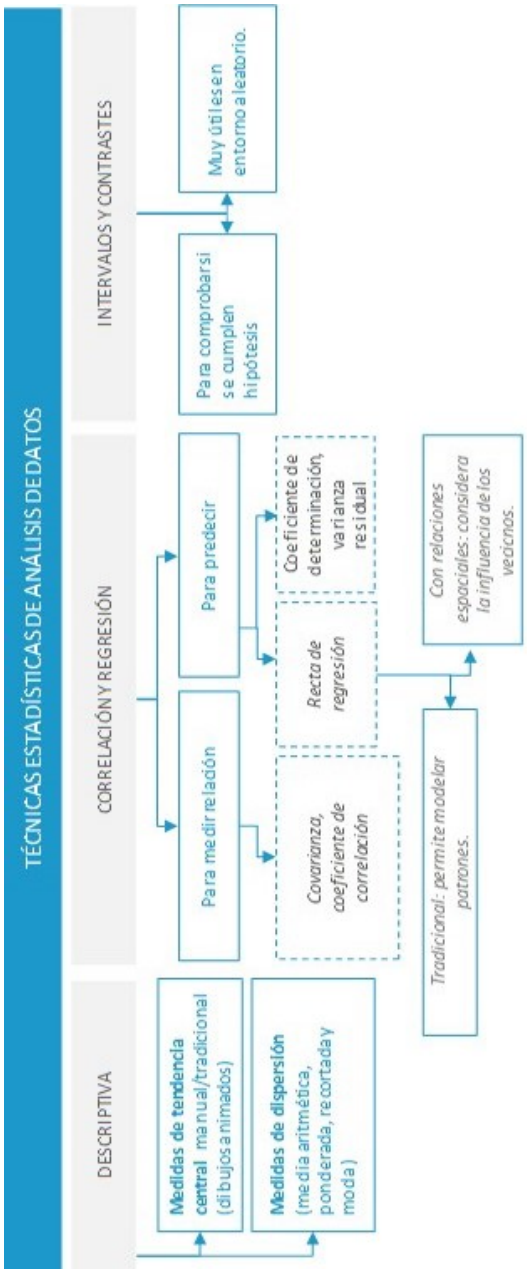
Análisis de Datos Masivos para el Negocio

Tema 3. Técnicas estadísticas de análisis de datos

Índice

Esquema



3.1. Introducción y objetivos

El gerente o administrador de una empresa debe de tomar decisiones de forma frecuente, muchas veces, estas decisiones son casi automáticas derivadas del funcionamiento continuo de la propia empresa, pero otras veces, se hace necesario un razonamiento o estudio mucho más profundo para tomar un camino u otro. Esta decisión suele venir determinada, tanto por la propia naturaleza de dicha decisión, como por multitud de factores internos y externos, por lo que es necesario contar con instrumentos para llegar a decisiones óptimas para la empresa.

Es por ello, que **la estadística y la econometría** pueden sernos de gran ayuda para esta toma de decisiones. La econometría puede entenderse como un híbrido entre la estadística y la economía. Mientras esta última es una ciencia que trata de dar soluciones a los agentes económicos a los problemas de escasez de forma racional, la econometría es una ciencia que desarrolla métodos basados en técnicas estadísticas para explicar los comportamientos de los individuos creando modelos que permitan describir los problemas económicos y aportar soluciones basadas en los datos disponibles.

El lector llegado a este punto puede pensar que la tarea de la econometría es colosal ya que existen una cantidad cuasi infinita de problemas, y desarrollar metodologías, herramientas o procedimientos para resolver dichos problemas puede ser una tarea imposible. Sin embargo, la econometría trata de utilizar metodologías que puedan ser aplicables a un amplio abanico de problemas ajustando las herramientas en función a las características de cada problema.

En este tema se hace una introducción y/o repaso por las técnicas estadísticas y econométricas de análisis de datos. En la primera parte del tema se define la descripción estadística de los datos, así como el análisis de correlación, útil para detectar si dos variables cuentan o no con una relación.

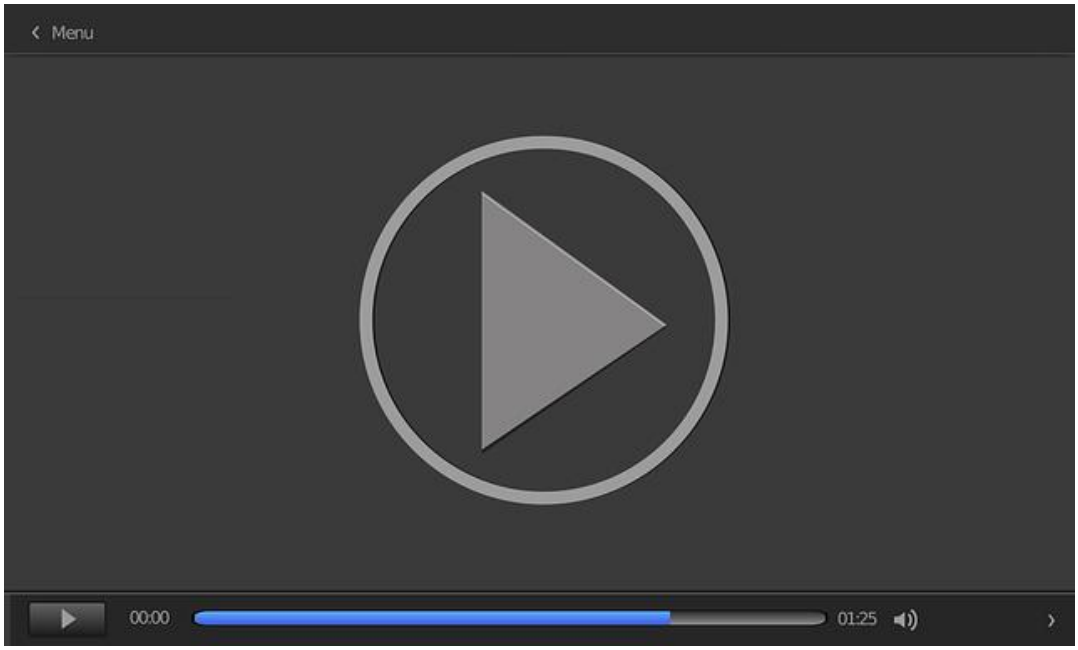
Ideas clave

Una vez que se conoce la descripción estadística de los datos, así como la existencia de relación entre ellos, es importante conocer el análisis de regresión, intervalos y contrastes para poder hacer predicciones fiables de una determinada variable. Todos estos aspectos son tratados en la segunda parte del tema.

En este contexto, con los conocimientos adquiridos en este tema, el alumno será capaz de:

- Conocer la descripción de un conjunto de datos, así como la existencia de correlación entre ellos.
- Predecir y contrastar hipótesis sobre una determinada variable.

Vídeo *Técnicas estadísticas de análisis de datos.*



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=bcd10264-123f-4e77-a7e1-b15d00aa2b13

Ideas clave

3.2. Estadística descriptiva

Antes de tratar los datos, resulta fundamental conocerlos. Para tener un conocimiento profundo, además de saber la naturaleza y estructura, como se ha expuesto en el apartado anterior, es muy importante describir los datos desde un punto de vista estadístico.

En este apartado se detallarán medidas de tendencia central y dispersión que aumentan el conocimiento de los datos y ayudan a elegir la metodología de tratamiento adecuada.

Medidas de tendencia central

La primera descripción estadística que se va a estudiar en este apartado está relacionada con la idea de centro de la distribución de los datos. Por tanto, estas medidas aportan información acerca de qué valor está situado en la mitad del conjunto de datos y persiguen señalar un valor que sea el representante de todo el conjunto.

La medida de tendencia central más sencilla y estudiada es la **media aritmética,** que consiste en sumar todos los valores de la muestra y dividir el resultado entre el número de datos recogidos.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Cuando los datos están recogidos en distintos contextos, la media aritmética no aporta una información certera de la medida central de los datos, pues esta considera que la importancia o peso que tiene cada dato en la muestra es idéntica. Para solucionar este problema surge la **media ponderada** que, como su nombre indica, pondera o asigna una importancia distinta a cada dato, en función al contexto.

$$\overline{x} = \frac{\sum w_i x_i}{\sum w_i}$$

donde los wi son los pesos o ponderaciones de cada xi.

Media aritmética y ponderada

Supongamos que una determinada empresa dedicada a la venta de cosméticos en Andalucía occidental quiere conocer el consumo anual medio de uno de sus productos, la barra de labios roja, para lanzar promociones a sus clientes.

Para ello, se hace una consulta al programa de ventas, que genera la siguiente tabla con información acerca del consumo anual medio en cada provincia, así como el peso que tiene cada una en el volumen de facturación total que genera la barra de labios roja:

	Barra de labios roja	
	Consumo	Peso
Huelva	2	75%
Sevilla	7	10%
Cádiz	3	10%
Córdoba	10	5%

Tabla 1. Datos de consumo y peso.

La media aritmética del conjunto de datos viene dada por:

$\bar{x} = \sum xin = 224 = 5{,}5$ *barras de labios*

Por su parte, la media ponderada del mismo conjunto de datos en función al volumen de facturación es:

$$\overline{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{300}{100} = 3\,\text{barras de labios}$$

Teniendo en cuenta que la mayor parte del negocio de la compañía se encuentra en Huelva, el consumo medio de barras de labios no es 5,5, sino 3.

La media aritmética y ponderada cuenta con una limitación importante, pues están muy condicionadas por valores atípicos u *outliers*. Por ello, en algunos casos resulta más certero utilizar otras medidas de posición central que no se vean afectadas por datos que las distorsionen. Aunque existen algunas más, en este tema estudiaremos la media recortada y la moda.

La media aritmética y ponderada se ven desvirtuadas por la presencia de valores atípicos u *outliers.*

La **media recortada** consiste en hacer la media aritmética a un subconjunto central del conjunto de datos. De esta manera, los valores *outliers* quedan a los extremos y no influyen en el resultado final obtenido. Por lo general, la forma de denominarla es*media recortada al y %* donde *y* indica el porcentaje de datos que debemos dejar de lado por cada extremo. Por ejemplo, si tenemos 100 datos y calculamos una media recortada al 10 %, debemos obviar 10 datos a la izquierda y 10 a la derecha calculándose la media únicamente sobre los dos valores centrales. Cabe destacar que cuando se calcula la media recortada al 25 %, esta se denomina **centrimedia.**

Por su parte, **la moda** es el valor más frecuente del conjunto de datos. Si el conjunto contiene dos datos con la misma frecuencia diremos que la distribución es bimodal por tener dos modas. Además, aunque no es habitual, nada impide que un conjunto de datos cuente con más de dos modas.

Media recortada y moda

Una determinada empresa dedicada a la venta de cigarrillos ha calculado el índice se sustitución en cada una de las provincias españolas. Este índice muestra cuántos consumidores de cigarrillos migran al tabaco de liar cuando abandonan los cigarrillos. En este sentido, cuando el índice es igual a 1 se produce una sustitución perfecta, es decir, todos los consumidores que abandonan el mercado de cigarrillos migran al tabaco de liar. Valores superiores a 1 indican una migración fuera del mercado legal, ya sea por dejar de fumar o por acudir al comercio ilícito de tabaco.

	IS
Cádiz	4,02
Huesca	3,58
Sevilla	2,75
Tarragona	2,20

Tabla 2. Índice de sustitución de cigarrillos ordenados de mayor a menor.

Rocío Galindo, directora de estrategia de la compañía, analiza los datos

de las 48 provincias y observa que existen outliers que desvirtúan la muestra por tratarse de provincias en las que el índice de sustitución es demasiado elevado. Por ello, calcula la media ponderada, que asciende a 3,43, y la compara con la media recortada al 15 %, que toma un valor de 1,55. Es decir, si se eliminan los 7 valores atípicos más grandes y más pequeños, el índice medio nacional es menos de la mitad de lo que arroja la media ponderada. Por último, con el objetivo de dar robustez a los resultados, calcula la moda y esta también toma valor de 1,55, es decir, el valor más repetido en las provincias coincide con la media recortada.

Dados los resultados, Rocío decide asimilar como índice de sustitución medio nacional el valor de 1,55.

Medidas de dispersión

Las medidas de dispersión nos indican cuánto se separan los datos de la media o unos de los otros. Este es un aspecto fundamental para conocer cómo se distribuyen los datos y poder predecir su comportamiento.

La primera medida de dispersión, la más básica, consiste en hacer la diferencia entre el valor máximo y mínimo para conocer cómo están los extremos de la distribución. Esta diferencia se denomina **rango**.

$$\text{rango} = x_{\max} - x_{\min}$$

Aunque el rango aporta información útil,en algunas ocasiones estos los puntos máximo y mínimo pueden estar muy separados por ser puntos atípicos u *outliers*. Por ello, otras medidas incluyen información sobre el resto de los valores, no solo del máximo y el mínimo. En esta línea surge la **varianza,** que es un promedio de las

desviaciones de los datos a su media. El hecho de que esas desviaciones se eleven al cuadrado viene motivado porque lo que importa es el valor absoluto de la dispersión, independientemente al signo que tome.

$$\text{Varianza} = s^2 = \frac{\sum (x_i - \overline{x})^2}{n}$$

Por otro lado, cabe destacar que otra forma de mostrar la dispersión que muestra un conjunto de datos es la **desviación típica,** entendida como la raíz cuadrada de la varianza.

$$\text{Desviación típica} = s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$$

Tanto la varianza como la desviación típica son magnitudes que por sí solas no aportan información relevante sin ser tratadas. Por ello, para realizar comparaciones entre muestras de datos, resulta aconsejable utilizar el **coeficiente de variación.** Esta es una medida que cuantifica el peso que tiene la desviación típica sobre la media.

$$\text{CV} = \frac{s}{\overline{x}} \times 100\%$$

3.3. Análisis de correlación

Cuando se representa una nube de puntos con los valores que toman dos variables, dicha representación puede indicarnos relaciones entre ellas. Esta relación no necesariamente tiene que ajustarse a una recta, por ejemplo, el gráfico que vemos a continuación expresa una relación entre las variables de carácter exponencial.

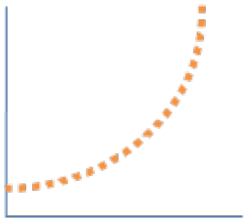


Figura 1. Relación entre variables de carácter exponencial

Sin embargo, en términos estadísticos, y sobre todo como base en posteriores aplicaciones, se analiza mayoritariamente la existencia de relación lineal entre variables.

Dos variables están en relación lineal cuando el aumento o disminución de una de ellas implica un aumento o disminución proporcional en la otra.

Existen muchas magnitudes que ayudan a medir el grado de relación lineal entre dos variables, entre las que se pueden destacar:

Figura 2. Parámetros para medir el grado de relación lineal entre dos variables.

Los dos primeros serán analizados en este apartado, siendo la recta de regresión, el coeficiente de determinación y la varianza residual objeto de análisis en el apartado siguiente.

Covarianza

Se llama covarianza de una variable (*X,Y*) a la media aritmética de los productos de las desviaciones de cada una de las variables en relación con sus medias respectivas. Una covarianza distinta de cero entre dos variables, indica la existencia de relación entre ellas.

La covarianza se representa por S_{XY}

La covarianza viene dada por las expresiones equivalentes:

$$S_{\mathrm{XY}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{p}\left(x_i - \overline{X}\right)\cdot\left(y_i - \overline{Y}\right)\cdot n_{\mathrm{ij}}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{p} x_i \cdot y_j \cdot n_{\mathrm{ij}}}{N} - \overline{X}\cdot\overline{Y}$$

La fórmula más habitual y sencilla que se suele utilizar es la que se muestra aquí abajo, donde

es la media de una nueva variable, la cual tiene como valores la multiplicación de los valores de

e $\overline{x \cdot y}$ $XY : x_i \cdot x_j; \forall i,\ j$

$$s_{XY} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y}$$

> Cálculo de la covarianza desde una tabla de contingencia
>
> Los datos a partir de los que vamos a trabajar son los que se muestran en la siguiente tabla. Además, en dicha tabla se encuentran las frecuencias marginales de cada variable y el producto de la variable por la frecuencia absoluta, como cálculo necesario para obtener la media de las variables.

X/Y	1	2	3	n_i	$x_i \cdot n_i$
2	1			1	2
3	2	2		4	12
4		1		1	4
5		2	2	4	20
n_j	3	5	2	$N = 10$	38
$y_j \cdot n_j$	3	10	6	19	

Tabla 3. Ejemplo de cálculos para obtener la covarianza.

Son, por tanto, necesarios los siguientes sumatorios:

$$\sum_{i=1}^{N} x_i = 19$$

$$\sum_{i=1}^{N} y_i = 38$$

A partir de estos se obtienen las medias de cada una de las variables:

$$\overline{X} = \frac{38}{10} = 3{,}8; \overline{Y} = \frac{19}{10} = 1{,}9$$

Para calcular la media del producto utilizamos una tabla similar, en la que colocaremos los productos cruzados $\mathrm{X} \cdot \mathrm{Y} \cdot \mathrm{n}_i$

X/Y	1	2	3	
2	$1 \cdot 1 \cdot 2 = \mathbf{2}$			**2**
3	$2 \cdot 1 \cdot 3 = \mathbf{6}$	$2 \cdot 2 \cdot 3 = \mathbf{12}$		**18**
4		$1 \cdot 2 \cdot 4 = \mathbf{8}$		**8**
5		$2 \cdot 2 \cdot 5 = \mathbf{20}$	$2 \cdot 3 \cdot 5 = \mathbf{30}$	**50**
	8	**40**	**30**	**78**

Así, la media del producto es:

$$\overline{\mathrm{XY}} = \frac{78}{10} = 7{,}8$$

La covarianza es, entonces: S_XY=7,8-3,8·1,9=0,58

Algunas de las propiedades de la covarianza son:

- $\mathrm{Cov}(X,Y) = E(X,Y) - E(X) \cdot E(Y)$
- $\mathrm{Cov}(X,Y) = \mathrm{Var}(X)$.
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.
- $\mathrm{Cov}(\mathrm{a} \cdot \mathrm{X},Y) = \mathrm{a} \cdot \mathrm{Cov}(X,Y)$, siendo $a \in R$.
- $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$.
- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X,Y)$.
- $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2 \cdot \mathrm{Cov}(X,Y)$.
- Si e son independientes, $\mathrm{Cov}(X,Y) =$ .La anterior propiedad no se verifica en sentido contrario.
- Si e son dependientes, $\mathrm{Cov}(X,Y) \neq$.

Coeficiente de correlación

El coeficiente de correlación sería similar a la covarianza en cuanto a su uso, pues mide la existencia de relación entre dos variables, pero tiene una ventaja: **no tiene unidades.** En este sentido, nos va a permitir realizar comparaciones entre parejas de variables.

El coeficiente de correlación lineal ρ se calcula:

Este coeficiente siempre estará entre –1 y 1. Su signo viene dado por el signo de la covarianza (dado que las desviaciones típicas siempre toman valores positivos):

- Si hay relación lineal positiva, $\rho >$ y próximo a 1.
- Si hay relación lineal negativa, $\rho <$ y próximo a –1.

- Si no hay relación lineal, ρ será próximo a 0 (incorreladas).

Cuando las variables X e Y son independientes, la covarianza y, por tanto, ρ=0. No podemos asegurar lo mismo en sentido contrario: si dos variables tienen covarianza cero, no podemos decir que son independientes. Sabemos que linealmente no tienen relación, pero podrían tener otro tipo de relación y no ser independientes.

3.4. Análisis de regresión

El análisis de regresión tiene como objetivo la generación demodelos, y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones. Se usa para modelar patrones en los datos y extraer inferencias acerca de la población bajo estudio.

La estadística inferencial pretende modelar patrones para poder deducir el comportamiento de una población completa a partir de los datos recabados en una muestra.

El modelo de regresión lineal

La técnica tradicional para modelar patrones y deducir comportamientos por excelencia es el modelo de regresión lineal. Esta metodología, ampliamente estudiada en asignaturas como Estadística II y Econometría, permite estimar la sensibilidad que tiene una determinada variable ante cambios en otras.

Se trata de explicar el comportamiento de una variable endógena (y) en función a otras exógenas (x). La forma funcional del modelo de regresión lineal viene dada por:

$$y = \beta + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Dada la complejidad de la realidad económica y empresarial y el entorno *big data,* está claro que nuestro modelo econométrico no es el verdadero, pues estamos haciendo una simplificación de la realidad para llevar a cabo nuestro análisis. Por este motivo el modelo incluye el término u, al que denominamos «error». Este término *error u* representa todos los demás factores que pueden influir en la variable explicada y que no se han considerado explícitamente en el modelo.

Los coeficientes de las variables en el modelo de regresión (β ,β_1,β_2,. . .) son

constantes desconocidas y se les denomina «coeficientes del modelo». Conocer el valor de estos es el primer interés de un análisis empírico, ya que indican la intensidad con la que su correspondiente variable explicativa afecta a la variable explicada.

El primer interés de un modelo de regresión es conocer los valores de los coeficientes β ,β_1,β_2,. . .,β_k. Para ello, tomamos una muestra de las variables dependientes e independientes y estimamos, por medio de las herramientas de inferencia estadística que conocemos de Estadística II, el valor de dichos coeficientes.

Con afán de no repetir cuestiones ya estudiadas en otras asignaturas, además de lo expuesto en este apartado, las características fundamentales de un análisis basado en el modelo de regresión lineal son:

- Permite **medir la sensibilidad** de una variable ante cambios en otras.
- Facilita el **contraste sobre si una determinada condición que se le presupone** a la sensibilidad de una variable frente a otra es compatible con la muestra utilizada para el estudio.
- Posibilita la elaboración de **intervalos** que muestran entre qué valor mínimo y máximo se moverá la sensibilidad de la variable explicada ante cambios en las explicativas.
- Aporta una **predicción puntual y por intervalo** sobre qué valor tomará la variable explicada en base a unos valores prefijados de las variables explicativas.

Residuos o errores de estimación

Sabemos que $\widehat{y} = a + \mathrm{b}\cdot\mathrm{x}$, y que dichas predicciones no son exactas, pues no todas ellas caerán sobre la nube de puntos. Por tanto, podemos definir los residuos o errores de estimación como la diferencia entre los valores reales de Y y los valores ajustados o predichos de $\widehat{Y}$:

$$e_i = y_i - \widehat{y}_i = y_i - (\mathrm{a} \cdot x_i + b)$$

Bondad de ajuste

Una vez definida la recta de regresión, es importante disponer de una medida que calcule la bondad del ajuste realizado, y que permita decidir si el ajuste lineal es suficiente o se deben buscar modelos alternativos. Dicha medida es el **coeficiente de determinación.**

El coeficiente de determinación va a ser el porcentaje de varianza de Y, que se puede explicar por X porque valora lo cerca que está la nube de puntos de la recta de regresión.

Así, como medida de bondad del ajuste de la recta de regresión se utiliza el coeficiente de determinación, definido como el cuadrado del coeficiente de correlación:

$$R^2 = \rho^2 = \left(\frac{S_{XY}}{S_X \cdot S_Y}\right)^2$$

Verifica siempre que 0<R2<1

Es habitual expresar esta medida en tanto por ciento, multiplicándola por cien.

Varianza residual

Se define como la varianza de los residuos o errores de estimación $(e1, e_2, \ldots e_N)$.

Podemos comprobar que:

La raíz cuadrada de la varianza residual (desviación típica de la varianza residual) es conocida como «error típico».

- Si la varianza residual es grande, los residuos serán grandes y la dependencia será pequeña; el ajuste será malo.

- Si la varianza residual es pequeña (cerca de cero), la dependencia será grande; el ajuste será bueno.

Nota: puedes profundizar sobre la interpretación de tablas y gráficos bidimensionales y de relación de variables de la mano del artículo introducido en el recurso «Aprendiendo a interpretar tablas y gráficos estadísticos».

El modelo de regresión lineal con relaciones espaciales

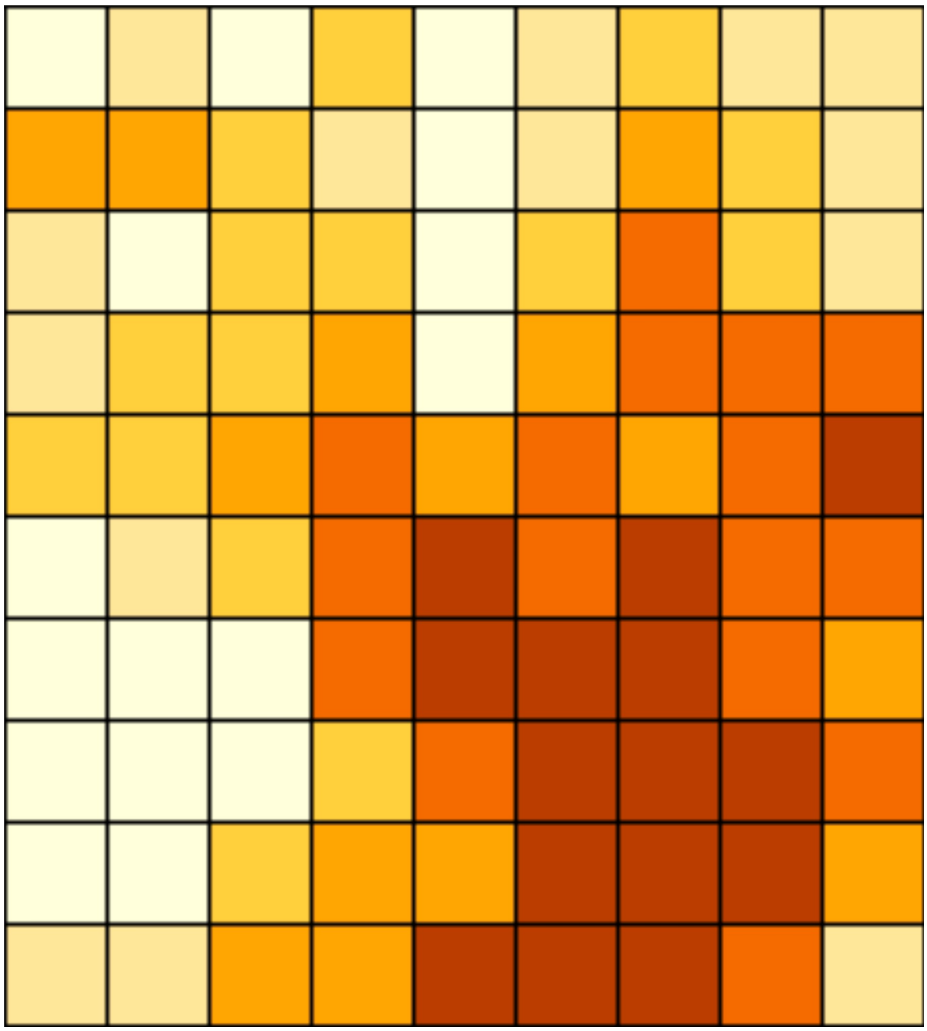
El análisis de datos espaciales es una disciplina que, aunque tiene su origen en Anselin (1988), ha adquirido una especial importancia en la actualidad debida, principalmente, al avance de la tecnología en las comunicaciones y la globalización de la economía. Si bien en 1988 tener datos geolocalizados estaba al alcance solo de grandes compañías que contaban con una tecnología muy avanzada, actualmente la implantación de sistemas GPS en sensores que recogen datos es algo muy común y barato.

Los sucesos que generan datos en un lugar específico tienen influencia tanto sobre sus vecinos directos como, en ocasiones, sobre vecinos aparentemente remotos. Imaginemos que tomamos datos sobre la venta de cigarrillos en Cádiz y Sevilla. Dado que son provincias fronterizas, no es descabellado pensar que parte del comportamiento observado en las ventas de Sevilla se debe a la proximidad que tiene esta zona geográfica con Cádiz, y viceversa. Por ello, en el estudio de cualquier fenómeno social o económico la localización de los datos debe ser tenida en cuenta por si existiera la denominada **autocorrelación espacial,** que consiste en una especie de efecto contagio, es decir, lo que le ocurre a mi vecino tiene influencia sobre mi territorio. Esa correlación espacial puede ser de dos tipos:

- **Autocorrelación espacial positiva:** como puede observarse en la figura 3a, este tipo de autocorrelación espacial supone que haya grupos de regiones con comportamientos similares. Llevado a la práctica, si en Huelva el número de ciudadanos con estudios superiores es muy elevado, en Sevilla y Cádiz (provincias

fronterizas) también lo será.

- **Autocorrelación espacial negativa:** en este caso, como puede observarse en la figura 3a, este tipo de autocorrelación espacial supone que haya grupos de regiones rodeadas de otras con comportamientos distintos. Llevado a la práctica, si en Huelva el número de ciudadanos con estudios superiores es muy elevado, en Sevilla y Cádiz (provincias fronterizas) no lo será.



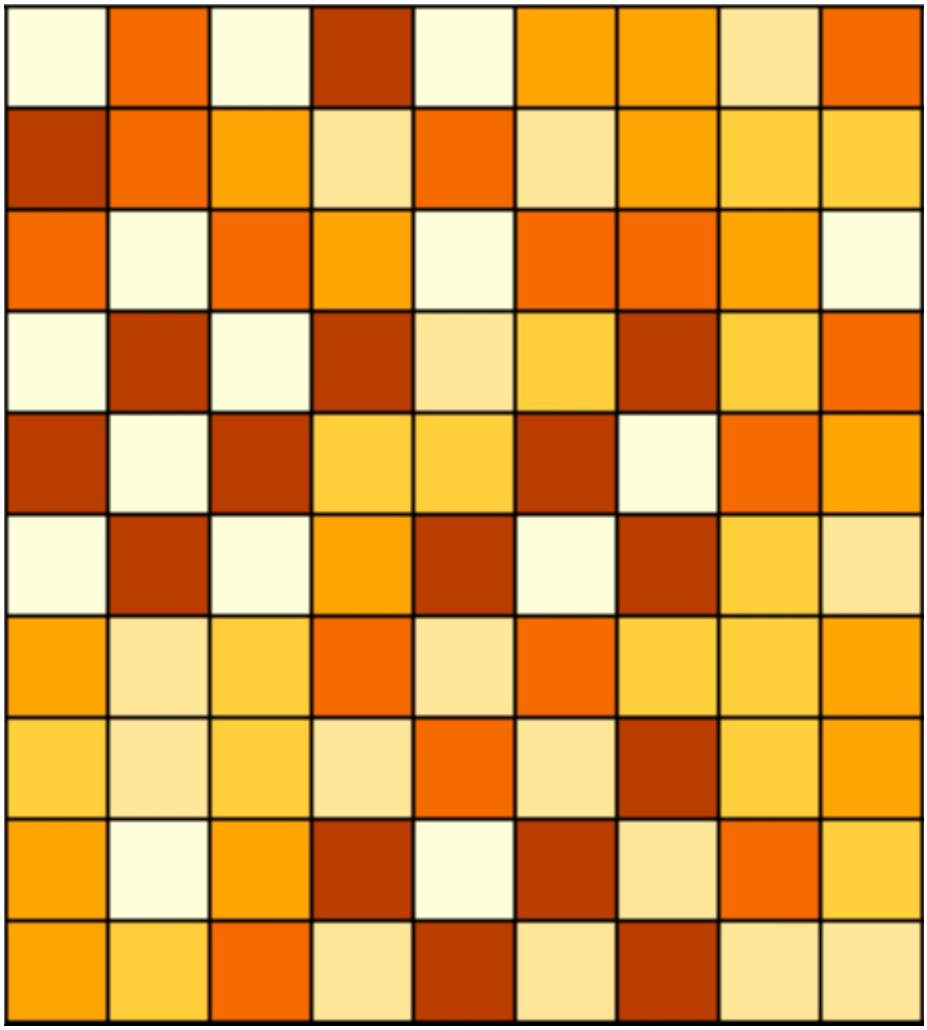


Figura 3. Tipos de autocorrelación espacial. Fuente: Anselín (1988)

Aunque, como ya se ha mencionado, en la actualidad las organizaciones cuentan con muchos datos georreferenciados, estos suelen ser tratados con herramientas de análisis tradicionales sin usar técnicas adecuadas para el análisis estadístico espacial. En este contexto, y gracias al desarrollo tecnológico de los sistemas de georreferenciación de datos, surge la necesidad de contar con métodos y herramientas apropiadas para el procesamiento, descripción y análisis de los datos, pues los métodos tradicionales de la estadística descriptiva e inferencial no tienen en cuenta la localización geográfica de los datos.

¿Cómo se incluye la información espacial en un modelo de regresión lineal?

Tradicionalmente, como se ha expresado con anterioridad, la forma funcional del modelo de regresión lineal viene dada por:

$$y = \beta + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Además, la expresión matricial de dicho modelo viene dada por:

$$y = X + u$$

En esta línea, si se quiere incorporar en el modelo de regresión lineal el efecto espacial de la variable dependiente; es decir, si quiero explicar el consumo de cigarrillos de una región en función a una serie de variables explicativas, pero, además, en función al consumo de cigarrillos de sus vecinos, basta con especificar dicho modelo como:

$$y = X + \mathrm{Wy} + u$$

De este modo, volviendo a la forma funcional no matricial, el modelo quedaría especificado como:

$$y = \beta + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + y_{\text{vecinos}} + u$$

Como puede observarse en la ecuación, los parámetros β_{1-k} miden cómo afectan las variables explicativas al consumo de cigarrillos de las regiones, es decir, lo tradicional. Por su parte, el parámetro ρ mide cómo afecta el consumo de cigarrillos de los vecinos al consumo de cigarrillos de las regiones.

Además, en el modelo de regresión lineal también se puede incluir el efecto espacial de las variables independientes; es decir, si quiero explicar el consumo de cigarrillos de una región en función a una serie de variables explicativas, pero, además, en función a las variables explicativas de sus vecinos, basta con especificar dicho modelo como:

$$y = X + \mathrm{WX} + u$$

De este modo, volviendo a la forma funcional no matricial, el modelo quedaría especificado como:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + P_1 x_{1\,\text{vecinos}} + P_2 x_{2\,\text{vecinos}} + \ldots P_k x_{k\,\text{vecinos}} + u$$

En este caso, los parámetros β_{1-k} miden cómo afectan las variables explicativas al consumo de cigarrillos de las regiones, es decir, lo tradicional. Por su parte, los parámetros ρ_{1-k} miden cómo afectan las variables explicativas de los vecinos al consumo de cigarrillos de las regiones.

En este tipo de modelos, es decir, en los que se considera que el valor que tomen las variables explicativas de los vecinos tiene influencia sobre la variable explicada, los parámetros se definen como:

- **Efecto directo;** es el que representan los parámetros β_{1-k} y se refiere al efecto tradicional, al que provocan sobre la variable explicada las variables explicativas observadas en la propia región.
- **Efecto indirecto;** es el que representan los parámetros β_{1-k} y se refiere al efecto espacial, al que provocan sobre la variable explicada las variables explicativas observadas en las regiones vecinas.

Cuando un modelo de regresión lineal incluye relaciones espaciales, es posible conocer qué parte del efecto total que tienen las variables explicativas en la explicada se debe a la propia región y qué parte proviene de la influencia de sus vecinos.

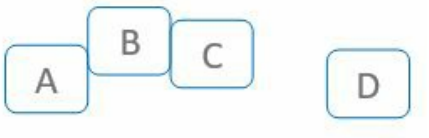
Llegado este punto, cabe destacar que la matriz W se denomina **matriz de contigüidad** y representa la relación que tiene cada una de las regiones con las demás regiones del espacio en estudio, tal y como se vería en un mapa. Aunque

existen multitud de formas en que la matriz de contigüidad puede ser constituida, la más sencilla es utilizando una notación binaria, en la que 0 representa ausencia de contigüidad y 1 la presencia de contigüidad entre dos regiones.

Aunque la matriz de contigüidad más básica implica dar valor 1 y 0 en función a si las regiones son fronterizas o no, *«la construcción de W está basada en un criterio de conectividad que define qué unidades pueden ser consideradas vecinas entre sí»*. Ahumada et al. (2018).

Construcción de matriz de contigüidad

Supongamos que queremos incluir el componente espacial en un modelo de regresión lineal. La muestra de datos que se recoge pertenece a un conjunto de cuatro regiones localizadas de la siguiente manera:



Para construir la matriz de contigüidad siguiendo el criterio expuesto, basta con aportar valores 0 y 1 en función a si las regiones son vecinas o no:

	A	B	C	D
A	0	1	0	0
B	1	0	1	0
C	0	1	0	0
D	0	0	0	0

Cálculo de elasticidad precio y elasticidad renta con efectos espaciales

Una determinada empresa dedicada al comercio de cigarrillos está muy preocupada por la fuerte caída que han sufrido las ventas de sus productos, por lo que quiere conocer qué efecto tiene el incremento en el precio, así como el aumento en la renta disponible, por lo que decide

plantear un modelo de regresión en logaritmos.

No obstante, al representar gráficamente las ventas de cigarrillos provinciales se observa una cierta dependencia espacial, por lo que la compañía decide incluir ese efecto en el modelo. Por ello, plantea la siguiente especificación:

$$In\left(cig\right) = \beta_o + \beta_1\, In\left(precio\right) + \beta_2\, In\left(PIB\right) + \rho_1\, In\left(precio\right)_{vecinos} + \rho_2\, In\left(PIB\right)_{vecinos} + u$$

Tras la estimación de los parámetros, los resultados obtenidos fueron:

	β_1	ρ_1	β_2	ρ_2
Parámetros	-0,45	-0,55	0,60	0,34

Tabla 5. Parámetros con información espacial.

Dados los resultados, el aumento de un 1 % en el precio de los cigarrillos a nivel nacional, tendrá un efecto total en cada región de -1 % en las ventas de cigarrillos, pues, aunque el efecto directo es de -0,45 %, existe un efecto indirecto provocado por los vecinos que asciende a -0,55 %. Por su parte, cuando la renta crece a nivel nacional en un 1 %, el efecto total sobre el mercado de cigarrillos en cada región es de un 0,94 %, del que 0,60 % corresponde a las propias regiones y 0,34 % al efecto que les provocan sus vecinos.

3.5. Intervalos de confianza y contrastes de hipótesis

El intervalo de confianza y contrastes de hipótesis se utilizan en economía y empresa, fundamentalmente, para comprobar si se cumplen determinados valores en la relación entre variables expuestas en el apartado anterior. Vendrá dado por la distribución del parámetro $\widehat{\beta_J}$ y será construido teniendo en cuenta el nivel de confianza definido en la sección anterior.

En esta línea, los límites estadísticos del intervalo de confianza para un parámetro vienen dados por:

$$\underbrace{\left[\hat{\beta}_j - SE(\hat{\beta}_j)\cdot t_{N-k-1,\frac{\alpha}{2}}\right.}_{\text{límites estadísticos}};\ \underbrace{\left.\hat{\beta}_j + SE(\hat{\beta}_j)\cdot t_{N-k-1,\frac{\alpha}{2}}\right]}_{\text{límites estadísticos}}$$

límites estadísticos

Una vez que sustituimos *β^j* y *SE(β^j)* por los valores correspondientes de la muestra, estos dejan de ser variables aleatorias y se convierten en estimaciones, valores numéricos.

Gasto en publicidad de una empresa y la de la competencia y su repercusión en el porcentaje de ventas

Se quiere estudiar el efecto del gasto en campañas publicitarias sobre el total de ventas (en miles de unidades) por la empresa A en un determinado período de tiempo de 173 días. Las estimaciones se realizan para las ventas de la empresa A, para la cual también se introduce el tamaño de su cuota de mercado en porcentaje.

Como se observa en el siguiente modelo las variables de gasto en publicidad se incluyen en logaritmos, por lo que deberemos llevar cuidado con esta cuestión a la hora de interpretar los coeficientes.

$$\text{ventas}\, A_t = \beta + \beta_1 \log(\text{expend}\, A_t) + \beta_2 \log(\text{expend}\, B_t) + \beta_3\, \text{cuota}\, A_t + u_i$$

De una muestra de 180 días, para las cuales se han tomado datos de las variables indicadas, se obtienen los siguientes resultados:

$$\widehat{\text{ventasA}}_t = \underset{(1.12)}{45.1} + \underset{(0.382)}{6.08} \log(\text{expend}\, A_t) - \underset{(0.379)}{6.62}\, \beta_2 \log(\text{expend}\, B_t) + \underset{(0.062)}{0.152}\, \text{cuota}\, A_t$$

Se pide obtener un intervalo de confianza para β_1 (el efecto sobre las unidades vendidas de incrementar en un 1 % el gasto en publicidad de la empresa A) para un nivel de confianza del 95 %.

Aplicando los límites del intervalo de confianza obtenidos se obtiene que:

$$\widehat{\beta}_j - \text{SE}\left(\widehat{\beta}_j\right) t_{N-k-1,\frac{\alpha}{2}} = 6.08 - 0.382 * 1.97 = 5.327$$

$$\widehat{\beta}_j + \text{SE}\left(\widehat{\beta}_j\right) t_{N-k-1,\frac{\alpha}{2}} = 6.08 + 0.382 * 1.97 = 6.833$$

Ya que puede obtenerse de Gretl que los puntos críticos a utilizar para un 95 % de confianza son -1.97 y 1.97, dado que:

$$P\left(t_{N-k-1} > 1.97\right) = 0.025$$

Por tanto, el intervalo resultante es [5.327;6.833] lo que indica que un incremento del 1 % en el gasto en publicidad provocará un incremento de las ventas de entre 0.05327 y 0.06833 miles de unidades (lo que equivale a un incremento de 53.27 y 68.33 unidades), rango de valores que se cumplirá con un nivel de confianza del 95 %.

Contraste de hipótesis para un coeficiente

Dado el intervalo obtenido y la forma en la que este ha sido interpretado, podemos realizar contraste de hipótesis de forma que se realicen afirmaciones sobre el valor de β_1. Dichas afirmaciones serán no rechazas siempre y cuando el valor llevado a contraste se encuentre en los valores límites del intervalo. Dicho contraste estará siendo calculado bajo un nivel de significatividad del α·100 %.

Existe así una relación entre los intervalos de confianza y los contrastes de hipótesis bilaterales. Si queremos contrastar la siguiente hipótesis:

$$H: \beta_j = \beta_j$$
$$H_1: \beta_j \neq \beta_j$$

Entonces, si β_j^0 no está contenido en el intervalo, rechazamos la hipótesis nula con un $\alpha\%$ nivel de significación, y si β_j^0 está contenido en el intervalo, no rechazamos la hipótesis nula con un $\alpha\%$ nivel de significación.

Gasto en publicidad de una empresa y la de la competencia y su repercusión en el porcentaje de ventas (continuación)

El director del departamento de marketing indica en una reunión que es necesario y rentable invertir en publicidad ya que dicha inversión repercutirá positivamente sobre la variable ventas. En concreto por cada incremento del 1 % en el gasto en publicidad, la variable ventas se verá incrementada en 0.050 miles de unidades (50 unidades).

Por tanto, el valor, dado que la variable gasto en publicidad está medida en logaritmos, llevado a contraste es el 5 %.

$$H: \beta_j = 5$$
$$H_1: \beta_j \neq 5$$

Como dicho valor no se encuentra dentro del intervalo [5.327;6.833] se rechaza la hipótesis nula al 5 % de significatividad.

Por lo que no hay evidencias en la muestra de que el número de unidades vendidas se incremente en 0.05 miles de unidades al variar un 1 % el gasto en publicidad.

Nótese que en realidad el efecto de la variable gasto en publicidad es mayor que el indicado por el valor llevado a contraste, pero es que el problema de realizar contraste haciendo uso de los intervalos de confianza es que trabajamos con contrastes bilaterales, no pudiendo así identificar si la hipótesis nula ha sido rechazada hacia un lado u otro de la desigualdad.

3.6. Referencias bibliográficas

Ahumada, H., Gabrielli, M. F., Herrera, M. y Escudero, W. S. (2018). *Una nueva Econometría. Automatización, big data, econometría espacial y estructural.* Bahía Blanca.

Anselin, L. (1988*). Spatial econometrics: methods and models* (vol. 4). Studies in Operational Regional Science. Dordrecht: Springer Netherlands.

El mapa del voto en toda España, calle a calle

Andrino, B., Llaneras, K., Grasso, D. y Sevilllano, E. (3 mayo 2019). El mapa del voto en toda España, calle a calle. *El País.*

Los datos con referencia espacial están a la orden del día, tanto en variables económicas como en otras de cualquier naturaleza.

En este recurso encontrarás el mapa de voto de las elecciones celebradas en España el 28 de abril de 2019, con un nivel de detalle geográfico que incluyo llega hasta cada calle, marcando cuál ha sido el partido más votado en cada una de ellas.

Accede al documento a través del aula virtual o desde la siguiente dirección web: https://elpais.com/politica/2019/05/01/actualidad/1556730293_254945.html

Modelo espacial con datos de panel

Ahumada, H., Gabrielli, M.F., Herrera, M. y Escudero, W.S. (2018). Una nueva Econometría. Automatización, big data, *econometría espacial y estructural.* Bahía Blanca.

En las páginas 145-148 de este manual podrás encontrar un modelo estimado e interpretado en el que existe dependencia espacial y esta ha sido tratada con una muestra de datos de panel.

Accede al documento a través del aula virtual o desde la siguiente dirección web:
https://www.researchgate.net/publication/331907304_Una_nueva_econometria_Automatizacion_big_data_econometria_espacial_y_estructural

1. Juan Boss, CEO de una compañía dedicada a la educación superior, está pensando en plantear un modelo de econometría espacial... Este hecho es una primera aproximación de existencia de:

A. Autocorrelación espacial positiva.

B. Autocorrelación espacial negativa.

C. Ausencia de autocorrelación espacial.

D. Ninguna de las anteriores es correcta.

2. En los modelos de econometría espacial, el efecto que tienen las variables explicativas de los vecinos se denomina:

A. Efecto directo.

B. Efecto indirecto.

C. Efecto positivo.

D. Ninguna de las anteriores es correcta.

3. Ubur S. L., empresa de transporte... ¿Qué medida debe utilizar para comparar la dispersión de los datos?

A. Debe utilizar el coeficiente de variación.

B. Debe utilizar la varianza.

C. Debe utilizar la desviación típica.

D. Ninguna de las anteriores es correcta.

4. Los *outliers* o valores atípicos:

A. Afectan drásticamente a la media aritmética y algo menos a la media ponderada.

B. También se ve muy afectada la dispersión media que mide la desviación típica.

C. Puede alterar la forma de la distribución de datos al realizar representaciones gráficas.

D. Todas las anteriores son correctas.

5. El análisis de regresión es útil para:

A. Generar modelos que relacionan variables.

B. Predecir los valores de una variable en función de los que tome la otra.

C. Hacer inferencia acerca de la población objeto de estudio.

D. Todas las anteriores son correctas.

6. ¿Cuál de las siguientes es una medida de tendencia central menos sensible a *outliers*?

A. Media aritmética.

B. Media ponderada.

C. Media recortada.

D. Varianza.

7. ¿Qué valor indica la proporción de variabilidad de la variable dependiente explicada por un modelo de regresión?

A. Covarianza.

B. Coeficiente de determinación (R^2).

C. Desviación típica.

D. Error estándar.

8. Si dos variables tienen un coeficiente de correlación cercano a 0, esto indica que:

A. No existe ninguna relación entre ellas.

B. No hay relación lineal entre ellas.

C. Son independientes.

D. Son proporcionales.

9. En el contexto de un intervalo de confianza al 95 %, si un valor hipotético cae fuera del intervalo:

A. Aceptamos la hipótesis nula.

B. No podemos tomar ninguna decisión.

C. Rechazamos la hipótesis nula con un 5 % de significación.

D. El intervalo no es válido.

10. En el análisis de regresión lineal, ¿qué representa el término de error (u)?

A. El error del modelo de correlación.

B. El valor que predice el modelo.

C. Factores no incluidos en el modelo que afectan a la variable dependiente.

D. La variación explicada por los regresores.