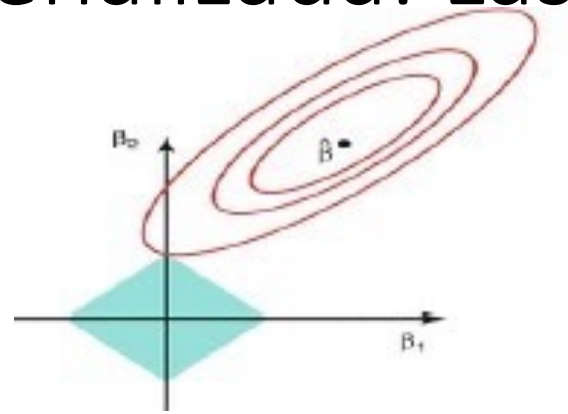
$$\underbrace{\sum_{i=1}^{n}(y_i - \beta_o - \sum_{j=1}^{p}\beta_j x_{ij})^2}_{\text{Mínimos cuadrados}} + \underbrace{\lambda \sum_{j=1}^{p}\beta_j^2}_{\text{Penalización ridge}}$$

- Es útil cuando existe colinearidad entre las variables utilizadas para el entrenamiento.
- El factor λ (lambda) sirve para controlar la intensidad de la regularización. Se utiliza la norma de regularización L2.
- La elección de este parámetro involucra un balance entre los componentes de sesgo y varianza del error cuadrático medio al estimar β.

PROGRAMA DE ESPECIALIZACIÓN ANALÍTICA
ADVANCED DATA SCIENCE

Online | DiiC Perú

Modelo de regresión penalizada: Lasso



$$\underbrace{\sum_{i=1}^{n}(y_i - \beta_o - \sum_{j=1}^{p}\beta_j x_{ij})^2}_{\text{Mínimos cuadrados}} + \underbrace{\lambda \sum_{j=1}^{p}|\beta_j|}_{\text{Penalización lasso}}$$

- Permite reducir a valores cercanos a cero los coeficientes de las variables menos relevantes.
- Se utiliza para la reducción de la dimensionalidad.
- El factor λ (lambda) controla el nivel de regularización.
- Lasso es una técnica de regresión lineal regularizada, como Ridge, con la leve diferencia en la penalización. (Norma L1 en lugar de L2)

Modelo de regresión penalizada: Lasso

- Para valores crecientes de λ , los coeficientes βj se contraen hacia cero como en Ridge (shrinkage), con la diferencia de que algunos de ellos se anulan.

- Esto es, Lasso produce estimación y selección de variables en forma continua y simultánea, siendo especialmente útil en el caso p>=n.

- En los últimos años se han presentado algunas generalizaciones y extensiones de las técnicas presentadas anteriormente, especialmente diseñadas para ciertas situaciones particulares.

Modelo de regresión penalizada: elastic net

Es una combinación de Ridge y Lasso. Se decide, que peso se le da a cada método de penalización y se implementa la regresión.

$$\underbrace{\sum_{i=1}^{n}(y_i - \beta_o - \sum_{j=1}^{p}\beta_j x_{ij})^2}_{\text{Mínimos cuadrados}} + \underbrace{\lambda_1 \sum_{j=1}^{p}\beta_j^2}_{\text{Ridge}} + \underbrace{\lambda_2 \sum_{j=1}^{p}|\beta_j|}_{\text{Lasso}}$$

$$\hat{\beta}^e = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \frac{1-\alpha}{2}\lambda\sum_{i=1}^{p}\beta_i^2 + \alpha\lambda\sum_{i=1}^{p}|\beta_i|$$

donde λ es un parámetro de precisión

- Si $\lambda = 0$, regresión lineal tradicional ($\hat{\beta}^e = \hat{\beta}$).
- Si $\lambda = \infty$, $\hat{\beta}^e = 0$
- Si $\alpha = 0$, entonces $\hat{\beta}^e = \hat{\beta}^{ridge}$
- Si $\alpha = 1$, entonces $\hat{\beta}^e = \hat{\beta}^{lasso}$

PROGRAMA DE
ESPECIALIZACIÓN ANALÍTICA

ADVANCED
DATA SCIENCE

Online | DMC Perú