Estadística

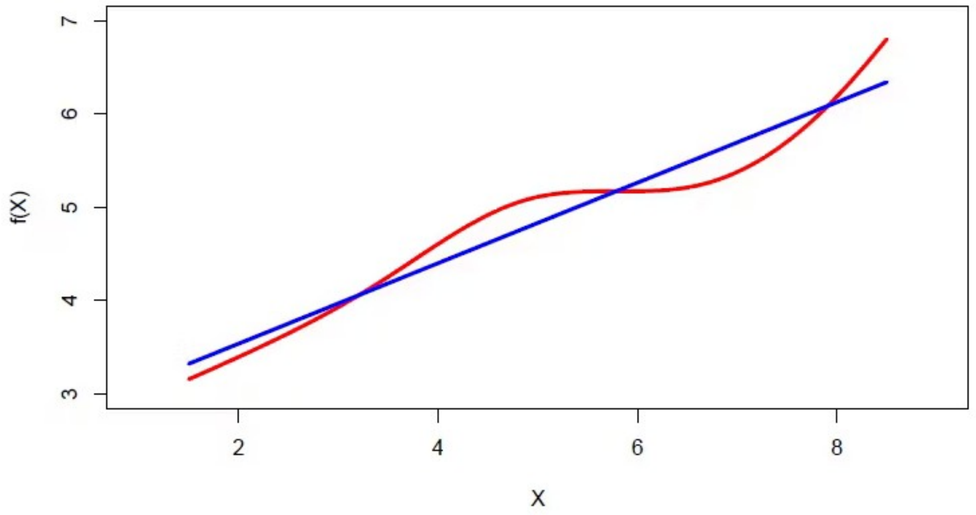
Regresión Lineal Simple

Instructions

Regresión Lineal

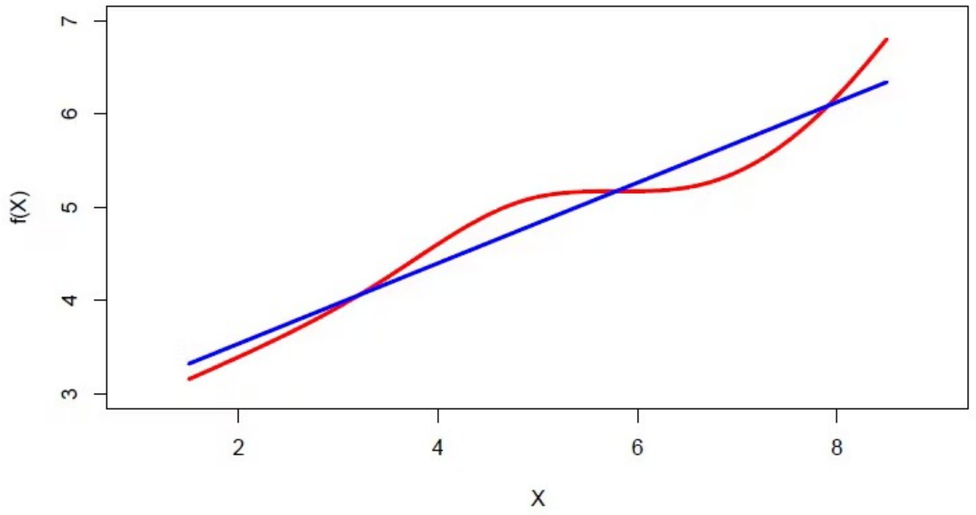
Simplista pero útil

3



Regresión Lineal

- La regresión lineal es un enfoque simple para el aprendizaje supervisado
- Asume que la dependencia de Y respecto a $X_1, X_2, \ldots, X_p$ es lineal
- Las funciones en la naturaleza rara vez son lineales



Regresión Lineal

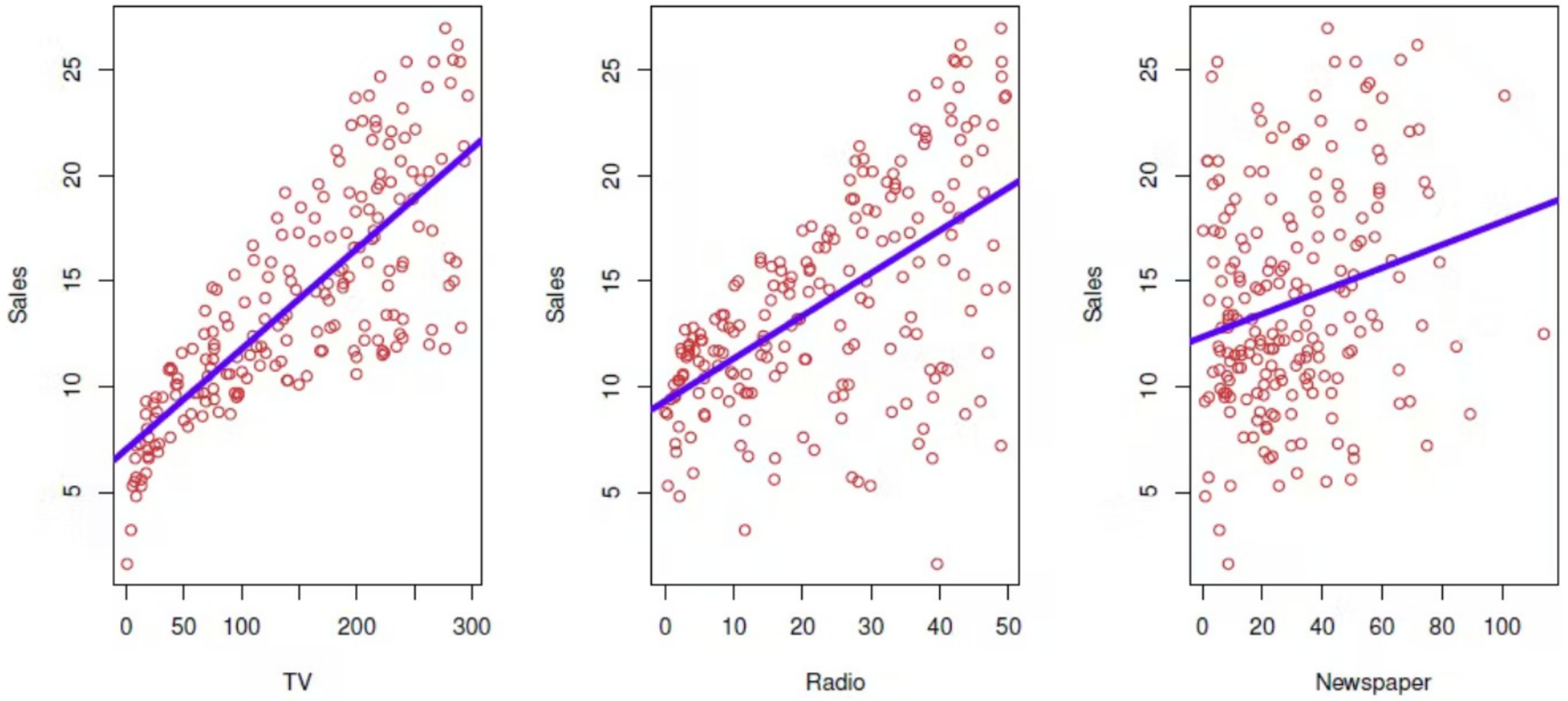
- Las verdaderas funciones de regresión nunca son lineales
- Aunque pueda parecer demasiado simplista, es extramadamente útil tanto conceptual como en la práctica

Estudio de datos publicitarios

Los datos de publicidad muestran las ventas (en miles de unidades) de un producto en particular como una función de los presupuestos de publicidad (en miles de dólares) para los medios de televisión, radio y periódicos.

Suponga que en nuestro papel como consultores estadísticos, se nos pide que sugieramos, con base en estos datos, un plan de marketing para el próximo año que dará como resultado altas ventas de productos.

¿Qué información sería útil para proporcionar tal recomendación?



Estudio de datos publicitarios

Preguntas que podríamos hacernos:

- ¿Existe una relación entre el presupuesto de publicidad y las ventas?
- ¿Qué tan fuerte es la relación entre el presupuesto de publicidad y las ventas?
- ¿Qué medios contribuyen a las ventas?
- ¿Con qué precisión podemos predecir las ventas futuras?
- ¿Es la relación lineal?
- ¿Existe sinergia entre los medios publicitarios?

RL: un único predictor numérico X

Asumimos un modelo

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

donde

β_0: intercepto

β_1: pendiente

ε: término de error

Dadas las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ (o también denotados b_0 y b_1) predecimos las ventas futuras usando $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Estimación por Mínimos Cuadrados

Sea $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ la predicción para Y basada en el i-ésimo valor de X. Entonces $\epsilon_i = y_i - \hat{y}_i$ representa el i-ésimo residual.

Definimos la suma residual de cuadrados (RSS) como:

$$RSS = \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_n^2,$$

o equivalentemente como:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

El enfoque de mínimos cuadrados elige $\hat{\beta}_0$ y $\hat{\beta}_1$ para minimizar el RSS.

Estimación por Mínimos Cuadrados

Se puede demostrar que los valores de minimización son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

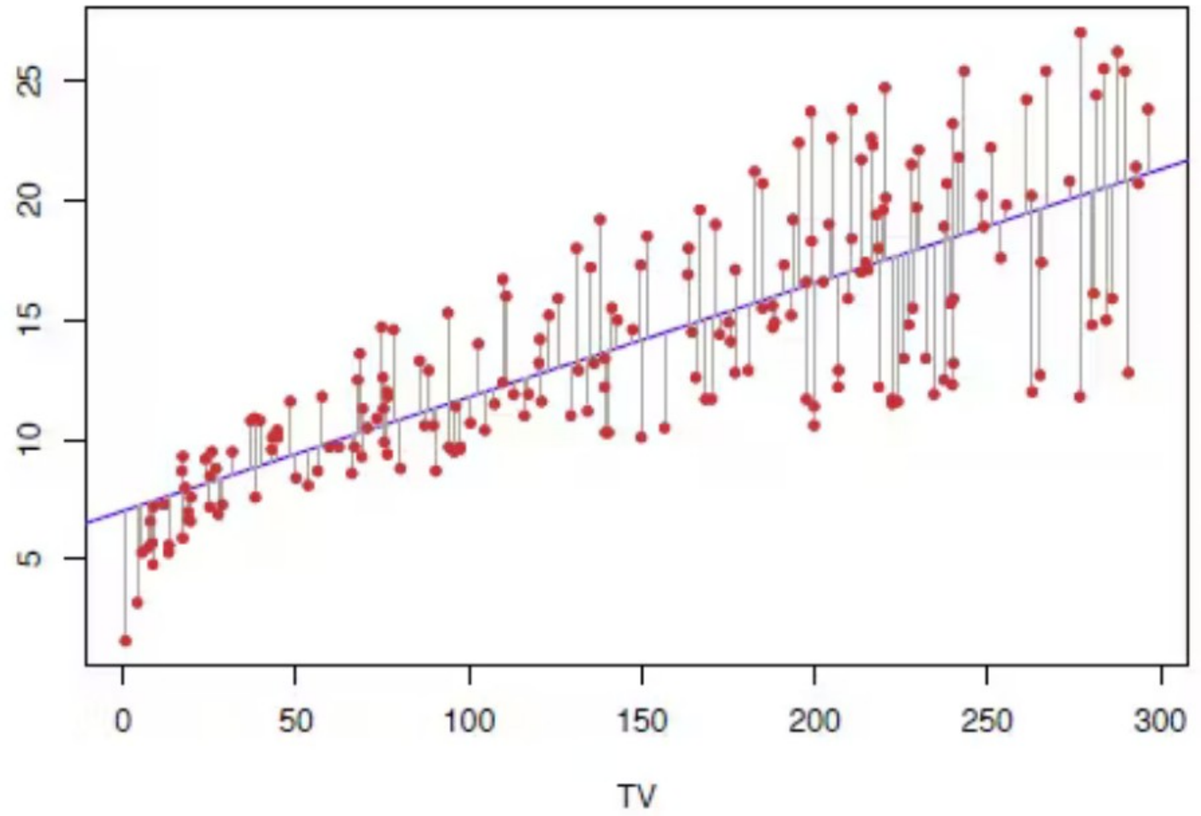
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Donde $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ son las medias muestrales.

Ejemplo: datos publicitarios

Los mínimos cuadrados se ajustan a la regresión de las ventas en TV.

En este caso, un ajuste lineal captura la esencia de la relación, aunque algo deficiente en la parte izquierda de la trama.



Evaluación de la precisión de las estimaciones del coeficiente

El error estándar de un estimador refleja como varía bajo muestreo repetido.

Tenemos

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right],$$

donde $\sigma^2 = Var(\varepsilon)$

Evaluación de la precisión de las estimaciones del coeficiente

- Con los errres estándar se construyen los Intervalos de Confianza (IC)
- Un IC es una variable aleatoria: a cada muestra se obtiene un nuevo intervalo
- IC: contiene al verdadero valor con una alta probabilidad
- Toman la forma $[\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)]$

Intervalos de confianza

Existe aproximadamente un 95% de probabilidad de que el intervalo

$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)]$$

contendrá el verdadero valor de β_1, es decir, el efecto de la inversión en TV sobre las ventas.

Para los datos publicitarios, el intervalo de confianza del 95% para β_1 es [0.042, 0.053].

Prueba de hipótesis

Los errores estándar también se pueden usar para realizar pruebas de hipótesis de los coeficientes.

La prueba de hipótesis más común consiste en probar las hipótesis

H_0 : No hay relación entre X y Y

vs.

H_1 : Hay alguna relación entre X y Y

Prueba de hipótesis

Matemáticamente, esto corresponde a probar

$$H_0 : \beta_1 = 0$$

vs.

$$H_1 : \beta_1 \neq 0,$$

ya que si $\beta_1 = 0$ entonces el modelo se reduce a formula $Y = \beta_0 + \varepsilon$.

En este caso, X no está asociado con Y

Prueba de hipótesis

Para probar la hipótesis nula, calculamos un estadístico T, dado por

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

Esta tendrá una distribución t de Student con $n - 2$ grados de libertad, suponiendo $\beta_1 = 0$ (bajo H_0 verdadera).

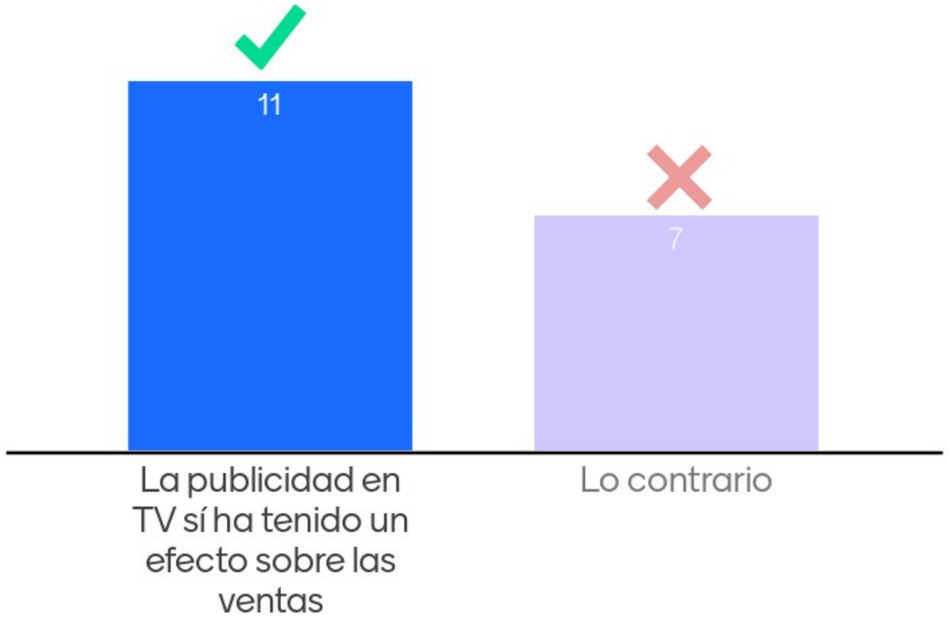
Usando software estadístico, es fácil calcular la probabilidad de observar cualquier valor igual a $|t|$ o más grande.

Llamamos a esta probabilidad el **valor p**.

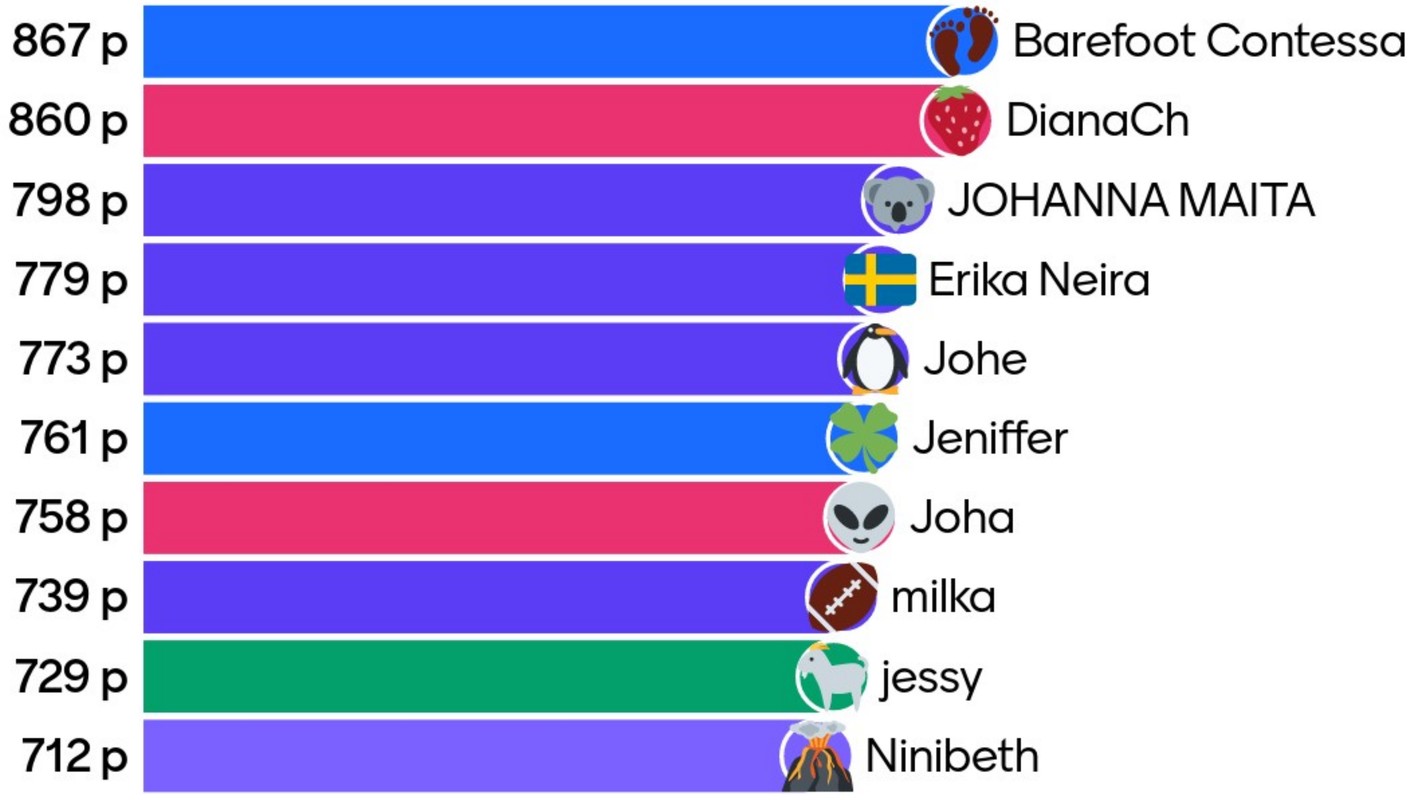
Valor p	Evidencia contra H_0	Conclusión
<0.01	evidencia muy fuerte	Se rechaza H_0
0.01-0.05	evidencia fuerte	Se rechaza H_0
0.05-0.10	evidencia débil	$\cdots$
>0.1	poca o ninguna evidencia	No se rechaza H_0

Toma de decisiones con base en el valor p

Para la prueba de hipótesis anterior, ¿qur quieres decir si se rechaza la hipótesis nula $H_0 : \beta_1 = 0$?

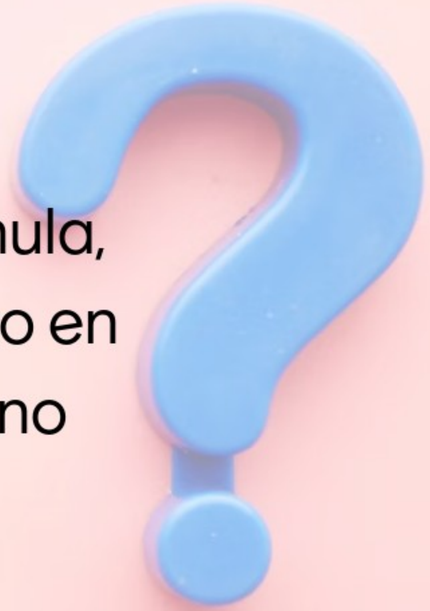


Leaderboard



Warning!

De no rechazarse la hipótesis nula, esto no quiere decir que el gasto en publicidad en TV y las ventas no estén relacionadas.



¿Por qué?

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Resultados del estudio de datos publicitarios

Evaluación de la precisión general del modelo

Calculamos el error estándar residual

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

donde la suma de cuadrados residual es

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

Esta es una medida de error:

mientras menor, mejor.

Evaluación de la precisión general del modelo

Ota medida importante es el coeficiente R^2
o fracción de varianza explicada es

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

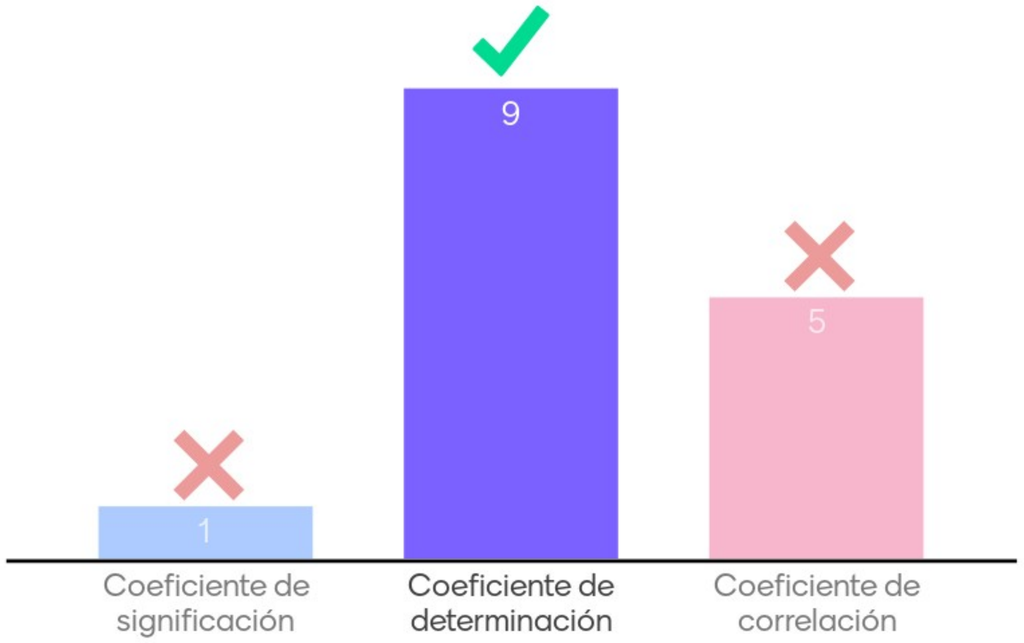
donde $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$,
es la suma total de cuadrados.

Evaluación de la precisión general del modelo

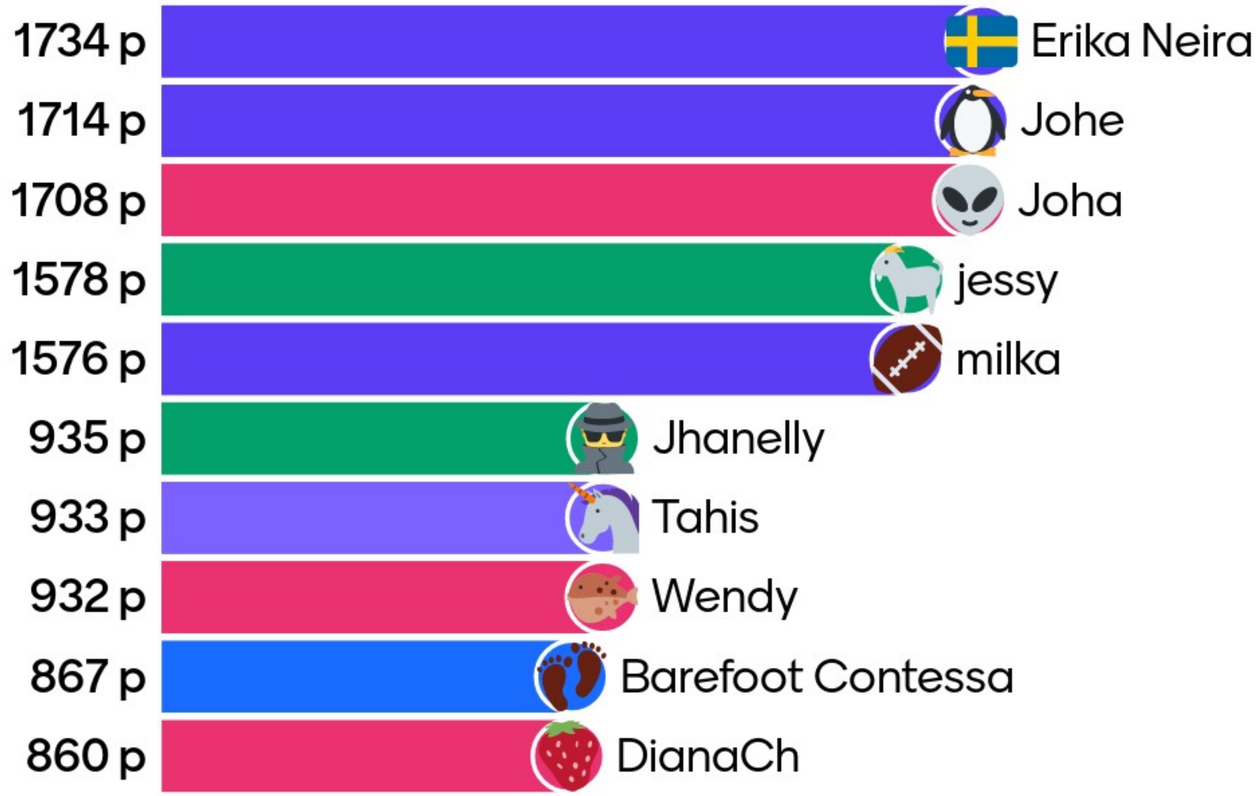
Se puede demostrar que en este ajuste de regresión lineal simple que $R^2 = r^2$, donde r es la correlación entre x y y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

¿Cómo también se lo conoce al coeficiente R^2?



Leaderboard



Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Resultados del estudio de datos publicitarios


EVERYONE
CAN CODE!