


REGRESIÓN LINEAL

IRONHACK

Intro

Hemos visto cómo estudiar la relación entre una variable cuantitativa (variable dependiente o respuesta) y dos o más variables cualitativas (variables independientes o factores). Lo hicimos comparando el valor medio poblacional de la variable cuantitativa en cada valor de la variable cualitativa (o en cada combinación de los valores de dos variables cualitativas).

Como normalmente no disponemos de los datos de toda la población, inferimos el valor de esta comparación aplicando métodos de estadística inferencial (pruebas de hipótesis y/o intervalos de confianza) a la información que obtenemos de la muestra o muestras que recogemos.



Intro

Pero ¿qué hacemos cuando queremos analizar la relación entre una variable cuantitativa (dependiente o respuesta) y otra variable cuantitativa (independiente, explicativa o predictiva)?

Si esta relación puede describirse mediante una línea recta, utilizamos modelos lineales para estudiarla.

Cuando la relación no está bien descrita por una línea recta, también podemos intentar ajustar un modelo lineal, pero obtendremos mejores resultados (y más fiables) utilizando modelos no lineales (principalmente, polinomial y segmentado).



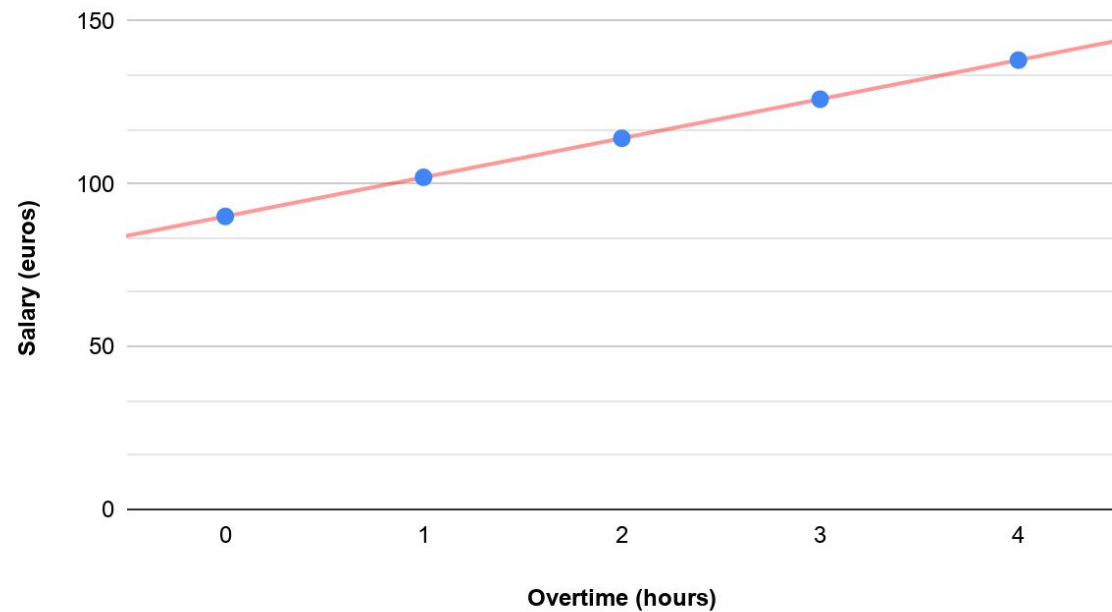
Intro

¿Serías capaz de hacer una fórmula matemática que relacione el dinero que ganas en un trabajo en un día (y , en euros) con las horas extra que haces (x , en horas)? Supongamos que tienes un sueldo fijo de 90 euros al día (jornada laboral estándar) y que ganas 12 euros más por cada hora extra de trabajo que hagas.

Esto se llama una relación determinista. Si construyéramos un modelo que hipotetizara una relación exacta entre las variables, se llamaría modelo determinista.



Salary and overtime



Overtime work (in hours, x)	Salary (in euros, y)
0	90
1	102
2	114
3	126
4	138

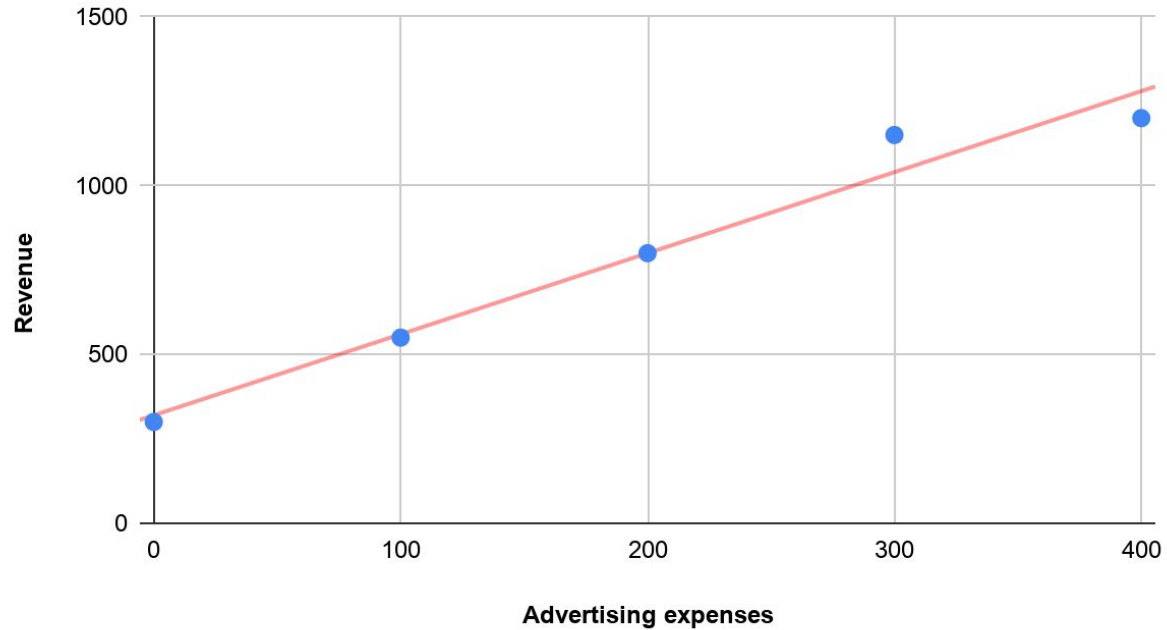
Intro

Ahora imagina que tienes un comercio electrónico y quieres hacer un modelo que relacione los ingresos por ventas mensuales (y) con los gastos mensuales en publicidad (x). Supongamos que en los 2 primeros meses has ganado 2,5 euros por cada euro gastado en publicidad (cada mes). ¿Crees que puedes hipotetizar un modelo que relacione los ingresos de un mes con el dinero que inviertes en publicidad en ese mes?

También sabes que, si no gastas dinero en publicidad, ganarás, aproximadamente, 300 euros al mes.



Revenue and Advertising exp (in a month)



Revenue (in euros, y)	Advertising expenses (in euros, x)
300	0
550	100
800	200
1150	300
1200	400

Intro

Un modelo matemático es una expresión matemática de algún fenómeno. A menudo describe relaciones entre variables.

- Un modelo puede ser un Modelo Determinista (hipotetiza relaciones exactas, sin error)
- O un Modelo Probabilístico (hipotetiza dos componentes: el componente Determinista y el error Aleatorio)



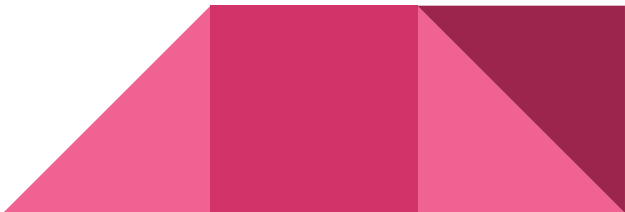
Intro

Forma general de los modelos probabilísticos

- $y = \text{Componente determinista} + \text{Error aleatorio}$

Donde y es la variable de interés.

Siempre suponemos que el valor medio del error aleatorio es igual a 0. Esto equivale a suponer que el valor medio de y , $E(y)$, es igual al componente determinista del modelo:

- $y = \text{Componente determinista} + \text{Error aleatorio}$
 - $E(y) = \text{Componente determinista} + 0$
 - $E(y) = \text{Componente determinista o línea de medias}$
- 

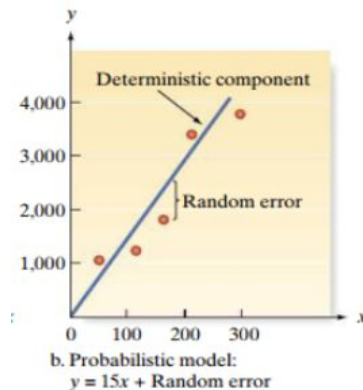
REGRESIÓN LINEAL SIMPLE

Componentes

Un modelo probabilístico de primer orden (línea recta) con una variable independiente cuantitativa:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- $y \rightarrow$ Variable dependiente o de respuesta (variable cuantitativa a modelar)
- $x \rightarrow$ Variable independiente o predictora (variable cuantitativa utilizada como predictor de y)
- $E(y) = \beta_0 + \beta_1 x \rightarrow$ Componente determinista o línea de medias (error = 0)
- β_0 (beta cero) \rightarrow **intercepción de la recta**, es decir, el punto en el que la recta intercepta o corta el eje y . Para simplificar, la intersección es el valor medio de y cuando x es igual a 0 $\rightarrow E(y) = \beta_0 + \beta_1(0) \rightarrow E(y) = \beta_0$
- β_1 (beta uno) \rightarrow pendiente de la línea, es decir, el cambio (cantidad de aumento o disminución) en el valor medio de y por cada aumento de 1 unidad en x .
- ε (épsilon) \rightarrow es un componente de error aleatorio que mide lo lejos que está por encima o por debajo de la línea de regresión (línea de medias) la observación real de y . La media de ε es cero.



Parámetros

Un modelo de regresión lineal simple contiene 3 **parámetros desconocidos**; β_0 (intercepción de la línea), β_1 (pendiente de la línea) y la varianza de la ε (si la línea está bien ajustada ya sabemos que la media del error es 0).

Tendremos que inferir estos parámetros (o características de la población), como siempre, utilizando la información de nuestra muestra.

Target parameters	Estimators (from sample)
β_0	$\hat{\beta}_0$
β_1	$\hat{\beta}_1$
σ^2 of the ε	s^2 of the ε



Método mínimos cuadrados

La primera condición para encontrar una recta que se ajuste es que la suma de los errores sea igual a 0. Hemos encontrado una recta bien ajustada si la suma de los errores (o residuos) es igual a cero. Sin embargo, es posible encontrar muchas rectas que cumplan esta condición (diapositiva anterior).

Según el método de los mínimos cuadrados, la mejor recta es aquella cuya SSE es la mínima:

- La suma de los errores es la suma de las diferencias entre los valores de los puntos de datos de la muestra (lo que obtenemos) y los valores que esperamos (con la recta).
- SSE es la suma de errores al cuadrado. Esta métrica es más fiable porque hace visibles los errores negativos (entre otras cosas).



Método mínimos cuadrados

Por lo tanto, nuestra misión es obtener la ecuación que define la línea que mejor se ajusta. Siguiendo el método de los mínimos cuadrados, esta línea se denomina línea o recta de mínimos cuadrados.

- La recta de mínimos cuadrados \hat{y} es el estimador muestral de la componente determinista $E(y)$.
- Es decir, la línea de mínimos cuadrados es el mejor estimador de la verdadera línea de medias $\rightarrow E(y) = \beta_0 + \beta_1 x$

Por lo tanto, a través de este método podemos estimar β_0 y β_1 con los datos de la muestra.



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

# Ordinary Least Squares (OLS) model
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:	Tue, 15 Mar 2022	Prob (F-statistic):	7.93e-74			
Time:	20:28:48	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.			

$\hat{\beta}_0$

$\hat{\beta}_1$

Métricas - Coeficientes

Intercepción: $\hat{\beta}_0$ representa el valor previsto de y cuando $x = 0$. (Precaución: Este valor no será significativo si los valores $x = 0$ no tienen sentido o están fuera del rango de los datos de la muestra)

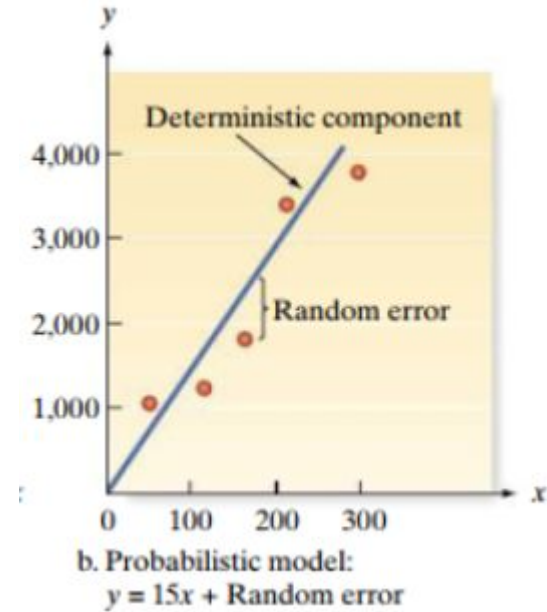
Pendiente: $\hat{\beta}_1$ representa el aumento (o disminución) esperado de y por cada aumento de 1 unidad en x . (Precaución: esta interpretación sólo es válida para valores de x dentro del rango de los datos de la muestra).



Métricas - Distribución del error

Habrán algunos puntos de datos que no coincidirán con el valor esperado de y indicado por el componente determinista estimado del modelo. Entre estos puntos de datos y el valor esperado de y (dado un valor de x) hay una distancia que hemos llamado error aleatorio.

En la recta de mejor ajuste que hemos trazado según nuestra muestra, los valores estimados de la pendiente y el intercepto serán siempre los mismos, independientemente del punto de datos (componente determinista estimada), mientras que el valor del error puede cambiar para cada punto de datos (como se muestra en la imagen).



Métricas - Distribución del error

Así, el modelo probabilístico que intenta explicar la relación entre dos variables cuantitativas tiene que tener en cuenta todos los valores posibles del error aleatorio.

Por tanto, **el siguiente paso es definir la distribución del error en nuestro modelo**. Para ello, seguiremos una serie de supuestos y comprenderemos parte de la información que ya tenemos.

1. Lo primero que vamos a suponer es que el error aleatorio ε se distribuye normalmente.

Para definir cómo es esta distribución normal tenemos que especificar el valor de su media y varianza (o desviación estándar) $\rightarrow N(\mu, \sigma^2)$

Pero, ¿cómo obtenemos estos valores?



Métricas - Distribución del error

Haber utilizado el método de los mínimos cuadrados para trazar la recta ya nos ha dado cierta información. Recuerda que la recta de mejor ajuste cumple dos condiciones:

- La recta garantiza que la suma de los errores es aproximadamente cero.
- Tiene la menor SSE de todas las rectas posibles.

Con una de estas condiciones podemos especificar el valor de la media de la distribución de errores. ¿Serías capaz de decirme cuál y por qué?



Métricas - Distribución del error

Con esto ya sabemos que:

1. La distribución del error ε es normal. (Test de Jarque Bera, entre otros)
2. La media de la distribución del error ε es 0 para todos los ajustes de la variable independiente x .

Lo que no sabemos es cómo estimar la varianza de esta distribución, que dependerá de la variabilidad de los errores (o residuos) de los datos con los que estamos trabajando.

Sin embargo, sí sabemos que esta varianza tiene que permanecer constante para todos los valores de x , de lo contrario no podríamos definir la distribución del error ε y nuestro modelo no sería válido. Esta propiedad se llama **homocedasticidad**.

3. La varianza de la distribución del error ε es constante para todos los valores de x .

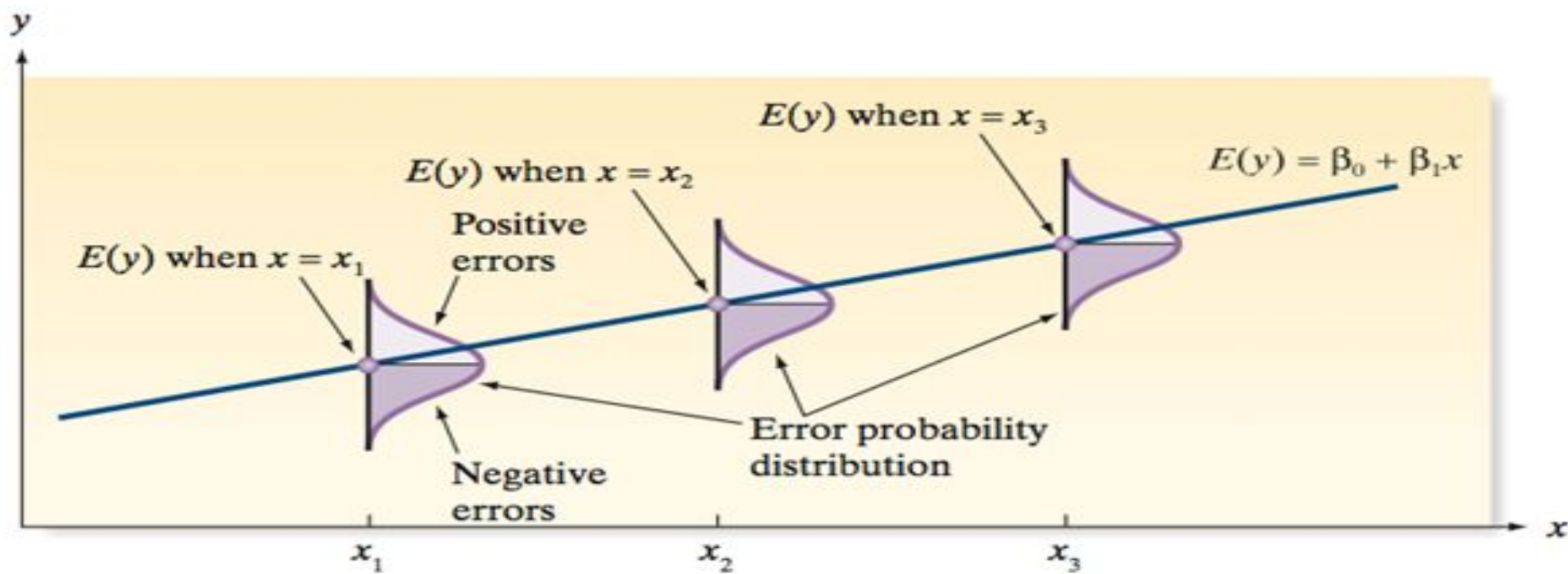
Métricas - Distribución del error

El último supuesto:

4. Los valores de ε asociados a dos valores observados de y son independientes. Es decir, el valor de ε asociado a un valor de y no tiene efecto en ninguno de los valores de ε asociados a cualquier otro valor de y . (**Test de Durbin Watson, entre otros**)



Basic Assumptions of the Probability Distribution



Métricas - Distribución del error

- **Skew, Kurtosis:** nos permiten imaginar la forma de la distribución. (Slides de estadística descriptiva).
- **Durbin Watson:** estudia la autocorrelación de los errores. El valor siempre está entre 0 y 4. Si el estadístico de Durbin-Watson es sustancialmente menor que 2, hay evidencia de correlación serial positiva.
- **Jarque Bera:** H_0 : la distribución es normal, H_a : la distribución no es normal



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

# Ordinary Least Squares (OLS) model
model = ols('Sales ~ Advertising', data=datasales).fit()
model.summary()
```

OLS Regression Results

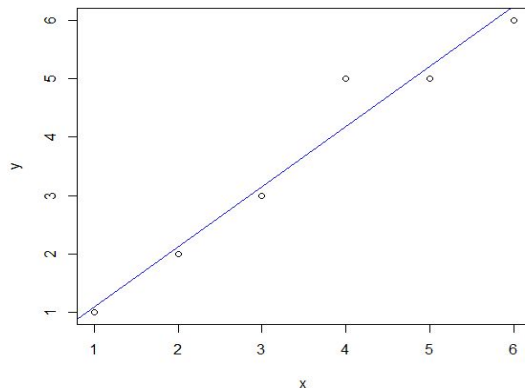
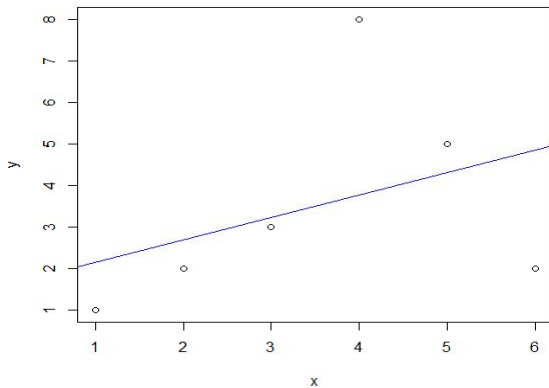
Dep. Variable:	Sales	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.811
Method:	Least Squares	F-statistic:	856.2
Date:	Tue, 15 Mar 2022	Prob (F-statistic):	7.93e-74
Time:	20:28:48	Log-Likelihood:	-448.99
No. Observations:	200	AIC:	902.0
Df Residuals:	198	BIC:	908.6
Df Model:	1		
Covariance Type:	nonrobust		

	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.

Métricas - Distribución del error

Parece razonable suponer que cuanto mayor sea la variabilidad del error aleatorio ε (que se mide por su varianza σ^2), mayores serán los errores en la estimación de los parámetros del modelo β_0 y β_1 y en el error de predicción cuando se utiliza \hat{y} para predecir y para algún valor de x .



Métricas - Distribución del error

La varianza del error o MSE, nos sirve para ver cómo de preciso es nuestro modelo. Vemos esto a partir de la desviación estándar del error o RMSE (que no es más que la raíz cuadrada del MSE).

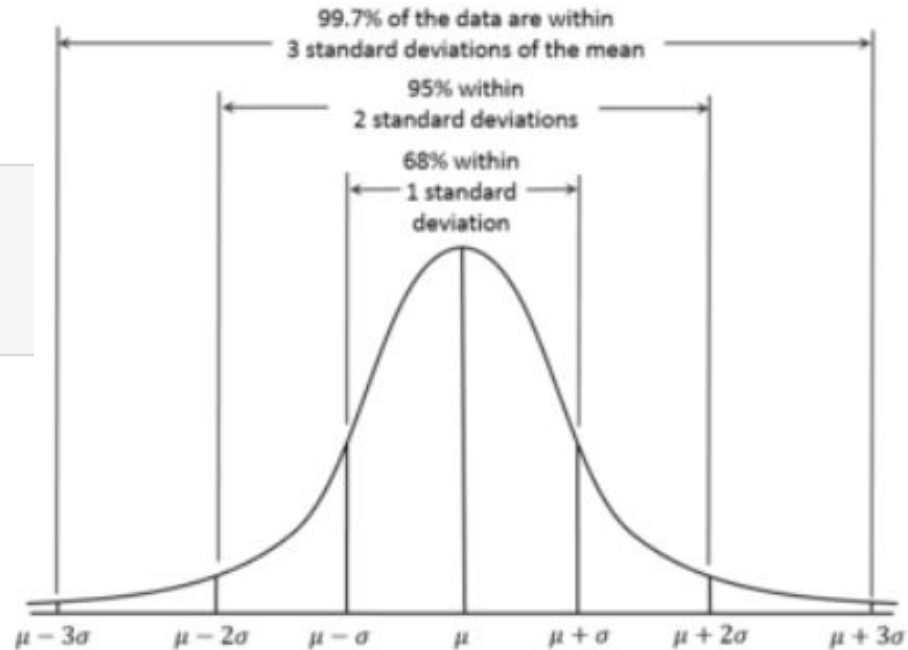
Podemos obtener el valor del RMSE e interpretarlo, de esta forma:

- Esperamos que la mayoría (95%) de los valores de tiempo de reacción observados se sitúen dentro de $2s = 2RMSE$ de sus respectivos valores predichos.
- Es decir, según nuestra muestra, esperamos que el 95% de los valores estén dentro de este intervalo $(\hat{y} - 2RMSE, \hat{y} + 2RMSE)$.

Por tanto, cuanto mayor sea el RMSE, menos preciso será nuestro modelo.



```
from statsmodels.tools.eval_measures import rmse  
  
ypred = model.predict(datasales['Advertising'])  
  
rmse = rmse (datasales['Sales'], ypred)  
rmse  
  
2.2842381438447106
```



Métricas - Inferencias sobre la pendiente real

En las secciones anteriores hemos estimado el valor de la pendiente a partir de la información de la muestra, pero recuerde que se trata de un estimador, no del valor de la verdadera pendiente β_1 .

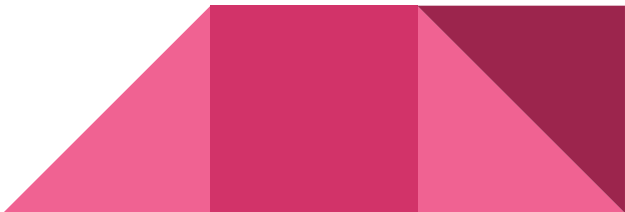
Para asegurarnos de que nuestro modelo es adecuado tenemos que demostrar estadísticamente que la verdadera pendiente β_1 no es igual a 0, porque, si no somos capaces de hacerlo, puede que el valor de la verdadera pendiente sea 0 y esto significa que NO existe una relación lineal (línea recta) entre la variable independiente (x) y la variable dependiente (y).



Métricas - Inferencias sobre la pendiente real

¿Qué se podría decir entonces sobre los valores de la pendiente y el intercepto en el modelo probabilístico hipotético ($y = \beta_0 + \beta_1 x + \varepsilon$) si x no aporta ninguna información para la predicción de y ?

La implicación es que la media de y (es decir, la parte determinista del modelo $E(y) = \beta_0 + \beta_1 x$) no cambia cuando cambia x . En el modelo lineal, esto significa que la verdadera pendiente, β_1 , es igual a 0. Por lo tanto, para probar la hipótesis nula de que el modelo lineal no aporta ninguna información para la predicción de y contra la hipótesis alternativa de que el modelo lineal es útil para predecir y , testeamos:

- $H_0: \beta_1 = 0$
 - $H_a: \beta_1 \neq 0$
- 

Métricas - Inferencias sobre la pendiente real

Para hacer este test podemos ver el output de nuestro modelo en Python e interpretar el estadístico, el pvalue o trazar un intervalo de confianza para estimar el valor de la pendiente real.



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

# Ordinary Least Squares (OLS) model
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:	Tue, 15 Mar 2022	Prob (F-statistic):	7.93e-74
Time:	20:28:48	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	0.322	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.

Métricas - Utilidad del modelo

Nos quedaría ver cómo de útil es el modelo, es decir, cómo de buena es nuestra variable x como predictora. Podemos ver esto a través del coeficiente de determinación y/o el coeficiente de correlación.

- ¿Cómo de fuerte es la relación entre la variable dependiente(y) y la variable independiente (x)? **Coeficiente de correlación (r)**
- ¿En qué medida la información proporcionada por x contribuye a describir o predecir y ? **Coeficiente de determinación (r^2)**

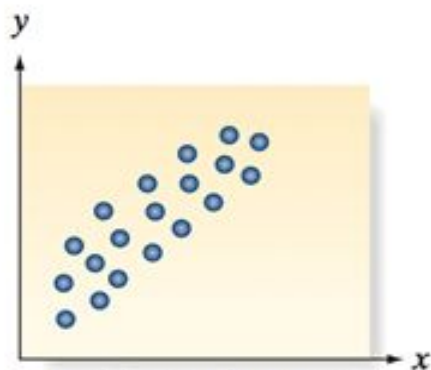


Métricas - Utilidad del modelo. Correlación

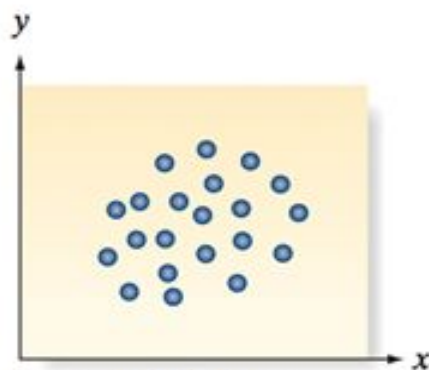
Responde a: "Según nuestra muestra, ¿qué intensidad tiene la relación lineal entre dos variables?"

- El coeficiente de correlación mide el grado de asociación entre la variable independiente (x) y la variable dependiente (y).
- Los valores posibles de este coeficiente van de -1 a 1, dependiendo de la fuerza y la dirección de la relación entre las dos variables.
- ATENCIÓN: este coeficiente no indica una relación causa-efecto, es decir, por muy fuerte que sea la relación entre dos variables, no podemos indicar que una causa la otra utilizando este coeficiente.
- Si la correlación entre x e y es débil, es posible que queramos encontrar otra variable independiente (x) si queremos obtener buenos resultados en la predicción de la variable dependiente (y).

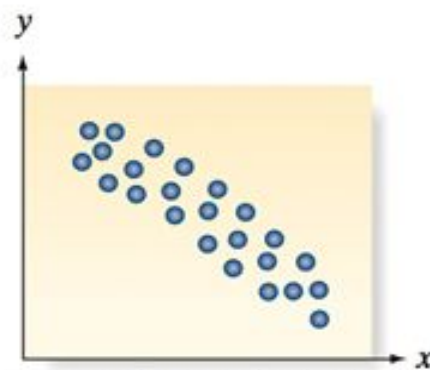
Coefficient of Correlation



a. Positive r : y increases
as x increases



b. r near 0: little or
no relationship
between y and x



c. Negative r : y decreases
as x increases

Métricas - Utilidad del modelo. Coeficiente de determinación

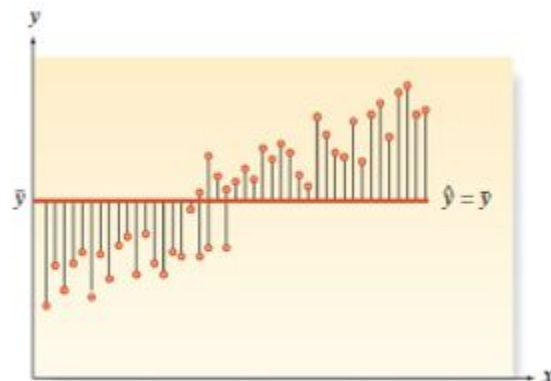
Otra forma de medir la utilidad de un modelo lineal es medir la **contribución de x en la predicción de y**. Para ello, calculamos cuánto se redujeron los errores de predicción de y al utilizar la información proporcionada por x (en comparación con los errores que obtendríamos si utilizáramos sólo la media de y para predecir el valor de las nuevas observaciones)

r^2 representa la proporción de la variabilidad total de la muestra en torno a \bar{y} que se explica por la relación lineal entre x e y.

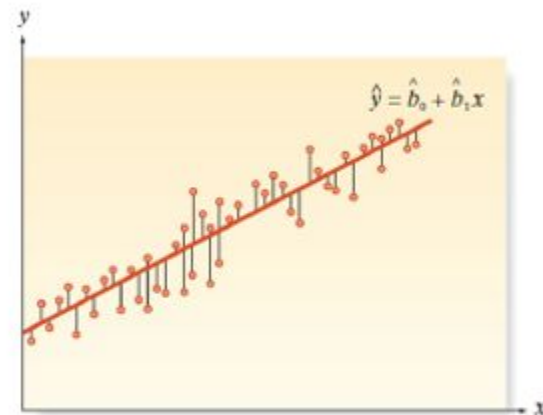




a. Scatterplot of data



b. Assumption: x contributes no information for predicting y , $\hat{y} = \bar{y}$



c. Assumption: x contributes information for predicting y , $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

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

# Ordinary Least Squares (OLS) model
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:	Tue, 15 Mar 2022	Prob (F-statistic):	7.93e-74
Time:	20:28:48	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.

Coeficiente de determinación

```
correlation_coef = 0.812**(0.5)
correlation_coef
```

0.9011104260855048

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

# Ordinary Least Squares (OLS) model
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:	Tue, 15 Mar 2022	Prob (F-statistic):	7.93e-74
Time:	20:28:48	Log-Likelihood:	-448.99
No. Observations:	200	AIC:	902.0
Df Residuals:	198	BIC:	908.6
Df Model:	1		
Covariance Type:	nonrobust		

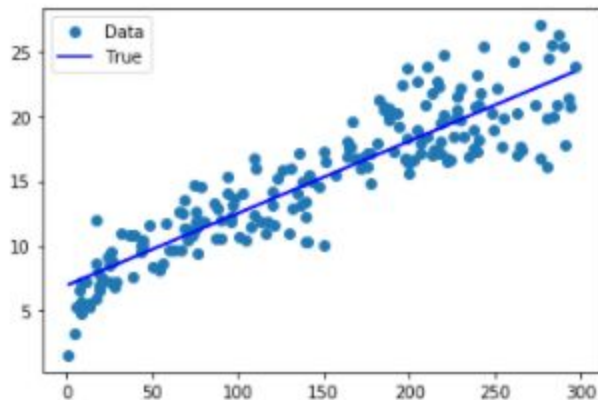
	coef	std err	t	P> t	[0.025	0.975]
Intercept	6.9748	0.323	21.624	0.000	6.339	7.611
Advertising	0.0555	0.002	29.260	0.000	0.052	0.059

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.

Utilizar el modelo para predecir

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

x = datasales['Advertising']
y = datasales['Sales']
xln = np.linspace(20.5, 25, 10)
fig, ax = plt.subplots()
ax.plot(x, y, 'o', label="Data")
ax.plot(x, ypred, 'b-', label="Model")
ax.legend(loc="best");
```



```
ypred = model.predict(datasales['Advertising'])
print(ypred)
```

```
0      19.737265
1       9.443004
2       7.928816
3      15.377734
4      17.002852
...
195     9.093576
196    12.199603
197    16.792086
198    22.704630
199    19.848195
Length: 200, dtype: float64
```

```
newdata = pd.DataFrame({'Advertising': [200]})  
newdata
```

Advertising	
0	200

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

0 18.067776

dtype: float64



Vuestro turno

Sales Dataset

Eres un analista de negocios de una empresa que vende ordenadores. Tu empresa quiere ajustar el presupuesto de publicidad que se gasta en un producto pero, antes de tomar una decisión, necesita cuantificar la eficacia de la publicidad en términos de volumen de ventas.

Para entender mejor la relación entre estas dos cosas, 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 en 200 mercados diferentes, junto con el presupuesto publicitario en cada uno de estos mercados. **Las ventas están en miles de unidades (variable dependiente y) y el presupuesto publicitario está en miles de dólares (variable independiente x).**

Sales Dataset

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

1. Según este modelo, ¿cuánto esperamos que se reduzca el volumen de ventas (en miles de unidades) si reducimos la inversión en publicidad en 10 puntos (10000 dólares)?
2. Para el próximo mes, su empresa tiene previsto invertir 45 (45.000 dólares) en uno de los 200 mercados del conjunto de datos. Estima el volumen de ventas (en miles de unidades).

