

# Estadística en Analítica

2023-2

*Pablo A. Saldarriaga*  
[psaldar2@eafit.edu.co](mailto:psaldar2@eafit.edu.co)

UNIVERSIDAD  
**EAFIT**

# ¿Dudas del Taller 1?

# Aprendizaje Supervisado VS No Supervisado

## *Supervisado*

El proceso de modelado se realiza sobre un **conjunto de ejemplos** formado por **entradas** al sistema y la **respuesta** que debería dar para cada entrada

Se tiene conocimiento a priori de las observaciones

### **Objetivo:**

Replicar/aprender el comportamiento y patrones ya conocidos de los datos

## *No Supervisado*

El proceso de modelado se realiza sobre un **conjunto de datos** formado por solo **entradas** al sistema.

No se tiene conocimiento a priori de las observaciones.

### **Objetivo:**

La comprensión y el estudio de los datos.

# Metodología Aprendizaje Supervisado



Análisis exploratorio



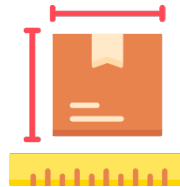
Calidad de datos



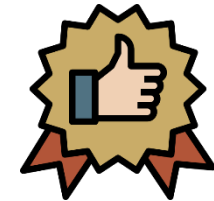
Partición de  
conjunto de  
datos



Definición de métricas



Estandarización



Selección de variables

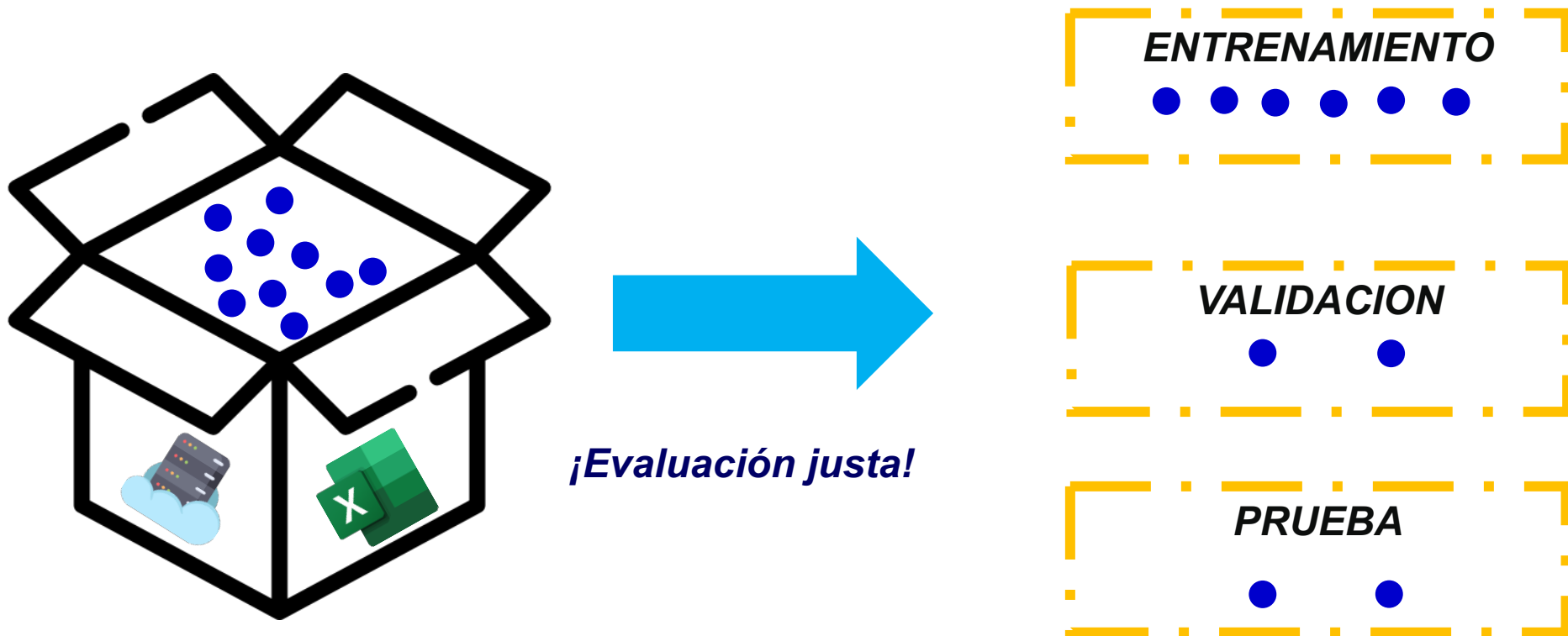


Ajuste y selección  
de modelos



Evaluación de modelos

# Particionamiento del Conjunto de Datos



# Técnicas supervisadas: Regresión

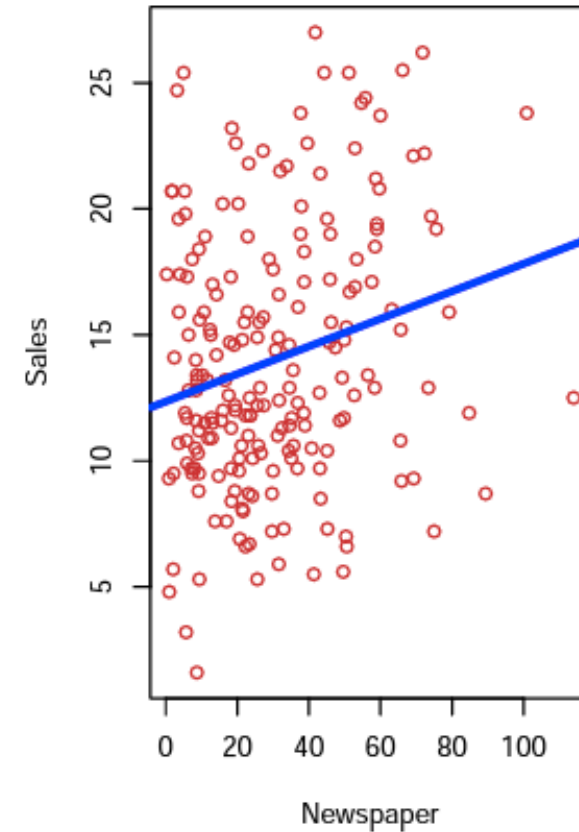
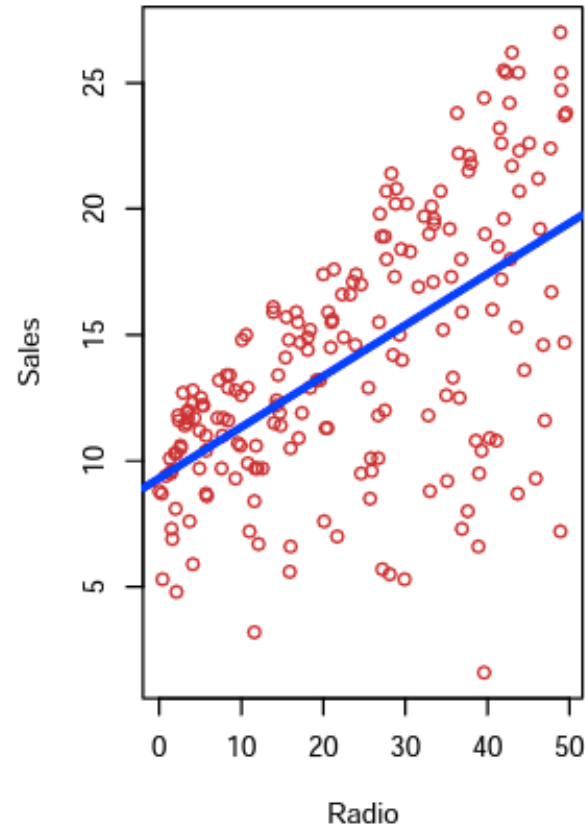
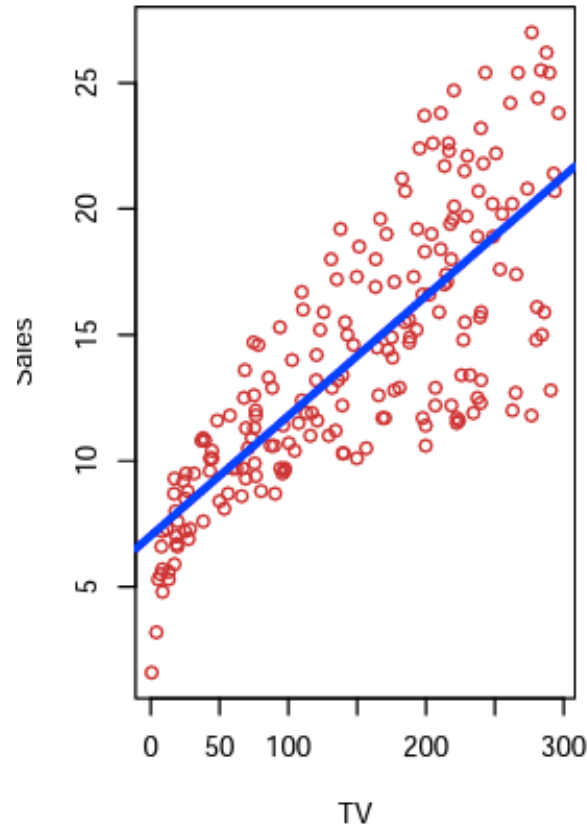
# ¿Qué es Aprendizaje Estadístico?

*Busca lidiar con el problema de la inferencia estadística, buscando encontrar la función predictiva basado en la información disponible*

## **Considera aspectos como:**

- ✓ Significancia estadística de parámetros
- ✓ Busca ajustes (en algunos casos) basados en niveles de confianza
- ✓ Analiza patrones distribucionales en la información

# ¿Qué es Aprendizaje Estadístico?



*¿Será posible predecir las ventas utilizando esas 3 variables?*



# ¿Qué es Aprendizaje Estadístico?

***Acá debemos identificar:***

- ✓ Cual es la variable respuesta u objetivo que se desea predecir
- ✓ Determinar las variables predictoras disponibles (Vector X)

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

$$Y = f(X) + \boxed{\epsilon}$$

*Captura medidas de  
error e información no  
explicada*

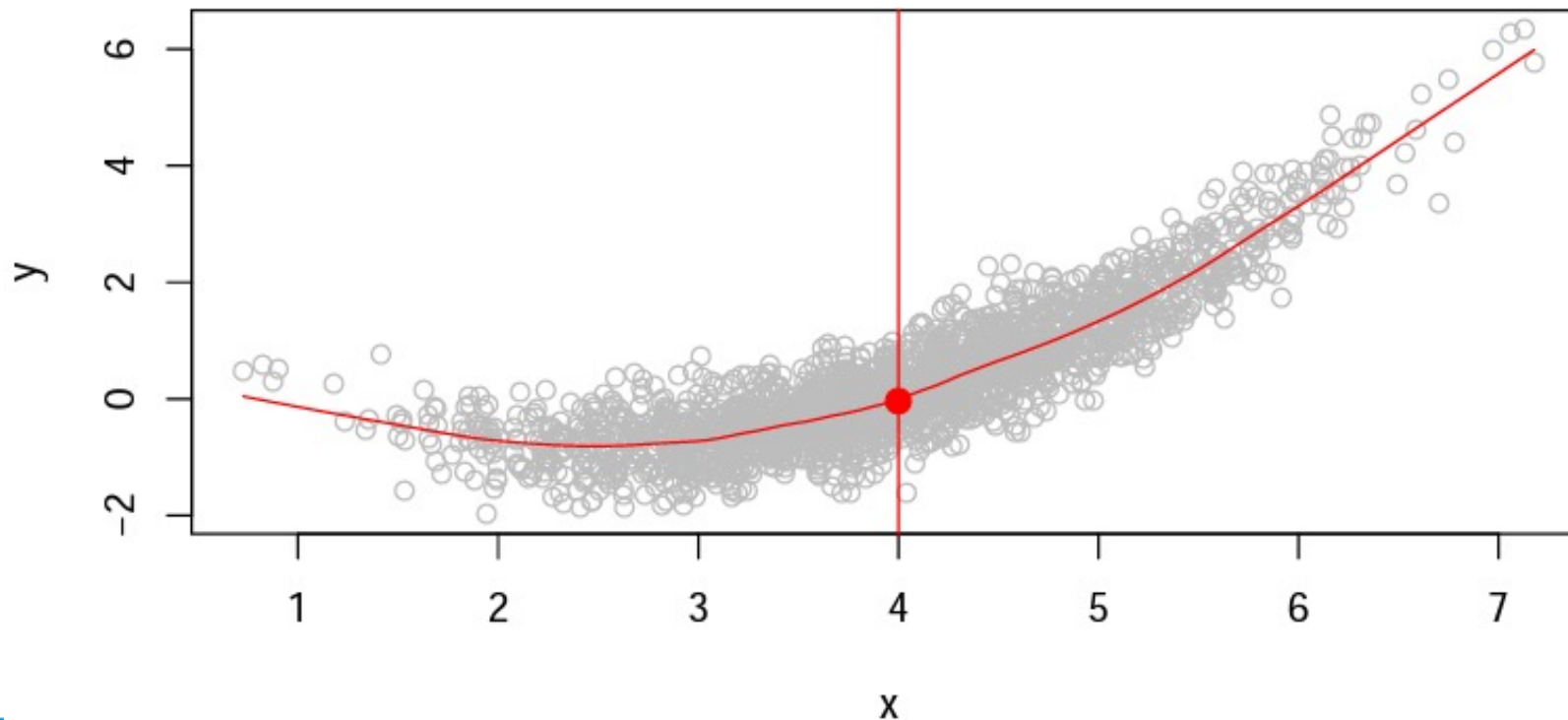
# ¿Para qué estimar $f$ ?

- ✓ Podemos realizar predicciones para la  $Y$  en nuevos puntos de  $X$
- ✓ Podemos entender qué componentes de  $X$  son los más importantes para explicar  $Y$ , además de cuales son poco relevantes
- ✓ Dependiendo de la complejidad de  $f$ , se puede entender como cada componente de  $X$  afecta la  $Y$

# ¿Para qué estimar $f$ ?

¿Existirá alguna función  $f(X)$  ideal para estos datos?

$$f(x) = E[Y \mid X = x]$$



# ¿Para qué estimar $f$ ?

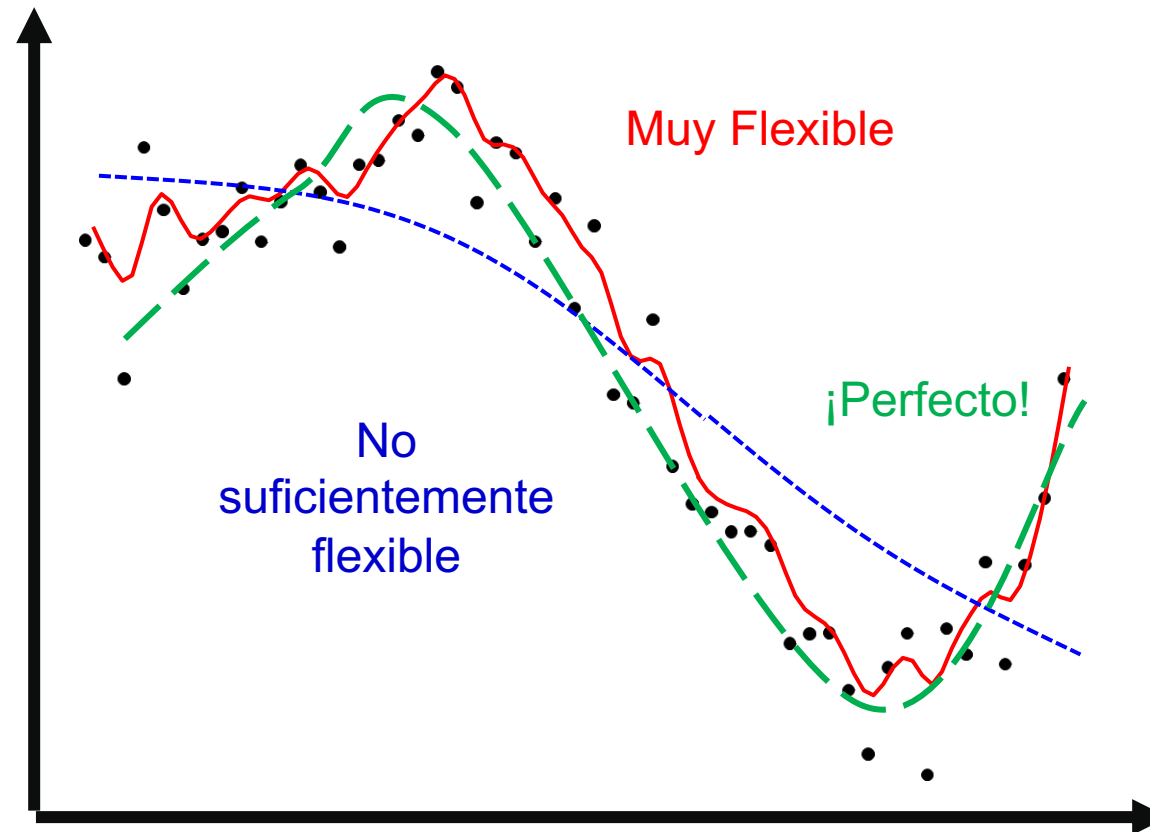
$$f(x) = E[Y | X = x]$$

## **Consideramos que**

- ✓  $f(x)$  es el predictor óptimo de  $Y$  cuando se desea minimizar el error cuadrático medio
- ✓  $\epsilon = Y - f(x)$  es el error irreducible, incluso si conociéramos el valor verdadero de  $f$
- ✓ Podemos representar el error cuadrático medio como:

$$E[(Y - \hat{f}(X))^2 | X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}$$

# Complejidad del modelo



# Métricas de desempeño: Mod. de regresión

$$error_j = y_{real_j} - y_{pred_j}$$

Error cuadrático medio (**RMSE**)

$$\sqrt{\frac{1}{N} \sum_{j=1}^N error_j^2}$$

Error medio absoluto (**MAE**)

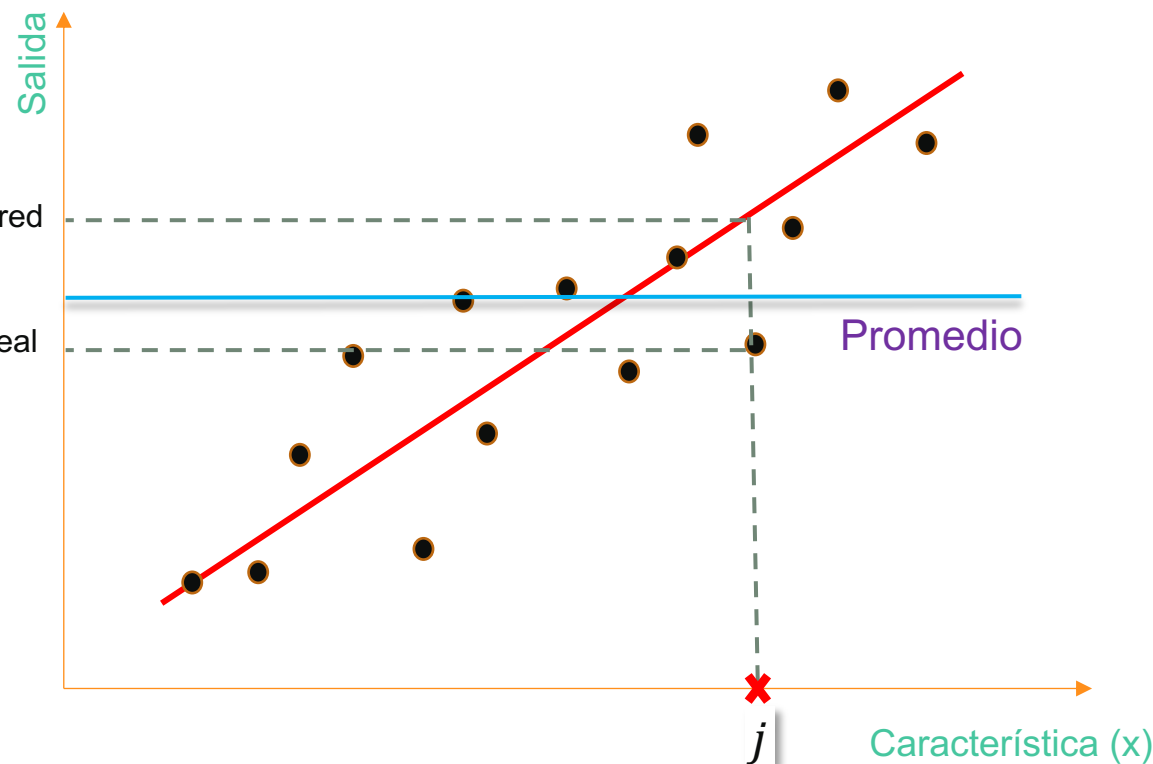
$$\frac{1}{N} \sum_{j=1}^N |error_j|$$

Error cuadrático medio (**MSE**)

$$\frac{1}{N} \sum_{j=1}^N error_j^2$$

Error medio absoluto (**MAPE**)

$$\frac{1}{N} \sum_{j=1}^N \left| \frac{error_j}{y_{real_j}} \right|$$



R cuadrado (**R<sup>2</sup>**)

$$1 - \frac{\sum_{j=1}^N error_j^2}{\sum_{j=1}^N (y_j - \bar{y}_j)^2}$$

# Noción de distancia

Sea  $X$  un conjunto no vacío, definimos una función de distancia  $d: X \times X \rightarrow \mathbb{R}$  que cumple:

i.  $d(x, y) \geq 0$

ii.  $d(x, y) = 0 \leftrightarrow x = y$

iii.  $d(x, y) = d(y, x)$

iv.  $d(x, z) \leq d(x, y) + d(y, z)$

**Dist. Euclidea**

$$d(X, Y) = \|X - Y\|_2$$

**Dist. Manhattan**

$$d(X, Y) = \|X - Y\|_1$$

**Dist. Chebyshev**

$$d(X, Y) = \|X - Y\|_\infty$$

# Distancia de Mahalanobis

$$d(X, Y) = \sqrt{(X - Y)^T \Sigma^{-1} (X - Y)}$$

- ✓ Dentro del cálculo de la Distancia, considera la matriz de varianzas y covarianzas
- ✓ Tiene en cuenta la correlación/estructura entre las variables

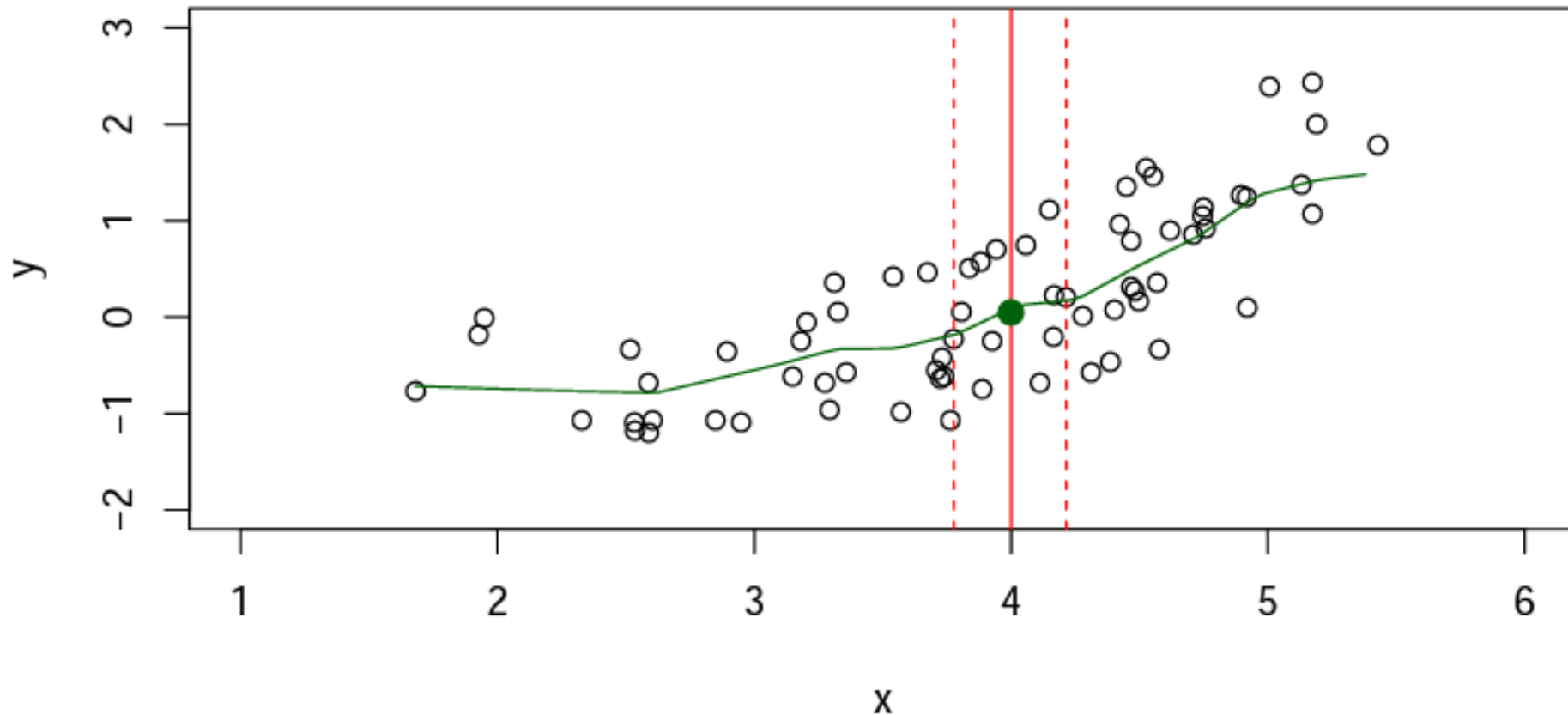
*Las distancias calculadas utilizando Mahalanobis, siguen un comportamiento Chi-Cuadrado con  $p$  grados de libertad. ( $p$  asociado a la cantidad de variables)*



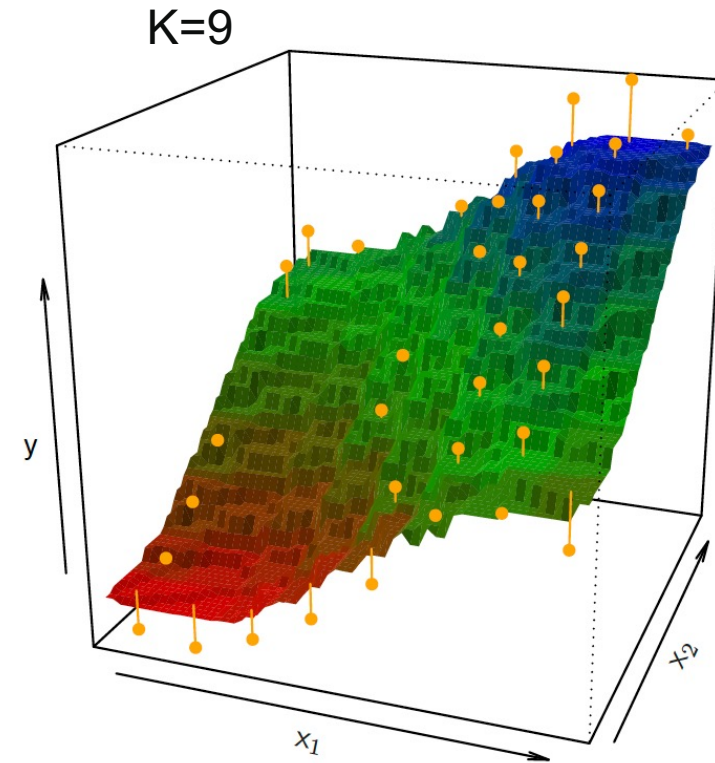
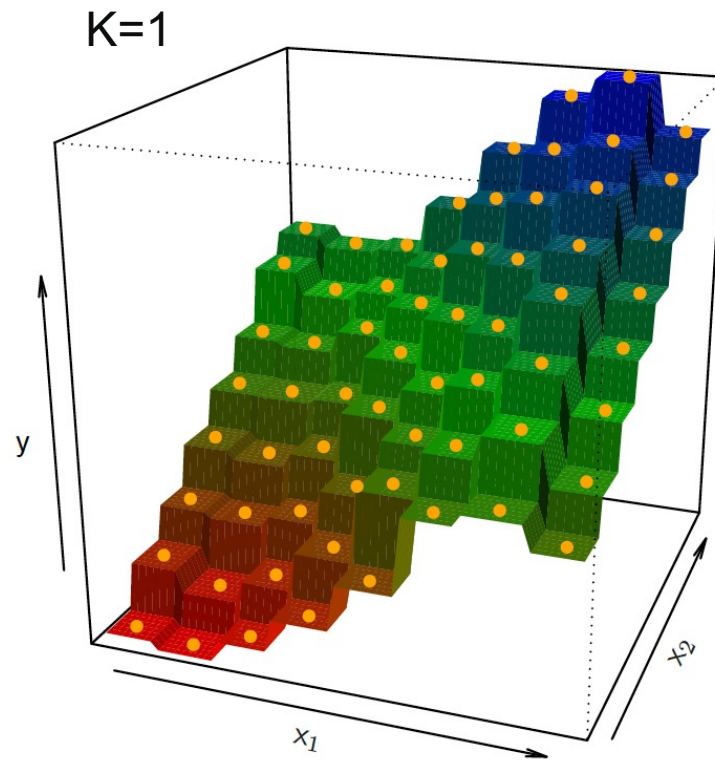
# ¿Cómo estimar $f$ ?

$$\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x))$$

*Versión relajada de la  
esperanza*



# K- Nearest Neighbors (Regresión)



$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

# K- Nearest Neighbors (Regresión)

- ✓ Es un método No Paramétrico, ya que no asume de antemano como es el comportamiento de  $f$
- ✓ Valores grandes de  $K$ , generan un comportamiento más suavizado, pero con menor ajuste

# Regresión Lineal Simple

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Diagram illustrating the components of the Simple Linear Regression equation:

- $Y$ : Variable Dependiente (Dependent Variable)
- $\beta_0$  and  $\beta_1$ : Coeficientes (Coefficients)
- $X$ : Regresor (Regressor)
- $\varepsilon$ : Error term

$$Error = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \longrightarrow \text{Error Cuadrático Medio (ECM)}$$

**Estimación de los Coeficientes:**

$$\hat{\beta}_1 = \frac{COV(X, Y)}{VAR(X)} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

**Medida de Ajuste:**

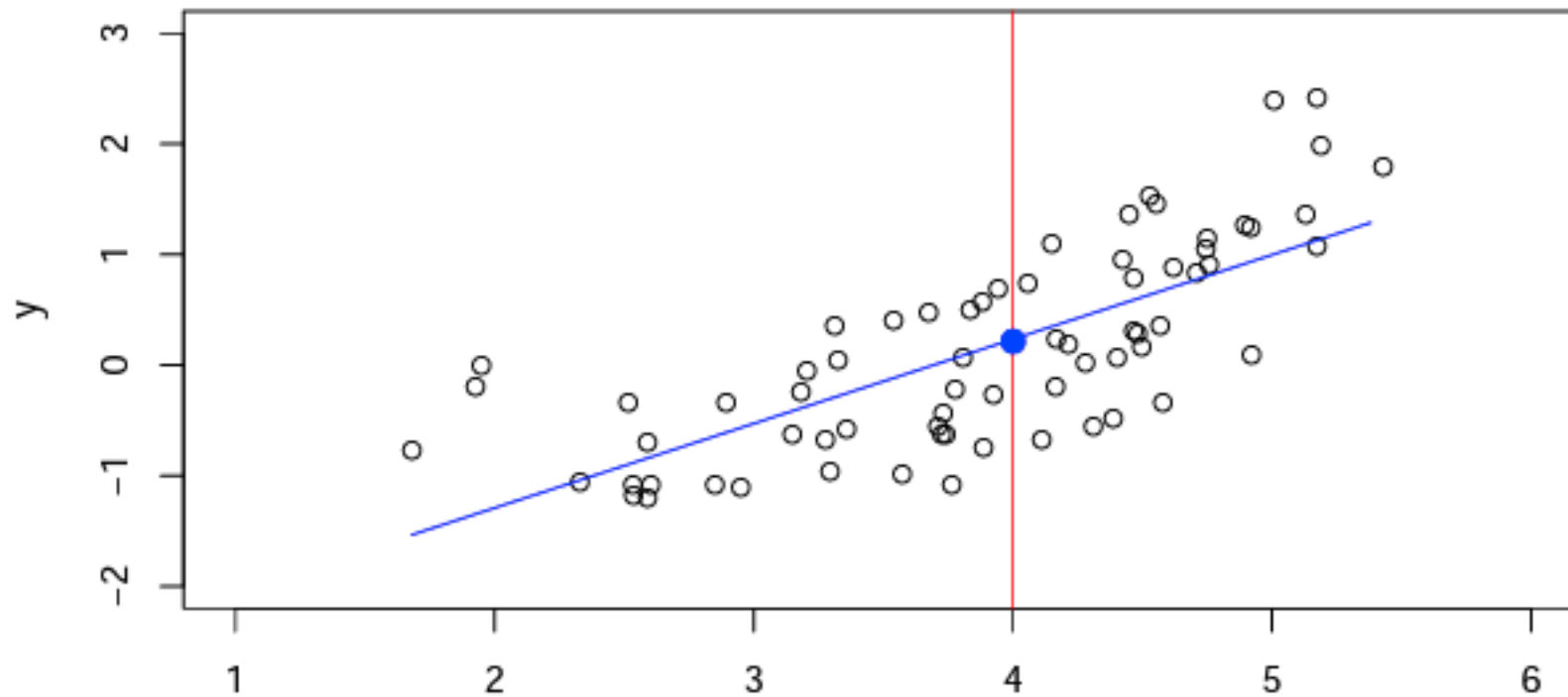
$$R^2$$

Este índice corresponde al *porcentaje de variabilidad* que puede ser explicada una variable en términos de las otras

# Regresión lineal

*¿Qué tal un ajuste lineal?*

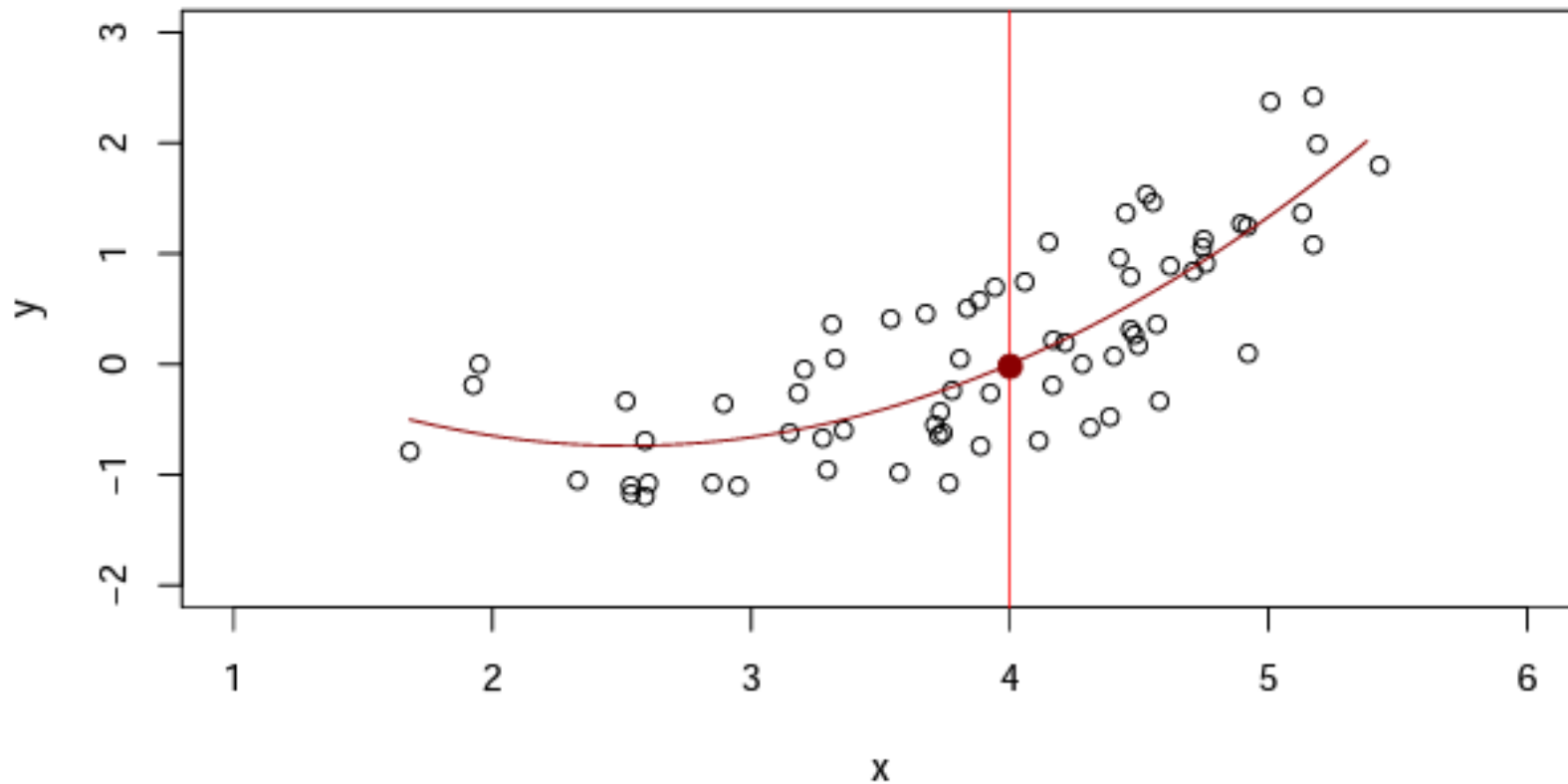
$$f(x) = \hat{\beta}_0 + \hat{\beta}_1 X_1$$



# Regresión lineal

*¿Qué tal un ajuste cuadrático?*

$$f(x) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_1^2$$



# Regresión lineal

## *Ejercicio – Ajuste Polinómico*

# Regresión Lineal Multiple

$$Y = X\beta + \varepsilon$$

Busca minimizar el error cuadrático medio  $(Y - X\beta)'(Y - \beta X)$

$$\text{Error} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \longrightarrow \text{Error Cuadrático Medio (ECM)}$$

$$\hat{\beta} = (X'X)^{-1}X'Y \longrightarrow \Sigma_{xx}^{-1}\Sigma_{xy}$$



# Regresión Lineal

Dep. Variable:	Peso	R-squared (uncentered):	0.991
Model:	OLS	Adj. R-squared (uncentered):	0.987
Method:	Least Squares	F-statistic:	298.4
Date:	Mon, 07 Feb 2022	Prob (F-statistic):	8.18e-19
Time:	23:07:51	Log-Likelihood:	-88.182
No. Observations:	27	AIC:	190.4
Df Residuals:	20	BIC:	199.4
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
sexo	4.5551	5.564	0.819	0.423	-7.052	16.162
Estatura	-0.4681	0.538	-0.870	0.395	-1.591	0.654
Long_pie	1.6018	1.670	0.959	0.349	-1.881	5.084
Long_brazo	0.3297	0.715	0.461	0.650	-1.161	1.820
Anchura_espalda	1.7685	0.744	2.376	0.028	0.216	3.321
Diam_craneo	-0.8285	0.801	-1.034	0.313	-2.500	0.843
Long_rodilla_tobillo	0.4787	0.989	0.484	0.634	-1.585	2.542

# Supuestos del modelo de regresión

1. No existe una relación lineal entre las variables explicativas  $X_j$ ,  $j = 1, \dots, k$ . (Ausencia de multicolinealidad)
2. Las  $x_{ij}$  son números fijos o realizaciones de las v.a.'s  $X_j$ ,  $j = 1, \dots, k$ ; las cuales son independientes de los términos de error  $\varepsilon_i$ ;  $i = 1, \dots, n$ .
3. El valor esperado de la v.a.  $Y$  es una función lineal de las variables independientes  $X_j$ ,  $j = 1, \dots, k$ .
4. Los términos de error son v.a. que siguen una distribución normal, tienen media cero y la misma varianza  $\sigma^2$  (supuesto de homocedasticidad o varianza uniforme):
$$E[\varepsilon_i] = 0 \quad \text{y} \quad \text{Var}[\varepsilon_i] = \sigma^2; \quad i = 1, \dots, n$$
5. Los términos de error aleatorios,  $\varepsilon_i$ , no están correlacionados entre sí, por lo que  $E[\varepsilon_i, \varepsilon_j] = 0$ ;  $\forall i \neq j$ .

# Propiedades del estimador de Mínimos Cuadrados

- La recta estimada pasa por el punto  $(\bar{x}_1, \bar{x}_2, \dots, \bar{y})$ .

- La suma de los residuos mínimo-cuadráticos es cero:

$$\sum_{i=1}^n e_i$$

- Los residuos mínimo-cuadráticos están incorrelacionados con las variables explicativas:

$$\sum_{i=1}^n x_{ik} e_i = 0 ; \forall j = 1, \dots, k$$

- La suma de los productos cruzados entre los valores ajustados (predichos) y los residuos es igual a cero:

$$\sum_{i=1}^n \hat{y}_i e_i = 0 ; \forall j = 1, \dots, k$$

# Inferencia sobre los coeficientes

1. Se desea probar:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0 \text{ vs. } H_1: \text{al menos un } \beta_j \neq 0; j = 1, 2, \dots, k$$

2. Estadístico de prueba : Se obtiene completando la siguiente tabla de análisis de varianza(ANAVA)

Fuente de variación	Suma de Cuadrados	Grados de libertad	Cuadrados medios	Estadístico $F_C$
Modelo	$SCR = SCT - SCE$	$k$	$CMR = \frac{SCR}{k}$	$\frac{CMR}{CME}$
Error	$SCE = \sum_{i=1}^n e_i^2$	$n - (k + 1)$	$CME = \frac{SCE}{n - (k + 1)}$	
Total	$SCT = \sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$		

# Inferencia sobre los coeficientes

3. Se rechaza  $H_0$  si  $F_C > F_{(\alpha, k, n-(k+1))}$
4. Conclusión:
  - Si no se rechaza  $H_0$ , se puede afirmar que no hay una relación lineal útil entre la variable  $Y$  y cualquiera de las  $k$  variables explicativas, con un nivel de significancia  $\alpha$
  - Si se rechaza  $H_0$ , se puede afirmar que al menos una de las  $k$  variables explicativas está relacionada linealmente con  $Y$ , con un nivel de significancia  $\alpha$ .



# Inferencia sobre los coeficientes

Consiste en realizar las pruebas de hipótesis para  $\beta_j, j = 1, 2, \dots, k$ :

## Prueba de hipótesis para $\beta_j, j = 1, 2, \dots, k$

1. Se desea probar:  $H_0: \beta_j = 0$  vs.  $H_1: \beta_j \neq 0$ .
2. Estadístico de prueba :  $t_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$   
donde  $s_{\hat{\beta}_j}$  es el error estándar (desviación estándar) del estimador  $\hat{\beta}_j$ .
3. Se rechaza  $H_0$  si  $|t_j| > t_{\left(\frac{\alpha}{2}, n-(k+1)\right)}$

- Si  $X_j$  influye en el modelo, es decir si  $\beta_j \neq 0$ , entonces debe estar en él.

**Nota:** Otra forma es construyendo el I.C. al  $(1 - \alpha)100\%$  para  $\beta_j, j = 1, 2, \dots, k$

$$\hat{\beta}_j \pm t_{\left(\frac{\alpha}{2}, n-(k+1)\right)} s_{\hat{\beta}_j}$$

Y verificando que el IC construido no contenga al cero.

# Inferencia sobre los coeficientes

3.2 Intervalo de confianza al  $(1 - \alpha)100\%$  para  $\mu_{Y|x_1^*, \dots, x_k^*} = E[Y|x_1^*, \dots, x_k^*]$

$$\hat{y} \pm t\left(\frac{\alpha}{2}, n-(k+1)\right) s_{\hat{Y}}$$

donde  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \dots + \hat{\beta}_k x_k^*$  es el valor calculado (estimado) del estadístico  $\hat{Y}$ , y  $s_{\hat{Y}}$  es el error estándar del estimador  $\hat{Y}$ .

# Interpretación de los coeficientes

Variable dependiente	Regresor $X_j$	$\beta_j$	Interpretación
Niveles	Niveles	Efecto mg en $Y$ ante un cambio unitario en $X_j$	$Y$ aumenta o disminuye $\beta_j$ veces cuando aumenta $X_j$
Logaritmo	Logaritmo	Elasticidad $X_j$ de $Y$	Un incremento en 1 % en $X_j$ genera un incremento o disminución de $\beta_j$ % en $Y$
Logaritmo	Niveles	Tasas de crecimiento o retorno	Un incremento en una unidad de $X_j$ genera un incremento o disminución en $\beta_j * 100$ % en $Y$
Niveles	Logaritmo	Respuesta de $Y$ ante una variación de $X_j$	Un incremento en 1 % en $X_j$ genera un incremento o disminución en $\beta_j / 100$ en $Y$



# Interpretación de los coeficientes

Source	SS	df	MS	Number of
Model	1179.73204	1	1179.73204	F(1, 524)
Residual	5980.68225	524	11.4135158	Prob > F
Total	7160.41429	525	13.6388844	R-squared
				Adj R-squa
				Root MSE

\* Un año adicional de educación hace que el salario por hora aumente 54 centávos US\$ por hora

\* Debido al caracter lineal, cada año adicional de educación hace que el salario aumente en una misma cantidad, independiente del nivel inicial de educación

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.5413593	.053248	10.17	0.000	.4367534	.6459651
_cons	-.9048516	.6849678	-1.32	0.187	-2.250472	.4407687

# Interpretación de los coeficientes

Source	SS	df	MS	Number of obs	=	526
Model	27.5606288	1	27.5606288	F(1, 524)		
Residual	120.769123	524	.230475425	Prob > F		
Total	148.329751	525	.28253286	R-squared		
				Adj R-squa		
				Root MSE		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0827444	.0075667	10.94	0.000	.0678796	.0976091
_cons	.5837727	.0973358	6.00	0.000	.3925563	.7749891

\* El salario por hora aumenta 8.3 % por cada año adicional de educación

\* En este caso, el salario aumenta en un porcentaje constante

*¿Cómo puedo encontrar la mejor combinación de datos?*



Evaluando todas las combinaciones y mantener la de mejor resultado

*¿Si tengo  $N$  variables, cuantas posibles combinaciones hay?*




$N!$

*¿Es factible evaluarlas todas?*



No

# Selección de variables

- 
- ✓ Realizar regresión con todas las variables
  - ✓ Eliminar la que tenga el P-valor más alto
  - ✓ Volver a realizar la regresión

# Ventajas y Desventajas del modelo de regresión

## Ventajas:

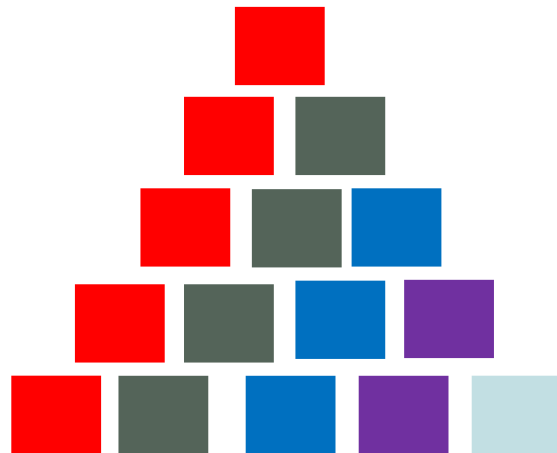
- ✓ Interpretabilidad
- ✓ Simple y fácil de usar

## Desventajas:

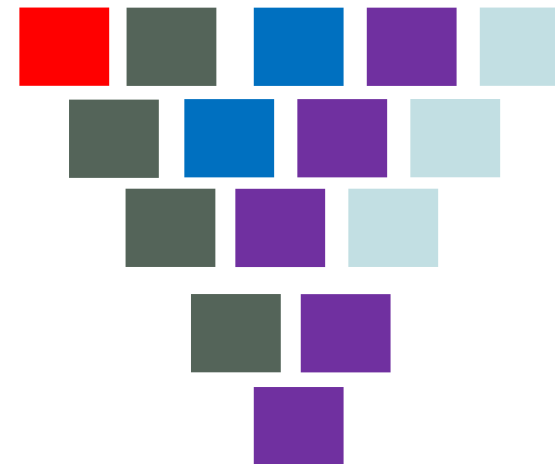
- ✓ Alta varianza cuando hay muchas variables
- ✓ La presencia de colinealidad afecta el modelo
- ✓ No se puede usar cuando hay mas variables que registros
- ✓ Requiere trabajo extra para seleccionar variables

# Selección de Variables

## *Forward Selection*



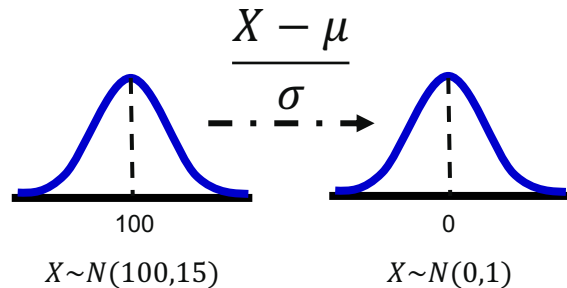
## *Backward Selection*



# Estandarización de los datos - Importancia

- ✓ Ayuda a evitar sesgos en la construcción de modelos
- ✓ Las operaciones que se realizan ocupan menos memoria
- ✓ Reduce/evita errores computacionales
- ✓ Permite obtener análisis que dependan del comportamiento estadístico de las variables libres de una unidad de medida

# Estandarización de los datos



*Normalización*

$$\frac{X - \mu}{\sigma} \equiv \frac{X(\text{unidad}) - \mu(\text{unidad})}{\sigma(\text{unidad})} \longrightarrow \text{Adimensional}$$

*Tipificación*

$$0 \leq \frac{X - X_{\min}}{X_{\max} - X_{\min}} \leq 1$$

*Escalamiento Min-Max*

$$\frac{X}{X_{\max}} \leq 1$$

*Escalamiento con Max*

$$\frac{X}{X_{\text{prom}}}$$

*Escalamiento con Promedio*



# Estandarización de los datos

## Regresión Lineal:

Ayuda a ver la variable que más impacta al modelo

### *Modelo con variables sin estandarizar*

$$\text{ValorSeguro} = 500 + 0.5\text{ingresos} + 20\text{edad} - 200\#\text{hijos}$$

### *Modelo con variables estandarizadas*

$$\text{ValorSeguro} = 5 + 0.8\text{ingresos} + 1.2\text{edad} - 0.5\#\text{hijos}$$

## PCA:

Evita sesgos en la selección de componentes.

Considere variables como: Ingresos, edad, # de hijos

Mayor variabilidad

Menor variabilidad

¡Debido a las unidades de medida!

# Regresión lineal con penalización

La regresión tradicional busca minimizar:

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

La regresión penalizada busca minimizar:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \text{Penalty}(\beta)$$

# Regresión lineal con penalización

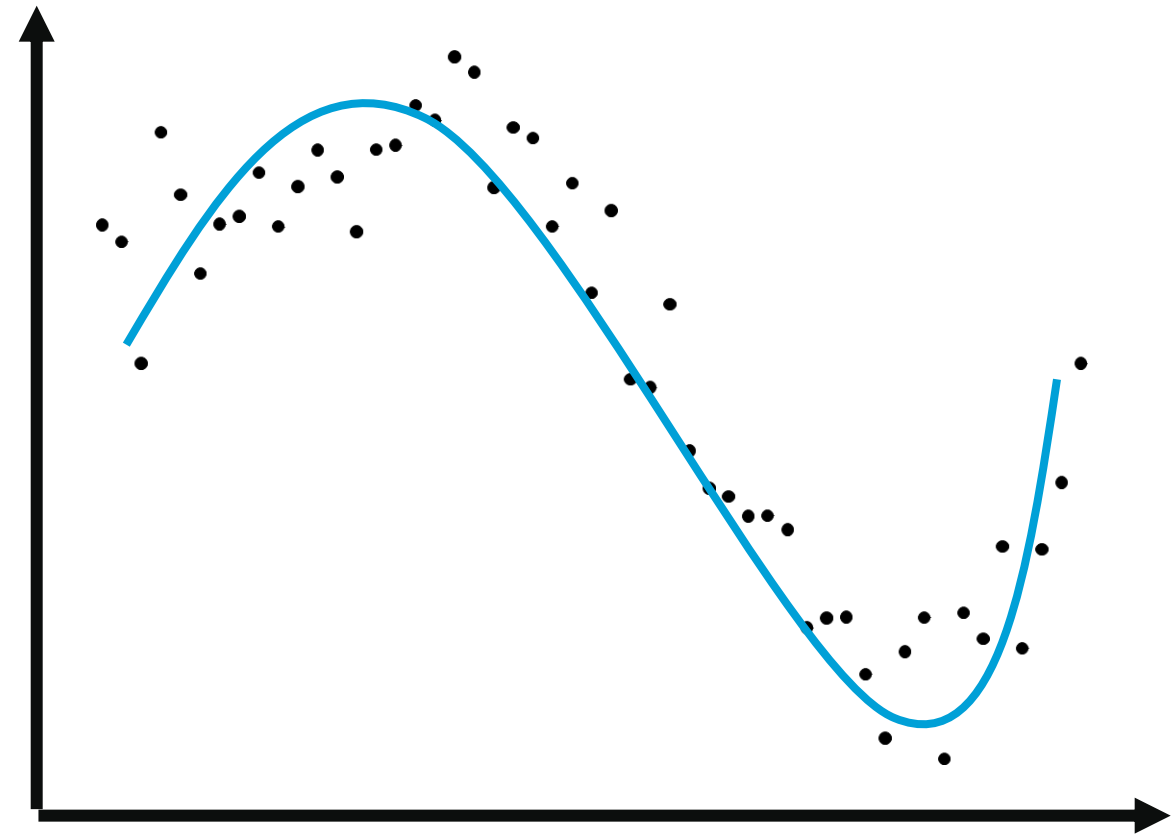
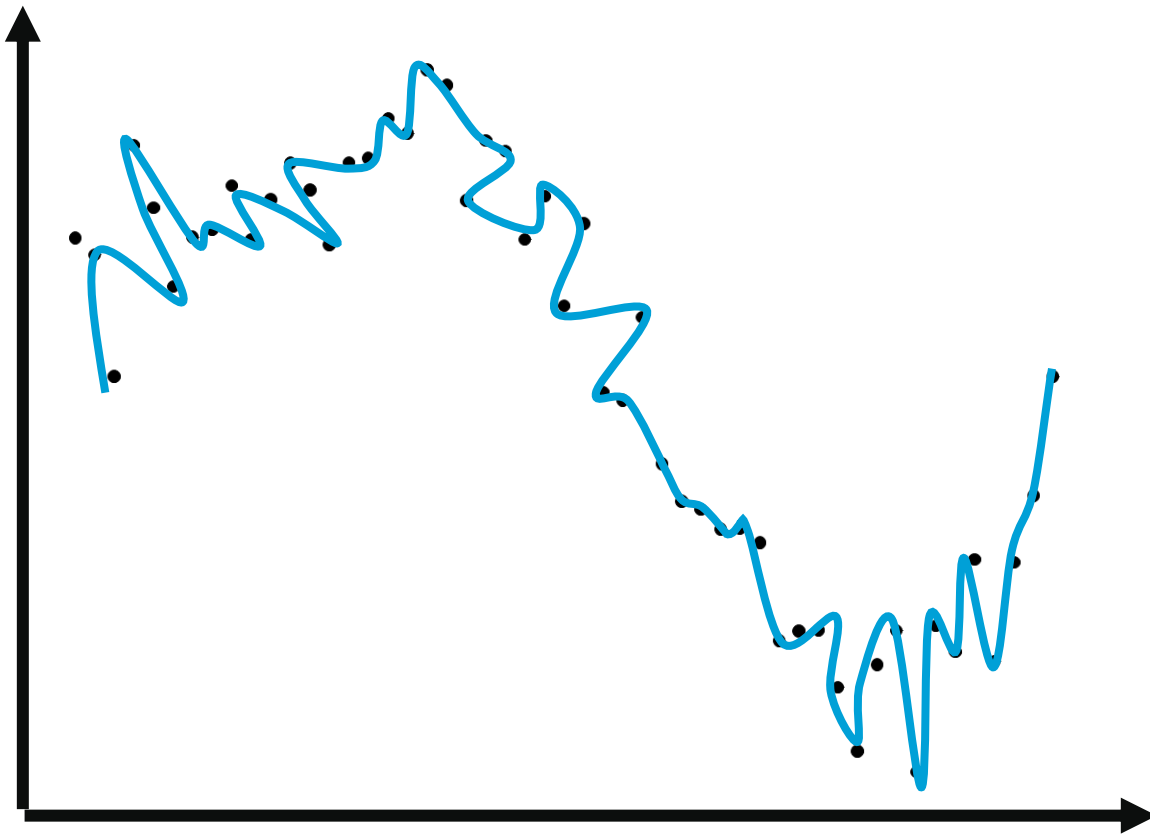
## *Opciones de penalización*

$$\text{Ridge: Penalty}(\beta) = \sum_{j=1}^p \beta_j^2.$$

$$\text{Lasso: Penalty}(\beta) = \sum_{j=1}^p |\beta_j|.$$

$$\text{Naive elastic net: Penalty}(\beta) = \alpha \sum_{j=1}^p \beta_j^2 + (1 - \alpha) \sum_{j=1}^p |\beta_j|.$$

# Regresión lineal con penalización



# Regresión Ridge

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{n-1} X_{n-1} + \beta_n X_n$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_{n-1} X_{n-1} + \hat{\beta}_n X_n$$

$$Error = \sum_{i=1}^m (y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^n \beta_j^2$$

# Regresión Ridge

- ✓  $\lambda \geq 0$  actua como un parametro a ser tuneado
- ✓ Cuando  $\lambda = 0$ , la regresion ridge es equivalente a la de minimos cuadrados
- ✓ Mientras mas grande sea  $\lambda$ , los coeficientes seran llevados a 0 (esto es el encogimiento)
- ✓  $\lambda$  se puede encontrar usando validacion cruzada
- ✓ La estandarizacion es importante ya que se aplica el mismo factor de penalización a las variables

$$\hat{\beta}^{\text{ridge}} = (A^T A + \lambda I)^{-1} A^T y$$

# Regresión Lasso

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{n-1} X_{n-1} + \beta_n X_n$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_{n-1} X_{n-1} + \hat{\beta}_n X_n$$

$$Error = \sum_{i=1}^m (y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^n |\beta_j|$$

# Regresión Lasso

- ✓  $\lambda \geq 0$  juega un rol similar que en la regresión Ridge
- ✓ Cuando  $\lambda = 0$ , la regresión ridge es equivalente a la de mínimos cuadrados
- ✓ Permite realizar selección de variables



# Regresión Ridge VS Lasso

- ✓ Lasso lleva a coeficientes a ser más dispersos (varios ceros), mientras que Ridge lleva los coeficientes a ser más densos (no ceros)
- ✓ Lasso se utiliza para selección de variables, lo que lleva a:
  - ✓ Interpretabilidad
  - ✓ Eficiencia computacional en las predicciones
- ✓ Es más "sencillo" resolver una regresión Ridge que una Lasso
- ✓ La regresión lasso tiene un buen desempeño solo si un número pequeño de predictores es necesario

# Elastic Net

- ✓ Es una combinación entre Ridge y Lasso
- ✓ El parametro alfa controla el mix entre ridge y lasso
- ✓ El parámetro lambda juega un papel similar a la regresión ridge y lasso
- ✓ Es un problema de optimización convexa

$$\text{Penalty}(\beta) = \alpha \sum_{j=1}^p \beta_j^2 + (1 - \alpha) \sum_{j=1}^p |\beta_j|.$$

**Inspira  
Crea  
Transforma**

[www.eafit.edu.co](http://www.eafit.edu.co)

VIGILADA | MINEPUCACIÓN

**UNIVERSIDAD  
EAFIT**