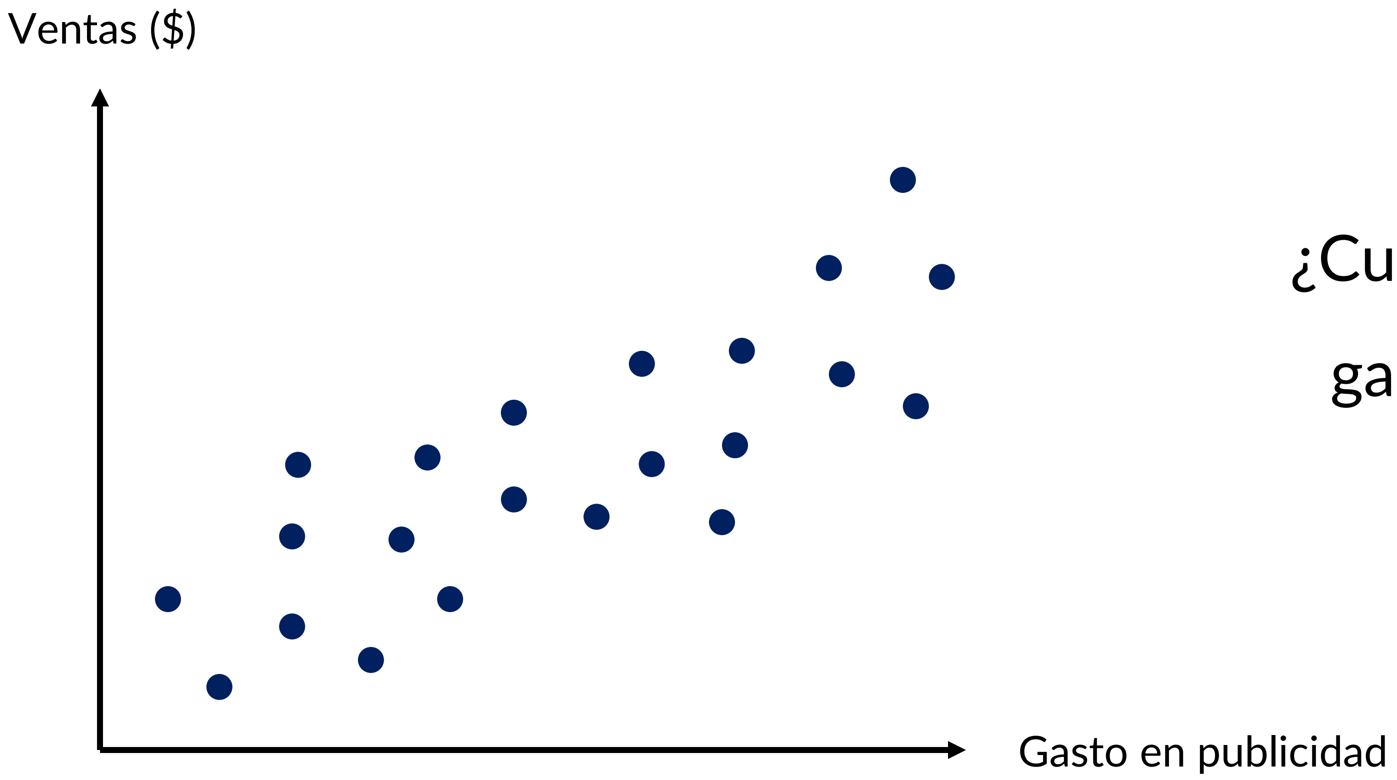
Bienvenidos

Regresión: Modelos lineales

Contenido

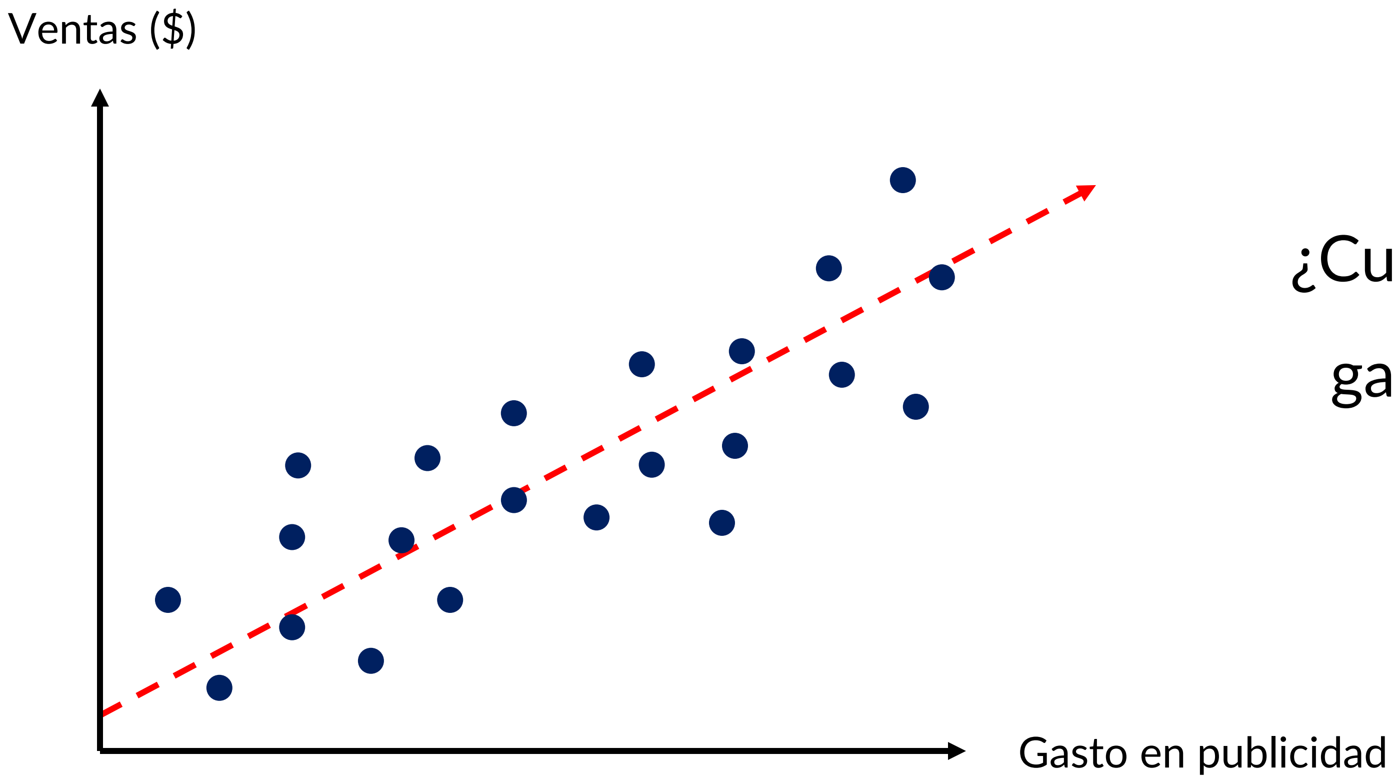
- Regresión lineal (MCO)
- Regresión polinomial
- Métricas de regresión
- Regularización: L1 Lasso
- Regularización: L2 Ridge
- Validación cruzada

Estimación Regresión Lineal



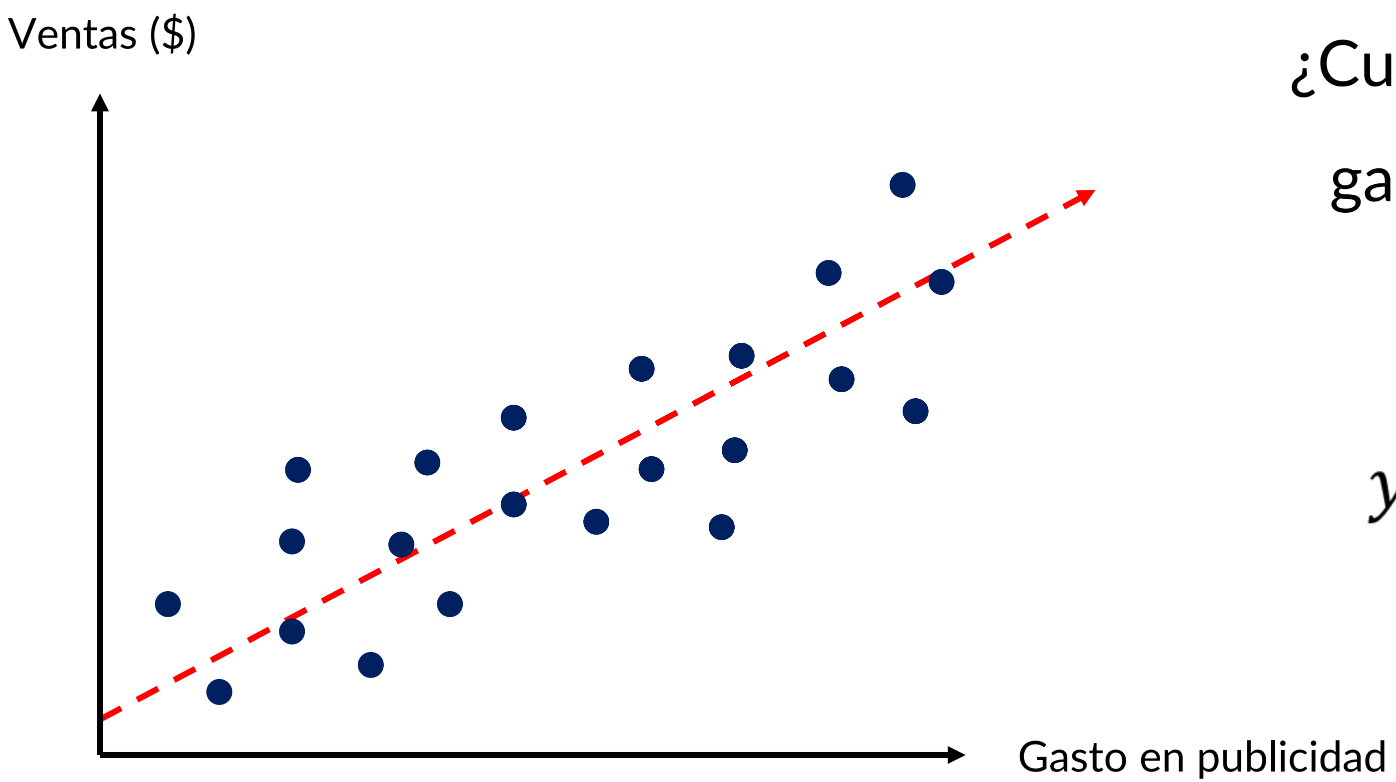
¿Cuál es la relación entre el gasto en publicidad y las ventas?

Estimación Regresión Lineal



¿Cuál es la relación entre el gasto en publicidad y las ventas?

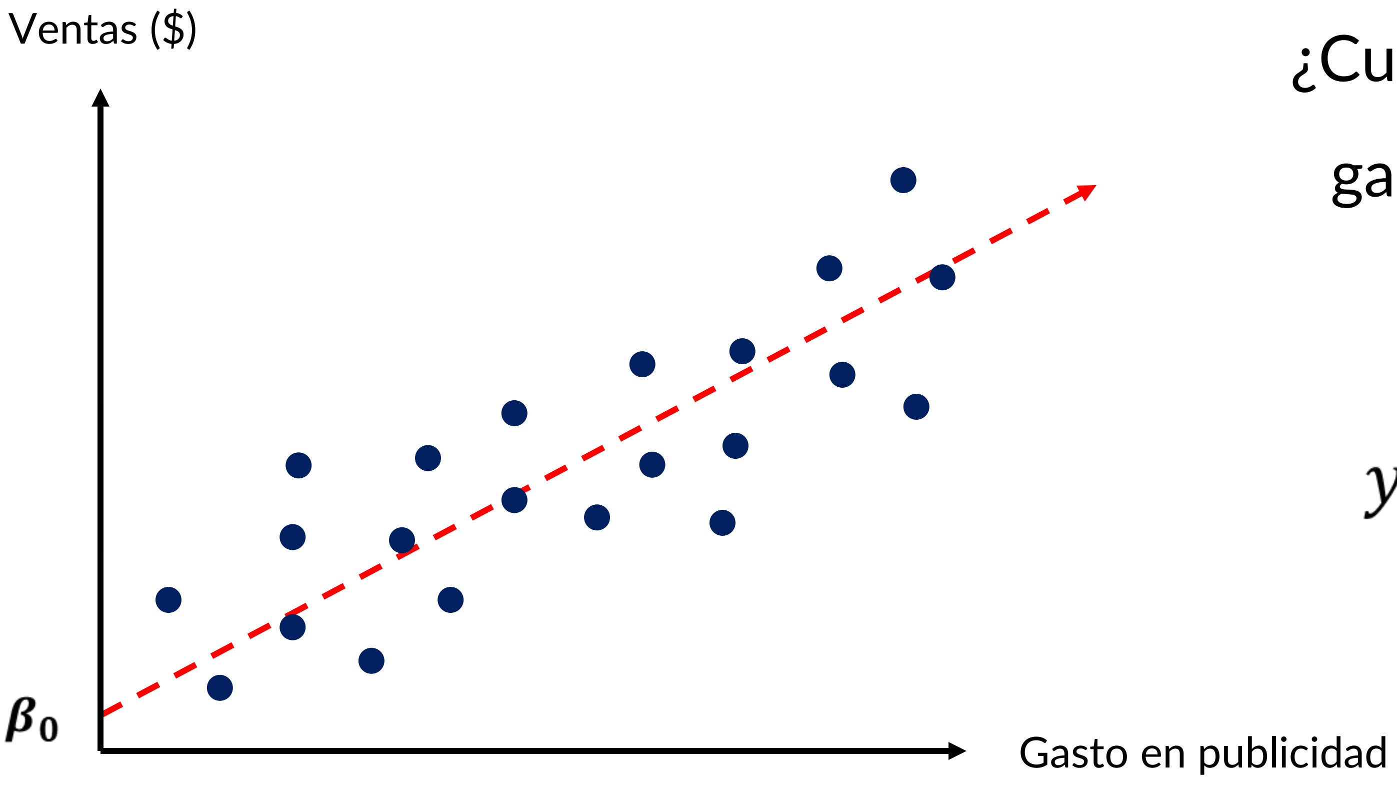
Estimación Regresión Lineal



¿Cuál es la relación entre el gasto en publicidad y las ventas?

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

Estimación Regresión Lineal

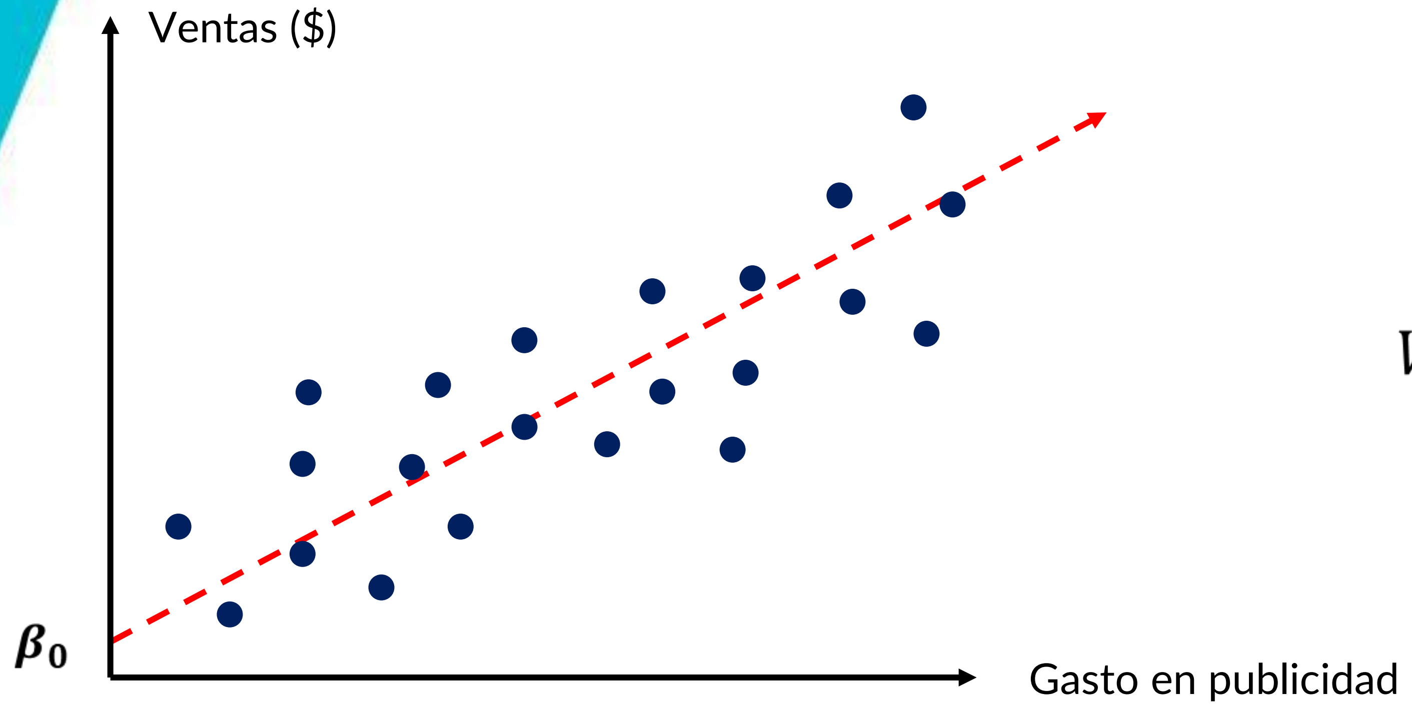


¿Cuál es la relación entre el gasto en publicidad y las ventas?

$$y_i = \boldsymbol{\beta_0} + \beta_1 x_{1i} + \varepsilon_i$$

Intercepto: Cuando x_{1i} es 0 ¿Cuánto es y_i?

Estimación Regresión Lineal

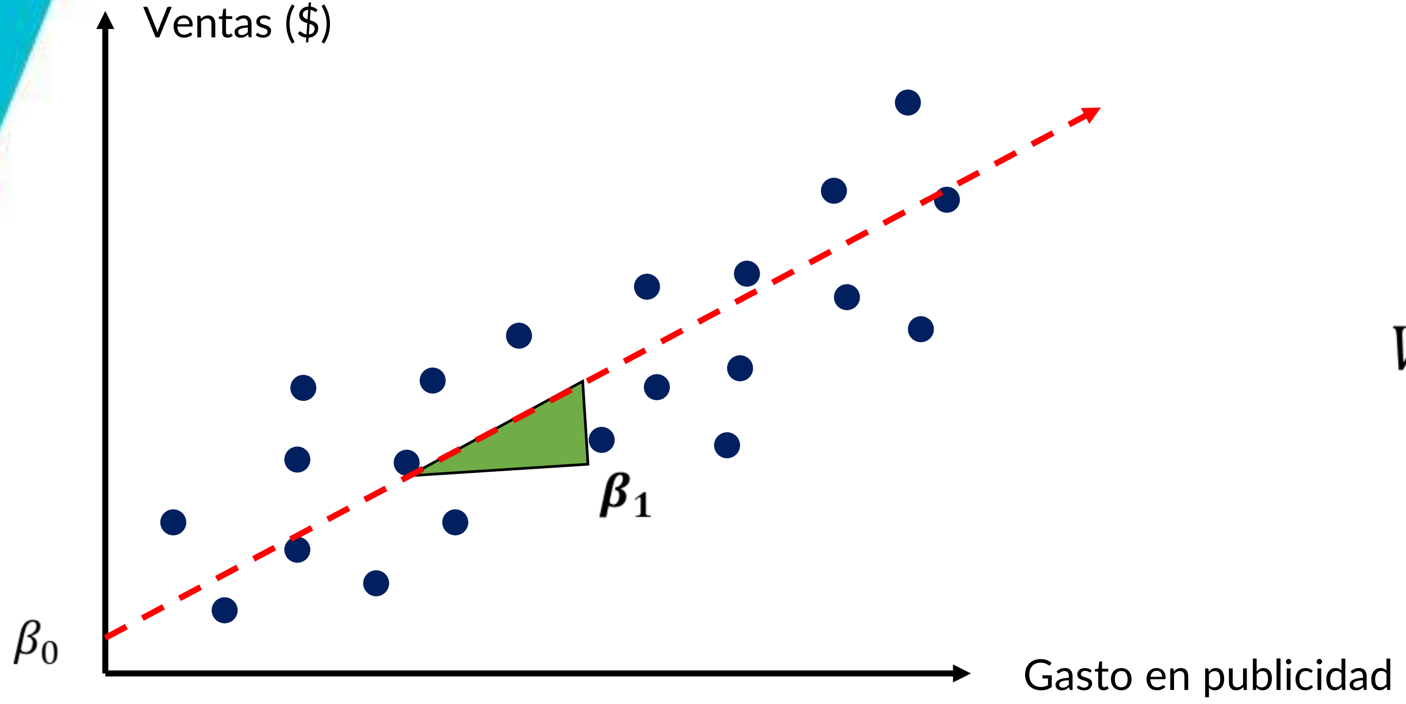


¿Cuál es la relación entre el gasto en publicidad y las ventas?

$$Ventas_i = \boldsymbol{\beta_0} + \beta_1 Publicidad_i + \varepsilon_i$$

Intercepto: Cuando no hay gasto en publicidad ¿Cuánto son las ventas en promedio?

Estimación Regresión Lineal



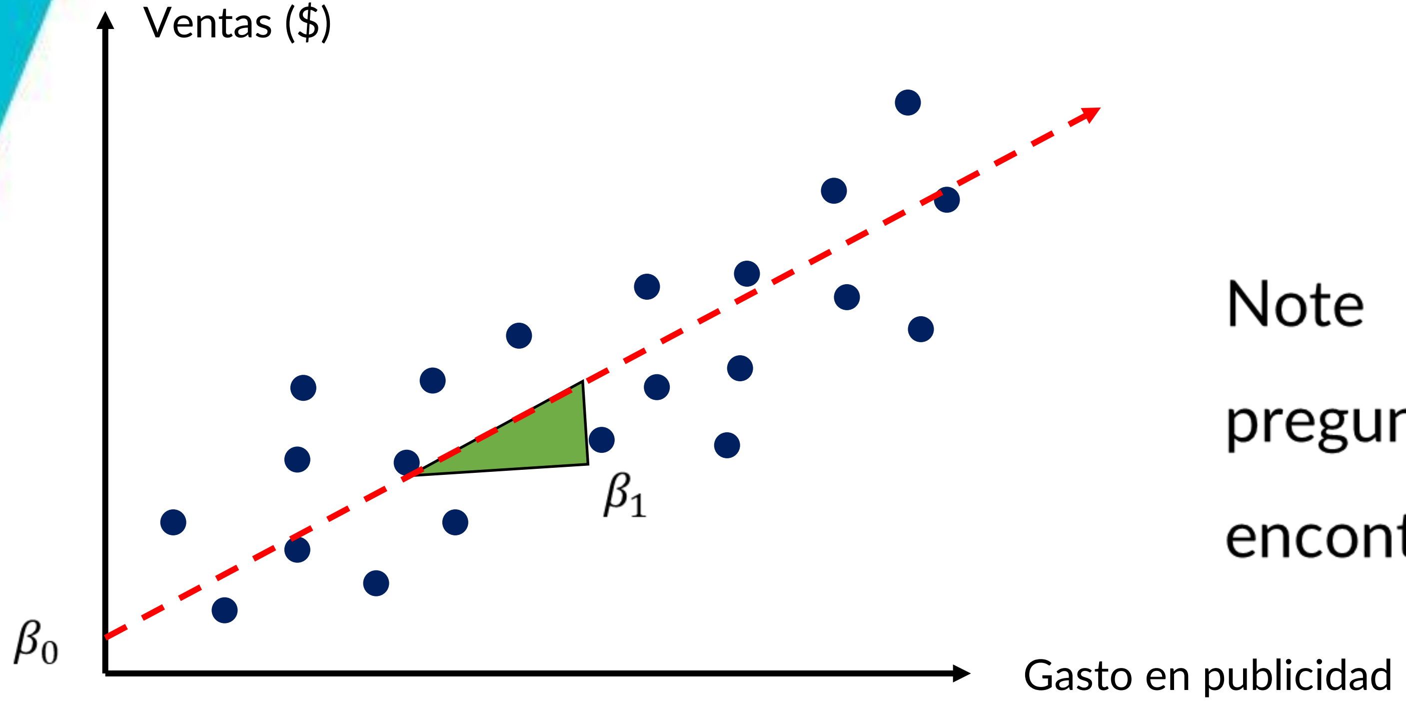
¿Cuál es la relación entre el gasto en publicidad y las ventas?

$$Ventas_i = \beta_0 + \boldsymbol{\beta_1} Publicidad_i + \varepsilon_i$$

Pendiente: Cuándo aumento la publicidad en un peso ¿En cuánto aumentan mis ventas en promedio?

Universidad de los Andes | Educación Continua
Vicerrectoría Académica

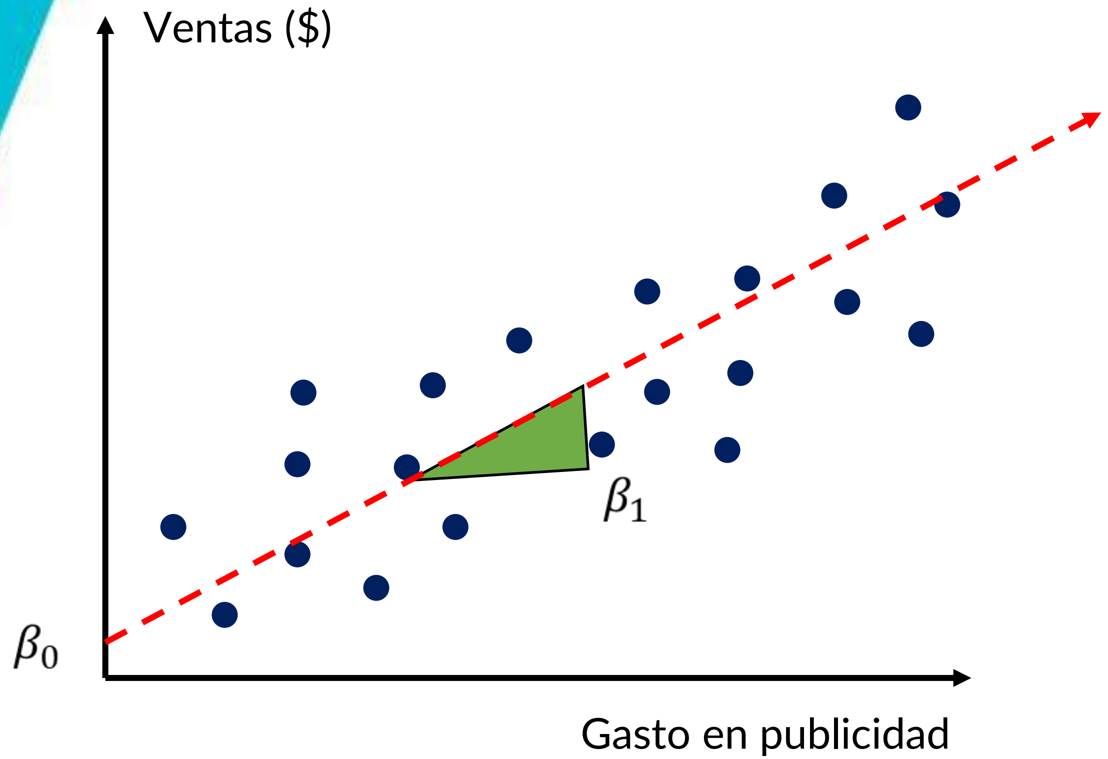
Estimación Regresión Lineal



¿Cuál es la relación entre el gasto en publicidad y las ventas?

Note que para responder dicha pregunta, solo es necesario encontrar dos números: β_0, β_1

Estimación Regresión Lineal



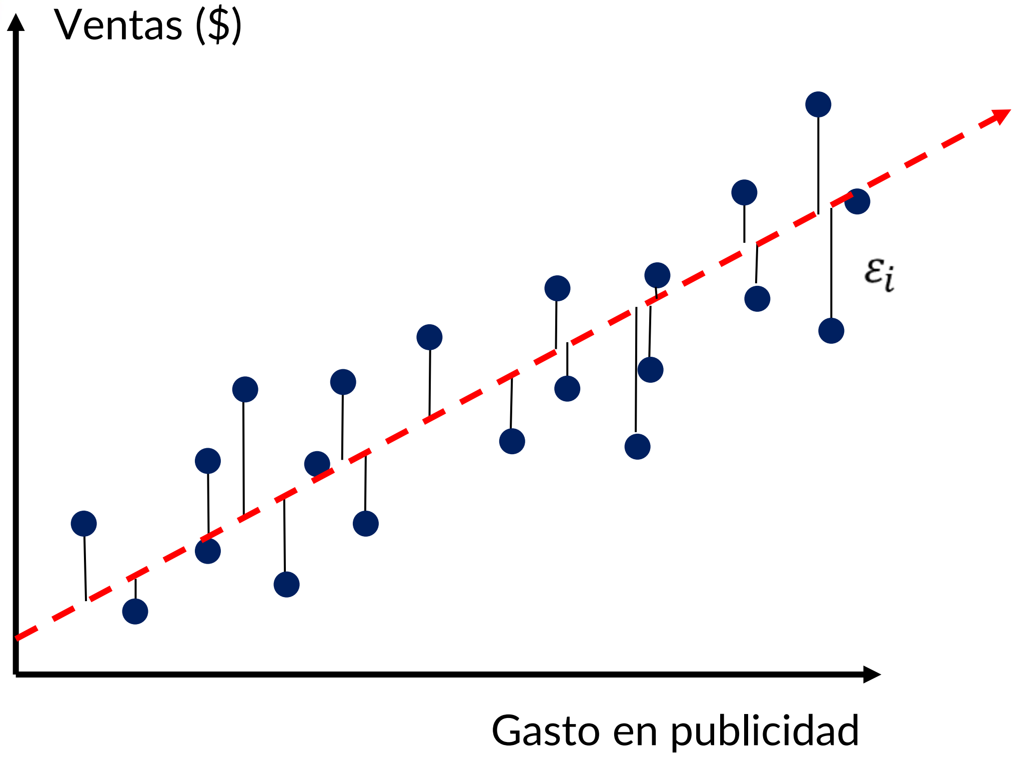
¿Cuál es la relación entre el gasto en publicidad y las ventas?

Note que para responder dicha pregunta, solo es necesario encontrar dos números: β_0, β_1

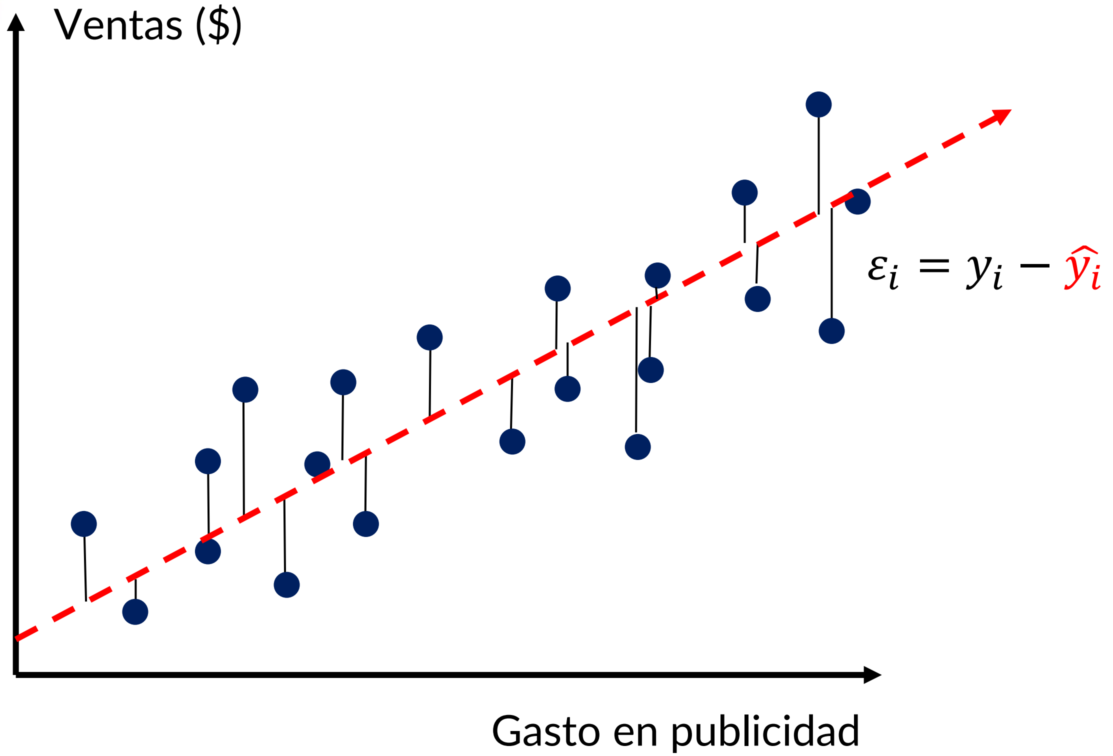
Llamamos estimación al procedimiento para encontrar esos números de manera correcta

Universidad de los Andes | Educación Continua Vicerrectoría Académica

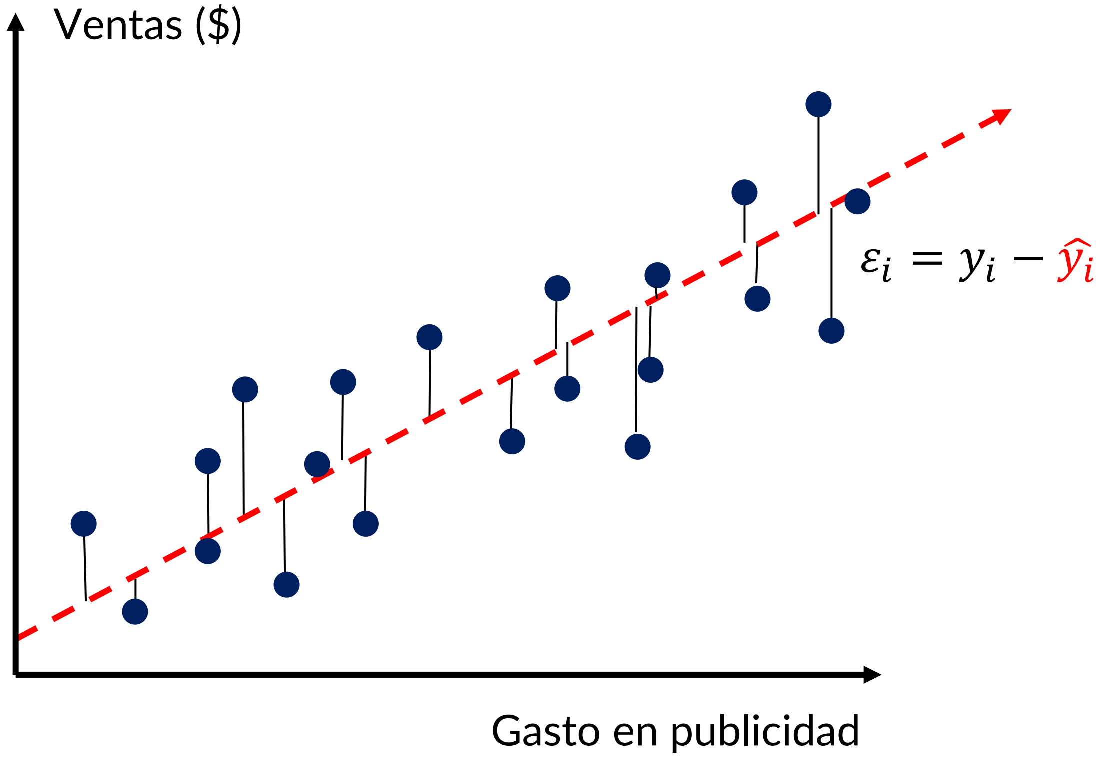
Mínimos Cuadrados Ordinarios



Mínimos Cuadrados Ordinarios



Mínimos Cuadrados Ordinarios



El método consiste en encontrar β_0, β_1 tal que la suma de todos los errores al cuadrado sea lo más pequeña posible.

$$RSS = \sum_{i=1} \varepsilon_i^2 = \sum (y_i - \hat{y}_i)^2$$

$$\widehat{\beta_0}, \widehat{\beta_1} = \min_{\beta_0, \beta_1} \sum (y_i - \beta_0 - \beta_1 x_{1i})^2$$

Esto se soluciona… ¡optimizando!

Generalicemos!

- Imaginémonos ahora que queremos predecir cuál es el precio promedio de vivienda en un vecindario con base en el ingreso medio de un hogar, la vejez del hogar y el número de habitaciones en promedio de las casas.
- En este caso, haremos la predicción con una línea recta cuya función se puede describir como:

$$\mathbf{Y} \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	Precio
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

Regresión lineal

- Para este nuevo problema nuestra "ecuación a estimar" será:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- Como en al caso anterior, los parámetros óptimos se van a encontrar minimizando un nuevo error que se definirá como:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3)^2$$

Regresión lineal – Solución General

- Para este nuevo problema nuestra "ecuación a estimar" será:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- Como en al caso anterior, los parámetros óptimos se van a encontrar minimizando un nuevo error que se definirá como:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3)^2$$

- Si utilizamos notación matricial queremos estimar

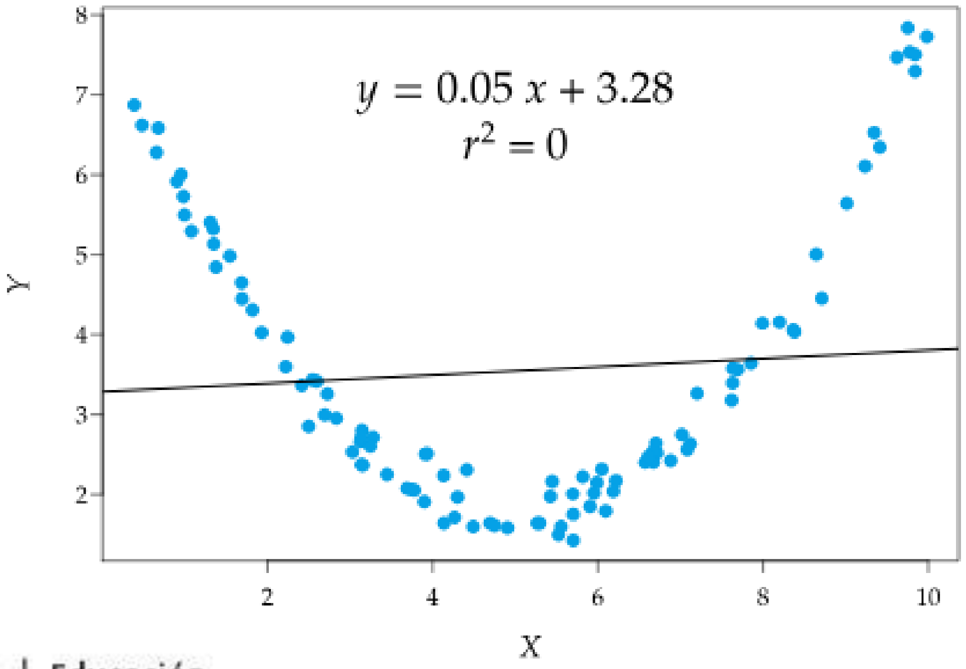
$$\vec{Y}_{1\times n} = \vec{\beta}_{1\times d} X_{d\times n} + \vec{\varepsilon}_{1\times n}$$

- La solución general para el es entonces.

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

Regresión polinomial

- Hasta ahora, nuestra regresión es de varias variables, pero estamos calculando únicamente relaciones lineales de covariables



En este ejemplo, la relación que hay entre la variable x y y NO ES lineal en covariables.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Debemos incluir efectos de orden polinomial superior.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

Universidad de los Andes | Educación Continua Vicerrectoría Académica

Regresión polinomial

- Cómo hacemos para capturar efectos de órdenes superiores?
- Construimos nuevas variables que den cuenta de estos cambios
  - x_1
  - x_2
  - x_1x_2
  - $x_1x_1 = x_1^2$
  - $x_2x_2 = x_2^2$
  - 1

En general, para *n* variables y un polinomio (interacciones) de grado *d* tendremos una cantidad de variables de tamaño:

$$\binom{n+d}{d} = \frac{(n+d)!}{d!n!}$$

- ¿Cómo sabemos si requerimos una regresión polinomial?
  - Contexto del problema
    - Business Sense - Intuición
  - Inspección visual
  - Métricas del problema

Métricas en Regresión

Mean Squared Error (MSE)

- Se utiliza para muchas tareas de regresión
- Se penaliza bastante los errores grandes y poco los pequeños

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

Root Mean Squared Error (RMSE)

- Similar al anterior pero misma escala que la variable dependiente

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

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i|$$

- Es robusto a los datos atípicos
- Es una medición más intuitiva que el anterior

Mean Absolute Percent Error (MAPE)

- Similar al anterior pero en escala porcentual

$$MAPE = \frac{1}{n}\sum \left|1 - \frac{\hat{y}_i}{y_i}\right| \times 100$$

Métricas en Regresión

R-Squared (R^2)

- Evalúa que tan bueno es el ajuste (bondad de ajuste)
- Muestra el porcentaje de la varianza de los datos que se explica por la varianza del modelo!
  - Nuestro caballito de guerra (bastante robusto y comparable entre distintos modelos)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Adjutes R-Squared (R2 adj)

- Similar al anterior pero toma en cuenta el número de predictores

$$adj\,R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

Pausa *epistemológica*… Econometría vs ML

Nuestra ecuación característica a estimar es:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$$

En un curso de inferencia causal (econometría), el objeto de estudio relevante se concentra siempre en los parámetros que encontramos:

- Nos preguntamos si caracterizan una relación causal de las covariables a la variable a estimar
- Requiere un estudio a profundidad sobre propiedades estadísticas deseables
    - Se discute ampliamente sobre la validez de ciertos supuestos de identificación

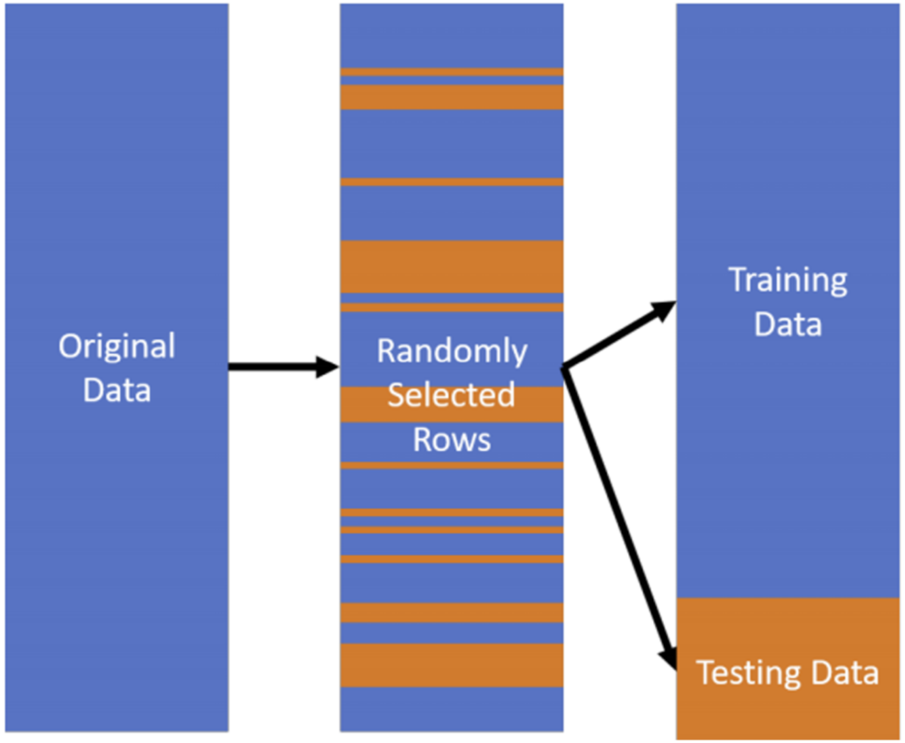
En nuestra aproximación (Machine Learning) Esta discusión pasa a un segundo plano.

- El método es relevante por sí mismo, más no se cuestiona (necesariamente) su validez.
- En este sentido, querer encontrar formas que "maximicen" o "minimicen" las métricas no es per se un problema.
    - Aunque va en contravía del componente "normativo" del análisis causal

Universidad de los Andes | Educación Continua Vicerrectoría Académica

El rol de la predicción en ML y los modelos lineales – Paradigma train-test

Ya que nos interesa la capacidad predictiva del modelo, tenemos que "simular datos" no observados



Para esto, vamos a usar el paradigma del train-test

- Separamos la muestra en 2 conjuntos
  - Muestra train, donde se entrena el modelo de regresión (Se recuperan los valores de los betas)
  - Muestra test, donde se ponen a prueba métricas del modelo con datos **NO OBSERVADOS**
- *Consideraciones:*
  - Estandarizar toda la muestra antes de separar
    - Esto puede ser un problema ante nuevos datos no observados
  - Usar una semilla para replicabilidad de los resultados

Universidad de los Andes | Educación Continua Vicerrectoría Académica

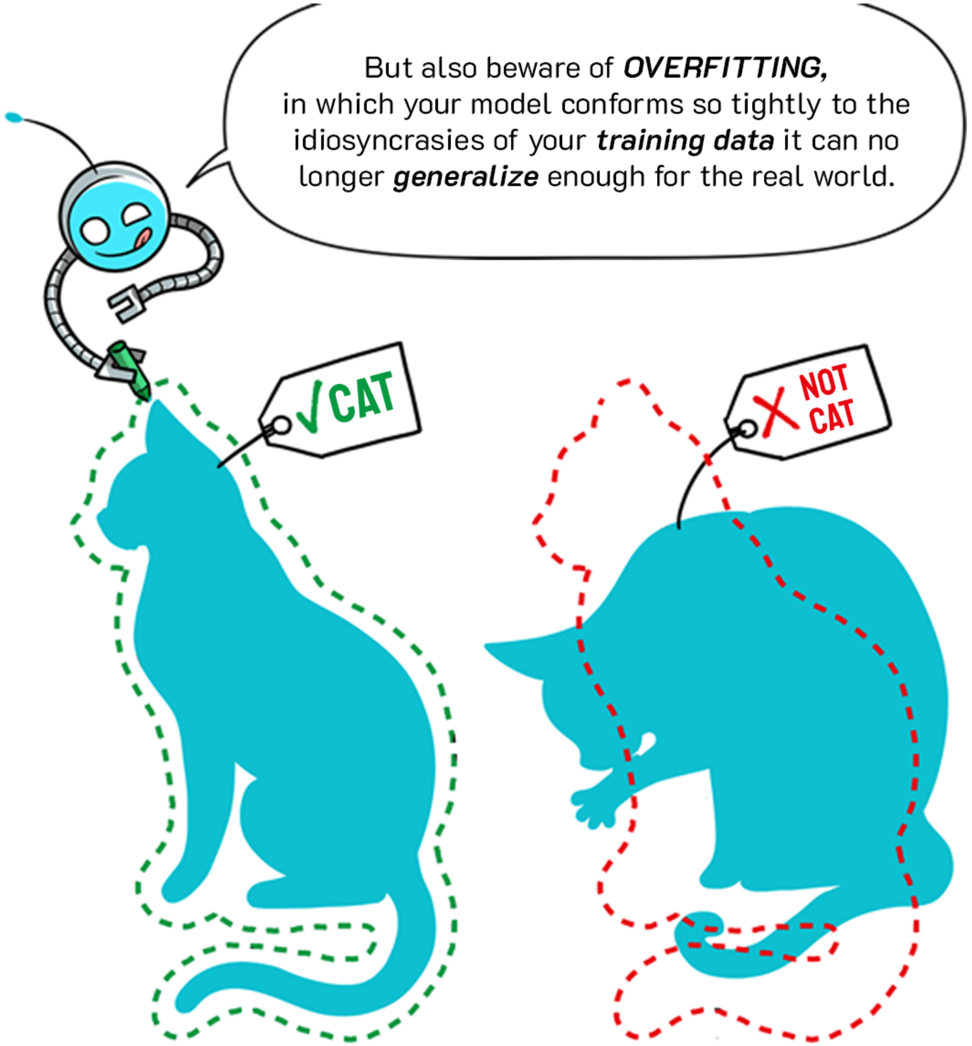
Break!

Regularización

- ¿Cómo prevenimos overfitting en nuestros modelos de regresiones lineales?
  - Pensemos en los retos de tener tantas covariables después de introducir interacciones de grado polinomial alto

<u>Regularización</u>

- La idea es limitar o restringir el tamaño de nuestros coeficientes β



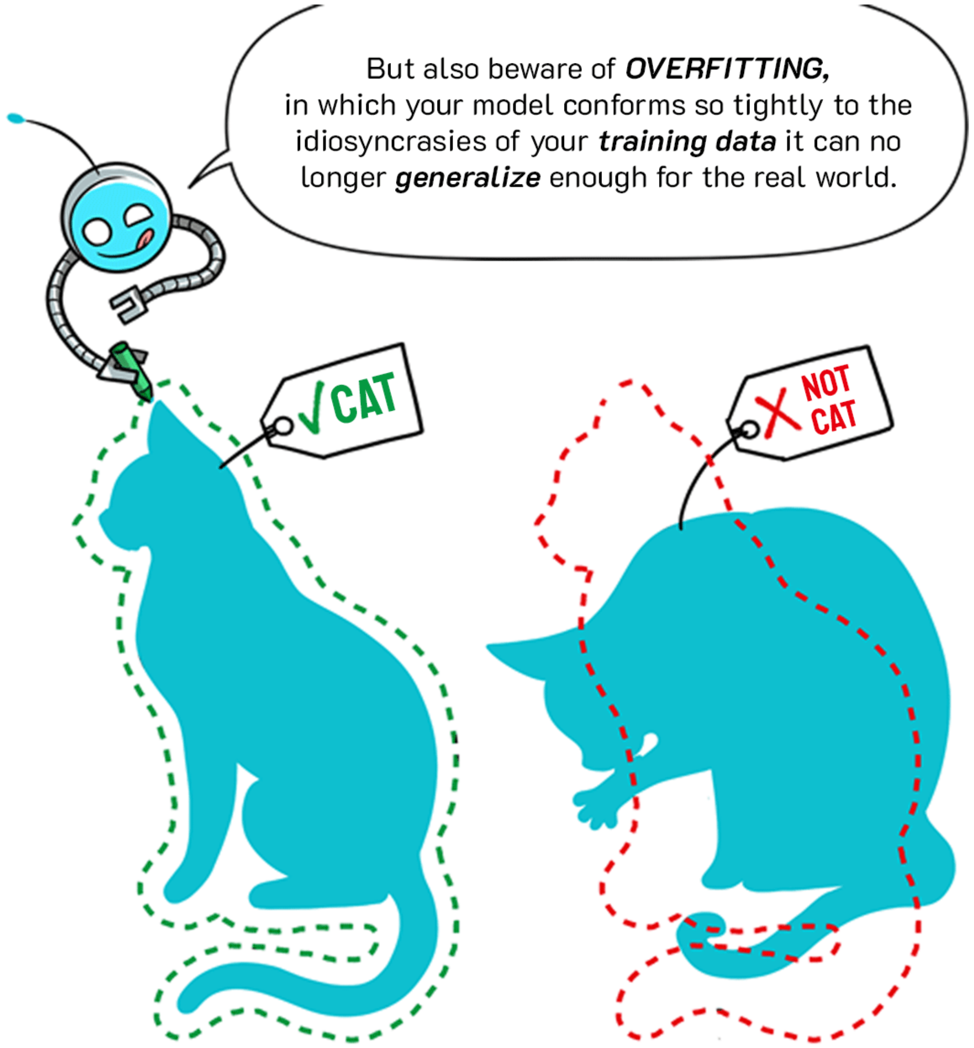
Regularización - ¿Para qué? - Consideraciones

¿Por qué regularizar?

- Prevenir el sobreajuste
- Tener mayor interpretabilidad
- Estabilizar los modelos
- Reducir multicolinealidad

Consideración Importante:

- **Todas las variables deben estar en la misma escala**
  - Estandarización



Regularización L1

- Idea general: introducir un término de penalización para la minimización de error

$$RSS_{L1} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \alpha \sum_{p=1}^{P} |\beta_p|$$

- En ML, se conoce como **Regresión LASSO**
- Enorme ventaja: puede volver los coeficientes 0
    - Nos sirve para hacer selección de variables!

Regularización L2

- Idea general: introducir un término de penalización para la minimización de error

$$RSS_{L2} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{p=1}^{P} {\beta_p}^2$$

- En ML, se conoce como **Regresión RIDGE**

Regularización

Regresión Lineal	Regularización L1 (Lasso)	Regularización L2 (Ridge)
$\hat{y} = \beta_0 + \sum_{p=1}^{P} \beta_p x_p$	$\hat{y} = \beta_0 + \sum_{p=1}^{P} \beta_p x_p$	$\hat{y} = \beta_0 + \sum_{p=1}^{P} \beta_p x_p$

Regularización

Regresión Lineal	Regularización L1 (Lasso)	Regularización L2 (Ridge)
$\hat{y} = \beta_0 + \sum_{p=1}^{P} \beta_p x_p$	$\hat{y} = \beta_0 + \sum_{p=1}^{P} \beta_p x_p$	$\hat{y} = \beta_0 + \sum_{p=1}^{P} \beta_p x_p$
$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$	$RSS_{L1} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \boxed{\alpha \sum_{p=1}^{P} \lvert\beta_p\rvert}$	$RSS_{L2} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \boxed{\alpha \sum_{p=1}^{P} {\beta_p}^2}$

$$\min_{\beta_0, \beta_1, \ldots, \beta_P} RSS \longrightarrow \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_P$$

Universidad de los Andes | Educación Continua Vicerrectoría Académica

Regularización

L1 (Lasso)

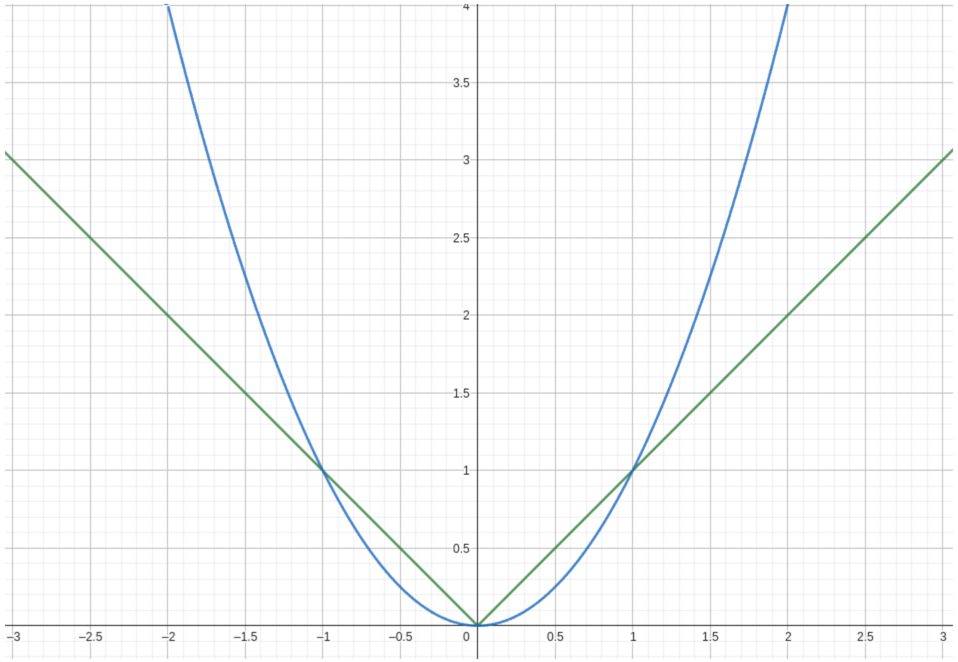
- Los coeficientes pequeños se penalizan más

$$RSS_{L1} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{p=1}^{P} \boxed{|\beta_p|}$$

L2 (Ridge)

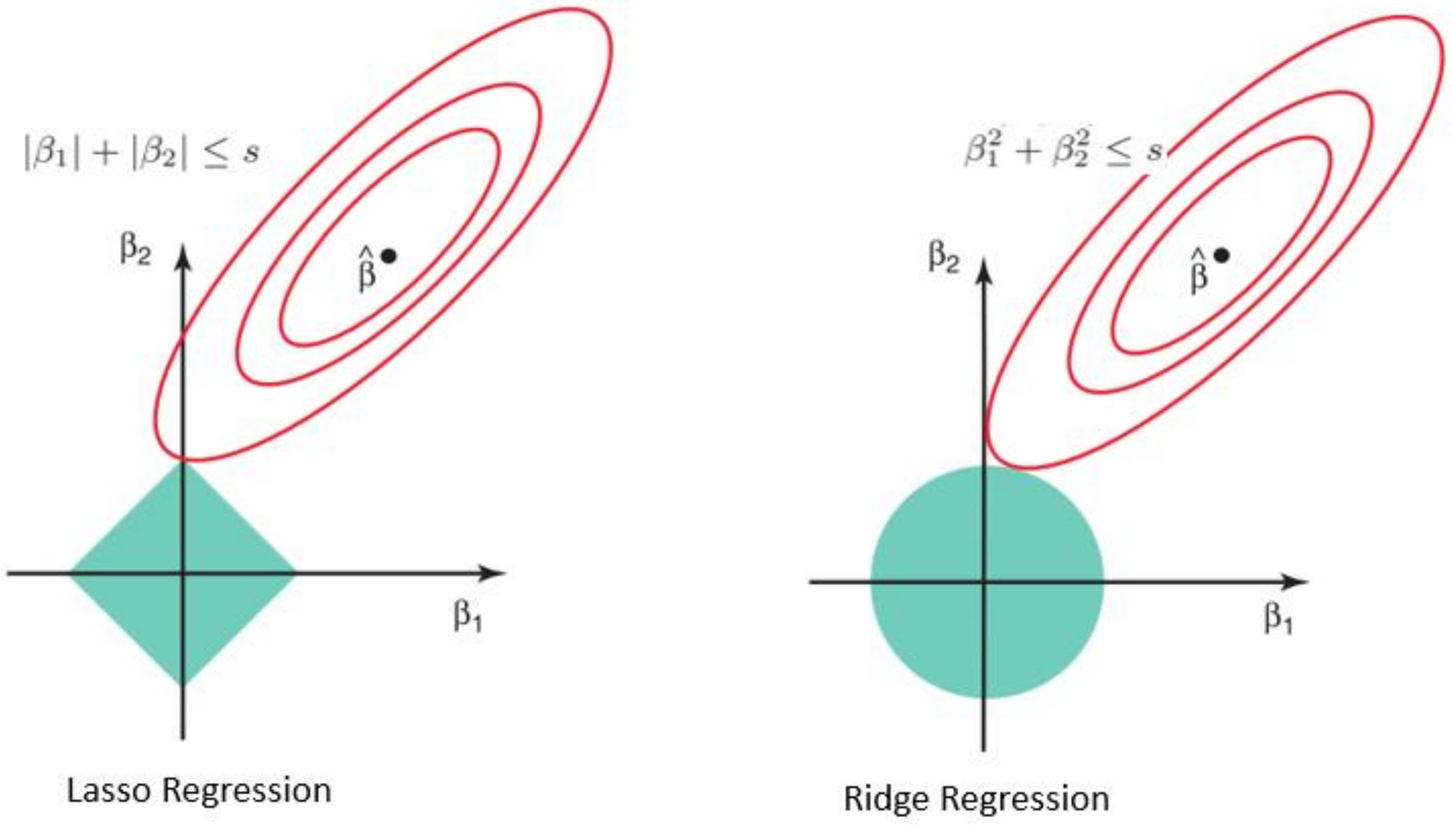
- Los coeficientes grandes se penalizan más

$$RSS_{L2} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{p=1}^{P} \boxed{\beta_p^2}$$



Universidad de los Andes | Educación Continua Vicerrectoría Académica

Regularización



- **Regularización Lasso** lleva coeficientes a cero, por lo tanto, permite hacer selección de variables

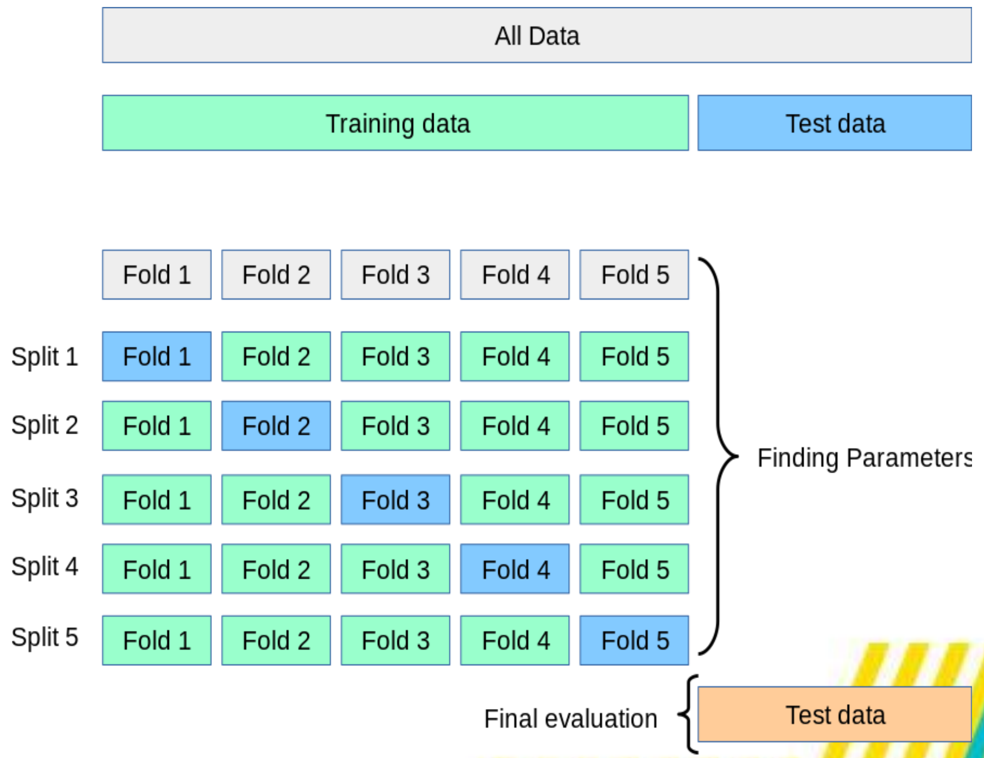
$$RSS_{L1} = RSS + \alpha(|\beta_1| + |\beta_2|)$$

$$RSS_{L2} = RSS + \alpha({\beta_1}^2 + {\beta_2}^2)$$

Regresando al paradigma Train-test - Crosvalidación

La escogencia del train – test puede llegar a ser arbitraria! Para tener resultados más robustos utilizamos cross-validación

1. El K-fold Crossvalidation separa la muestra train en K componentes y corre el modelo K veces
2. En cada una de esas corridas escoje de manera única y excluyente el sub-train y el sub-test
3. Saca un modelo estimado, evalúa una métrica en cuestión
4. Encuentra parámetros “ponderados” de todas las corridas
5. Evalúa en la muestra test



Discusión Final – Regresiones Lineales

<u>**Ventajas**</u>

- Simplicidad y facilidad de interpretación
- Rapidez y eficiencia
- Herramienta estrella para modelar relaciones lineles
- Regularización es una herramienta extra:
    - Limita el sobreajuste
    - Permite hacer selección de variables
    - Aumenta la capacidad de generalización del modelo

<u>**Desventajas**</u>

- Limitaciones cuando las relaciones son no-lineales (Lo veremos más adelante con random forest)
- Sensibilidad a los outliers
- Multicolinealidad
- Supuestos de gauss-markov
    - El fantasma de la econometría

¡Gracias!

Aprendiendo juntos a lo largo de la Vida

educacioncontinua.uniandes.edu.co

Siguenos en **EdcoUniandes**