Pontificia Universidad JAVERIANA Bogotá

REGRESIÓN LINEAL

ING. HERIBERTO FELIZZOLA JIMENEZ

RUTA DE CIENCIA DE DATOS

Educación **Continua**

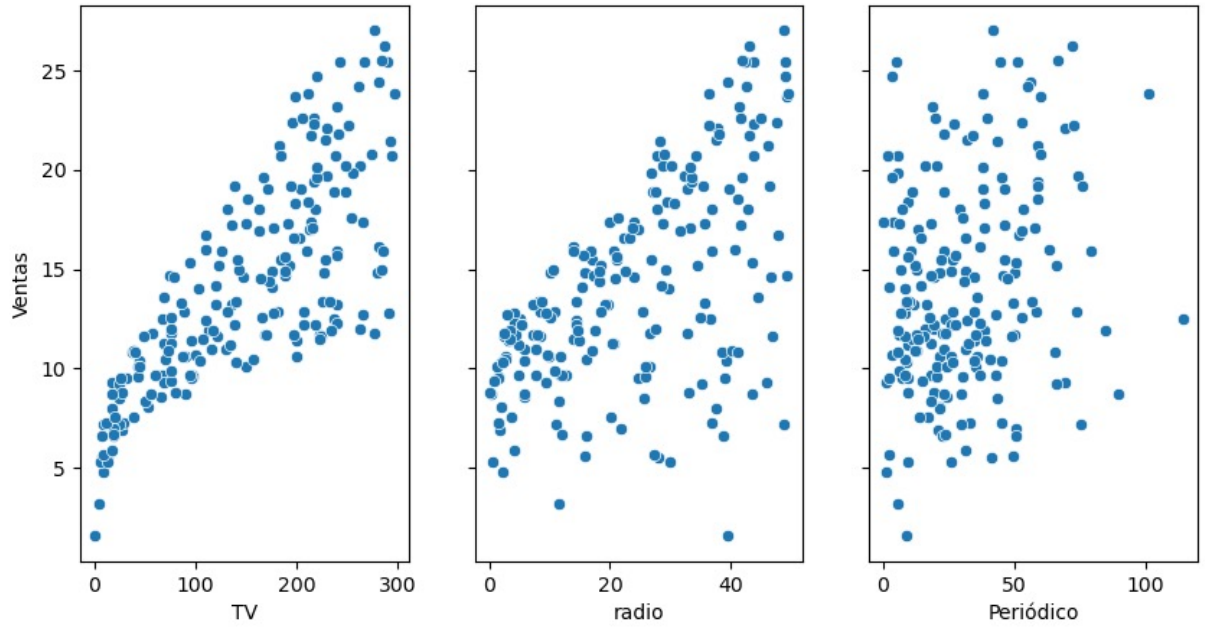
Generamos experiencias educativas

CONTENIDO

1. Motivación
2. Modelo de regresión
3. Conceptos básicos
4. Mínimos cuadrados
5. Salidas de un modelo de regresión
6. Interpretación del modelo de regresión
7. Análisis de errores
8. Variables categóricas

MOTIVACIÓN

- Supongamos que somos consultores estadísticos contratados por un cliente para investigar la asociación entre el gasto en publicidad y las ventas de un producto en particular.
- Nuestro cliente no puede incrementar directamente las ventas del producto pero puede controlar el gasto publicitario.
- Por lo tanto, si determinamos que existe una asociación entre publicidad y ventas, entonces podemos instruir a nuestro cliente para que ajuste los presupuestos publicitarios.



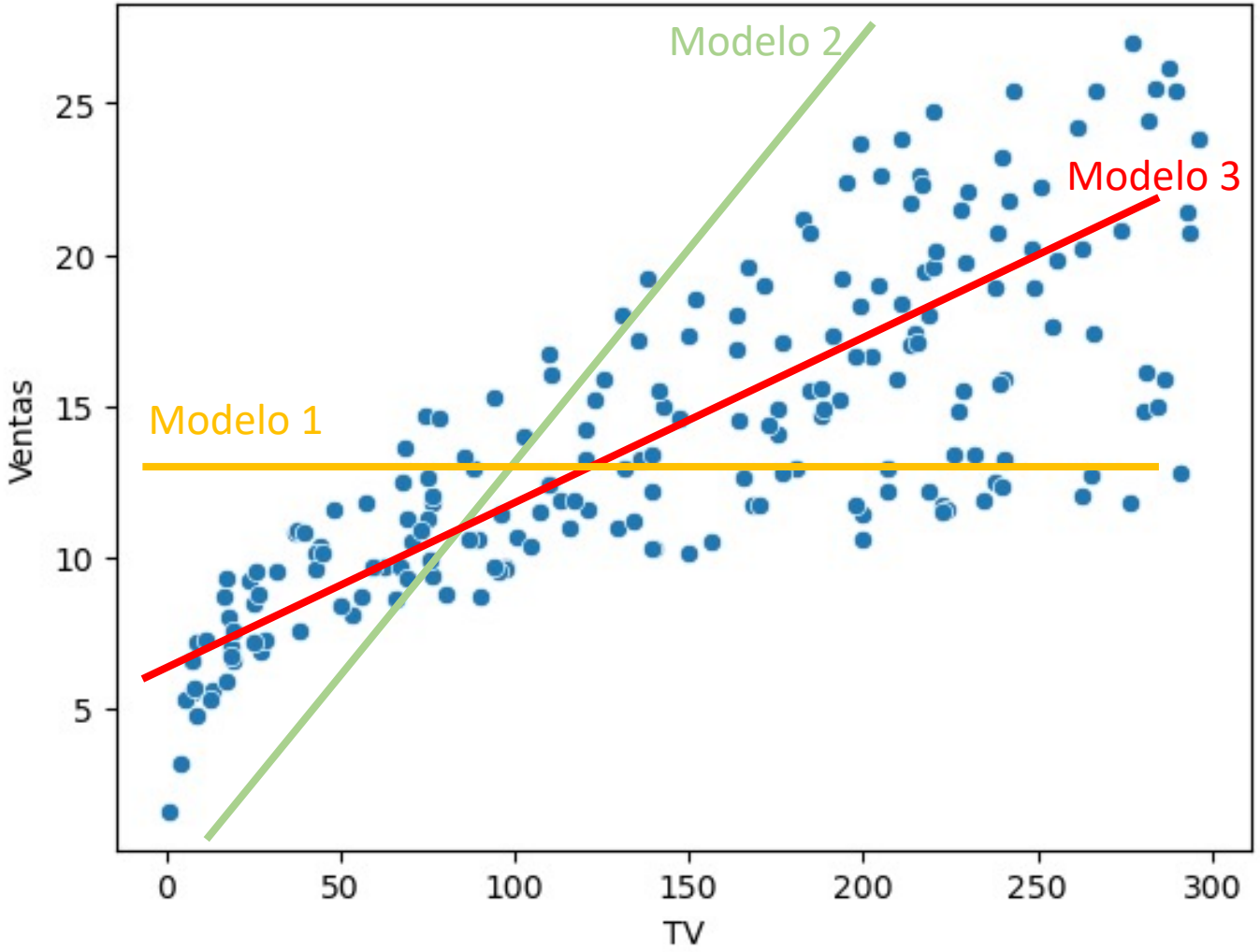
En otras palabras, nuestro objetivo es desarrollar un modelo preciso que se pueda utilizar para predecir las ventas sobre la base de los tres presupuestos de medios (TV, radio y periódico).

PREGUNTAS CLAVES…

- Con el modelo de regresión para predecir las ventas en función del presupuesto de publicidad en los diferentes medios, 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 la comportamiento de las ventas?
  - ¿Con qué precisión podemos predecir las ventas futuras?
  - ¿Existe correlación entre los medios publicitarios?

MODELO DE REGRESIÓN

- ¿Es posible trazar una línea que capture de forma perfecta todos los puntos?
- De forma intuitiva podemos decir que construir una modelo de predicción es encontrar la línea que nos permita capturar de la mejor manera la relación entre dos variables.
- En la gráfica podemos ver que el modelo 3, es el que mejor representa la relación entre las ventas y el presupuesto para TV.
- Observe que este modelo no captura de forma perfecta todos los puntos, por tanto la línea de tendencia central es el valor esperado de las ventas dado un presupuesto de TV.
- Debemos ser conscientes de que cualquier predicción de las ventas a partir del presupuesto para TV tendrá un error.
- No existe una predicción perfecta.



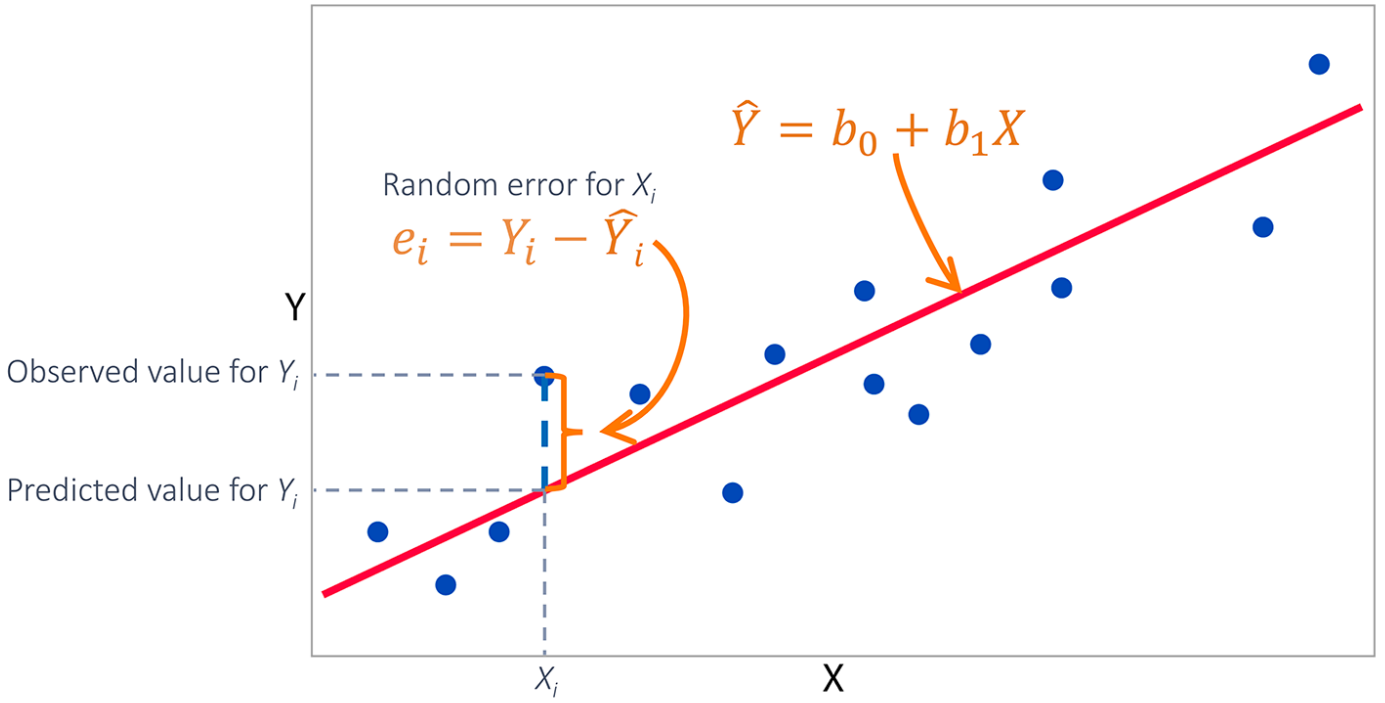
CONCEPTOS BÁSICOS

- 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

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

TV Radio Periódico

Contribuciones de las variables

$$, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

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 LA REGRESIÓN

Pontificia Universidad JAVERIANA Bogotá

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

- Esta pregunta se responde evaluando la significancia de la regresión.
- Esta prueba permite determinar si existe una relación lineal entre la variable de respuesta y las variables de entrada.

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

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

- Para esto se revisa el valor P de la regresión, en la tabla el Prob (F-statistic)
- Cuando este valor 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.

INTERPRETACIÓN DE LA REGRESIÓN

Pontificia Universidad JAVERIANA Bogotá

¿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.

INTERPRETACIÓN DE LA REGRESIÓN

Pontificia Universidad JAVERIANA Bogotá

¿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:

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

INTERPRETACIÓN DE LA REGRESIÓN

Pontificia Universidad JAVERIANA Bogotá

¿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

INTERPRETACIÓN DE LA REGRESIÓN

Pontificia Universidad JAVERIANA Bogotá

¿Con qué precisión podemos predecir las ventas futuras?

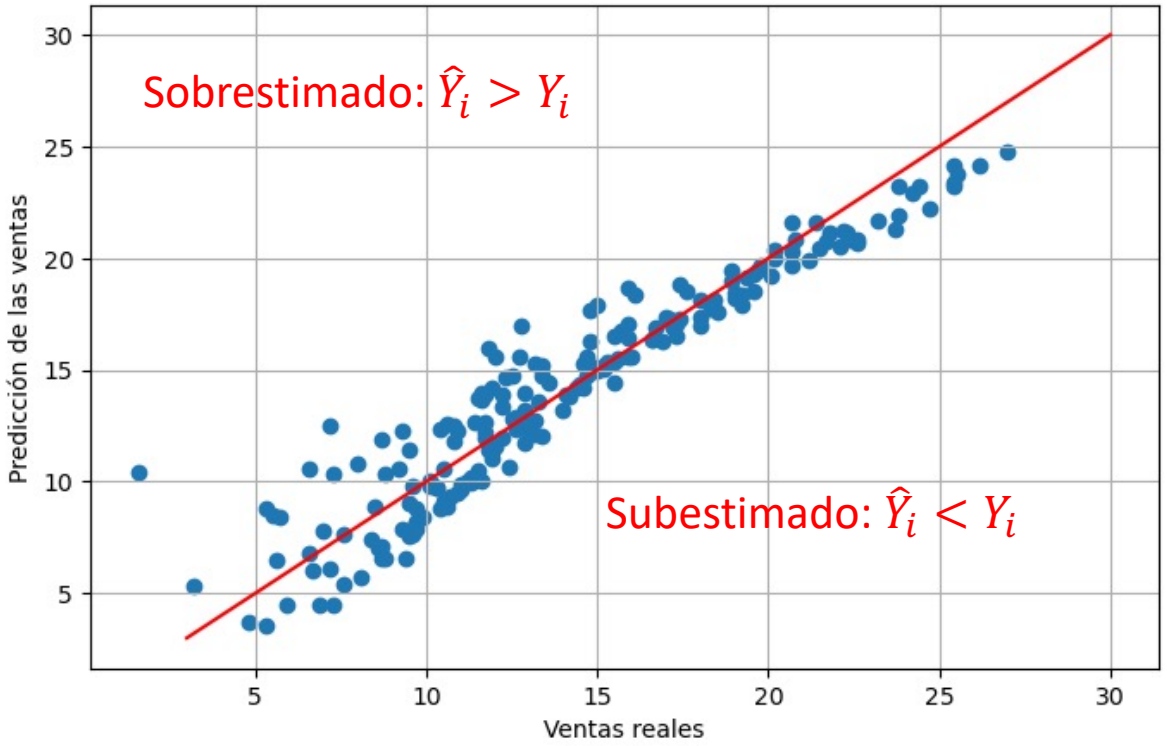
- Otra forma de evaluar la precisión del modelo es comparar el valor real de Y con la predicción.
- Además, se pueden utilizar los siguientes indicadores:
  - Error cuadrático medio MSE:
    - $MSE = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$
  - Raíz del error cuadrático medio RMSE:
    - $RMSE = \sqrt{MSE}$
  - Porcentaje de error:
    - $CV = \frac{RMSE}{|\bar{Y}|} = \frac{1.67}{|14.02|}$
    - $\text{CV} = 0.12$
- El CV indica un porcentaje de error de aproximadamente el 12%

Y Real	Predicción
22.1	20.523974
10.4	12.337855
9.3	12.307671
18.5	17.597830
12.9	13.188672
...	...
7.6	5.370342
9.7	8.165312
12.8	12.785921
25.5	23.767321
13.4	15.173196

$MSE = 2.78$

$RMSE = 1.67$

$\bar{Y} = 14.02$



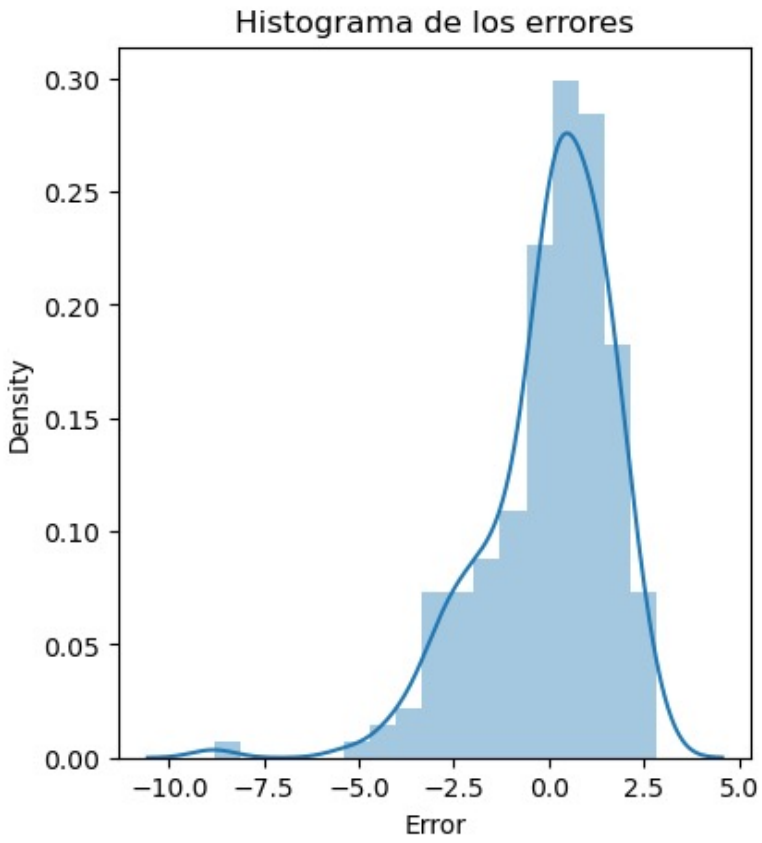
La gráfica y el CV indican una buena capacidad predictiva del modelo.

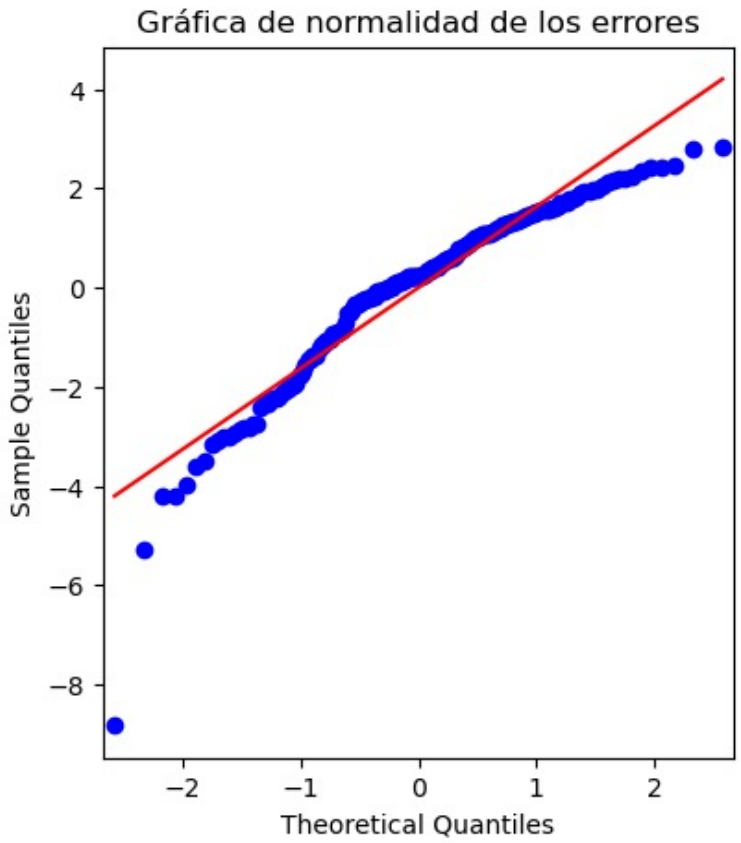
POTENCIALES PROBLEMAS

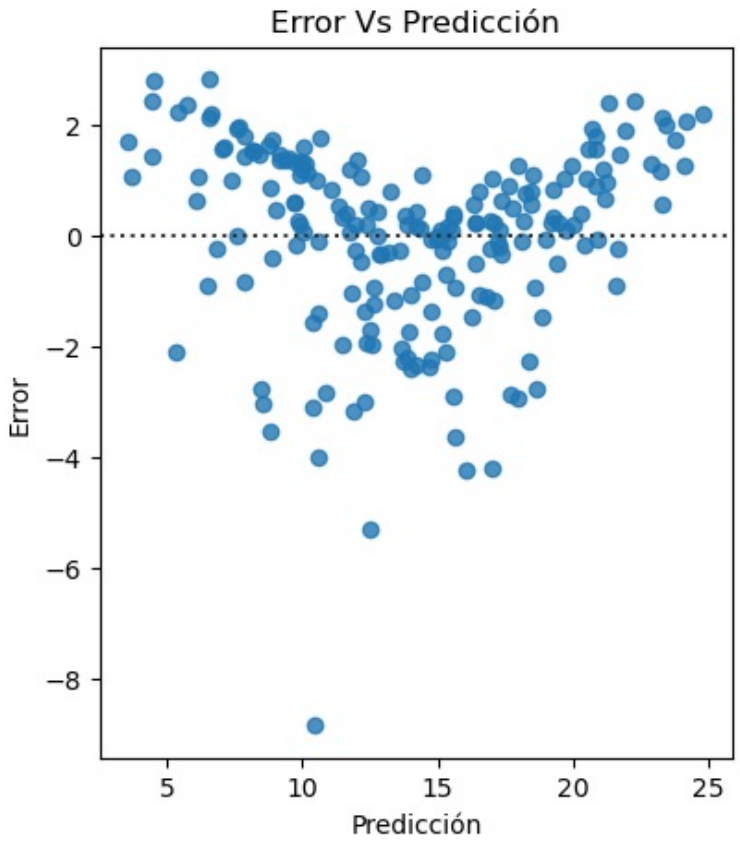
1. Falta de normalidad de los errores
2. Varianza no constante de los términos de error.
3. Valores atípicos.
4. Puntos de alto apalancamiento.
5. Multicolinealidad.
6. No linealidad de las relaciones respuesta-predictor.

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 la gráficas de los errores se pueden observar algunas 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.

VARIABLES CATEGÓRICAS

- Hasta el momento hemos asumido que todas las variables en nuestro modelo de regresión lineal son cuantitativas.
- Pero en la práctica, este no es necesariamente el caso; a menudo, 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 ficticia

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 Ficticias

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

¿PREGUNTAS?