

# REGRESIÓN LINEAL MÚLTIPLE

IRONHACK

# Intro

La mayoría de las aplicaciones prácticas del análisis de regresión utilizan modelos más complejos que la regresión lineal simple.

Imagina que acabas de ser ascendido a director de recursos humanos y crees que la forma en que se paga a cada trabajador no es justa. Por ello, decides crear un método más objetivo que el establecido. Esta es una buena oportunidad para crear un modelo de regresión múltiple.



# Intro

Lo primero que harías es consultar las bases de datos de la empresa y recoger la información que consideres importante a la hora de determinar el salario de un trabajador (el salario sería la variable dependiente  $y$ ). Por ejemplo, la antigüedad en la empresa ( $x_1$ ), el nivel de especialización ( $x_2$ ) y las horas de trabajo ( $x_3$ ) de todos los empleados. Estas serían nuestras variables independientes.

A continuación, habría que hacer un análisis exploratorio para entender cómo se comportan estas variables, de qué tipo son y cómo se relacionan con la variable dependiente. Este análisis le permite hipotetizar un modelo (o modelos) que pueda describir bien la relación entre estas variables. Este es el primer paso de un análisis de regresión múltiple.



# Intro

Con un análisis de regresión lineal múltiple puedes utilizar esta información para estimar el salario que debería pagarse a cada empleado de una forma más justa, puedes ver los casos que están por encima o por debajo del salario estimado (y ajustarlos) y puedes estimar el salario de los nuevos empleados.

Si este modelo no funciona bien puede que tengas que añadir más variables como el nivel de producción de la empresa ( $x_4$ ), o bien, eliminar algunas de las que has elegido, creando, así, nuevos modelos. Entonces sólo tendrías que comparar los resultados de cada modelo para elegir el mejor.



# Intro

En este módulo, vamos a hablar de los modelos de regresión lineal múltiple de primer orden. Son los modelos más sencillos, ya que es básicamente lo mismo que hemos visto en la regresión lineal simple, pero con más variables independientes.



# Intro

Componente determinístico

Error aleatorio

Los modelos probabilísticos que incluyen más de una variable independiente se denominan modelos de regresión múltiple. La forma general de estos modelos es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + \varepsilon$$

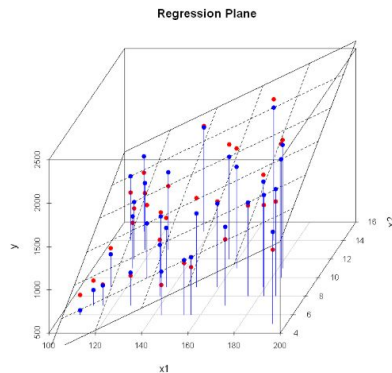
Donde:

- $y$  → variable dependiente (o de respuesta) (variable cuantitativa a modelar)
- $x_1, x_2, x_3, x_4$  y así sucesivamente → son las variables independientes (explicativas/ predictoras) (variables cuantitativas utilizadas como predictores de  $y$ )
- $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k$  → Componente determinista o línea de medias (error = 0)
- $\beta_0$  es el intercepto del modelo
- $\beta_1$  es la pendiente de la recta que relaciona  $y$  y  $x_1$ ,  $\beta_2$  es la pendiente de la recta que relaciona  $y$  y  $x_2$ ,  $\beta_3$  es la pendiente de la recta que relaciona  $y$  y  $x_3$ ,...y así sucesivamente.
- $\varepsilon$  es el **error aleatorio**

# Intro

Como puedes ver, la única diferencia con los modelos de regresión lineal simple es que hay más variables independientes ( $x$ ) y, por tanto, más pendientes, una pendiente por cada variable independiente.

Esto se debe a que **cada una de las variables independientes tiene una relación lineal con la variable independiente  $y$ , por tanto, al representar el modelo completo, ya no tendremos una sola línea, tendremos múltiples líneas**. Muchas líneas que acaban formando visualizaciones difíciles de interpretar, como ésta (con SOLAMENTE dos variables independientes):



# Intro

En la regresión lineal múltiple, seguiremos los mismos pasos: Construir el modelo, Evaluar el modelo y Utilizar el modelo.

Solo cambiarán algunas de las métricas que utilizaremos y algunas interpretaciones como consecuencia de que ahora trabajaremos con más variables independientes, pero el proceso es muy similar.



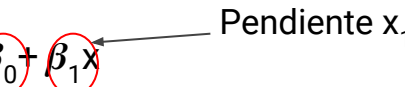


# Regresión Lineal Simple vs Regresión Lineal Múltiple

# Construir el modelo. Estimar $E(y)$

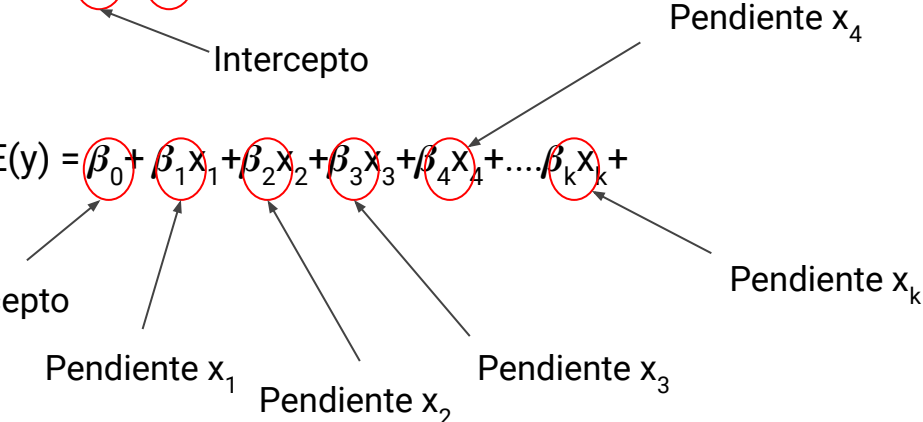
Al ajustar el modelo también utilizaremos el método de mínimos cuadrados para obtener la recta de mínimos cuadrados. La diferencia es que ahora estimamos más parámetros, ya que tenemos una pendiente para cada variable independiente.

RLS:  $E(y) = \beta_0 + \beta_1 x_1$



vs.

RLM:  $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \dots + \beta_k x_k +$



# Construir el modelo. Estimar E(y)

## Regresión Lineal Simple

```
model = ols('Sales ~ Advertising', data=datasales).fit()  
model.summary()
```

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.812			
Model:	OLS	Adj. R-squared:	0.811			
Method:	Least Squares	F-statistic:	7.105			
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	7.105e-05			
Time:	12:30:08	Log-Likelihood:	-448.99			
No. Observations:	200	AIC:	902.0			
Df Residuals:	198	BIC:	908.6			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	6.9748	0.323	21.624	0.000	6.339	7.611
Advertising	0.0555	0.002	29.260	0.000	0.052	0.059
Omnibus:	0.013	Durbin-Watson:	2.029			
Prob(Omnibus):	0.993	Jarque-Bera (JB):	0.043			
Skew:	-0.018	Prob(JB):	0.979			
Kurtosis:	2.938	Cond. No.	338.			

$\hat{\beta}_1$

## Regresión Lineal Múltiple

```
model = ols('Sales ~ Radio + TV + Newspaper', data=datasales2).fit()  
model.summary()
```

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.903			
Model:	OLS	Adj. R-squared:	0.901			
Method:	Least Squares	F-statistic:	60.1			
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	8.13e-05			
Time:	12:29:05	Log-Likelihood:	-383.34			
No. Observations:	200	AIC:	774.7			
Df Residuals:	196	BIC:	787.9			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	4.8251	0.308	15.641	0.000	4.019	5.232
Radio	0.1070	0.008	12.604	0.000	0.090	0.124
TV	0.0544	0.001	39.592	0.000	0.052	0.057
Newspaper	0.0003	0.006	0.058	0.954	-0.011	0.012
Omnibus:	16.081	Durbin-Watson:	2.251			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.655			
Skew:	-0.431	Prob(JB):	9.88e-07			
Kurtosis:	4.805	Cond. No.	454.			

$\hat{\beta}_0$

$\hat{\beta}_1$

$\hat{\beta}_2$

$\hat{\beta}_3$

# Interpretaciones

## Slope ( $\hat{\beta}_k$ )

Estimated  $y$  changes by  $\hat{\beta}_k$  for each 1 unit increase in  $x_k$   
***holding all other variables constant***

## y – intercept ( $\hat{\beta}_0$ )

Average value of  $y$  when  $x_k = 0$

# Evaluación del modelo. Supuestos

## Assumptions about Random Error $\varepsilon$

For any given set of values of  $x_1, x_2, \dots, x_k$ , the random error  $\varepsilon$  has a probability distribution with the following properties:

1. The mean is equal to 0.
2. The variance is equal to  $\sigma^2$ .
3. The probability distribution is a normal distribution.
4. Random errors are independent (in a probabilistic sense).

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Ordinary Least Squares (OLS) model
model = ols('Sales ~ Radio + TV + Newspaper', data=datasales2).fit()
model.summary()
```

#### OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.901
Method:	Least Squares	F-statistic:	805.4
Date:	Fri, 18 Mar 2022	Prob (F-statistic):	8.13e-99
Time:	11:26:04	Log-Likelihood:	-383.34
No. Observations:	200	AIC:	774.7
Df Residuals:	196	BIC:	787.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	4.6251	0.308	15.041	0.000	4.019	5.232
Radio	0.1070	0.008	12.604	0.000	0.090	0.124
TV	0.0544	0.001	38.592	0.000	0.052	0.057
Newspaper	0.0003	0.006	0.058	0.954	-0.011	0.012
Omnibus:	16.081	Durbin-Watson:	2.251			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.655			
Skew:	-0.431	Prob(JB):	9.88e-07			
Kurtosis:	4.605	Cond. No.	454.			

valor DW cercano a 2: no evidencia de correlación

valor pvalue JB test muy alto: no rechazamos la H0. No hay evidencia para demostrar que la distribución NO es normal

#### Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
import matplotlib.pyplot as plt
%matplotlib inline

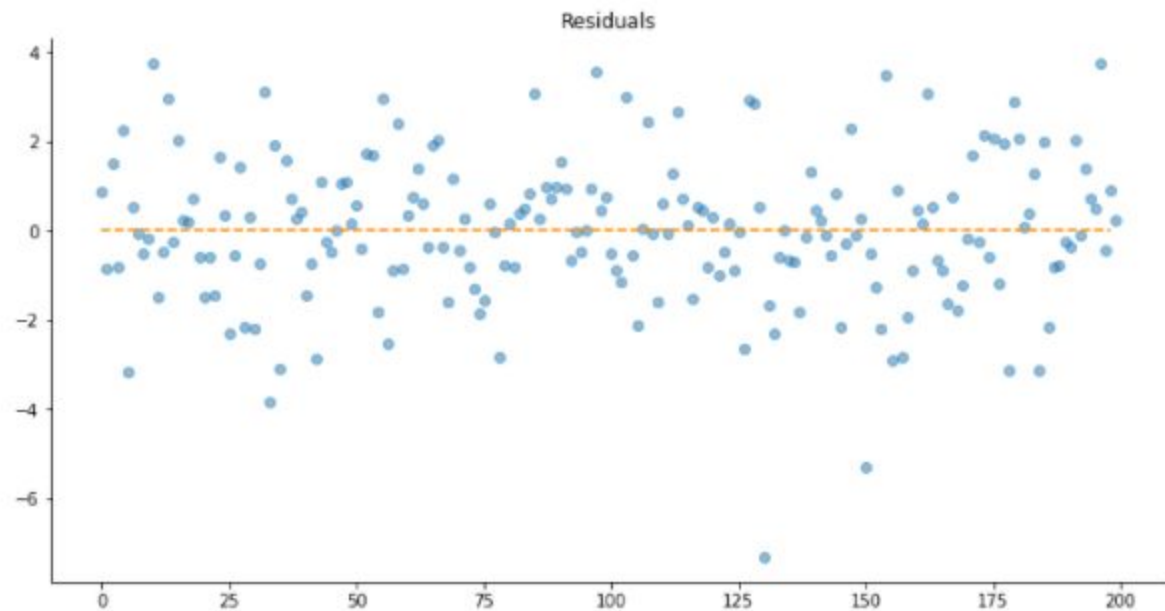
def homoscedasticity_assumption(model):
    """
    Homoscedasticidad: Varianza constante en el error
    """
    print('Los residuos deben tener una varianza relativamente constante')

    # Calculating residuals for the plot
    df_results = model.resid

    # Plotting the residuals
    plt.subplots(figsize=(12, 6))
    ax = plt.subplot(111) # To remove spines
    plt.scatter(x=df_results.index, y=df_results, alpha=0.5)
    plt.plot(np.repeat(0, df_results.index.max()), color='darkorange', linestyle='--')
    ax.spines['right'].set_visible(False) # Removing the right spine
    ax.spines['top'].set_visible(False) # Removing the top spine
    plt.title('Residuals')
    plt.show()
```

```
homoscedasticity_assumption(model)
```

Los residuos deben tener una varianza relativamente constante





# Evaluación del modelo

Al evaluar el modelo también mediremos su precisión (variabilidad del error), adecuación y utilidad.

	Simple linear Reg	Multiple Linear Reg	
<b>Precisión</b>	s or RMSE	s or RMSE	<i>The lower, the better</i>
<b>Adecuación</b>	p value t test for the slope	p value f test for all the slopes	<i>The lower, the better</i>
<b>Utilidad</b>	$r^2$ (Coeff. of determination) or $R^2$	Adjusted $r^2$ or $R^2_a$	<i>The higher, the better</i>

*\*Nota: utilizamos el  $r^2$  ajustado en lugar del  $r^2$  de la reg. lineal simple porque tiene en cuenta el tamaño de la muestra y el número de parámetros  $\beta$  en el modelo y, por tanto, es más fiable*

# Evaluación del modelo

## Regresión Lineal Simple

```
model = ols('Sales ~ Advertising', data=datasales).fit()  
model.summary()
```

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.811
Method:	Least Squares	F-statistic:	856.2
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	7.93e-74
Time:	12:30:08	Log-Likelihood:	-448.99
No. Observations:	200	AIC:	902.0
Df Residuals:	198	BIC:	908.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	6.9748	0.323	21.624	0.000	6.339	7.611
Advertising	0.0555	0.002	29.260	0.000	0.052	0.059

Omnibus:	0.013	Durbin-Watson:	2.029
Prob(Omnibus):	0.993	Jarque-Bera (JB):	0.043
Skew:	-0.018	Prob(JB):	0.979
Kurtosis:	2.938	Cond. No.	338.

p value test t bilateral  
pendiente

$r^2$  Coef  
determinación

## Regresión Lineal Múltiple

```
model = ols('Sales ~ Radio + TV + Newspaper', data=datasales2).fit()  
model.summary()
```

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.901
Method:	Least Squares	F-statistic:	605.4
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	8.13e-99
Time:	12:29:05	Log-Likelihood:	-383.34
No. Observations:	200	AIC:	774.7
Df Residuals:	196	BIC:	787.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	4.6251	0.308	15.041	0.000	4.019	5.232
Radio	0.1070	0.008	12.604	0.000	0.090	0.124
TV	0.0544	0.001	39.592	0.000	0.052	0.057
Newspaper	0.0003	0.006	0.058	0.954	-0.011	0.012

Omnibus:	16.081	Durbin-Watson:	2.251
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.655
Skew:	-0.431	Prob(JB):	9.88e-07
Kurtosis:	4.605	Cond. No.	454.

Adj  $r^2$  Coef  
determinación

p value test f  
pendientes

# Evaluación del modelo

## Regresión Lineal Simple

```
from statsmodels.tools.eval_measures import rmse  
  
ypred = model.predict(datasales['Advertising'])  
  
rmse = rmse (datasales['Sales'], ypred)  
rmse
```

2.2842381438447106

RMSE  
Se mide a través de  
la diferencia entre  
lo **observado** y lo  
**esperado**

predict y en función de los valores  
de x almacenados en la columna  
Advertising

## Regresión Lineal Múltiple

```
from statsmodels.tools.eval_measures import rmse  
  
ypred = model.predict(datasales2)  
rmse = rmse (datasales2['Sales'], ypred)  
rmse
```

1.6449942697855562

predict y en función de los valores  
de todas las variables  
independientes ( $x_{,k}$ ) almacenadas  
en en data set

# Interpretaciones

**Adj. r<sup>2</sup>** → 100(Adj. r<sup>2</sup>)% de la variación de la muestra en y puede explicarse utilizando las variables independientes utilizadas como predictores.

**s o RMSE** → Esperamos que la mayoría (95%) de los valores observados de y se sitúen dentro de  $\pm 2s$  de sus respectivos valores predichos.

**P value del test f para las pendientes** → si el valor p es menor que alfa, rechazamos la H<sub>0</sub> de que todas las pendientes son iguales a cero. Esto significa que, al menos una variable ind. del modelo es útil para predecir y, porque están relacionadas linealmente.

*\*Nota: en el test f testearnos→*

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  (All model terms are unimportant for predicting y.)  
 $H_a: \text{At least one } \beta_i \neq 0$  (At least one model term is useful for predicting y.)

# Utilizar el modelo para predecir y (Sales)

```
ypred = model.predict(datasales2)
print(ypred)
```

```
0      21.220972
1      11.268248
2      10.496209
3      17.312447
4      15.644137
...
195     7.105490
196    10.280941
197    15.259287
198    24.582220
199    18.185120
Length: 200, dtype: float64
```

```
newdata = pd.DataFrame({'TV': [200], 'Radio': [50], 'Newspaper': [30]})
newdata
```

	TV	Radio	Newspaper
0	200	50	30

```
model.predict(newdata)
```

```
0      20.874411
dtype: float64
```



# Regresión lineal múltiple con variables cuantitativas y cualitativas

# Proceso de creación de variables dummies

Los modelos de regresión múltiple también pueden escribirse para incluir variables independientes cualitativas (o categóricas). Las variables cualitativas, a diferencia de las cuantitativas, no pueden medirse en una escala numérica. Por lo tanto, debemos codificar los valores de la variable cualitativa (llamados niveles) como números antes de poder ajustar el modelo.



# Proceso de creación de variables dummies

Estas variables cualitativas codificadas se denominan variables dummy (o indicadoras), ya que los números asignados a los distintos niveles se seleccionan arbitrariamente. Por ejemplo, si queremos incluir el género en nuestro modelo como predictor, primero tenemos que "dummificar" la variable:

- Género: variable cualitativa con dos categorías: mujer (nivel A) y hombre (nivel B).
- Género = x (variable dummy)

$$x = \begin{cases} 1 & \text{if level A} \\ 0 & \text{if level B} \end{cases} \longrightarrow \text{base level}$$





# Proceso de creación de variables dummies

Si queremos dummificar un factor con más categorías tenemos que crear más variables dummy. Utilice siempre un número de variables dummy que sea uno menos que el número de niveles de la variable cualitativa. Así, para una variable cualitativa con  $k$  niveles, utilizas  $k - 1$  variables dummy.

Estado civil: variable cualitativa con cuatro categorías: casado, soltero, divorciado y viudo.

Estado civil =  $x_1, x_2, x_3$  (tres variables dummy)

- $x_1 \rightarrow 1$  si está casado, 0 si no lo está
- $x_2 \rightarrow 1$  si está soltero, 0 si no lo está
- $x_3 \rightarrow 1$  si está divorciado, 0 si no lo está
- $x_1, x_2$  y  $x_3 = 0$  si es viudo (base level)

Entonces, de  $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ :

- $E(y)$  si viudo  $\rightarrow E(y) = \beta_0$
- $E(y)$  si casado  $\rightarrow E(y) = \beta_0 + \beta_1(1)$
- $E(y)$  si soltero  $\rightarrow E(y) = \beta_0 + \beta_2(1)$
- $E(y)$  si divorciado  $\rightarrow E(y) = \beta_0 + \beta_3(1)$

# Nuevas interpretaciones

Lo único que cambiaría es la forma de interpretar las pendientes asociadas con las variables dummy.

La interpretación de estas estimaciones es la siguiente. Esperamos que el valor de  $E(y)$  aumente o disminuya en (el valor del estimador beta  $i$ ) si utilizamos (el nivel del factor  $i$ ) en comparación con si utilizamos (el nivel del factor elegido como nivel base), manteniendo constantes todas las demás variables independientes.



```
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Ordinary Least Squares (OLS) model
model = ols('Sales ~ Radio + TV + Newspaper+ AdType + Season + Country', data=datasales3).fit()
model.summary()
```

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.944			
Model:	OLS	Adj. R-squared:	0.942			
Method:	Least Squares	F-statistic:	359.0			
Date:	Fri, 18 Mar 2022	Prob (F-statistic):	3.49e-114			
Time:	12:49:39	Log-Likelihood:	-327.15			
No. Observations:	200	AIC:	674.3			
Df Residuals:	190	BIC:	707.3			
Df Model:	9					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	10.1141	0.853	11.863	0.000	8.432	11.796
AdType[T.AdType2]	-2.7067	0.310	-8.729	0.000	-3.318	-2.095
AdType[T.AdType3]	2.8056	0.654	4.289	0.000	1.515	4.096
AdType[T.AdType4]	0.5959	0.422	1.414	0.159	-0.236	1.427
Season[T.Standard]	-1.4981	0.681	-2.201	0.029	-2.841	-0.155
Season[T.Summer]	-0.6393	0.502	-1.274	0.204	-1.629	0.350
Radio	0.0573	0.008	7.037	0.000	0.041	0.073
TV	0.0292	0.003	10.321	0.000	0.024	0.035
Newspaper	0.0031	0.004	0.694	0.489	-0.006	0.012
Country	-0.0393	0.184	-0.214	0.831	-0.402	0.323
Omnibus:	39.543	Durbin-Watson:	2.293			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	112.183			
Skew:	-0.809	Prob(JB):	4.36e-25			
Kurtosis:	6.293	Cond. No.	2.34e+03			

## EJEMPLOS INTERPRETACIONES

Estimación  $\beta_1 \rightarrow$  Esperamos que el volumen medio de ventas disminuya en 2,7 (2706 unidades) si utilizamos el Tipo de Anuncio 2 en lugar del Tipo de Anuncio 1, manteniendo constantes todas las demás variables independientes..

Estimación  $\beta_9 \rightarrow$  Esperamos que el volumen medio de ventas disminuya en 0,039 (39 unidades) si la campaña se realiza en un País Extranjero en lugar de en España, manteniendo constantes todas las demás variables independientes..

Estimación  $\beta_4 \rightarrow$  Esperamos que el volumen medio de ventas disminuya en 1,498 (1498 unidades) si la campaña se realiza en primavera u otoño (Estándar) en lugar de en Navidad o cierre (invierno), manteniendo todas las demás variables independientes constantes.

```
ypred = model.predict(datasales3)
```

```
print(ypred)
```

```
0      22.014086
1       9.558415
2      11.920993
3      17.038241
4      16.107722
```

```
...
```

```
195     7.278867
196    11.671151
197    15.154124
198    23.767531
199    18.003891
```

```
Length: 200, dtype: float64
```

```
newdata = pd.DataFrame({'TV': [200], 'Radio': [50], 'Newspaper': [30], 'Country': 0, 'Season': 'Standard', 'AdType': 'AdType4'})
newdata
```

	TV	Radio	Newspaper	Country	Season	AdType
0	200	50	30	0	Standard	AdType4

```
model.predict(newdata)
```

```
0      18.004918
```

```
dtype: float64
```



Vuestro turno

# Sales Dataset MLR

Usted es un analista de negocios de una empresa que vende ordenadores. Su empresa quiere maximizar el volumen de ventas de un producto ajustando su inversión en publicidad, pero manteniendo una publicidad multicanal. Su empresa invierte en publicidad en televisión, radio y prensa.

Para entender mejor la relación entre estas variables, accedes a la base de datos de la empresa, ejecutas una consulta y obtienes un dataset.

Información del dataset: El conjunto de datos contiene estadísticas sobre las ventas de un producto ( $y$ , medido en miles de unidades) en 200 mercados diferentes, junto con el dinero gastado en publicidad en televisión ( $x_1$ ), el dinero gastado en publicidad en radio ( $x_2$ ) y el dinero gastado en publicidad en prensa ( $x_3$ ) en cada uno de estos mercados (todos ellos medidos en miles de dólares).

# Sales Dataset MLR

Replica el ejemplo de las slides con el dataset SALESADVMLR, interpreta todas las métricas y coeficientes del modelo y responde a estas preguntas.

1. Para el próximo mes, tu empresa tiene previsto invertir 40 (40.000 dólares) en publicidad en televisión, 15 (15.000 dólares) en publicidad en radio y 10 (10.000 dólares) en publicidad en prensa, en uno de los 200 mercados del conjunto de datos. ¿Qué volumen de ventas (en miles de unidades) esperas obtener?
2. Tu empresa te ha pedido que distribuyas el presupuesto de publicidad para maximizar el volumen de ventas y la única condición que te pone la empresa es que tiene que mantener la publicidad multicanal, es decir, tiene que invertir dinero en publicidad en televisión, radio y prensa. Observa los resultados de tu modelo, ¿cómo distribuirías el presupuesto para maximizar el volumen de ventas y por qué?
3. Observa los resultados de tu modelo. ¿Qué opinas del dinero invertido en la publicidad en los periódicos?