

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.

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 illustrates the components of a linear regression equation. The equation is $earn_i = \beta_0 + \beta_1 \cdot height_i + \varepsilon_i$. Labels in green boxes with arrows point to each part: 'Variable Dependiente' points to $earn_i$, 'Pendiente' points to β_1 , 'Variable Independiente' points to $height_i$, 'Intercepto' points to β_0 , and 'Error' points to ε_i .

Variable Dependiente

Pendiente

Variable Independiente

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

Intercepto

Error

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

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.

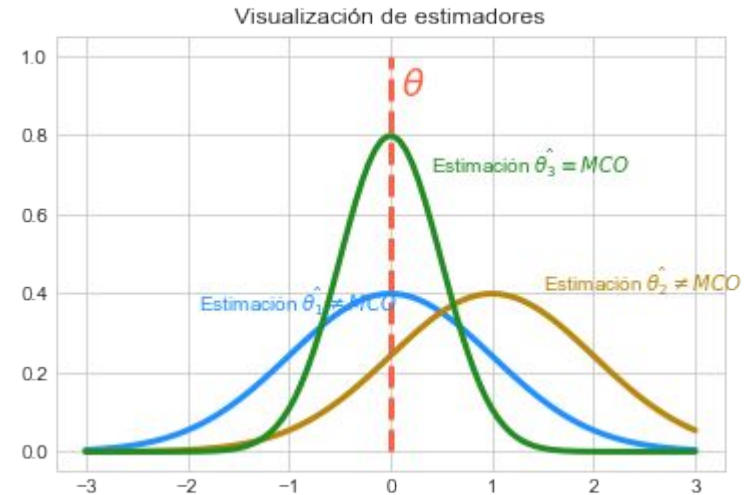
Validez de las Estimaciones

- 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

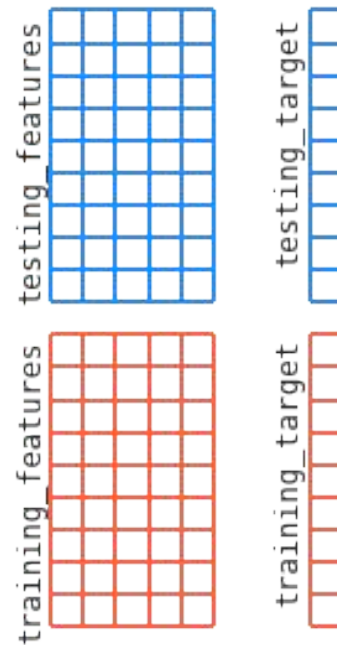
- Conocer los elementos:
 - Conocer qué representan.
- Determinar los objetivos de trabajo:
 - Los objetivos de trabajo determinan la arquitectura y modelos a implementar.
- Diseñar e implementar los Modelos:
 - ¿Qué esperamos como resultado?
 - ¿Qué parámetros estimaremos?
 - ¿Qué hiperparámetros consideraremos?

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.

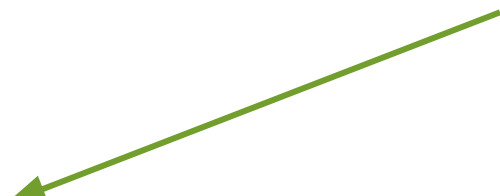


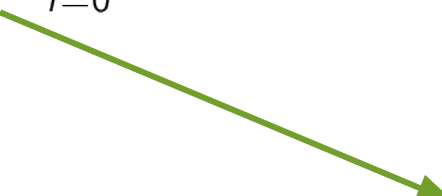
Generación de Predicciones

- 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

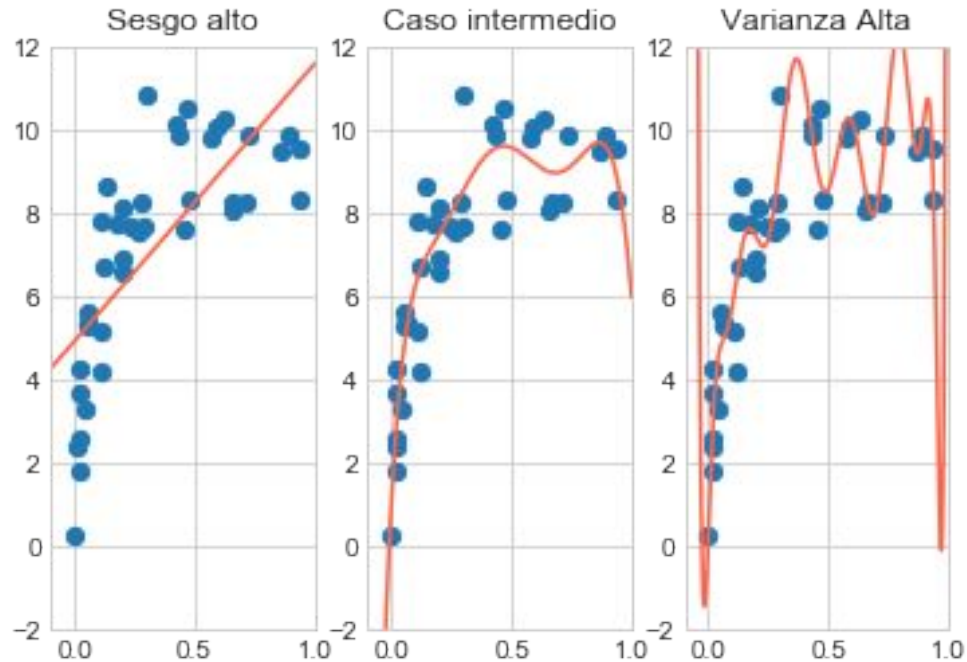
$$\text{MSE}(\hat{f}, \text{datos}) = \frac{1}{n} \sum_{i=0}^n \left(y_i - \hat{f}(\mathbf{x}_i) \right)^2$$


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


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

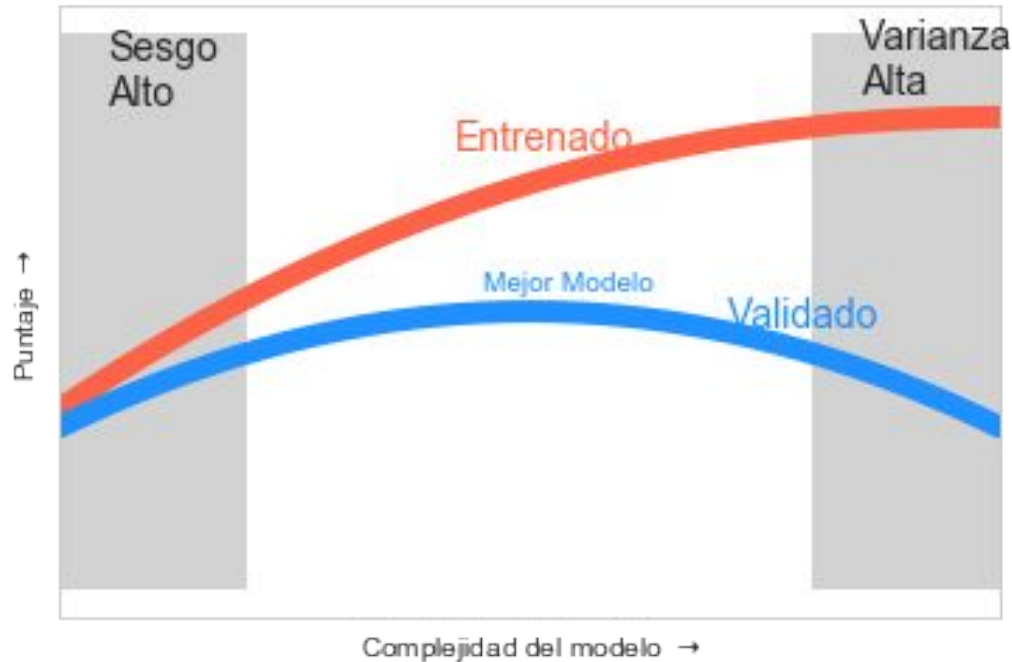
Trueque entre Sesgo y Varianza

- 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