Diplomado en Análisis de Información Geoespacial

Descripción bivariada

Autor:
M. en G. Alberto Porras Velázquez

Derechos Reservados © Centro de Investigación en Geografía y Geomática "Ing. Jorge L. Tamayo", A.C.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

Introducción

La descripción bivariada trata de estudiar la relación entre pares de atributos medidos simultáneamente en la misma muestra y comprende un conjunto de herramientas que se enfoca en el análisis de dos variables con el fin de determinar las relaciones empíricas entre ellas.

El análisis bivariado puede ayudar a determinar en qué medida es posible predecir el valor de una variable en el caso en que conozcamos el valor de otra variable. En específico, el estudio de la correlación entre dos variables se refiere a un conjunto de relaciones estadísticas que involucran una dependencia entre ellas. El estudio de la correlación es útil debido a que puede indicar relaciones predictivas que pueden ser utilizadas en la práctica.

Es pertinente mencionar que, aunque se encuentre una relación de dependencia entre dos variables, esto no es necesariamente un indicativo de una relación causal entre ellas, es decir, la correlación no implica causalidad.

La relación entre dos variables puede ser representada de manera visual o a través de un conjunto de medidas.

Diagrama de dispersión

Una herramienta gráfica sencilla para estudiar la relación entre dos variables es el diagrama de dispersión que consiste en un diagrama en coordenadas cartesianas en donde se despliegan los valores de dos variables para un conjunto de datos.

Los datos se grafican como una colección de puntos, cada uno con un valor de una variable en el eje horizontal x y el valor de la otra variable en el eje vertical (y).

Derechos Reservados © Centro de Investigación en Geografía y Geomática "Ing. Jorge L. Tamayo", A.C.

Los diagramas de dispersión que se presentan a continuación fueron generados a partir de un conjunto de datos muestreados en la región de Jura, en Suiza, por el Swiss Federal Institute of Technology at Laussane. La muestra consiste de 259 observaciones de siete metales y características del suelo.

Con fines didácticos, comenzaremos describiendo el diagrama de dispersión de una variable consigo misma. En la ilustración 1 se grafican en el eje x y en el eje y los mismos valores, en este caso, las concentraciones de cobre para las 259 observaciones realizadas en la región de Jura.

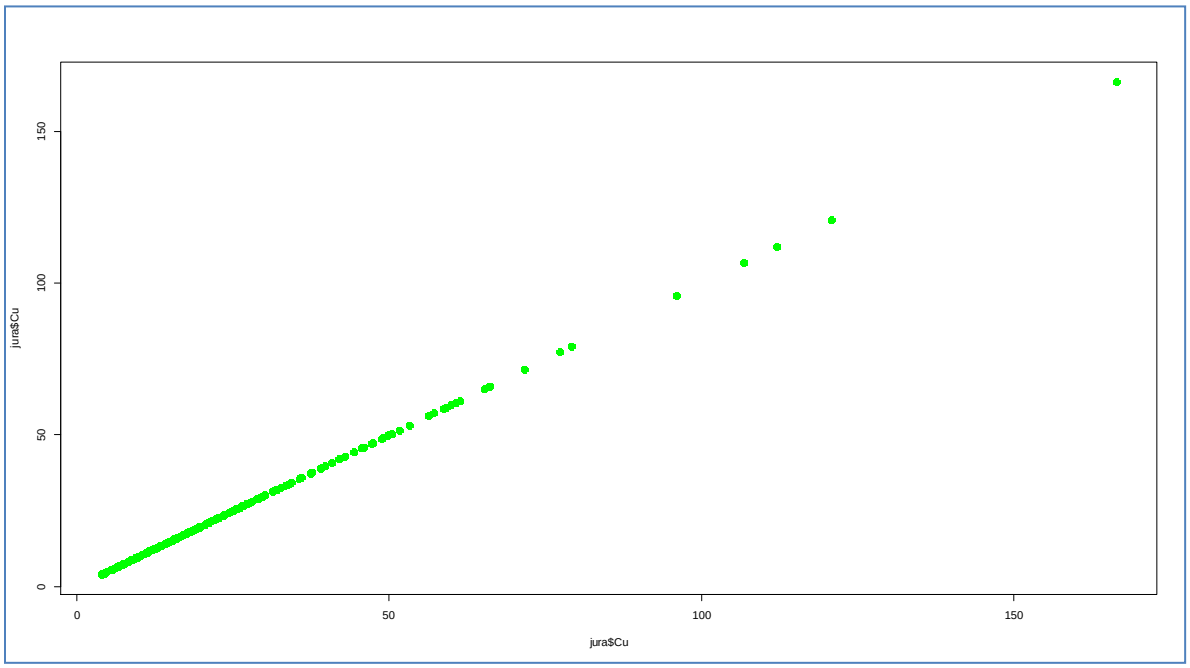

Ilustración 1 Diagrama de dispersión del cobre vs cobre.

Como se observa, los puntos caen sobre una línea recta. Esto representa una correlación lineal perfecta que cumple la condición de que los puntos se ubiquen cerca de una línea recta con una pendiente distinta de cero y no infinita. Es pertinente mencionar que este diagrama de dispersión no tiene sentido desde el punto de vista analítico, ya que una variable tiene una correlación lineal perfecta consigo misma.

Derechos Reservados © Centro de Investigación en Geografía y Geomática "Ing. Jorge L. Tamayo", A.C.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

Por lo regular, al realizar el diagrama de dispersión entre dos variables, el conjunto de puntos de la gráfica no adoptará el patrón exacto de una línea recta como en el caso anterior, sino la forma de una nube de puntos. Un conjunto de puntos más agrupado en torno a una línea nos indicará un mayor grado de correlación lineal entre las dos variables.

Por ejemplo, para los datos Jura se muestran en la ilustración 2 los diagramas de dispersión del plomo vs el cobre (izquierda) y el del cadmio vs el cobre (derecha). A cada uno se le agregó una línea de tendencia ajustada por el método de mínimos cuadrados.

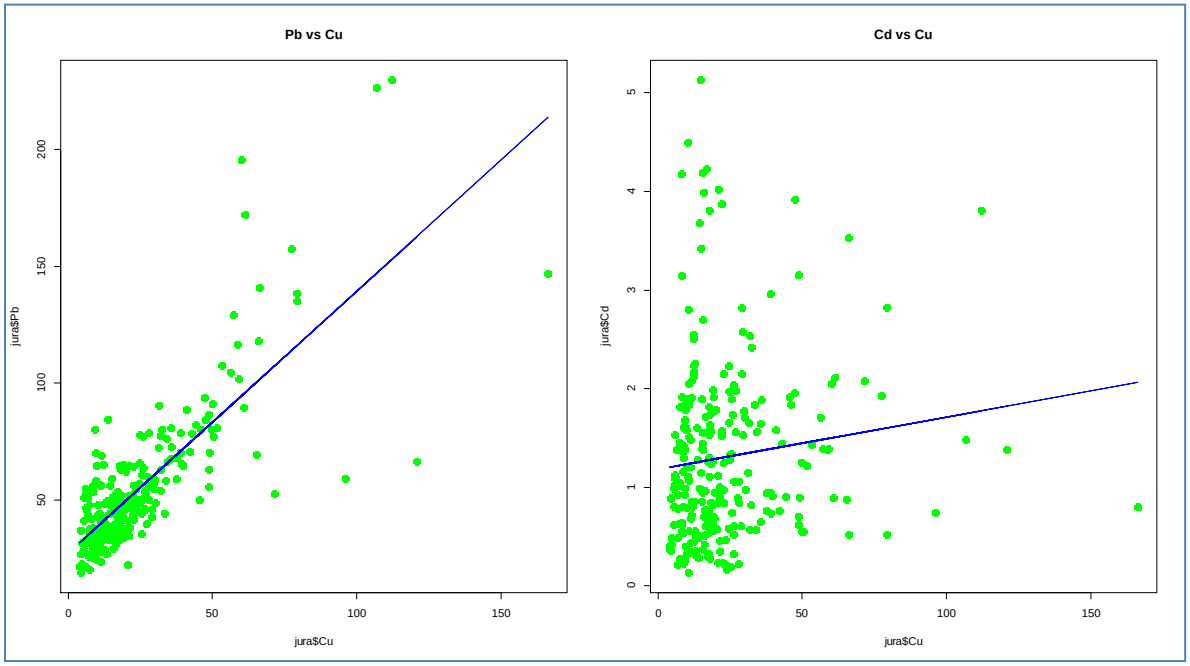

Ilustración 2. Diagramas de dispersión del Plomo vs Cobre (izquierda) y Cadmio vs Cobre (derecha)

A partir de los diagramas se identifica que en la gráfica del plomo contra el cobre la nube de puntos está más agrupada en torno a la línea en comparación con la gráfica del cadmio contra el cobre, indicándonos que el cobre tiene una mayor correlación lineal

Derechos Reservados © Centro de Investigación en Geografía y Geomática "Ing. Jorge L. Tamayo", A.C.

con el plomo que con el cadmio. Estos diagramas implican un primer análisis cualitativo de la relación entre las variables.

Covarianza

La covarianza es una de las medidas utilizadas para cuantificar la variación conjunta entre dos variables X y Y. La expresión matemática para la covarianza es:

$$C(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})$$

Esto es, el promedio de los productos generados que se obtiene al multiplicar la diferencia de cada observación de la variable Y, menos la media de la variable Y, por la diferencia entre la respectiva observación de la variable X, menos la media de la variable X; **n** es el tamaño de la muestra y las observaciones se numeran de 1 a n.
En la ecuación, la suma de n términos dividida entre n denota un promedio. En este caso se trata del promedio de los productos de las desviaciones con respecto a las medias de X y Y.

Observa que los productos de los residuales pueden tener valores positivos, siempre y cuando ambos residuales (que son las diferencias con respecto a la media) tengan el mismo signo.

Otra posibilidad es que los productos de los residuales posean valores negativos, lo cual es viable si un residual tiene un signo positivo y otro negativo.

Derechos Reservados © Centro de Investigación en Geografía y Geomática "Ing. Jorge L. Tamayo", A.C.

La covarianza es positiva si en promedio los residuales de las dos variables tienen mismo signo; en caso contrario, será negativa

Signo de los residuales

$$C(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})$$

Productos con signo positivo

Productos con signo negativo

Signo de los residuales

Observa que la covarianza de una variable consigo misma implica que la fórmula de la covarianza coincida con la fórmula para la varianza de la variable, es decir, la fórmula se convierte en el promedio de las diferencias con respecto a la media elevadas al cuadrado.

Para la covarianza el signo es el elemento que indica el tipo de correlación: positiva, negativa. No obstante, la magnitud de la covarianza no es una medida indicativa, pues su valor depende directamente del rango de valores que asumen las variables. Lo anterior dificulta una comparación de los valores obtenidos para las correlaciones calculadas entre pares distintos de variables.

Coeficiente de correlación lineal:

Para poder comparar el grado de correlación lineal entre dos variables se utiliza el coeficiente de correlación lineal, que es la forma normalizada de la varianza. Se calcula

Derechos Reservados © Centro de Investigación en Geografía y Geomática "Ing. Jorge L. Tamayo", A.C.

como la covarianza dividida entre el producto de la desviación estándar de la variable X por la desviación estándar de la variable Y.

$$r(X,Y) = \frac{C(X,Y)}{S_X S_Y} \; con - 1 \; \leq r(X,Y) \leq 1$$

Dividir entre el producto de las desviaciones estándar tiene como efecto que el coeficiente tenga valores entre uno y menos uno.

Un coeficiente cercano a uno indica una alta correlación positiva. Un valor del coeficiente cercano a menos uno indica una alta correlación negativa.
Por último, un valor del coeficiente cercano a cero nos indica una muy baja o nula correlación lineal entre las variables.

Un criterio cualitativo para clasificar una correlación como buena, regular o mala es el siguiente.

Cuando el resultado del coeficiente da un valor absoluto entre cero y 0.4, existe una mala correlación entre las variables. Un valor absoluto entre 0.4 y 0.7 se considera una correlación regular. Finalmente, un valor mayor a 0.7 es aceptado como signo de una buena correlación entre las variables.

Cabe mencionar que, como el nombre de indica, el coeficiente de correlación lineal cuantifica la relación lineal entre las variables. Si el coeficiente de correlación lineal entre dos variables tiene un valor cercano a cero tan solo significa que la correlación no es lineal; sin embargo podría existir algún otro tipo de correlación.

Derechos Reservados © Centro de Investigación en Geografía y Geomática "Ing. Jorge L. Tamayo", A.C.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.