Pontificia Universidad JAVERIANA Bogotá

Técnicas Predictivas para la Generación de Nuevo Conocimiento

ING. HERIBERTO FELIZZOLA JIMENEZ

Educación **Continua**

Generamos experiencias educativas

Contenido

1. Introducción al Aprendizaje de Máquina
2. Modelo de Regresión
    a) Introducción
    b) Conceptos Básicos
    c) Mínimos Cuadrados
    d) Interpretación de los Modelos de Regresión
    e) Análisis de las predicciones
    f) Potenciales Problemas
    g) Análisis de Errores
    h) Variables Categóricas
    i) Componentes No Lineales
3. Taller Grupal

Pontificia Universidad JAVERIANA Bogotá

1. Introducción al Aprendizaje de Máquina

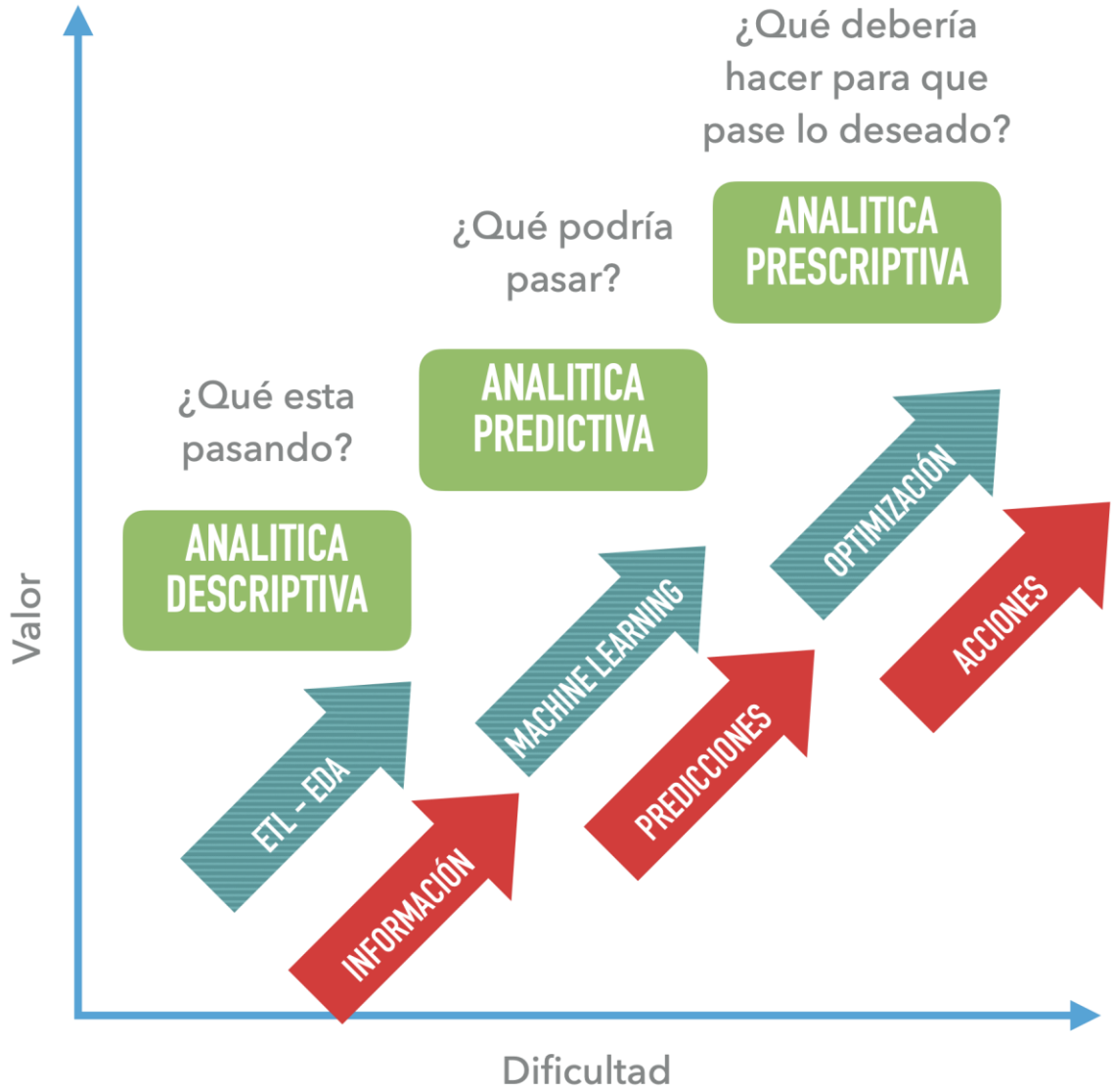
ING. HERIBERTO FELIZZOLA JIMENEZ

Educación **Continua**

Generamos experiencias educativas

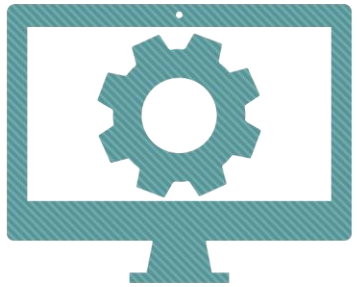
Analítica Predictiva

- Es el arte de construir modelos que permiten realizar predicciones a partir de los patrones encontrados en los datos.
- Algunas aplicaciones de las predicciones son:
  - Precios de productos, servicios y mercado de valores
  - Recomendación de productos
  - Sistemas de recomendación
  - Riesgo en proyectos
  - Diagnostico de enfermedades
  - Clasificación de imágenes, texto y video
  - Condiciones meteorológicas



Aprendizaje de Máquina

Datos de Entrenamiento

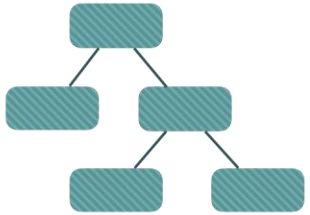


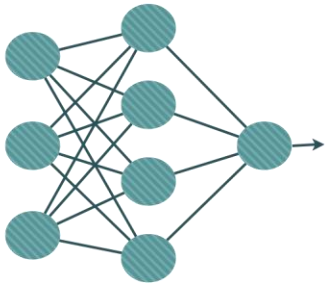
Algoritmos y Modelos de Aprendizaje de Maquina

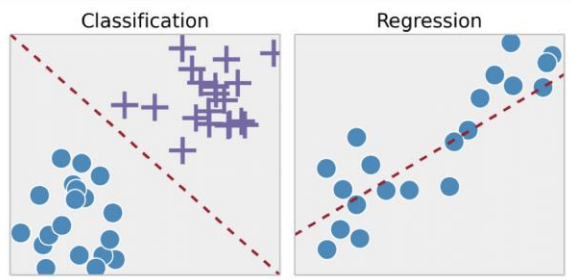
Predicciones

X1	X2	...	Xp	Y
X11	X12	...	X1p	Y1
X21	X22	...	X2p	Y3
...	...	...	...	...
Xn1	Xn2	...	Xnp	Yn

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$



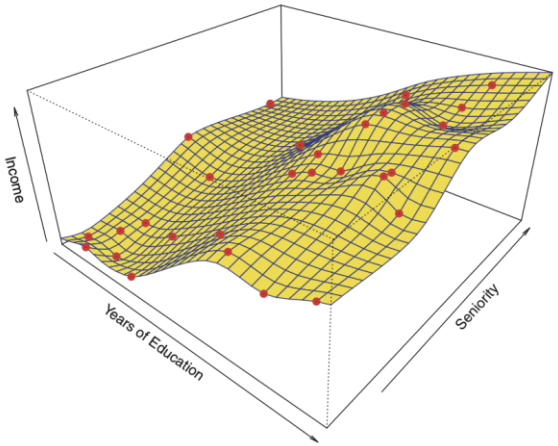




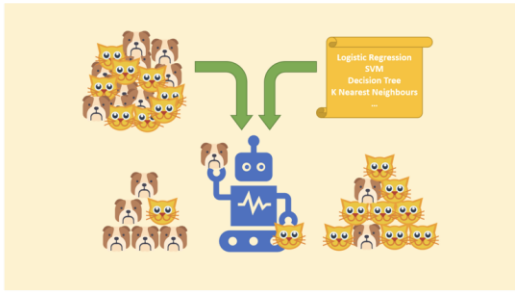
Tareas en Aprendizaje Automático

APRENDIZAJE SUPERVISADO

REGRESIÓN

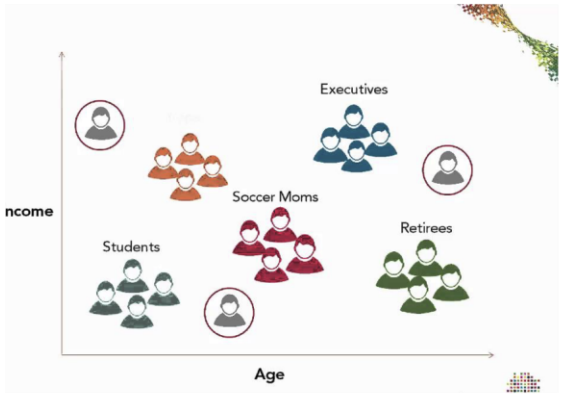


CLASIFICACIÓN

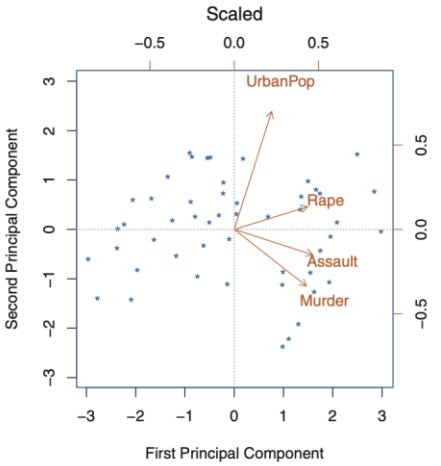


APRENDIZAJE NO SUPERVISADO

CLUSTERING



REDUCCIÓN DE DIMENSIONES



Instancia - Atributo - Salida

- La entrada a un esquema de aprendizaje automático es un conjunto de instancias.
- Estas instancias son las cosas que deben clasificarse, asociarse o agruparse.
- El conjunto de características de cada instancia es denominado atributos o variables de entrada (X).
- El resultado que se quiere predecir es denominado salida (Y).

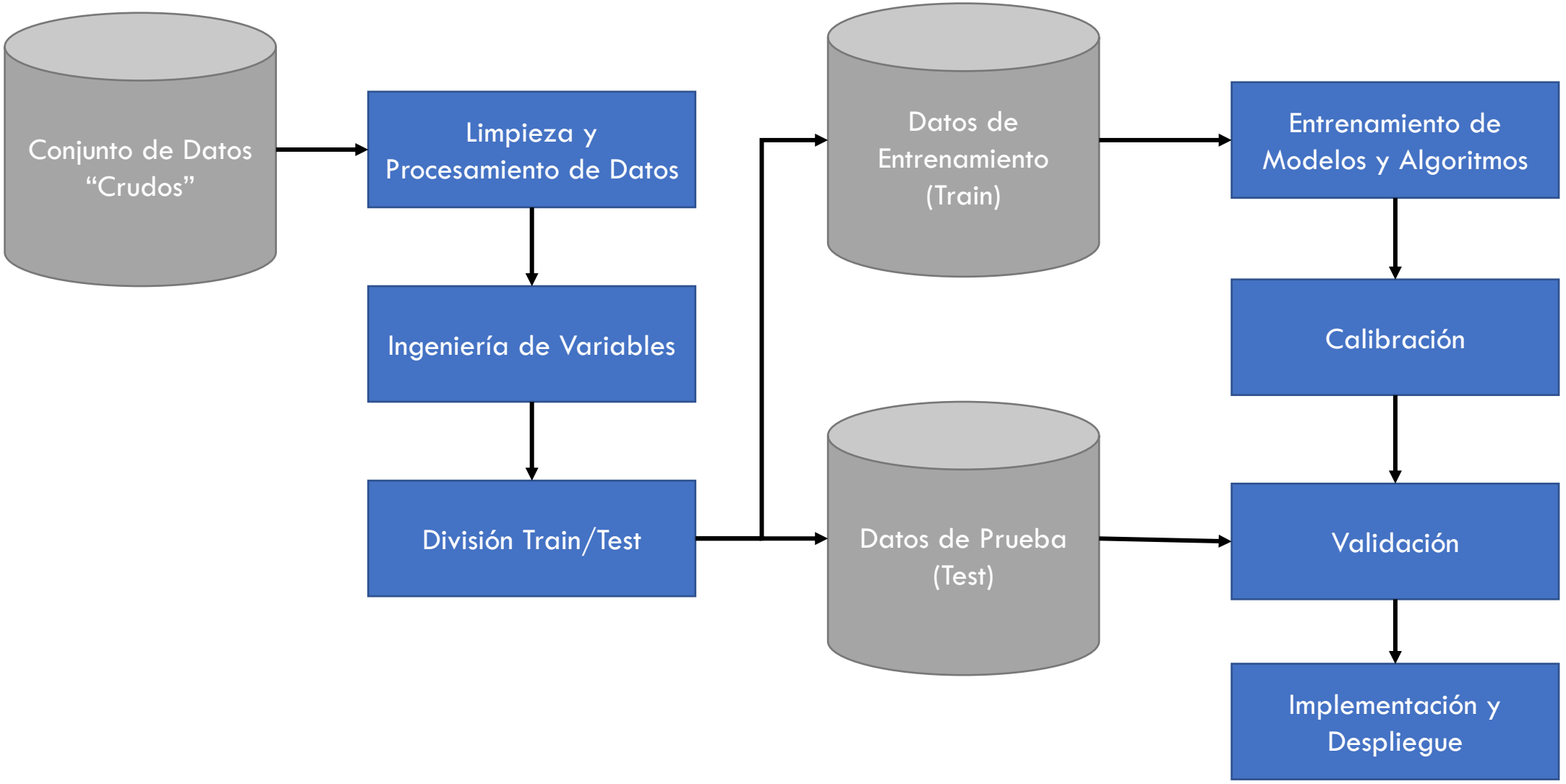
Y: Salida

X: Atributos - Variables de Entrada

Instancias

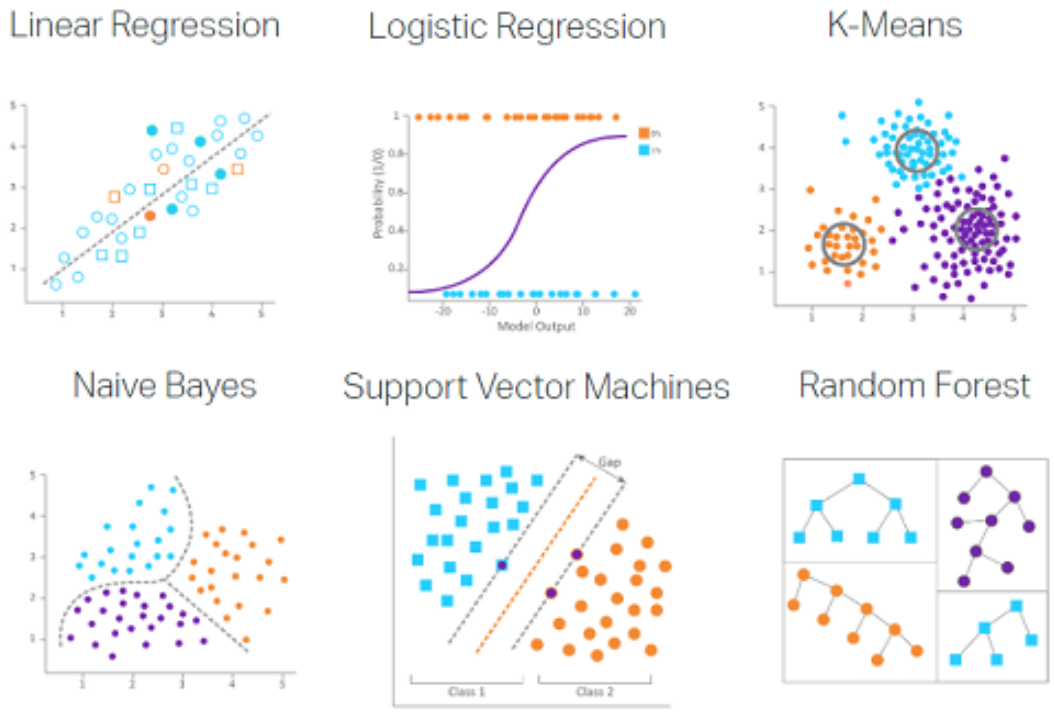
Outlook	Temperature	Humidity	Windy	Play Time
Sunny	85	85	false	5
Sunny	80	90	true	0
Overcast	83	86	false	55
Rainy	70	96	false	40
Rainy	68	80	false	65
Rainy	65	70	true	45
Overcast	64	65	true	60
Sunny	72	95	false	0
Sunny	69	70	false	70
Rainy	75	80	false	45
Sunny	75	70	true	50
Overcast	72	90	true	55
Overcast	81	75	false	75
Rainy	71	91	true	10

Proceso de Entrenamieto



Modelos y Algoritmos

- Aprendizaje supervisado:
  - Modelos lineales: Mínimos Cuadrados, Análisis discriminante, Regresión Logística.
  - Modelos de penalización: Ridge, Lasso, ElasticNet
  - Modelos no lineales: Splines, Regresión Local, Modelos Generales Aditivos (GAM), Kernels.
  - Arboles: CART, C5.0, Bagging, Random Rorest, Boosting
  - Máquinas de soporte vectorial SVM.
  - Naive-Bayes
  - Redes neuronales.
- Aprendizaje no supervisado:
  - Algoritmos de clustering: K-mean, Modelos Jerarquicos
  - Reducción de dimensiones: Componentes principales - PCA



Pontificia Universidad JAVERIANA Bogotá

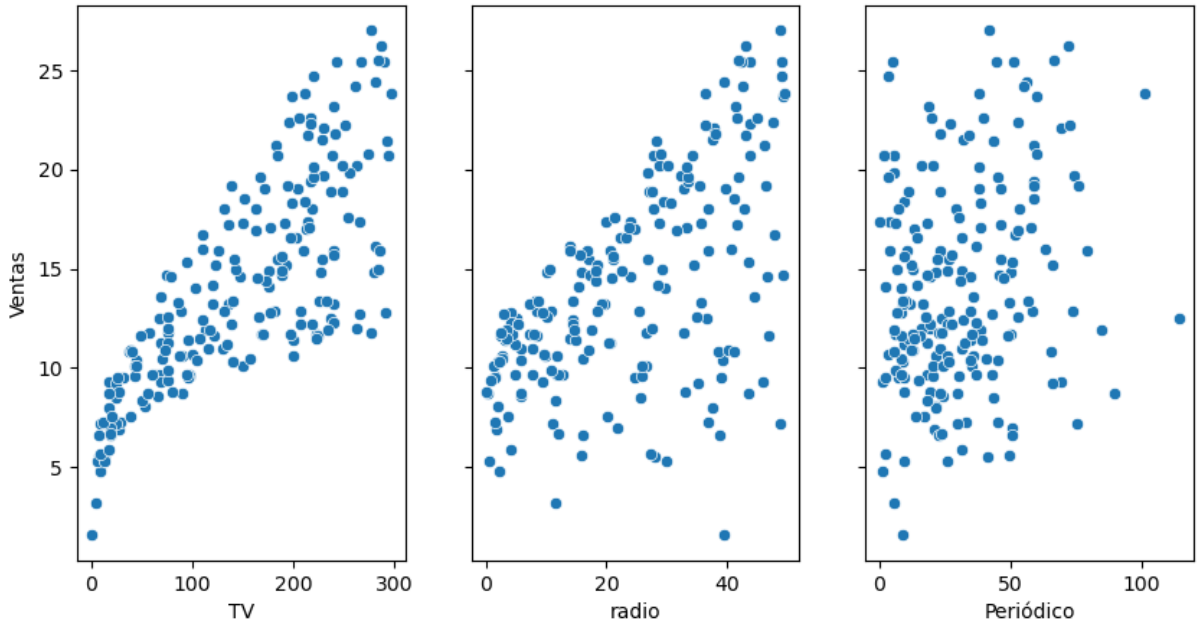
2.1 Regresión Lineal

Educación **Continua**

Generamos experiencias educativas

Ejemplo

- Escenario: una empresa de marketing quiere conocer la asociación entre el gasto en publicidad y las ventas de un producto.
- Objetivo: Determinar si existe una asociación entre el gasto en publicidad y las ventas.
- Beneficios: Si se determina una asociación, se pueden ajustar los presupuestos publicitarios para mejorar las ventas.
- Entregable: Modelo de regresión que permita analizar y predecir las ventas sobre utilizando los tres presupuestos de medios.



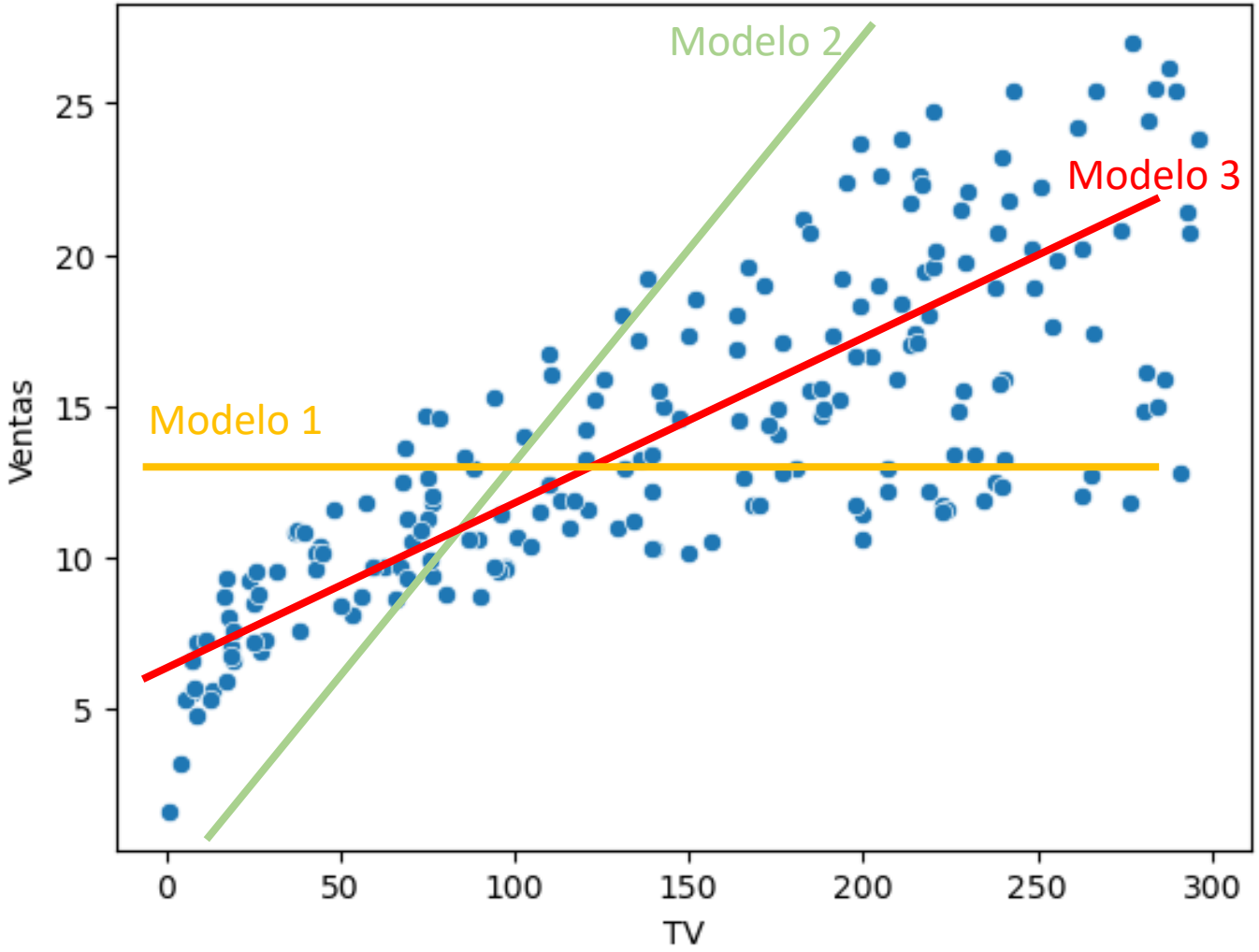
Preguntas Claves

Con el modelo de regresión buscamos resolver las siguientes preguntas:

- ¿Existe una relación entre el presupuesto publicitario y las ventas?
- ¿Qué medios contribuyen a las ventas?
- ¿Qué contribución hace cada medio para el comportamiento de las ventas?
- ¿Con qué precisión podemos predecir las ventas futuras?
- ¿Existe correlación entre los medios publicitarios?

Regresión Lineal

- La regresión lineal es una técnica para encontrar la línea que mejor representa la relación entre dos variables.
- No es posible encontrar una línea que capture de forma perfecta todos los puntos, siempre habrá algún error.
- El error de un modelo de regresión es la diferencia entre los datos reales y la línea de tendencia central.
- No existe una predicción perfecta, siempre habrá algún error.



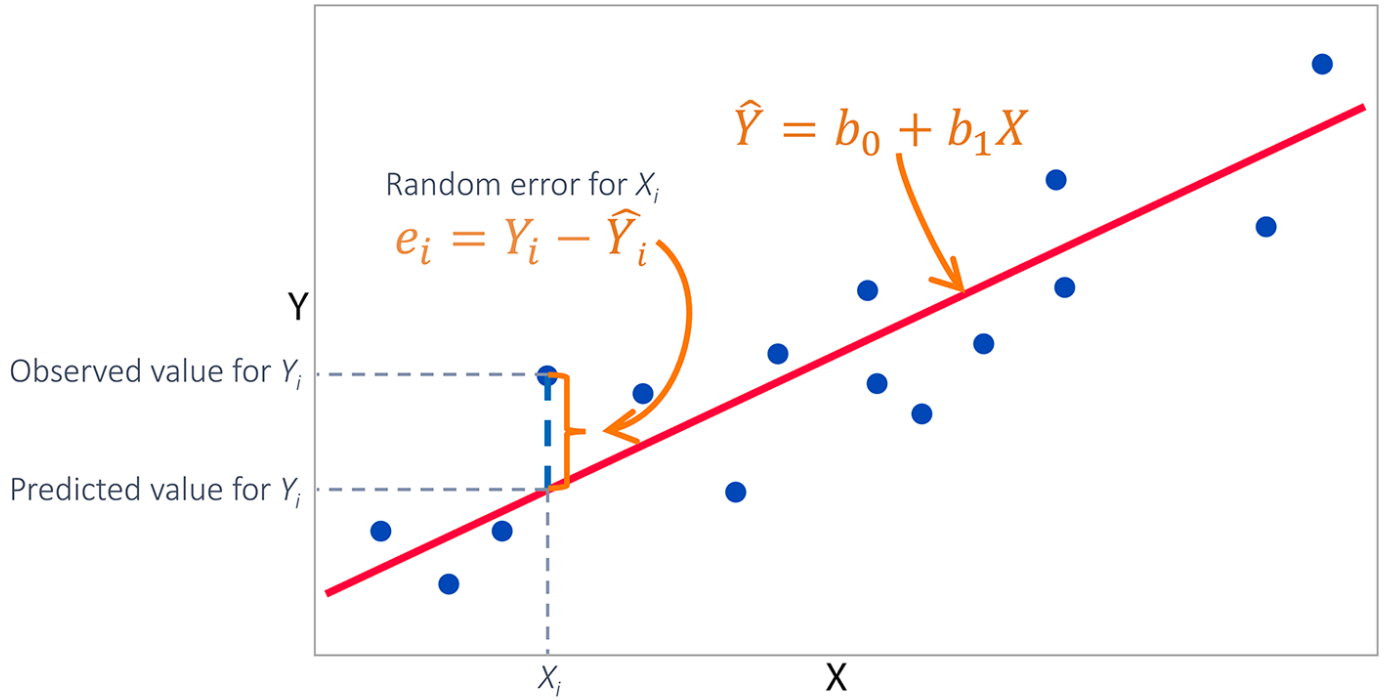
Modelo de Regresión

- Objetivo o respuesta: es la variable que deseamos predecir, en este caso las ventas. Generalmente nos referimos a la respuesta como la variable *Y*.
- Entrada o predictor: son las variables independientes o controlables que utilizamos para predecir la respuesta, nos referimos a estas como X's, este caso los presupuestos de:
  - TV (X_1)
  - Radio (X_2)
  - Periódico (X_3)
- Modelo de regresión: es una función que relaciona la variable de respuesta con los predictores, generalmente se escribe como:$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Donde, β_0 es el intercepto
- $\beta_1, \beta_2\ y\ \beta_3$ son las contribuciones individuales (cuando las demás permanecen constantes) de cada variable de entrada a la predicción de la respuesta Y.
- Error (ϵ): es la diferencia entre el valor real de Y y su predicción $\hat{Y}$, por tanto

$$\rightarrow \epsilon = Y - \hat{Y}$$

Mínimos Cuadrados

- Construir un modelo de regresión implica encontrar las contribuciones de las variables TV, Radio y Periódico $(\beta_1, \beta_2\ y\ \beta_3)$ que permiten predecir las ventas con el mínimo error.
- Para estimar las constantes de la regresión se utiliza el método de mínimos cuadrados (Least Square).
- El método busca elegir los valores para $\beta_0, \beta_1, \beta_2\ y\ \beta_3$ que minimicen la suma de cuadrados de los errores ϵ.
- En otras palabras:



$$\text{Min RSS} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

MÍNIMOS CUADRADOS

La solución al problema de optimización planteado con el método de mínimos cuadrados es:

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

Ventas

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

TV Radio Periódico

Contribuciones de las variables

SALIDA EN PYTHON

Pontificia Universidad JAVERIANA Bogotá

OLS Regression Results

Dep. Variable:	ventas	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Sat, 23 Oct 2021	Prob (F-statistic):	1.58e-96
Time:	19:37:49	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

Interpretación de los Parámetros

¿Qué contribución hace cada medio para la comportamiento de las ventas?

- Las contribuciones de cada medio están expresadas en sus coeficientes (*coef*):
  - TV: por cada peso que se gasta en TV las ventas aumentan 0.0458.
  - Radio: por cada peso que se gasta en Radio las ventas aumentan 0.1885.
  - Periódico: por cada peso que se gasta en periódicos las ventas disminuyen 0.0010
- Para cada contribución también se presentan intervalos de confianza del 95%.

OLS Regression Results

Dep. Variable:	ventas	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Sat, 23 Oct 2021	Prob (F-statistic):	1.58e-96
Time:	19:37:49	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

Intervalos de Confianza

	coef	std err	t	P>|t|	[0.025	0.975]
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

El modelo para predecir las ventas tiene la forma:

$$Ventas = 2.939 + 0.046TV + 0.189Radio - 0.0010Periodico$$

Inferencia de los Parámetros

¿Qué medios contribuyen a las ventas?

- Para responder a esta pregunta, podemos examinar los valores p (P > |t|) asociados con el estadístico t de cada predictor.
- Para cada variable del modelo se plantea la siguiente hipótesis:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

- Los valores p para televisión y radio son bajos ≈ 0, pero el valor p para periódicos (0.86) no lo es.

OLS Regression Results

Dep. Variable:	ventas	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Sat, 23 Oct 2021	Prob (F-statistic):	1.58e-96
Time:	19:37:49	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

Esto sugiere que solo la televisión y la radio están relacionadas con las ventas. En cambio, el presupuesto para periódico no afecta significativamente en las ventas.

Significancia del Modelo

¿Existe una relación entre el presupuesto publicitario y las ventas?

- Permite determinar si existe una relación lineal entre la variable de respuesta y las variables de entrada.
- Se evalúa a través del valor P de la regresión, que se encuentra en la tabla como Prob (F-statistic).

$$H_0 = \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$Ha: \beta_j \neq 0, \textit{para al menos una } j$$

- Si el valor P es muy bajo, por lo general menor a 0.05, entonces podemos inferir que al menos una variable de respuesta genera un efecto significativo sobre la variable de respuesta.

OLS Regression Results

Dep. Variable:	ventas	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Sat, 23 Oct 2021	Prob (F-statistic):	1.58e-96
Time:	19:37:49	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

En este caso, el valor p correspondiente al estadístico F en la tabla es muy bajo, lo que indica una clara evidencia de una relación entre publicidad y ventas.

Coeficiente de Determinación

¿Con qué precisión podemos predecir las ventas futuras?

- La precisión se puede evaluar por la cantidad de variabilidad observada que es explicada por las variables de entrada, algunas de los indicadores utilizados son:
  - R^2 (R-squared): Indica que proporción de la variabilidad total es absorbida por las variables regresoras.
  - $R^2_{Ajustado}$ (Adj. R-squared): No siempre aumenta con la inclusión de nuevas variables, pero si la diferencia entre R^2 y $R^2_{Ajustado}$ es considerable, entonces existe el riesgo de haber agregado términos no significativos.
- Estos indicadores se encuentran entre 0 – 1, entre más cercano mejor la precisión del modelo.

OLS Regression Results

Dep. Variable:	ventas	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Sat, 23 Oct 2021	Prob (F-statistic):	1.58e-96
Time:	19:37:49	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

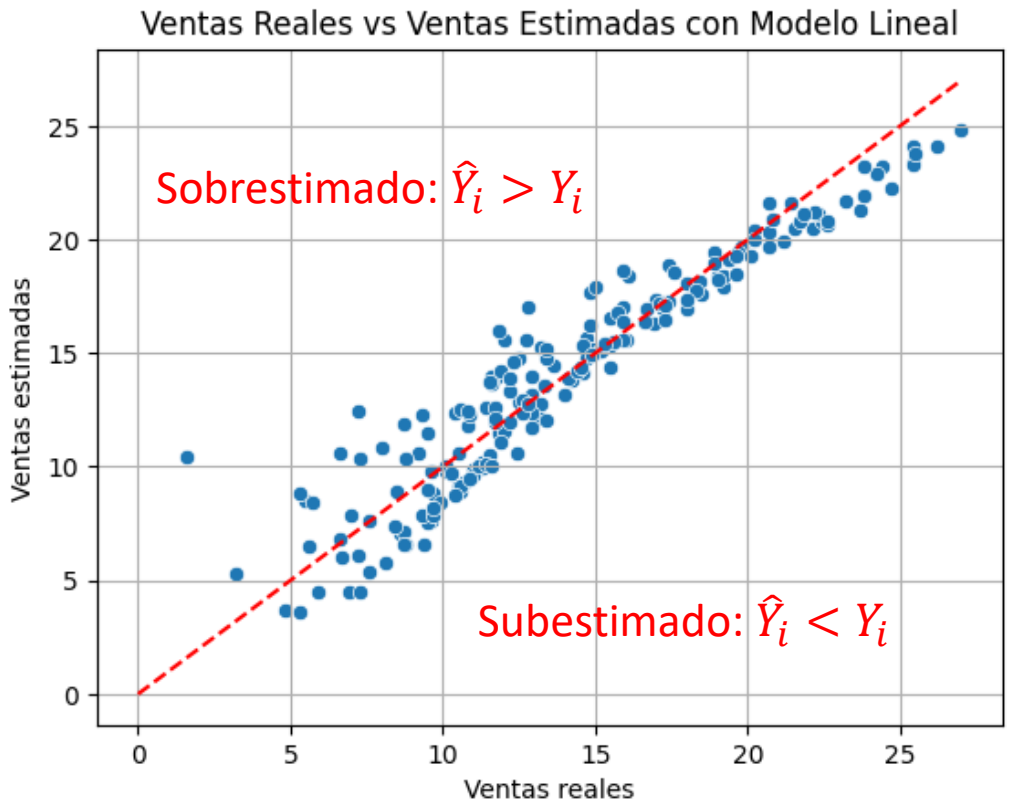
Los valores R^2 (0.897) y $R^2_{Ajustado}$ (0.896) indican una buena capacidad predictiva del modelo

Análisis de las Predicciones

En la columna diferencia se muestra la discrepancia entre las ventas y sus estimaciones, esto se denomina error o residual

Registro	Ventas Reales	Ventas Modelo Lineal	Diferencia
0	22.10	20.52	-1.58
1	10.40	12.34	1.94
2	9.30	12.31	3.01
3	18.50	17.60	-0.90
4	12.90	13.19	0.29
...	...	...	...
193	19.60	18.49	-1.11
194	17.30	16.50	-0.80
195	7.60	5.37	-2.23
196	9.70	8.17	-1.53
197	12.80	12.79	-0.01

La gráfica permite comparar las ventas reales vs las estimadas
Las predicciones "perfectas" se encuentran sobre la línea roja: ventas reales = estimadas



¿Cómo evaluar las predicciones?

- Para evaluar las predicciones se utilizan indicadores que comparan los valores reales de las ventas con las predicciones que realizan los modelos o algoritmos.
- Como base se toma las diferencias entre el valor real de las ventas ($Real_i$) y el valor estimado para las ventas con el modelo lineal ($Estimado_i$)
- A continuación, se listan algunas métricas utilizadas:
  - Error cuadrático medio MSE (Mean Square Error):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Real_i - Estimado_i)^2$$

  - Raíz del error cuadrático medio RMSE (Root Mean Square Error):

$$RMSE = \sqrt{MSE}$$

  - Error porcentual absoluto medio MAPE (mean absolute percentage error):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Real_i - Estimado_i}{Real_i}\right|$$

ESTIMACIÓN DE GASTOS MEDICOS (PARTE 1)

Pontificia Universidad JAVERIANA Bogotá

Para el caso de gastos médicos desarrollo las siguientes tareas:

1. Análisis exploratorio con variables numéricas
2. Modelo de regresión con variables numéricas
3. Análisis y evaluación de predicciones

Variables Categóricas

- Hasta el momento hemos asumido que todas las variables en nuestro modelo de regresión lineal son cuantitativas.
- Pero en la práctica algunos predictores son cualitativos.
- Si un predictor cualitativo (también conocido como factor), entonces incorporarlo a un modelo de regresión es muy simple, simplemente creamos un indicador o variable “dummy”

Factor

TV	radio	periodico	ventas	zona
230.1	37.8	69.2	22.1	sur
44.5	39.3	45.1	10.4	occidente
17.2	45.9	69.3	9.3	occidente
151.5	41.3	58.5	18.5	occidente
180.8	10.8	58.4	12.9	oriente
...	...	...	...	...
38.2	3.7	13.8	7.6	occidente
94.2	4.9	8.1	9.7	sur
177.0	9.3	6.4	12.8	occidente
283.6	42.0	66.2	25.5	norte
232.1	8.6	8.7	13.4	norte

Variables Dummies

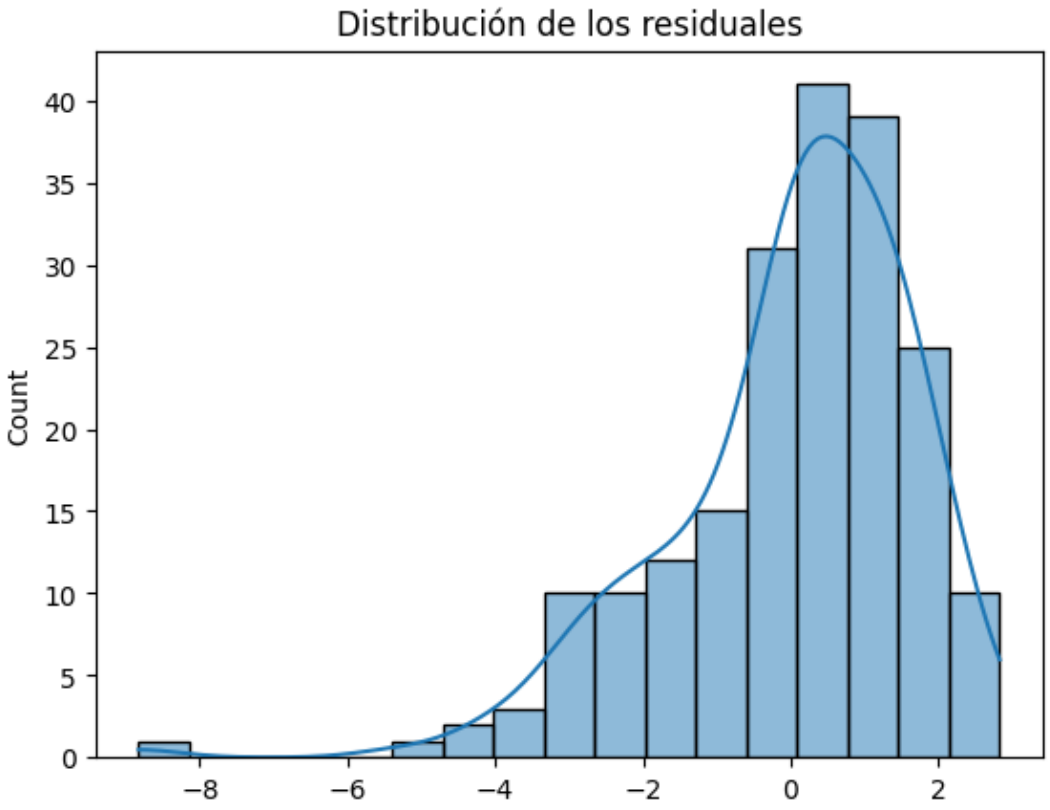
zona_norte	zona_occidente	zona_oriente	zona_sur
0	0	0	1
0	1	0	0
0	1	0	0
0	1	0	0
0	0	1	0
...	...	...	...
0	1	0	0
0	0	0	1
0	1	0	0
1	0	0	0
1	0	0	0

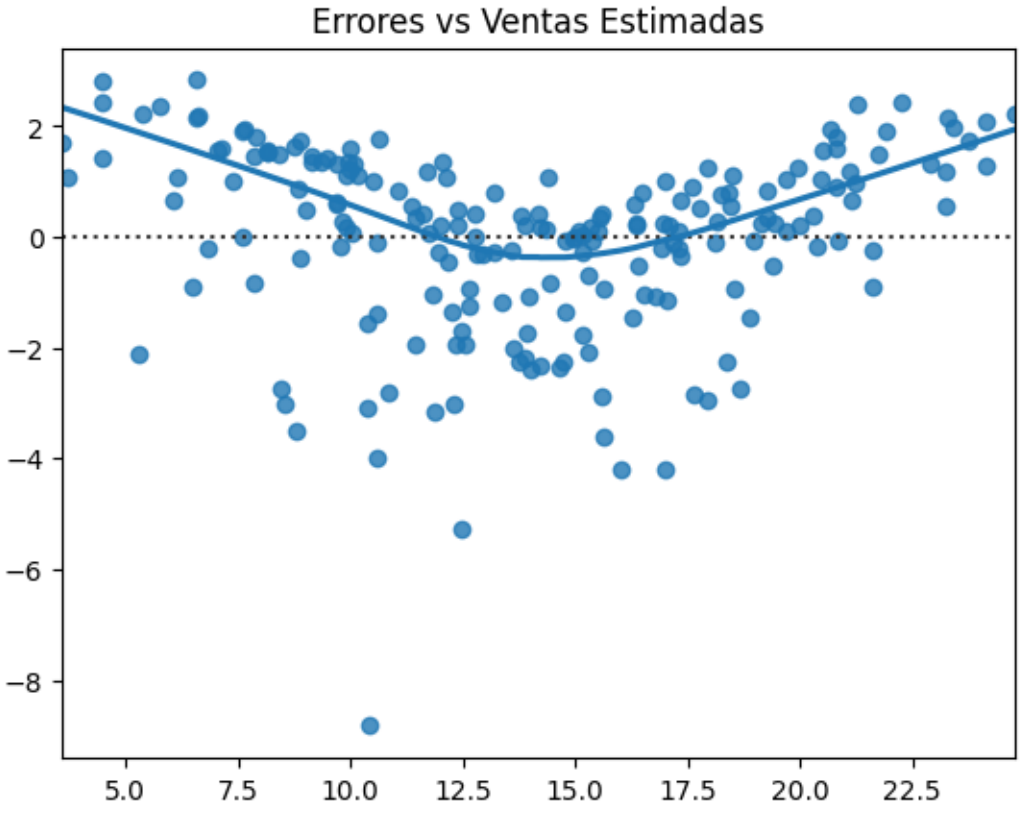
Potenciales Problemas

1. Falta de normalidad de los errores
2. Varianza no constante de los términos de error.
3. Valores atípicos.
4. Puntos de alto apalancamiento.
5. Multicolinealidad.
6. No linealidad de las relaciones respuesta-predictor.

ANÁLISIS DE LOS ERRORES

- En un modelo de regresión los errores ϵ_i deben distribuirse con media cero y varianza constante.
- Para verificar este se deben analizar las gráficas de normalidad y error vs predicción





En las gráficas de los errores se pueden observar algunos problemas tales como: sesgo (falta de normalidad), datos atípicos y heterocedasticidad (cambios en la varianza).

INTERPRETACIÓN DE LA REGRESIÓN

Pontificia Universidad JAVERIANA Bogotá

Prueba de Normalidad de los errores

- Para comprobar los problemas con la normalidad de los errores se pueden observar las 2 pruebas de normalidad:
  - Prob(Omnibus) con valor P ≈ 0,
  - Prob(JB) con valor P ≈ 1.44×10^{-33},
- Cuando el valor P < 0.05 indica que no se cumple con el supuesto de normalidad de los errores.
- Además, un sesgo (skew) de -1.327 confirma el sesgo (derecha) de la distribución de los errores.
- El Kurtosis > 3 indica también esa violación a la normalidad.

OLS Regression Results

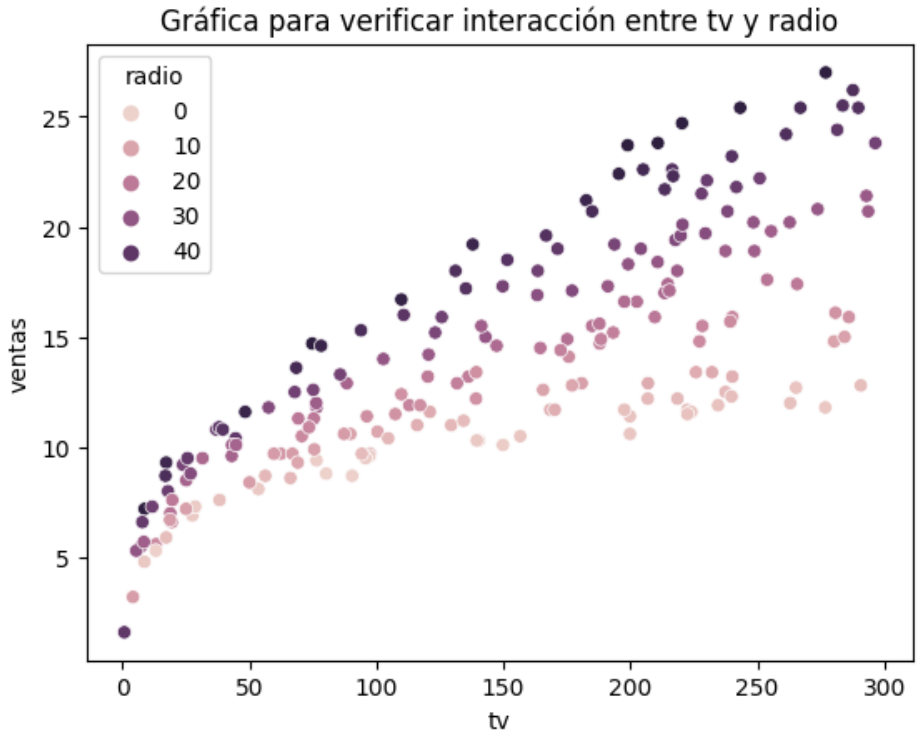
Dep. Variable:	ventas	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Sat, 23 Oct 2021	Prob (F-statistic):	1.58e-96
Time:	19:37:49	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
periodico	-0.0010	0.006	-0.177	0.860	-0.013	0.011
constante	2.9389	0.312	9.422	0.000	2.324	3.554

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

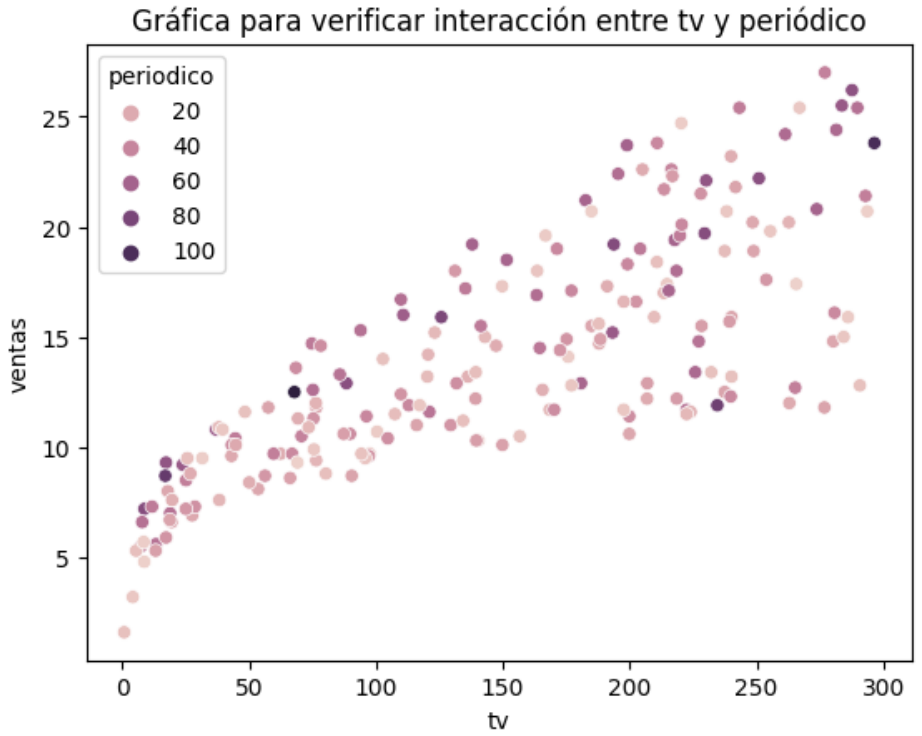
INTERACCIÓN

Observe que a medida que aumenta el gasto en radio (puntos más oscuros) aumenta el efecto de la tv (más crecen las ventas)c



Las Variables tv y radio interaccionan

Observe que esta gráfica el gasto en periódico no aumenta o disminuye el efecto de la tv.



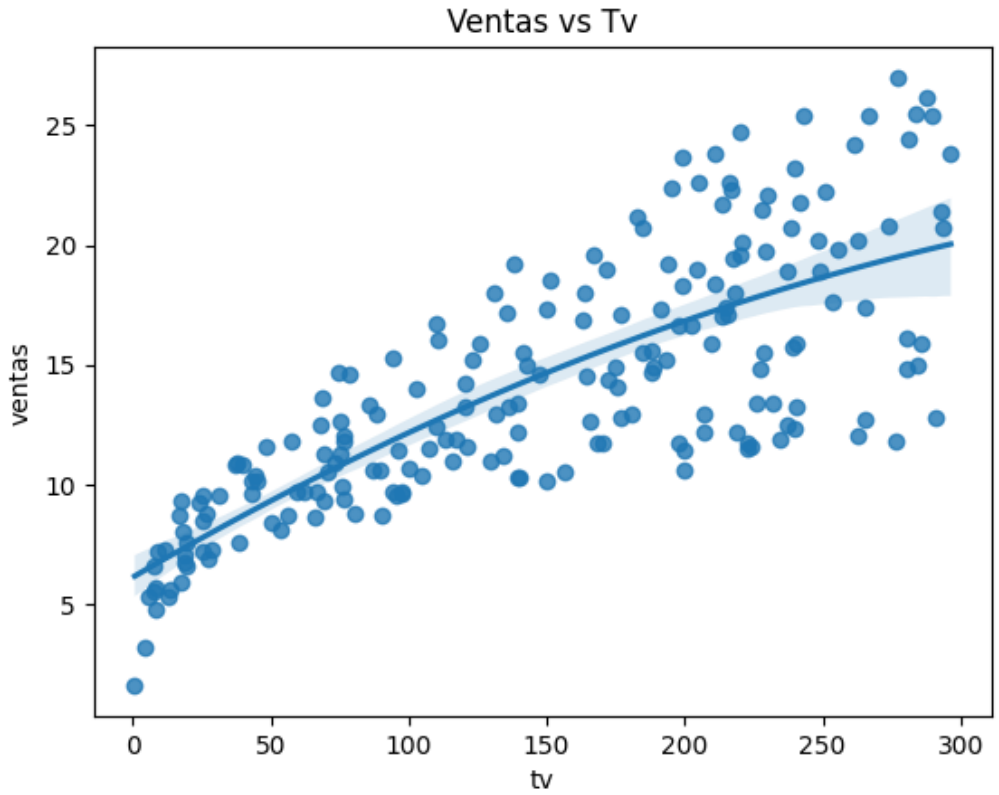
Las Variables tv y radio no interaccionan

El modelo para predecir las ventas debería tener la forma:

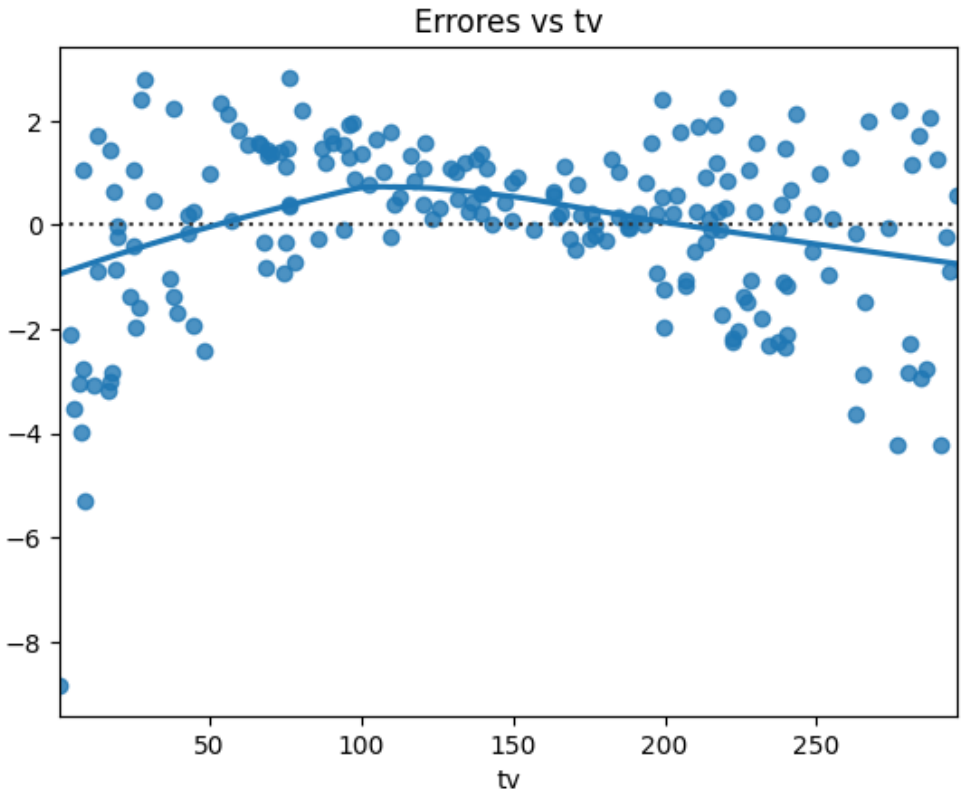
$$Ventas = 2.939 + 0.046TV + 0.189Radio - 0.0010Periodico + \beta_{TV \times Radio}(TV \times Radio)$$

RELACIONES NO LINEALES

En la gráfica se puede observar una relación no lineal



Esto se puede confirmar con el análisis de residuales vs el gasto en tv



El modelo para predecir las ventas debería tener la forma:

$$Ventas = 2.939 + 0.046TV + 0.189Radio \ - 0.0010Periodico + \beta_{TV^2}(TV^2)$$

ESTIMACIÓN DE GASTOS MEDICOS (PARTE 2)

Pontificia Universidad JAVERIANA Bogotá

Para el caso de gastos médicos desarrollo las siguientes tareas:

1. Análisis exploratorio con variables categóricas
2. Modelo de regresión con todas las variables
3. Análisis y evaluación de predicciones