AWAKELAB

BASECAMP

Ciencia de Datos

Análisis Exploratorio y Programación Estadística

Objetivo de la jornada

- Explicar los conceptos de correlación y regresión lineal para la caracterización de un conjunto de datos de una población.

Correlación

El análisis de correlación es el primer paso para construir modelos explicativos y predictivos más complejos.

A menudo nos interesa observar y medir la relación entre 2 variables numéricas mediante el análisis de correlación. Se trata de una de las técnicas más habituales en análisis de datos y el primer paso necesario antes de construir cualquier modelo explicativo o predictivo más complejo.

¿Qué es la correlación?

La correlación es un tipo de asociación entre dos variables numéricas, específicamente evalúa la tendencia (creciente o decreciente) en los datos.

Dos variables están asociadas cuando una variable nos da información acerca de la otra. Por el contrario, cuando no existe asociación, el aumento o disminución de una variable no nos dice nada sobre el comportamiento de la otra variable.

Dos variables se correlacionan cuando muestran una tendencia creciente o decreciente.

¿Cómo se interpreta la correlación?

La correlación nos permite medir el signo y magnitud de la tendencia entre dos variables. En la figura que se muestra más abajo, se

ven diferentes valores del coeficiente de correlación y sus diagramas de dispersión correspondientes. Podemos ver que:

1. El signo nos indica la dirección de la relación, como hemos visto en el diagrama de dispersión.
    - un valor positivo indica una relación directa o positiva,
    - un valor negativo indica relación indirecta, inversa o negativa,
    - un valor nulo indica que no existe una tendencia entre ambas variables (puede ocurrir que no exista relación o que la relación sea más compleja que una tendencia, por ejemplo, una relación en forma de U).
2. La magnitud nos indica la fuerza de la relación, y toma valores entre -1 a 1. Cuanto más cercano sea el valor a los extremos del intervalo (1 o -1) más fuerte será la tendencia de las variables, o será menor la dispersión que existe en los puntos alrededor de dicha tendencia. Cuanto más cerca del cero esté el coeficiente de correlación, más débil será la tendencia, es decir, habrá más dispersión en la nube de puntos.
    - si la correlación vale 1 o -1 diremos que la correlación es “perfecta”,
    - si la correlación vale 0 diremos que las variables no están correlacionadas.

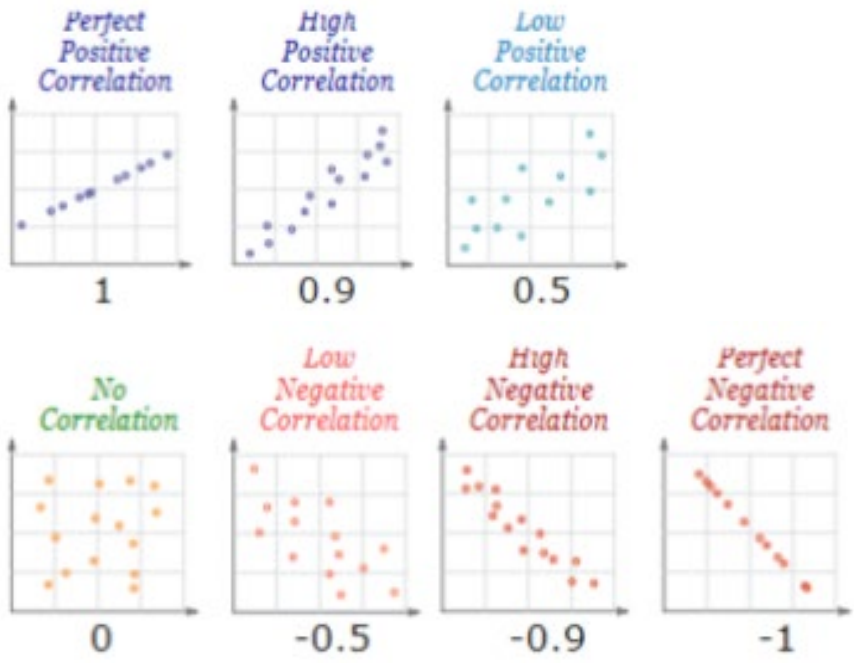


imagen:fuente propia

Tamaño del efecto

En estadística, **el tamaño del efecto es una medida de la fuerza o magnitud de un fenómeno**. El coeficiente de correlación es una medida del tamaño del efecto para la relación (lineal) entre dos variables numéricas.

Se trata de un dato esencial para interpretar los resultados de nuestro estudio y su ausencia en los artículos científicos se ha identificado como uno de los 7 fallos más comunes en investigación (según la APA 19961 , 20012).

Para interpretar qué tan fuerte es la correlación podemos utilizar el criterio de Cohen (1988)3, quien para valores absolutos indica que valores entre:

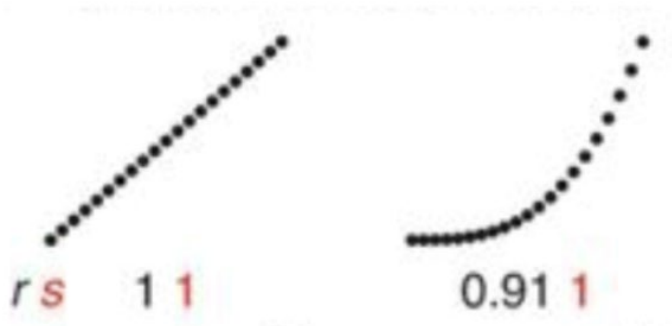
- 0.1 - 0.3 representan un efecto pequeño,
- 0.3 - 0.5 un efecto medio y
- 0.5 un efecto grande.

Son valores arbitrarios que te pueden servir de guía, pero te recomiendo interpretar la fuerza (o tamaño) de la correlación según el **contexto** de tu investigación. No es lo mismo analizar datos de un experimento físico controlado donde habrá poco ruido en los datos, que analizar datos sociales o biológicos donde se espera encontrar menores valores de correlación debido a la gran cantidad de dispersión o variabilidad de los datos.

¿Cómo se mide la correlación?

Veamos ahora los coeficientes de correlación más utilizados.

Tenemos el **coeficiente de correlación lineal de Pearson** que se sirve para cuantificar *tendencias lineales,* y el **coeficiente de correlación de Spearman** que se utiliza para tendencias de aumento o disminución, no necesariamente lineales pero sí *monótonas* (las variables tienden a moverse en la misma dirección relativa, pero no necesariamente a un ritmo constante; siguiente Figura).



En la figura anterior, se observa una Relación lineal y una Relación no lineal (monótona). Vemos representado con una "r" negra el coeficiente de Pearson y con una "s" en rojo el de Spearman. Cuando la relación es lineal, ambos coeficientes coinciden (valen 1), pero cuando la relación no es lineal el coeficiente de correlación de Spearman representa mejor la relación entre las variables.

El coeficiente de correlación lineal de Pearson mide una tendencia lineal entre dos variables numéricas.

Es el método de correlación más utilizado, pero asume que:

- La tendencia debe ser de tipo lineal.
- No existen valores atípicos (outliers).
- las variables deben ser numéricas Si las variables son de tipo ordinal (como las preguntas en escala de likert), no podremos aplicar la correlación de Pearson.
- tenemos suficientes datos (algunos autores recomiendan tener más de 30 puntos u observaciones).

Los dos primeros supuestos se pueden evaluar simplemente con un diagrama de dispersión, mientras que para los últimos basta con mirar los datos y evaluar el diseño que tenemos.

El coeficiente de correlación de Spearman mide una tendencia monótona (creciente o decreciente) entre dos variables. Está basado en los rangos de los valores.

En los casos donde no se cumplen los requisitos del coeficiente de correlación lineal de Pearson, es conveniente utilizar la correlación de Spearman. Es una prueba no paramétrica (no asume una distribución previa de los datos) y es más robusta frente a la presencia de outliers que la prueba paramétrica de Pearson

Valores ausentes

Por defecto, cuando tenemos en nuestros datos algún valor ausente o perdido (identificado en R por un "NA" de *Not Available*), la función cor() nos devuelve otro "NA". Si queremos evitar este error y obtener el valor de correlación, podemos especificar cómo queremos que se traten los valores ausentes en la función mediante el argumento:

- use = "pairwise.complete.obs": que calcula el coeficiente de correlación para aquellas observaciones en las que no falta ningún valor de "x" ni "y". Esto garantiza que pueda calcular la correlación para cada par de variables sin perder información debido a los valores perdidos en las otras variables.

Si bien hemos visto que la relación entre Air.Flow y stack.loss sigue una tendencia lineal. Veamos ahora, por motivos didácticos, cómo calcularemos el coeficiente de correlación de Spearman para el caso en que la tendencia fuera no lineal o monótona.

```
cor(x=Air.Flow, y=stack.loss, method="spearman")
## [1] 0.9180247
```

Utilizamos la misma función cor() que antes pero ahora especificamos que el método sea el de Spearman.

Obtenemos un valor de correlación positivo y alto, que no varía mucho de la anterior, vale r = .92. Los valores de correlación son similares debido a que si se cumplen las condiciones de la correlación de Pearson.

Graficando la correlación de variables

Al igual que los histogramas, los diagramas de correlación son una representación gráfica que muestra la relación de una variable con respecto a otra, aunque esta no tiene porque ser una relación causa-efecto.

Se relaciona el desempeño de una característica de interés con factores de causa potenciales, el objetivo es ayudar a entender las causas potenciales de variación como respuesta y explicar cómo cada factor contribuye a esa variación. Esto se consigue mediante relación estadística de la variación en la variable dependiente con una variación de la variable causa o independiente y obtener el mejor ajuste al minimizar la desviación entre lo predictivo y la respuesta real.

Este diagrama de dispersión se usa para estudiar la posible relación entre dos variables, y probar las posibles relaciones entre causa y efecto. No permite probar que una variable es causa de la otra, pero si consigue aclarar si se establecen relaciones y la intensidad que se establece entre ambas.

Los beneficios que aporta la representación de un diagrama de correlación es que puede proveer la relación entre varios factores y la respuesta de interés, por lo que nos va a ayudar a tomar decisiones con el proceso bajo estudio y finalmente una mejora del proceso.

Tiene habilidad para describir comportamientos en respuesta a datos consistentes, comparar diferentes grupos relacionados, y analizar posibles causa-efecto entre variables. Este tipo de análisis también puede contribuir a estimar la magnitud de relación entre variables y descubrir fuentes de influencia que no habían sido bien medidos u omitidos anteriormente, por lo que esta información ayuda a mejorar el sistema de medición o el proceso.

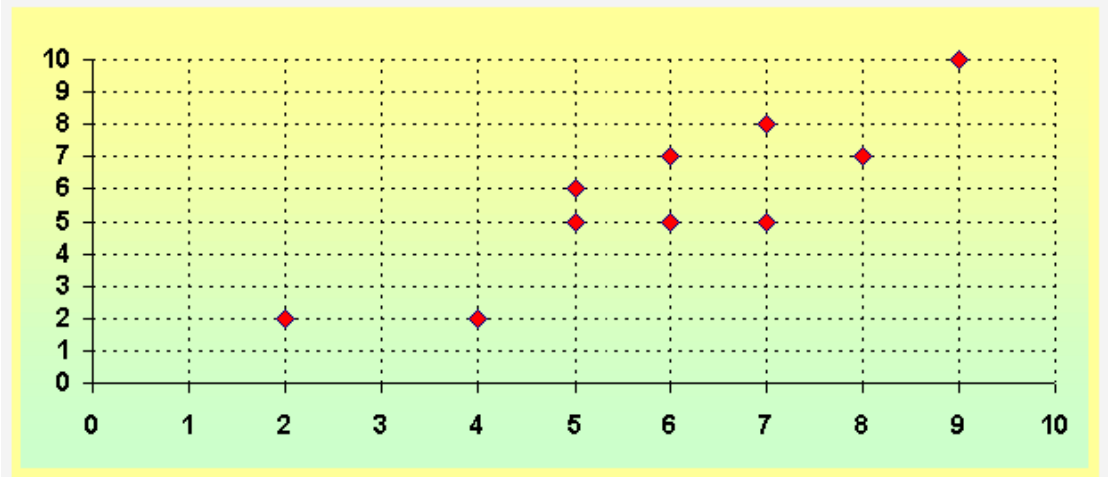


imagen:fuente propia

Aunque este tipo de estudios presenta limitaciones, pues para realizar dicho estudio se requieren conocimientos especiales para especificar el análisis de regresión adecuado. Algunas veces un problema encontrado al desarrollar los modelos de regresión es la presencia de datos los cuales su validez es cuestionable. Siempre y cuando sea posible la validez de tales datos debe ser investigada ya que pueden influir en los parámetros del método, y por tanto en los resultados finales del proceso. Es importante simplificar las variables explicativas o independientes, pues omitir una variable importante puede limitar seriamente el modelo y utilidad de los resultados.

Tablas de Contingencia

Una tabla de contingencia es una tabla que cuenta las observaciones por múltiples variables categóricas. Las filas y columnas de las tablas corresponden a estas variables categóricas.

Por ejemplo, después de una elección reciente entre dos candidatos, una encuesta de salida registró el sexo y el voto de 100 electores seleccionados de manera aleatoria y los datos se tabularon de la siguiente manera:

	Candidato A	Candidato B	Todos
Hombre	28	20	48
Mujer	39	13	52
Todos	67	33	100

Esta tabla de contingencia cuenta las respuestas según sexo y voto. El conteo en la intersección de la fila i y la columna j se denota como nij, y representa el número de observaciones que muestra esa combinación de niveles. Por ejemplo, n1,2 muestra el número de encuestados masculinos que votaron por el Candidato B.

La tabla también incluye los totales marginales para cada nivel de las variables. Los totales marginales para las filas muestran que 52 de los encuestados fueron mujeres. Los totales marginales para las columnas muestran que 67 encuestados votaron por el Candidato A. Además, el total general muestra que el tamaño de la muestra es 100.

Las tablas de contingencia también pueden revelar asociaciones entre las dos variables. Utilice una prueba de chi-cuadrada o una prueba exacta de Fisher para determinar si los conteos observados difieren significativamente de los conteos esperados bajo la hipótesis nula de que no existe asociación. Por ejemplo, usted podría probar si existe una asociación entre sexo y voto.

Las tablas de contingencia más simples son tablas de dos factores que cuentan las respuestas según dos variables. Usted puede categorizar las observaciones según tres o más variables al "cruzarlas". En el ejemplo de votación anterior, las respuestas también podrían clasificarse según el estatus de empleo de la manera siguiente:

	Candidato A	Candidato B	Total
Hombre / empleado	18	19	37
Hombre / desempleado	10	1	11
Mujer / empleada	33	10	43
Mujer / desempleada	6	3	9
Total	67	33	100

Un análisis de correspondencia simple puede detectar asociaciones en las tablas de contingencia que categorizan los datos por más de dos variables.

Gráficos Scatterplot

Los gráficos de dispersión se usan para averiguar la intensidad de la relación entre dos variables numéricas. El eje X representa la variable independiente, mientras que el eje Y representa la variable dependiente.

Los gráficos de dispersión permiten responder preguntas sobre los datos, por ejemplo: ¿cuál es la relación entre dos variables? ¿Cómo se distribuyen los datos? ¿Dónde están los valores atípicos?

Ejemplos:

Los siguientes ejemplos muestran gráficos de dispersión utilizando dos variables, tres variables y bins.

Dos Variables

Un departamento de obras públicas ha observado un aumento de fugas en la red de agua. El departamento quiere saber hasta qué punto afecta la longitud total de las tuberías al número de fugas, en comparación con el impacto de las propiedades de las tuberías, como antigüedad o circunferencia. Se puede utilizar un gráfico de dispersión para representar

el número total de fugas en comparación con la longitud total de las tuberías de cada zona.

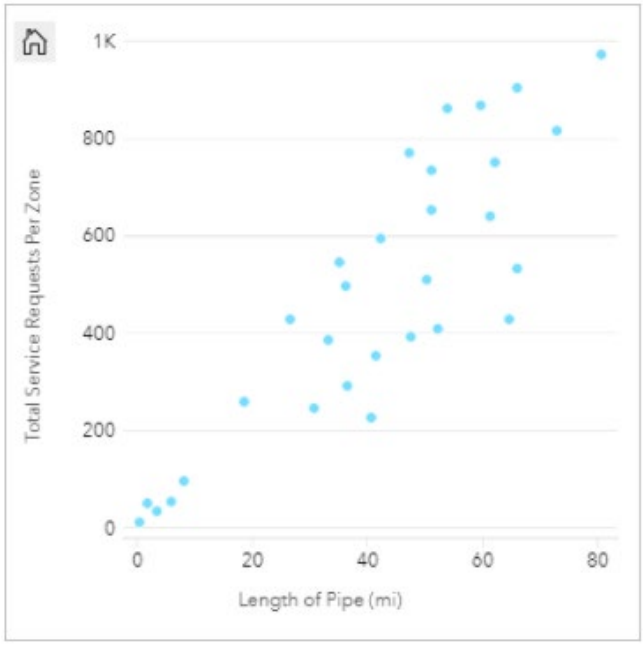
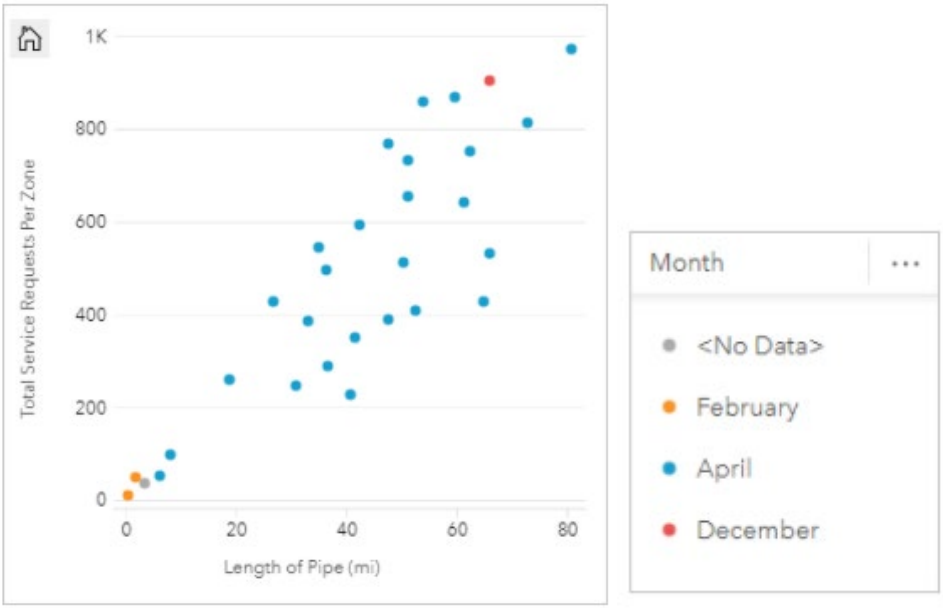


imagen:fuente propia

El departamento de obras públicas también quiere saber si hay alguna diferencia entre las tuberías inspeccionadas en distintas épocas del año. Al utilizar la opción Colorear por, el departamento puede dar estilo a los puntos con colores únicos para cada valor único en el campo especificado.



AWAKELAB

imagen:fuente propia

El gráfico de dispersión indica que la mayor parte de las inspecciones de las tuberías se realizaron en abril.

Un gráfico de dispersión puede utilizar un análisis de regresión para calcular la fuerza y la dirección de la relación entre las variables dependiente e independiente. Los modelos de estadísticas se ilustran con una línea recta o curva, dependiendo de la estadística del gráfico que se ha seleccionado. Se puede agregar el valor R2 para dar una medida del impacto de la longitud de las tuberías en el número de fugas.

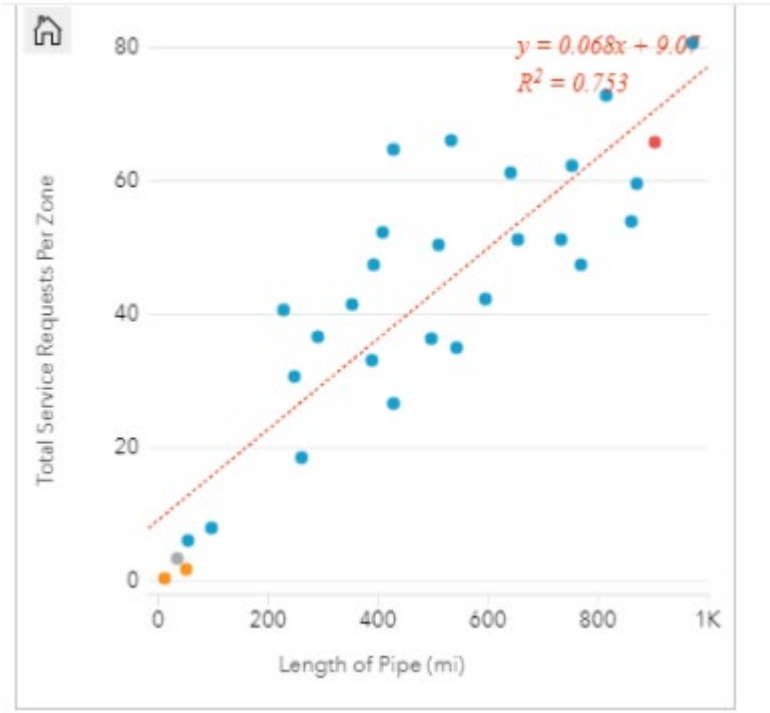


imagen:fuente propia

Agrega una tercera variable

Un departamento de obras públicas ha observado un aumento de fugas en la red de agua. El departamento quiere saber hasta qué punto afecta la longitud total de las tuberías al número de fugas, en comparación con

AWAKELAB

el impacto de las propiedades de las tuberías, como antigüedad o circunferencia. El departamento también desea saber si hay una relación entre el número de fugas o la longitud de las tuberías y el coste por día (incluidos la construcción, el mantenimiento y las reparaciones y la pérdida de recursos por las fugas). Se puede utilizar un gráfico de dispersión con símbolos proporcionales para representar el número total de fugas en comparación con la longitud total de las tuberías de cada zona; el tamaño de los puntos representaría el coste por día.

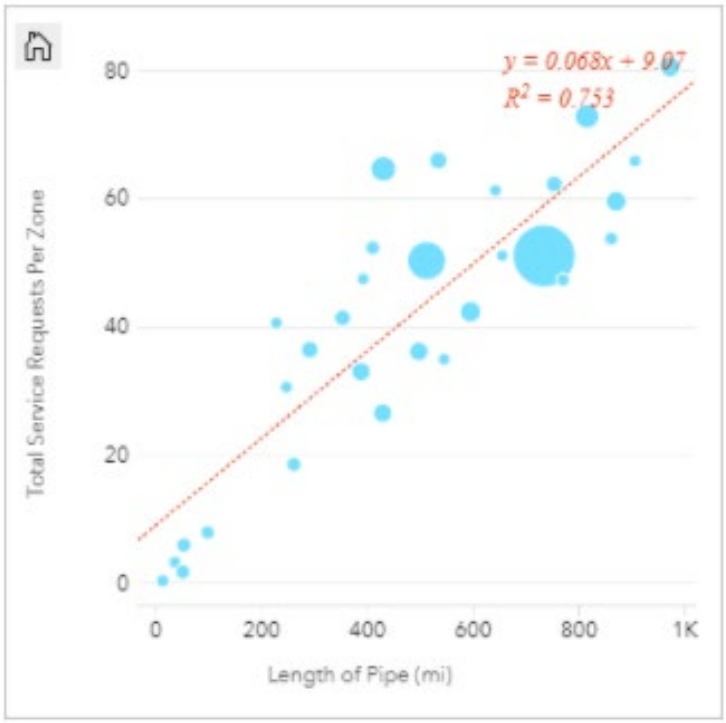


imagen:fuente propia

El departamento de obras públicas también quiere saber si hay alguna diferencia entre las tuberías inspeccionadas en distintas épocas del año. Al utilizar la opción Colorear por, puede dar estilo a los puntos con colores únicos para cada valor único en el campo especificado.

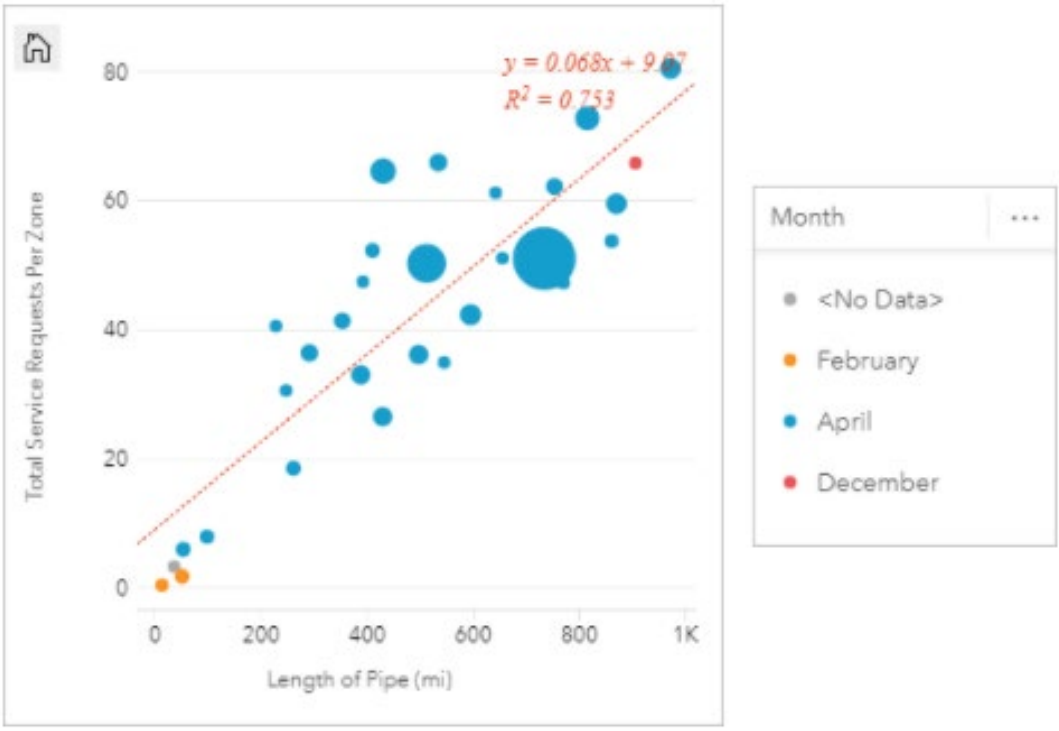


imagen:fuente propia

El gráfico de dispersión indica que la mayor parte de las inspecciones de las tuberías se realizaron en abril.

Visualizar con bins

Un analista de SIG que trabaja para un consorcio de universidades quiere averiguar qué estados tienen universidades de alto valor. El analista comienza su análisis creando un gráfico de dispersión que muestra el coste de las universidades y los ingresos medios tras la graduación. El gráfico de dispersión muestra una relación positiva, pero los puntos están distribuidos con una densidad excesiva para revelar patrones más específicos.

imagen:fuente propia

El analista puede cambiar el estilo del gráfico a Bins para ver la distribución de los puntos en el gráfico de dispersión. El patrón revela que la mayor concentración de universidades tiene un coste de en torno a los 20.000 dólares y da lugar a unos ingresos inferiores a los 50.000 dólares.

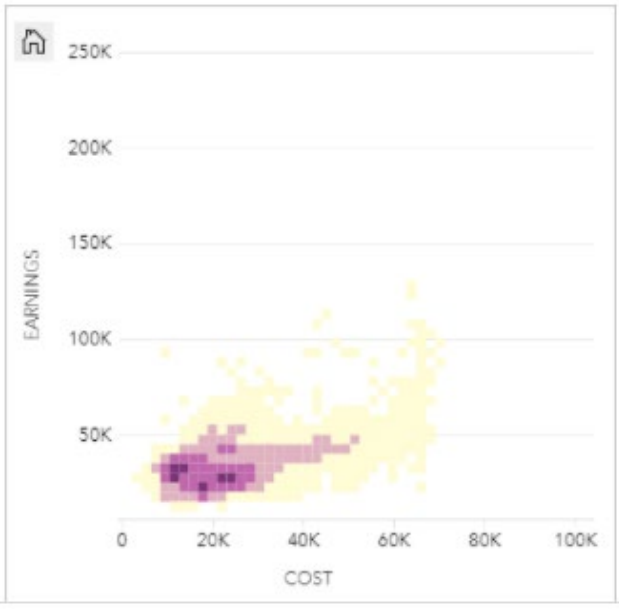


imagen:fuente propia

Midiendo la correlación de variables con el indicador Pearson

El coeficiente de correlación de Pearson es una prueba que mide la relación estadística entre dos variables continuas. Si la asociación entre los elementos no es lineal, entonces el coeficiente no se encuentra representado adecuadamente.

El coeficiente de correlación puede tomar un rango de valores de +1 a -1. Un valor de 0 indica que no hay asociación entre las dos variables. Un valor mayor que 0 indica una asociación positiva. Es decir, a medida que aumenta el valor de una variable, también lo hace el valor de la otra. Un valor menor que 0 indica una asociación negativa; es decir, a medida que aumenta el valor de una variable, el valor de la otra disminuye.

Para llevar a cabo la correlación de Pearson es necesario cumplir lo siguiente:

- La escala de medida debe ser una escala de intervalo o relación.
- Las variables deben estar distribuidas de forma aproximada.
- La asociación debe ser lineal.
- No debe haber valores atípicos en los datos.

Cálculo

La fórmula del coeficiente de correlación de Pearson es la siguiente:

$$r_{xy} = \frac{\sum z_x z_y}{N}$$

Donde:

“x” es igual a la variable número uno, “y” pertenece a la variable número dos, “zx” es la desviación estándar de la variable uno, “zy” es la desviación estándar de la variable dos y “N” es es número de datos.

Interpretación

El coeficiente de correlación de Pearson tiene el objetivo de indicar cuán asociadas se encuentran dos variables entre sí por lo que:

Correlación menor a cero: Si la correlación es menor a cero, significa que es negativa, es decir, que las variables se relacionan inversamente.

Cuando el valor de alguna variable es alto, el valor de la otra variable es bajo. Mientras más próximo se encuentre a -1, más clara será la covariación extrema. Si el coeficiente es igual a -1, nos referimos a una correlación negativa perfecta.

Correlación mayor a cero: Si la correlación es igual a +1 significa que es positiva perfecta. En este caso significa que la correlación es positiva, es decir, que las variables se correlacionan directamente.

Cuando el valor de una variable es alto, el valor de la otra también lo es, sucede lo mismo cuando son bajos. Si es cercano a +1, el coeficiente será la covariación.

Correlación igual a cero: Cuando la correlación es igual a cero significa que no es posible determinar algún sentido de covariación. Sin embargo, no significa que no exista una relación no lineal entre las variables.

Cuando las variables son independientes significa que estas se encuentran correlacionadas, pero esto significa que el resultado sea verdadero.

Ventajas y desventajas

Entre las principales **ventajas** del coeficiente de correlación de Karl Pearson se encuentran:

- El valor es independiente de cualquier unidad que se utiliza para medir las variables.
- Si la muestra es grande, es más probable la exactitud de la estimación.

Alguna de las **desventajas** del coeficiente de correlación son:

- Es necesario que las dos variables sean medidas a un nivel cuantitativo continuo.
- La distribución de las variables deben ser semejantes a la curva normal.

Casualidad v/s Correlación

La correlación examina la relación entre dos variables. Sin embargo, observar que dos variables se mueven conjuntamente no significa necesariamente que una variable sea la causa de la otra. Por eso solemos decir que "la correlación no implica causalidad".

Una correlación fuerte puede indicar causalidad, pero también es probable que existan otras explicaciones:

- Puede ser el resultado del azar: las variables parecen estar relacionadas, pero en realidad no hay una relación subyacente.
- Puede haber una tercera variable al acecho que haga que la relación parezca más fuerte (o más débil) de lo que realmente es.

En los datos observacionales, las correlaciones no pueden confirmar la causalidad...

Las correlaciones entre variables nos muestran que hay un patrón en los datos, que las variables tienden a moverse conjuntamente. Sin embargo, por sí mismas, las correlaciones no nos muestran si los datos se mueven juntos *porque una variable causa la otra*.

Es posible encontrar una correlación fiable y estadísticamente significativa entre dos variables que en realidad no tienen ninguna relación causal. ¡De hecho, estas correlaciones son comunes! A menudo esto se debe a que ambas variables están asociadas a una variable causal diferente, que tiende a darse junto a los datos que estamos midiendo.

Ejemplo: el ejercicio y el cáncer de piel

Vamos a verlo con un ejemplo. Imagine que está analizando datos sobre salud y observa una correlación positiva y estadísticamente significativa entre ejercicio y casos de cáncer de piel, esto es, que las personas que hacen más ejercicio tienden a sufrir cáncer de piel. La correlación parece significativa y fiable, y podemos observar en múltiples poblaciones de pacientes. Sin hacer más indagaciones, ¡se podría llegar a la conclusión de que el ejercicio causa cáncer! Basándose en estos resultados, incluso podría desarrollarse una hipótesis plausible: quizás el estrés del ejercicio causa que el cuerpo pierda parte de su capacidad para protegerse del daño del sol.

Pero en realidad, esta correlación podría estar presente en su conjunto de datos porque las personas que viven en lugares que tienen mucha luz solar todo el año son significativamente más activas en su vida diaria que las personas que viven en lugares con menos luz. Esto se refleja en los datos como un incremento del ejercicio. Al mismo tiempo, mayor exposición diaria a la luz solar significa que hay más casos de cáncer de piel. Ambas variables (la tasa de ejercicio y la de cáncer de piel) han sido

afectadas por una tercera variable causal (la exposición a la luz solar) pero no tenían una relación causal entre sí.

...pero con estudios empíricos bien diseñados, ¡*podemos* establecer la causalidad!

Es fundamental para el conocimiento de datos poder distinguir entre aquello que ofrece, o no, una evidencia causal. En el mundo real, la determinación de causalidad nunca es perfecta. Sin embargo, hay una variedad de técnicas experimentales, estadísticas y de diseño de estudios que sirven para encontrar evidencias de relaciones causales: p. ej., la aleatorización, los experimentos controlados y los modelos predictivos con múltiples variables. Más allá de las limitaciones intrínsecas de las pruebas de correlación (p. ej., la correlación no puede medir relaciones causales potenciales con tres variables), es importante entender que la evidencia de causalidad usualmente no procede de pruebas estadísticas individuales, sino de un diseño experimental cuidadoso.

Ejemplo: enfermedades de corazón, dieta y ejercicio

Por ejemplo, vamos a imaginar de nuevo que somos investigadores de salud y que estamos examinando un amplio conjunto de datos sobre las enfermedades cardíacas, dieta y otros hábitos de salud. Supongamos que encontramos dos correlaciones: un número mayor de enfermedades cardíacas está correlacionado con dietas más altas en grasa (correlación positiva) y una mayor cantidad de ejercicio está correlacionada con menos enfermedades cardíacas (correlación negativa). Ambas correlaciones son grandes y las encontramos de manera fiable. Seguro que esto nos da una pista sobre la causalidad, ¿verdad?

En este caso, la correlación puede sugerir una relación causal subyacente, pero sin un trabajo adicional, no la establece. Imaginemos que después de encontrar estas correlaciones, nuestro siguiente paso es diseñar un

estudio biológico que examine las maneras en las que el cuerpo absorbe la grasa y cómo afecta esto al corazón. Quizás encontremos un mecanismo a través del cual un mayor consumo de grasa se almacena de tal manera que somete al corazón a mayor presión. También podemos examinar en más profundidad el ejercicio y diseñar un experimento aleatorizado y controlado que descubra que el ejercicio interrumpe el almacenamiento de grasa, lo que reduce la presión sobre el corazón.

Todas estas evidencias encajan en una explicación: las dietas más altas en grasa realmente pueden causar enfermedades cardíacas. Y a medida que nos adentramos en el problema, las correlaciones originales siguen siendo válidas: ¡las dietas altas en grasa y las enfermedades cardíacas están relacionadas!

Pero en este ejemplo, observamos que la evidencia causal no fue facilitada por la prueba de correlación en sí, la cual simplemente estudia la relación entre datos observacionales (como el índice de enfermedades cardíacas y dieta y ejercicio reportados). En su lugar, usamos un estudio empírico para encontrar evidencias de esta asociación.

Por tanto, ¿cómo exploramos la causalidad? ¡Con un tipo de estudio adecuado!

Entender la causalidad es un tema complicado. En el mundo real, nunca podemos acceder a todos los datos que necesitaríamos para establecer todas las relaciones posibles entre variables. Pero hay algunas estrategias clave para ayudarnos a aislar y explorar los mecanismos entre diferentes variables. Por ejemplo, en un experimento controlado, podemos intentar crear dos grupos muy similares y aplicar al azar un tratamiento o intervención a uno solo de ellos.

El principio de la aleatorización es fundamental en el diseño experimental y entender este contexto puede cambiar lo que somos capaces de inferir de las pruebas estadísticas.

Vamos a ver de nuevo el primer ejemplo de arriba, que examinaba la relación entre ejercicio y tasas de cáncer de piel. Imaginemos que de alguna manera podemos seleccionar una muestra grande de personas distribuidas en todo el mundo y asignarles al azar que hagan ejercicio a diferentes niveles cada semana durante diez años. Al finalizar este periodo, también recopilamos las tasas de cáncer de piel de este gran grupo. Al final, tendríamos un conjunto de datos diseñado experimentalmente para probar la relación entre ejercicio y cáncer de piel. Como el ejercicio ha sido manipulado directamente en el experimento mediante asignación aleatoria, no estará sistemáticamente relacionado con ninguna otra variable que podría ser diferente entre estos dos grupos (asumiendo que todos los demás aspectos del estudio son válidos). Esto significa que en este caso, como nuestros datos se han obtenido a través de un diseño experimental sólido, una correlación positiva entre ejercicio y cáncer de piel *sí* que sería una evidencia significativa de causalidad.

AWAKELAB

Regresión Lineal

La regresión lineal es una de las técnicas más usadas en Machine Learning. Su fortaleza estriba en su simplicidad e interpretabilidad. La regresión polinómica, como ya veremos, es una extensión de la regresión lineal.

La regresión lineal es una técnica paramétrica de machine learning. Con «paramétrica» queremos decir que incluso antes de mirar a los datos, ya sabemos cuántos parámetros (o coeficientes) vamos a necesitar.

En el caso que estemos usando una sola variable, x, sabemos que una línea necesita 2 parámetros. La fórmula para la regresión lineal con una sola variable x es:

$$y = wx + b$$

El aprendizaje consiste en encontrar cuáles son los mejores parámetros (coeficientes) para los datos que tengamos. Los mejores coeficientes serán los que minimicen alguna medida de error. Para la regresión lineal usaremos el error cuadrático medio.

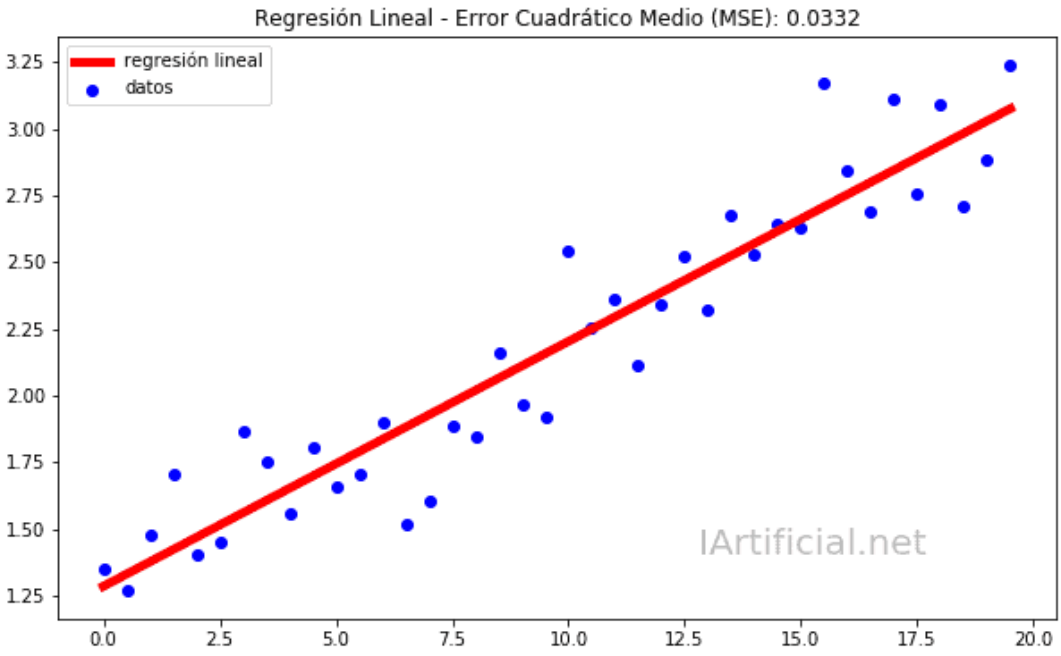


imagen:www.IArtificial.net

AWAKELAB

Hemos usado una regresión lineal para encontrar los parámetros de la línea que minimiza el error de los datos que tenemos. El proceso de aprendizaje consiste en estimar los parámetros *w* y *b*. Así nos queda que para estos datos, los mejores valores son:

$$w=0.0918b=1.2859$$

así que nos queda:

$$y = 0.0918x + 1.2859$$

Podemos usar este modelo de regresión lineal para estimar cuáles serán los resultados para otros valores de *x*. Por ejemplo, si queremos saber el resultado para $x = 5$, usaremos el modelo anterior y veremos que el resultado es 1.7449:

$$y=0.0918 \cdot 5+1.2859=1.7449$$

Este es un ejemplo muy simple. En realidad, los problemas de machine learning tienen muchas más variables. Sin embargo, he escogido este ejemplo porque es muy fácil de visualizar, explicar y entender. Espero que la intuición de este ejemplo sirva para entender lo que está pasando cuando haya más variables.

Notación

Antes de explicar el método de los mínimos cuadrados para resolver regresiones lineales, tenemos que expandir la notación. Debemos tener en cuenta que normalmente, tendremos muchas variables.

Con una variable, la ecuación para la regresión lineal es:

$$y = wx + b$$

Cuando tengamos un dato con *N* variables, llamaremos al dato *X*. También tenemos que expandir los parámetros *W* para que cada

parámetro vaya con una variable:

X = [x_0, x_1, x_2, ..., x_N] \\ W = [w_0, w_1, w_2, ..., w_N]

X=[x0,x1,x2,...,xN]W=[w0,w1,w2,...,wN]

Si hacemos que

y = b + wx = w_0 x_0 + w_1 x_1

y=b+wx=w0x0+w1x1

Para el caso general, la ecuación lineal quedaría:

y = WX

usando el producto matricial. Si tienes la multiplicación de matrices un poco oxidada, la versión intuitiva sería:

y = w_0 x_0 + w_1 x_1 + w_2 x_2 + ... + w_N x_N

y=w0x0+w1x1+w2x2+...+wNxN

Aprendizaje: El método de los mínimos cuadrados

El proceso de aprendizaje consiste en averiguar qué parámetros *W* minimizan el error cuadrático medio entre los resultados reales y los estimados.

El método de los mínimos cuadrados proporciona una solución analítica. Es decir, nos da una fórmula para obtener la mejor estimación posible de *W* para los datos de entrada y resultados que hemos proporcionado. La fórmula es la siguiente:

$$\widehat{W} = (X^T X)^{-1} X^T X \quad \& \quad \widehat{W} = (X^T X)^{-1} X^T y$$

En la práctica hay librerías numéricas que calculan automáticamente la mejor estimación de *W* por nosotros. Ya veremos algún ejemplo práctico del cálculo de regresión lineal.

De momento, sólo quería indicar que dependiendo de la cantidad de datos y atributos, puede ser una operación costosa computacionalmente hablando. Fíjate que hay que transponer matrices, multiplicar matrices e invertir matrices. Todo ello muy costoso computacionalmente para grandes cantidades de datos.

Otras formas de resolver el problema de la regresión lineal

El método de los mínimos cuadrados no es la única forma de estimar los mejores parámetros *W*. También podemos utilizar métodos de optimización numérica tales como el gradiente descendiente.

El gradiente descendiente va a servir no sólo para resolver regresiones lineales y polinómicas sino que es también fundamental para el aprendizaje automático de redes neuronales y aprendizaje profundo.

Ejemplo

Carl Friedrich Gauss es famoso, entre otras muchas contribuciones, por la distribución Gaussiana o el método de los mínimos cuadrados para resolver regresiones lineales.

Cuentan que la primera aplicación del método de los mínimos cuadrados fue la determinación de la posición de Ceres. Ceres es un asteroide que se «perdió» a los 40 días de descubrirse. Realmente no se perdió, sino que al acercarse a la claridad del Sol, dejó de verse.

Varios científicos y astrónomos intentaron localizar Ceres. La única información que tenían eran los datos de su observación durante 40 días. Gauss fue el único capaz de predecir dónde se encontraría el asteroide Ceres cuando abandonó la parte del firmamento tan iluminada por el Sol. Para ello, Gauss usó el método de los mínimos cuadrados.

A finales de 1801 los astrónomos encontraron el asteroide Ceres exactamente donde Gauss predijo que estaría.

Regresión Lineal en Python

Para hacer una regresión lineal en Python, vamos a usar scikit-learn, que es una librería de python para aprendizaje automático. En particular, la clase *LinearRegression* implementa la funcionalidad descrita en la parte teórica. Vamos a explicarlo con un ejemplo.

Ejemplo:

Primero vamos a generar unos datos que siguen una línea, y le añadimos ruido gaussiano. Para ello usaremos la librería de python NumPy. La fórmula que he usado para generar los datos es:

$$y = 0.1x + 1.25 + N(0, 0.2)$$

$$y=0.1x+1.25+N(0,0.2)$$

El código en python quedaría así:

```
import numpy as np #librería numérica

import matplotlib.pyplot as plt #para crear gráficos con matplotlib

%matplotlib inline # Si quieres hacer estos gráficos dentro de un jupyter notebook


from sklearn.linear_model import LinearRegression # Regresión lineal con scikit-learn


def f(x): # función f(x) = 0.1*x + 1.25 + 0.2*Ruido_Gaussiano

   np.random.seed(42) # para poder reproducirlo

   y = 0.1*x + 1.25 + 0.2*np.random.randn(x.shape[0])

   return y


x = np.arange(0, 20, 0.5) # generamos valores x de 0 a 20 en intervalos de 0.5

y = f(x) # calculamos y a partir de la función que hemos generado


# hacemos un gráfico de los datos que hemos generado

plt.scatter(x,y,label='data', color='blue')

plt.title('Datos')
```

Este código genera los datos que se ven en la siguiente imagen:

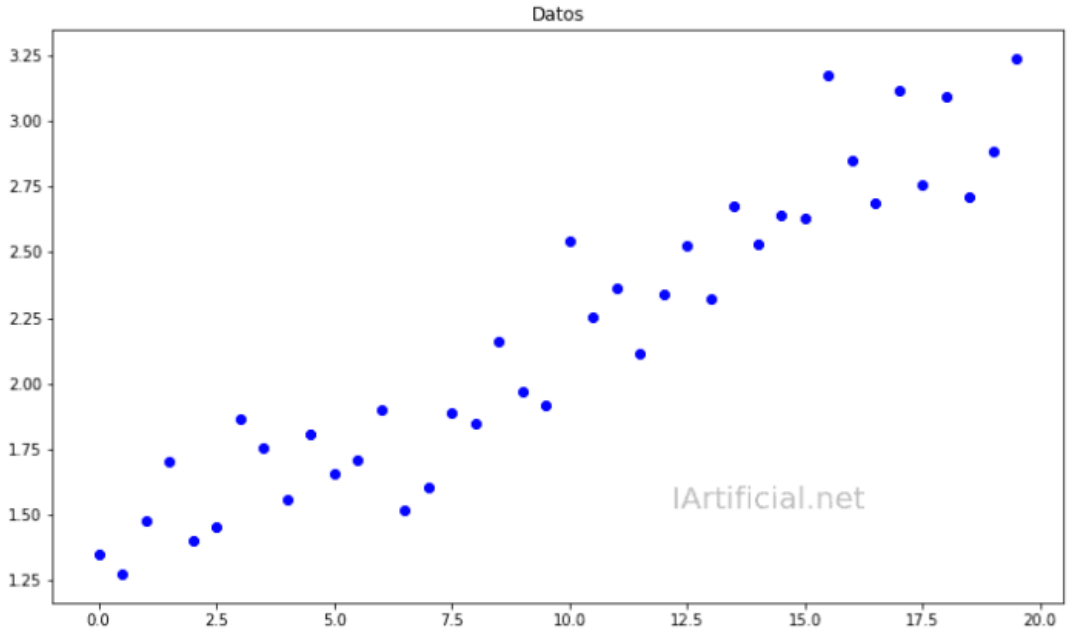


imagen:www.IArtificial.net

Datos de ejemplo para hacer una regresión lineal: y = 0.1*x + 1.25 + *N*(0, 0.2)

Para entrenar el modelo, simplemente tendremos que hacer uso de scikit-learn. El método *fit* se encarga de ajustar los parámetros de regresión lineal a los datos.

```
#importamos la clase de Regresion Lineal de scikit-learn

from sklearn.linear_model import LinearRegression


regresion_lineal = LinearRegression() # creamos una instancia de LinearRegression


# instruimos a la regresion lineal que aprenda de los datos (x,y)

regresion_lineal.fit(x.reshape(-1.1), y)


# vemos los parámetros que ha estimado la regresión lineal

print ('w = ' + str(regresion_lineal.coef_) + ', b =' + str(regresion_lineal.intercept_))


# resultado: w= [0.09183522], b = 1.2858792525736682
```

Como vemos, la regresión lineal casi ha averiguado cómo hemos generado los datos:

- Estima 0.092 en lugar de 0.1 para *w*
- Estima 1.286 en vez de 1.25 para *b*

Este pequeño error es normal debido a la cantidad de ruido gaussiano que hemos introducido y al hecho de que hay muy pocos datos.

Prediciendo regresión Lineal

Una vez que tenemos entrenado el modelo de regresión lineal, podemos hacer predicciones usando el método ***predict*** de la clase *LinearRegression*. Por ejemplo, si quisiéramos saber qué valor de *y* corresponde para *x=5* usamos este código.

```
# vamos a predecir y = regresion_lineal(5)

nuevo_x = np.array([5])

prediccion =
regresion_lineal.predict(nuevo_x.reshape(-1.1))


print(prediccion)


# resultado: [1.7449]
```

Vamos a evaluar la calidad del modelo aprendido usando solamente los datos de entrenamiento. Recuerda que en un problema real, hay que evaluar también la capacidad de generalización del modelo. Podemos evaluar la calidad del modelo midiendo el error cuadrático medio y el coeficiente de determinación R^2.

```
# importamos el cálculo del error cuadrático medio
(MSE)

from sklearn.metrics import mean_squared_error


# Predecimos los valores y para los datos usados en
el entrenamiento

prediccion_entrenamiento =
regresión_lineal.predict(x.reshape(-1,1))


# Calculamos el Error Cuadrático Medio (MSE= Mean
Squared Error)

mse = mean_squared_error(y_true = y, y_pred =
prediccion_entrenamiento)


@ La raíz cuadrada del MSE es el RMSE

rmse = np.sqrt(mse)


print('Error Cuadrático Medio (MSE) = ' + str(mse))

print('Raíz del Error Cuadrático Medio (RMSE) = ' +
str(rmse)
```

Nos da el siguiente resultado:

```
Error Cuadrático Medio (MSE) = 0.033
Raíz del Error Cuadrático Medio (RMSE) = 0.182
```

Coeficiente de Determinación R^2

El coeficiente de determinación R^2 determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo

El rango de R^2 está entre 0 y 1, siendo 1 lo mejor. Para medir el coeficiente de determinación R^2 de la regresión lineal usaremos el método ***score***.

```
# calculamos el coeficiente de determinación r2
r2 = regresion_lineal.score(x.reshape(-1,1), y)

print('Coeficiente de Determinación R2 = ' +str(r2))
```

Nos da el siguiente resultado, que es bastante bueno:

Siempre que podamos, es bueno visualizar los resultados, para saber cómo se está comportando el modelo. Este ejemplo que estamos haciendo es muy simple, y por tanto, muy fácil de visualizar. La línea que el modelo ha aprendido siguiendo el método de los mínimos cuadrados aparece en el siguiente gráfico en rojo.

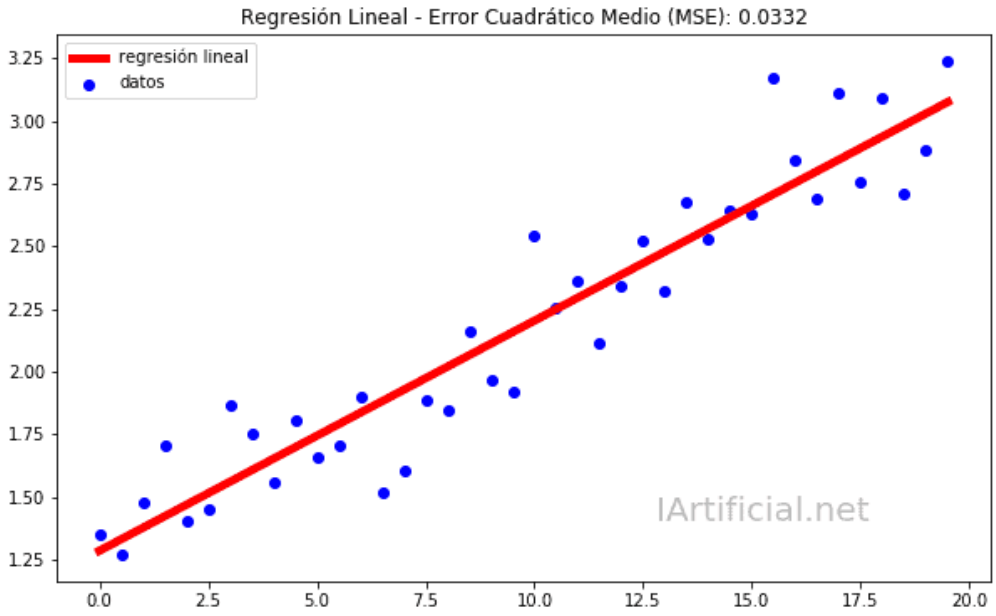


imagen:www.IArtificial.net

Librería statsmodels

statsmodels es un módulo de Python que proporciona clases y funciones para la estimación de muchos modelos estadísticos diferentes, así como para realizar pruebas estadísticas y exploración de datos estadísticos. Hay disponible una lista extensa de estadísticas de resultados para cada estimador. Los resultados se comparan con los paquetes estadísticos existentes para garantizar que sean correctos.

Cargando Módulos y Funciones

Después de instalar statsmodels y sus dependencias, cargamos algunos módulos y funciones:

```
from __future__ import print_function
import statmodels.api as sm
import pandas
from patsy import dmatrices
```

Pandas se basa en listas para proporcionar estructuras de datos enriquecidas y herramientas de análisis de datos. La función pandas.DataFrame proporciona matrices etiquetadas de datos.

La función pandas.read_csv se puede usar para convertir un archivo de valores separados por comas en un objeto DataFrame.

patsy es una biblioteca de Python para describir modelos estadísticos y construir matrices de diseño utilizando fórmulas tipo R.

Datos

Descargamos el conjunto de datos de Guerry, una recopilación de datos históricos utilizados en apoyo del Ensayo sobre las estadísticas morales de Francia de 1833 de Andre-Michel Guerry. El conjunto de datos se hospeda en línea en el formato de valores separados por comas (CSV) del repositorio de Rdatasets.

Podríamos descargar el archivo localmente y luego cargarlo usando read_csv, pero pandas se encarga de todo esto automáticamente por nosotros:

```
df = sm.datasets.get_rdataset("Guerry",
"HistData").data
```

Seleccionamos las variables de interés y miramos las 5 últimas filas:

```
vars = ['Department', 'Lottery', 'Literacy',
'Wealth', 'Region']
df = df[vars]
df[-5:]
```

Resultado

	Department	Lottery	Literacy	Wealth	Region
81	Vienne	40	25	68	W
82	Haute-Vienne	55	13	67	C
83	Vosges	14	62	82	E
84	Yonne	51	47	30	C
85	Corse	83	49	37	NaN

Observe que falta una observación en la columna Region. Lo eliminamos utilizando un método DataFrame provisto por pandas:

```
df = df.dropna()
df[-5:]
```

Resultado:

	Department	Lottery	Literacy	Wealth	Region
80	Vendee	68	28	56	W
81	Vienne	40	25	68	W
82	Haute-Vienne	55	13	67	C
83	Vosges	14	62	82	E
84	Yonne	51	47	30	C

Problema y Modelo

Queremos saber si las tasas de alfabetización en los 86 departamentos franceses están asociadas con las apuestas en la Lotería Real en la década de 1820. Necesitamos ver el nivel de riqueza en cada departamento, y también queremos incluir una serie de variables ficticias en el lado derecho de nuestra ecuación de regresión para controlar la heterogeneidad no observada debido a los efectos regionales.

El modelo se estima utilizando la regresión de mínimos cuadrados ordinarios o MCO (OLS en inglés).

Para ajustar la mayoría de los modelos cubiertos por statsmodels, deberá crear dos matrices de diseño. La primera es una matriz de variables endógenas (es decir, dependiente, respuesta, regresión y etc.). El segundo es una matriz de variables exógenas (es decir, independiente, predictor, regresor, etc.).

El módulo patsy proporciona una función conveniente para preparar matrices de diseño utilizando fórmulas tipo R.

Usamos la función dmatrices de patsy para crear matrices de diseño:

```
y, X = dmatrices('Lottery ~ Literacy + Wealth +
Region', data=df, return_type='dataframe')

y[:3]
```

Resultado:

```
   Lottery
0     41.0
1     38.0
2     66.0

   Intercept  Region[T.E]  Region[T.N]  Region[T.S]  Region[T.W]  Literacy  \
0        1.0          1.0          0.0          0.0          0.0      37.0
1        1.0          0.0          1.0          0.0          0.0      51.0
2        1.0          0.0          0.0          0.0          0.0      13.0

   Wealth
0    73.0
1    22.0
2    61.0
```

Ajustar el modelo y resumir

El ajuste de un modelo en statsmodels generalmente implica 3 pasos sencillos:

- Usa la clase modelo para describir el modelo
- Ajustar el modelo utilizando un método de clase.
- Inspeccionar los resultados utilizando un método de resumen.

Para regresión de mínimos cuadrados ordinarios, esto se logra mediante:

```
mod = sm.OLS(y, X)      # Describe el modelo
res = mod.fit()         # Ajusta el modelo
print(res.summary())    # Resumen el modelo
```

Resultado:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.338
Model:                            OLS   Adj. R-squared:                  0.287
Method:                 Least Squares   F-statistic:                     6.636
Date:                Mon, 14 May 2018   Prob (F-statistic):           1.07e-05
Time:                        21:48:07   Log-Likelihood:                -375.30
No. Observations:                  85   AIC:                             764.6
Df Residuals:                      78   BIC:                             781.7
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      38.6517      9.456      4.087      0.000      19.826      57.478
Region[T.E]   -15.4278      9.727     -1.586      0.117     -34.793       3.938
Region[T.N]   -10.0170      9.260     -1.082      0.283     -28.453       8.419
Region[T.S]    -4.5483      7.279     -0.625      0.534     -19.039       9.943
Region[T.W]   -10.0913      7.196     -1.402      0.165     -24.418       4.235
Literacy       -0.1858      0.210     -0.886      0.378      -0.603       0.232
Wealth          0.4515      0.103      4.390      0.000       0.247       0.656
==============================================================================
Omnibus:                        3.049   Durbin-Watson:                   1.785
Prob(Omnibus):                  0.218   Jarque-Bera (JB):                2.694
Skew:                          -0.340   Prob(JB):                        0.260
Kurtosis:                       2.454   Cond. No.                         371.
==============================================================================
```

imagen:fuente propia

El objeto res tiene muchos atributos útiles. Por ejemplo, podemos extraer estimaciones de parámetros y r-cuadrado escribiendo:

```
res.params
```

Escriba dir(res) para obtener una lista completa de atributos.

Diagnóstico y test

También puedes realizar una variedad de diagnósticos de regresión útiles y tests. Por ejemplo, aplica la prueba Rainbow para la linealidad (la hipótesis nula es que la relación se modela correctamente como lineal):

```
sm.stats.lineal_rainbow(res)
```

Resultado

```
(0.8472339976156913, 0.6997965543621643)
```

statsmodels también proporciona funciones gráficas. Por ejemplo, podemos dibujar una gráfica de regresión parcial para un conjunto de regresores por:

```
sm.graphics.plot_partregress('Lottery', 'Wealth',
['Region', 'Literacy'], data = df, obs_labels =
False)
```

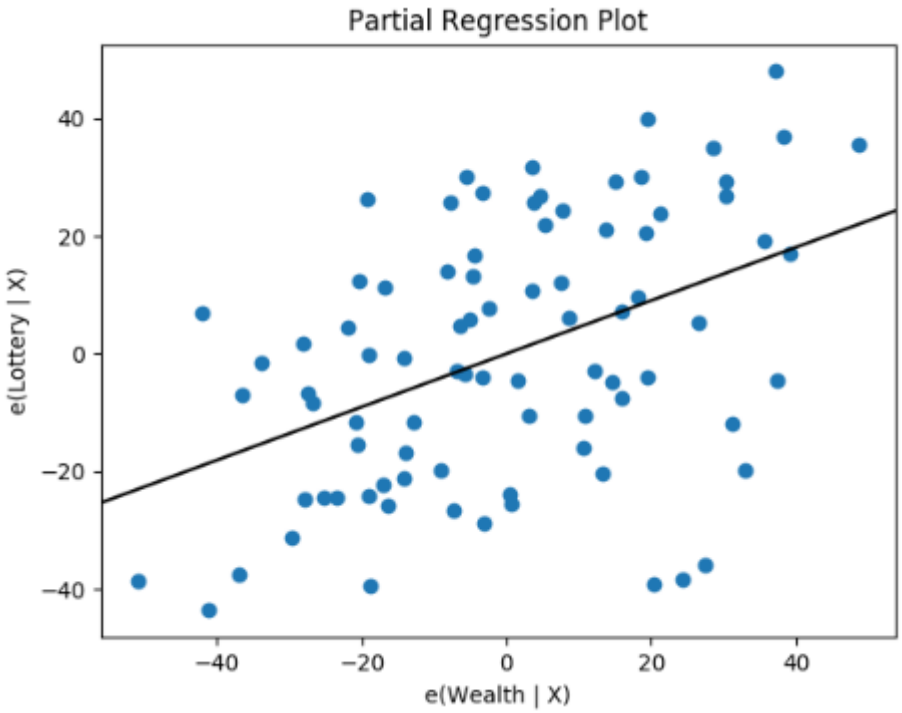


imagen: fuente propia

El concepto de Regresión

El análisis de regresión es una herramienta de frecuente uso en estadística. La cual permite investigar las relaciones entre diferentes variables cuantitativas. Esto, mediante la formulación de ecuaciones matemáticas.

Visto de otro modo, dicho análisis es un proceso o **modelo** que analiza el vínculo entre una variable dependiente y una o varias variables independientes. Así, a partir de dicho estudio, se halla una relación matemática.

Gracias a los procesos de regresión, es posible entender el modo en que la variable dependiente es afectada por cambios en los demás factores.

Aplicación

Una de las principales aplicaciones del análisis de regresión es la proyección con diferentes escenarios. Esto, teniendo en cuenta el grado de influencia (en estadística se conoce a esto como correlación) sobre la variable dependiente.

Es decir, el objetivo del análisis es construir una función que permita estimar el valor futuro de la variable de estudio.

Desde otro punto de vista, la regresión permite calcular una esperanza (promedio) condicional. Para ese fin, se toman como dados los valores de las variables independientes.

Cabe precisar que cuando se tiene en cuenta solo una variable independiente hablamos de regresión lineal simple. En cambio, si se incluyen más factores, se trataría de una regresión lineal múltiple.

El análisis de regresión tiene aplicaciones para la vida cotidiana. Esto, desde el estudio de accidentes de tráfico en una determinada zona geográfica hasta comprobar si un plan de estudios es recomendable según la tasa de abandono escolar, por ejemplo.

Crítica al análisis de regresión

Una crítica común a este tipo de modelo de predicción matemática es que no es óptimo, pues suele confundir **correlación con causalidad**.

Lo anterior quiere decir que se puede establecer, por ejemplo, una relación matemática entre el crecimiento económico y la frecuencia de lluvias en un país. Sin embargo, si no hay un fundamento teórico que vincule esas variables, el estudio carece de relevancia porque se trata de una **relación espuria**.

Ejemplo de análisis de regresión

Veamos un ejemplo muy simple del análisis de regresión. Supongamos que una empresa quiere calcular la demanda por una determinada mercancía.

Como variable independiente, tomaremos el precio del producto. Entonces, la compañía, en base a sus datos históricos, construye una ecuación como la siguiente:

$$D = ap + b$$

Donde:

D: Cantidad demanda

p: Precio del producto

Así, el análisis de regresión tiene como objetivo hallar los valores de a (**coeficiente de correlación lineal**) y b.

Determinación de los coeficientes de regresión

El **coeficiente de determinación** es la proporción de la varianza total de la variable explicada por la regresión. Es también denominado R cuadrado y sirve para reflejar la bondad del ajuste de un modelo a la variable que se pretende explicar.

El coeficiente de determinación puede adquirir resultados que oscilan entre 0 y 1. Así, cuando adquiere resultados más cercanos a 1, mayor resultará el ajuste del modelo a la variable que se pretende aplicar para el caso en concreto. Por el contrario, cuando adquiere resultados que se acercan al valor 0, menor será el ajuste del modelo a la variable que se pretende aplicar y, justo por eso, resultará dicho modelo menos fiable.

La **fórmula del coeficiente de determinación** es la siguiente:

Se trata de una fórmula de gran complejidad, por lo que lo mejor es ir por partes. En primer lugar, analizaremos el **numerador**, ya que la fórmula del coeficiente de determinación es una fracción. El numerador de esta fórmula es la expresión de la varianza, si bien existen dos diferencias de gran importancia. En primer lugar, la Y lleva un circunflejo, lo cual significa que la Y es la estimación de un modelo sobre lo que, de acuerdo con las variables aplicables al respecto, vale Y. Así, no se trata del valor real de Y, sino de una estimación del valor de Y. En segundo lugar, falta la división entre N (el número de observaciones en la fórmula a través de la cual se calcula la varianza).

Ahora analizaremos el **denominador** de la fracción que conforma la fórmula del coeficiente de determinación. En este caso, la única diferencia que existe respecto de la fórmula de la varianza es que no se aplica su denominador. Así, no hay una división entre N (número de observaciones).

Interpretación

Ya hemos dicho que el coeficiente de determinación adquiere unos valores que oscilan entre 0 y 1, pero ¿cómo se han de interpretar los resultados? ¿Cómo se representan en una gráfica? Vamos a ello.

Con un ejemplo estos términos se entienden mejor, por lo que vamos a poner uno. Supongamos que queremos analizar la cantidad de canasta que anota Pau Gasol según la cantidad de partidos de baloncesto en los que juega. Como resulta lógico, podemos suponer que cuántos más partidos juegue Pau Gasol, mayores serán las canastas que anote. Si representamos estos datos en una gráfica, su pendiente sería directa y ascendente, resultando, por tanto, una relación positiva. Así, cuántos más partidos Pau Gasol, más canastas anota. El ajuste, si tenemos en cuenta los datos se acercará bastante al valor de 1, lo que quiere decir que se trata de un modelo cuyas estimaciones se ajustan de forma bastante

correcta a la variable real. Así, si el resultado fuera de 0,8, podríamos decir -si bien no es técnicamente correcto utilizar esta expresión- que el modelo explica en un 80% la variable real.

Indicador de ajuste R2

El coeficiente de determinación es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también llamado R cuadrado, refleja la bondad del ajuste de un modelo a la variable que pretende explicar.

Es importante saber que el resultado del coeficiente de determinación oscila entre 0 y 1. Cuanto más cerca de 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que estamos intentando explicar. De forma inversa, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, menos fiable será.

$$R^2 = \frac{\sum_{t=1}^{T} (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2}$$

En la expresión anterior tenemos una fracción. Así pues, vayamos por partes. En primer lugar, analizaremos el numerador, es decir, la parte de arriba.

$$\sum_{t=1}^{T} (\hat{Y}_t - \bar{Y})^2$$

Para aquellos que no conozcan la expresión de la **varianza**, les recomiendo que lean el artículo sobre la misma. Para aquellos que sí la conozcan, podrán caer en la cuenta de que es la expresión de la varianza, pero con dos diferencias fundamentales.

La primera diferencia es que la Y lleva un circunflejo o lo que los profesores llaman de forma didáctica "sombrerito". Ese sombrerito lo que detalla es que esa Y es la estimación de un modelo sobre lo que según las variables explicativas vale Y, pero no es el valor real de Y, sino una estimación de Y.

En segundo lugar, faltaría dividir entre T. Que, en otros casos, se nota como N o número de observaciones. Sin embargo, dado que la fórmula del denominador también la llevaría, eliminamos los denominadores (parte de abajo) de ambas fórmulas para simplificar la expresión. De esta manera es más fácil trabajar con ella.

A continuación, vamos a realizar el mismo análisis con la parte del denominador (parte de abajo).

$$\sum_{t=1}^{T}(Y_t - \bar{Y})^2$$

En este caso, la única diferencia existente respecto a la fórmula original de la varianza es la ausencia de su denominador. Es decir, no dividimos entre T o N. De esta manera, una vez explicadas las dos partes de la expresión genérica del R cuadrado o coeficiente de determinación, vamos a ver un ejemplo.

Interpretación

Supongamos que queremos explicar la cantidad de goles que anota Cristiano Ronaldo según la cantidad de partidos que juega. Suponemos que, a mayor cantidad de partidos jugados, más goles meterá. Los datos pertenecen a las últimas 8 temporadas. De tal manera, tras extraer los datos, el modelo arroja la siguiente estimación:

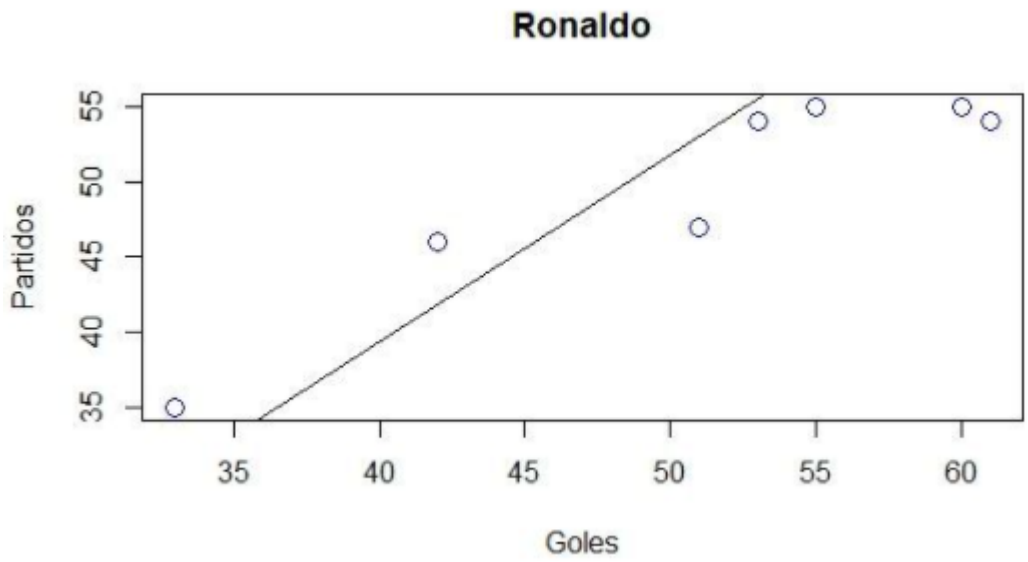


imagen: fuente propia

Cómo podemos ver en el gráfico, la relación es positiva. A más partidos jugados, como es lógico, más goles anota en la temporada. El ajuste, según el cálculo del R cuadrado, es de 0,835. Esto quiere decir que es un modelo cuyas estimaciones se ajustan bastante bien a la variable real. Aunque técnicamente no sería correcto, podríamos decir algo así como que el modelo explica en un 83,5% a la variable real.

Referencias

[1] Correlación

https://www.incibe-cert.es/blog/correlacion-herramientas-analisis-datos

[2] Diagrama de Dispersión y correlación lineal

https://www.youtube.com/watch?v=rv0Xy8edFRg

[3] Tablas de Contingencia

https://www.youtube.com/watch?v=CEWJZxFpC8w

https://conceptosclaros.com/que-es-tabla-contingencias-par-que-sirve/

[4] Coeficiente de Pearson

https://www.uv.es/webgid/Descriptiva/31_coeficiente_de_pearson.html

[5] Regresión Lineal

https://www.youtube.com/watch?v=SsFBnvkoZa4