

**{desafío}**  
**latam\_**

**Regresión \_**



# Regresión Lineal

# Objetivo

- Características de la regresión: **Marco analítico flexible para preguntas de asociación y causalidad.**
- Responde a la pregunta: ¿Cómo el cambio de una variable afecta el valor de otra?
- Conjetura básica de la regresión:

Variable Dependiente  $\xleftarrow{\text{Afecta}}$  Variable Independiente

# Algunas definiciones

- **Variable Dependiente:** Objeto de estudio medido en una variable
- **Variable Independiente:** Posibles factores explicativos de la variable dependiente
- **Error:** Término residual asociado a lo no explicado por el modelo.
- **Modelo:** Aproximación funcional a nuestro fenómeno.
- **Coeficientes:** Componentes estimados del modelo que permiten aproximar características de los datos en la variable dependiente.

# Codificación de Variables Categóricas

## One-Hot Encoding (OHE)

```
pd.get_dummies(df.region)
```

	Africa	Americas	Asia	Europe	Oceania
0	1	0	0	0	0
1	1	0	0	0	0
2	1	0	0	0	0
3	1	0	0	0	0
4	1	0	0	0	0
...	...	...	...	...	...
189	0	0	0	0	1
190	0	0	0	0	1
191	0	0	0	0	1
192	0	0	0	0	1
193	0	0	0	0	1

194 rows × 5 columns

## Binary Encoding

```
pd.get_dummies(df.region, drop_first = True)
```

	Americas	Asia	Europe	Oceania
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
...	...	...	...	...
189	0	0	0	1
190	0	0	0	1
191	0	0	0	1
192	0	0	0	1
193	0	0	0	1

194 rows × 4 columns

# Codificación de Variables Categóricas

## Label Encoding

```
region_coded = pd.get_dummies(df.region, drop_first = True)
pd.concat([df, region_coded], axis = 1)
```

	country	region	gdp	school	adfert	chldmort	life	pop	urban	femlab	literacy	co2	gini	Americas	Asia	Europe	Oceania
0	Algeria	Africa	7300.399902	6.716667	7.300000	34.75	72.316666	34172236	64.933334	0.4522	72.599998	15.00	NaN	0	0	0	0
1	Benin	Africa	1338.800049	3.100000	111.699997	122.75	54.733334	8237634	41.000000	0.8482	41.700001	1.20	NaN	0	0	0	0
2	Botswana	Africa	12307.400391	8.600000	52.099998	60.25	52.250000	1941233	59.250000	0.8870	84.099998	9.20	NaN	0	0	0	0
3	Burkina Faso	Africa	1063.400024	1.300000	124.800003	170.50	53.783333	15308383	23.583334	0.8584	23.600000	0.20	NaN	0	0	0	0
4	Burundi	Africa	349.200012	2.483333	18.600000	168.50	48.866665	7821783	10.250000	1.0344	66.599998	0.10	33.299999	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
189	Samoa	Oceania	4012.600098	10.300000	28.299999	26.75	71.533333	181600	20.666668	0.5010	98.800003	3.10	NaN	0	0	0	1
190	Solomon Islands	Oceania	2249.199951	4.500000	70.300003	36.00	66.500000	503617	17.766666	0.4858	NaN	1.40	NaN	0	0	0	1
191	Tonga	Oceania	4072.199951	10.133333	22.299999	19.25	71.833336	102550	23.266666	0.7150	99.000000	4.85	NaN	0	0	0	1
192	Tuvalu	Oceania	NaN	NaN	23.299999	36.50	66.033333	9767	49.233334	NaN	NaN	NaN	NaN	0	0	0	1
193	Vanuatu	Oceania	3809.800049	6.700000	54.000000	17.75	69.966667	225317	24.500000	0.8988	82.000000	1.50	NaN	0	0	0	1

194 rows × 17 columns

# Regresión Lineal desde la Econometría

# Conceptualizaciones de la Regresión

- Forma más simple: Tanto V.D como V.I son continuas.
- Resulta que cuando realizamos un diagrama de dispersión y agregamos esa recta de ajuste, estamos generando una regresión.
- Mediante la regresión, buscamos generar una explicación plausible de cómo V.I afecta los niveles de V.D, en promedio.



# Nuestra Primera Regresión

The diagram shows the linear regression equation  $earn_i = \beta_0 + \beta_1 \cdot height_i + \varepsilon_i$  with five labels in green boxes and arrows pointing to the corresponding parts of the equation:

- Variable Dependiente** points to  $earn_i$ .
- Pendiente** points to  $\beta_1$ .
- Variable Independiente** points to  $height_i$ .
- Intercepto** points to  $\beta_0$ .
- Error** points to  $\varepsilon_i$ .

$$earn_i = \beta_0 + \beta_1 \cdot height_i + \varepsilon_i$$

# Statsmodels

- Para implementar nuestra regresión utilizaremos el módulo ols de la librería statsmodels.
- Este genera un modelo de regresión mediante el método de mínimos cuadrados (Ordinary Least Squares).

```
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

# Bondad de Ajuste

- Métricas que informan sobre la capacidad explicativa y desempeño general del modelo.
  - **R-squared y Adj. R-squared:** ¿Cuál es la capacidad explicativa de nuestros regresores en la variabilidad de los puntajes de nuestro objetivo?
  - **F-Statistic y Prob(F-Statistic):** Prueba de rango de variabilidad entre partes explicadas y no explicadas.
  - **Log-Likelihood (Log-Verosimilitud):** Sirve para poder comparar el ajuste de nuestro modelo a los datos con respecto a un modelo sin predictores.
  - **IC (Criterio de información de Akaike):** Es una métrica de calidad relativa del ajuste de un modelo a los datos.
  - **BIC (Criterio de Información Bayesiano):** Métrica de ajuste relativo que debe ser comparada de entre los valores obtenidos para un conjunto de modelos candidatos.

# Coeficientes

- Interpretación descriptiva de los coeficientes: cómo los valores de una variable dependiente numérica varían en subpoblaciones definidas por una función lineal de atributos.
- Interpretación causal de los coeficientes: cómo el cambio en nuestra variable independiente causa cambios en nuestra variable dependiente.
- Problema de la interpretación causal: Muchos supuestos para hacerla válida.

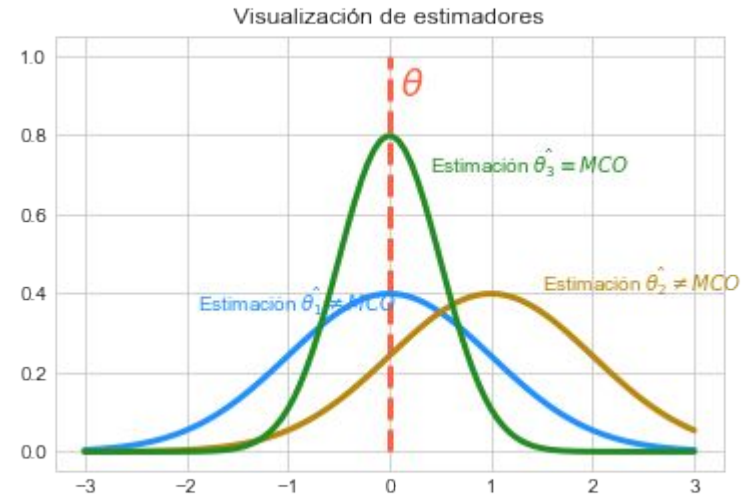
# Validez de las Estimaciones

- Método de Mínimos Cuadrados Ordinarios.
- Encontrar un estimador que reduzca la distancia residual entre los valores predichos y sus correlatos observados.

$$\begin{aligned}\beta &= \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}[(y_i - X^\top \beta)^2] \\ &= \sum_{i=0}^N (y_i - (\beta_0 + \beta_1 X))^2\end{aligned}$$

# Teorema de Gauss Markov

- La media del error es 0.
- El error es independiente de las variables explicativas.
- No existe correlación entre los residuos.
- El error debe ser constante.
- El error debe distribuirse de forma normal.



# Diagnósticos

- Una serie de diagnósticos de los errores nos permite determinar si el modelo satisface las condiciones de Gauss-Markov

# Variantes de la Regresión Lineal



# Variables Binarias

- Nuestra variable independiente toma dos valores.

$$\text{earn}_i = \beta_0 + \gamma_1 \times \text{male}_i + \varepsilon_i$$

# Términos Polinomiales

- Consideramos la posible no-linealidad de nuestras variables independientes.

$$\text{earn}_i = \beta_0 + \beta_1 \times \text{age}_i + \beta_2 \times \text{age}_i^2 + \varepsilon_i$$

# Múltiples Variables Independientes

- Se puede extender la cantidad de variables independientes a incluir en la ecuación, dando pie a una regresión lineal múltiple.

$$\text{earn}_i = \beta_0 + \beta_1 \times \text{age}_i + \gamma_2 \times \text{male} = 1_i + \varepsilon_i$$

# Regresión Lineal desde Machine Learning

# Estadística vs Machine Learning

Estadística	Machine Learning
Modelos	Redes, Grafos
Variable Dependiente	Vector Objetivo
Variable Independiente, Covariable	Atributo
Parámetros	Pesos
Ajuste	Aprendizaje
Desempeño en Entrenamiento	Generalización

# Pasos en el Flujo de Machine Learning

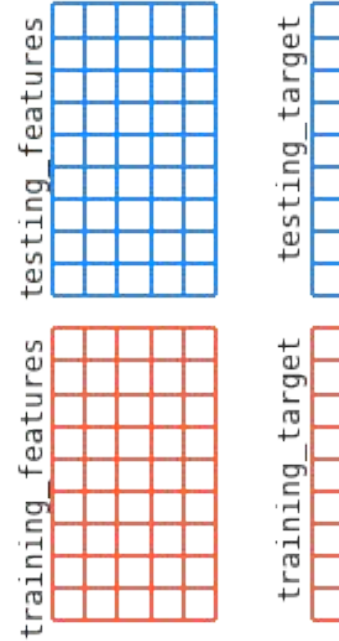
- Conocer los elementos:
  - Conocer qué representan.
- Determinar los objetivos de trabajo:
  - Los objetivos de trabajo determinan la arquitectura y modelos a implementar.
- Diseñar e implementar los Modelos:
  - ¿Qué esperamos como resultado?
  - ¿Qué parámetros estimaremos?
  - ¿Qué hiper parámetros consideraremos?

# Importación de Módulos

- Parte del flujo de trabajo de Machine Learning depende de scikit-learn.
- Se sugiere siempre importar cada componente de scikit-learn para reducir el overhead.
- Deben existir dos imports mínimos:
  - Uno de modelo.
  - Uno de métrica.

# División de la Muestra

- Se generan dos conjuntos de datos:
  - Training: Donde implementamos el modelo.
  - Test: Donde probamos el modelo.



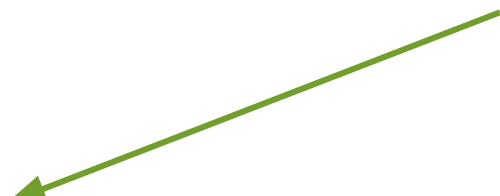


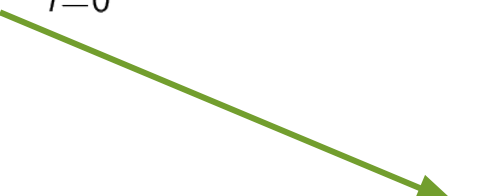
# Generación de Predicciones

- Con nuestro modelo entrenado, lo que evaluamos es su capacidad de generar explicaciones en un nuevo conjunto de datos no considerados anteriormente en el entrenamiento.
- Con ello, generamos una predicción de los valores en el conjunto de prueba que podemos contrastar posteriormente.

# Evaluación del desempeño

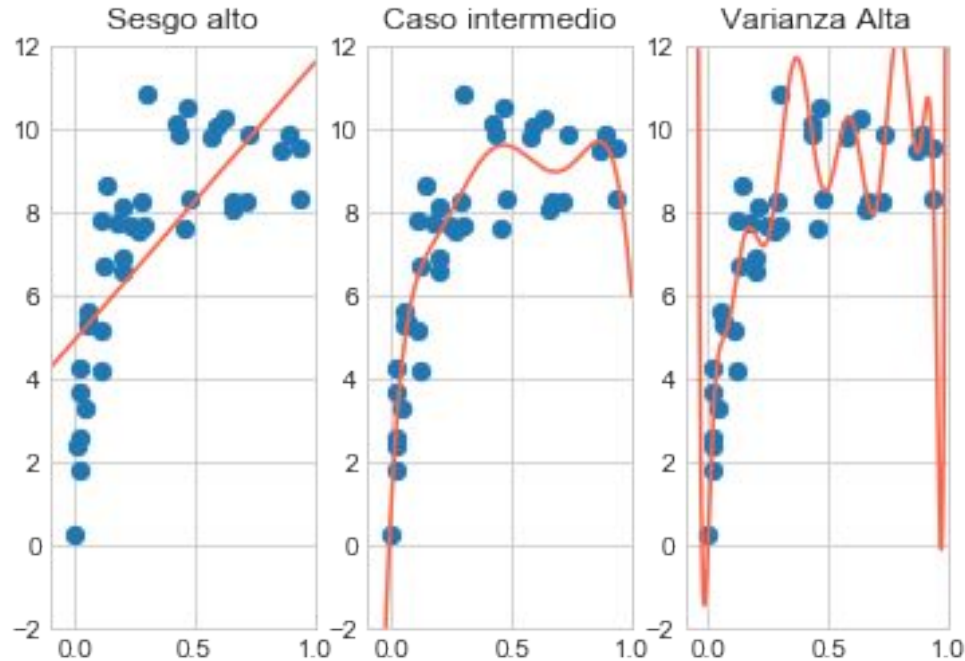
$$\text{MSE}(\hat{f}, \text{datos}) = \frac{1}{n} \sum_{i=0}^n \left( y_i - \hat{f}(\mathbf{x}_i) \right)^2$$


$$\text{MSE}_{\text{test}}(\hat{f}, \text{test}) = \frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} \left( y_i - \hat{f}(\mathbf{x}_i) \right)^2$$


$$\text{MSE}_{\text{train}}(\hat{f}, \text{train}) = \frac{1}{n_{\text{train}}} \sum_{i \in \text{train}} \left( y_i - \hat{f}(\mathbf{x}_i) \right)^2$$

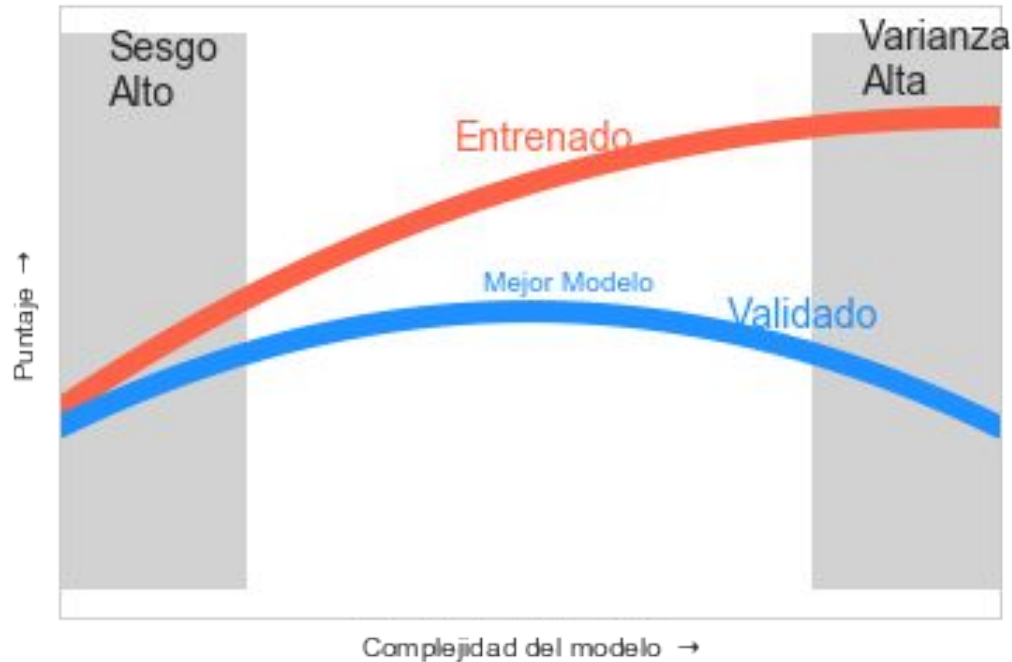
# Trueque entre Sesgo y Varianza

- Criterio de evaluación: capacidad de generalización del modelo



# Curva de Validación

- Evaluamos cómo se comporta el desempeño del modelo condicional a su complejidad.



# Curva de Aprendizaje

- Evaluamos cómo se desempeña el modelo, condicional a la cantidad de datos.



**{desafío}**  
**latam\_**

*Academia de  
talentos digitales*

[www.desafiolatam.com](http://www.desafiolatam.com)