

1. Correlación

La regresión lineal, de la que tratan estas notas, se ocupa de investigar la relación entre dos o más variables continuas. En esta sección, comenzaremos tratando de describir el vínculo observado y luego nos sofisticaremos resumiendo en un valor numérico nuestra conclusión.

¿Con qué datos contamos para llevar a cabo un análisis? Disponemos de n observaciones de dos variables aleatorias medidas en los mismos individuos, como describimos en la Tabla 1.

Tabla 1: Observaciones a nuestra disposición. Aquí X_1 quiere decir, la variable X medida en el individuo 1, etc.

| Individuo | Variable X | Variable Y |
|-----------|--------------|--------------|
| 1 | X_1 | Y_1 |
| 2 | X_2 | Y_2 |
| : | : | : |
| n | X_n | Y_n |

En estas notas, estamos pensando en que medimos ambas variables en la misma unidad: puede tratarse de un individuo, un país, un animal, una escuela, etc. Comencemos con un ejemplo.

1.1. Gráficos de dispersión (o scatter plots)

Ejemplo 1.1 Queremos investigar la relación entre el porcentaje de niños que ha sido vacunado contra tres enfermedades infecciosas: difteria, pertusis (*tos convulsa*) y tétanos (*DPT*, que se suele denominar, *triple bacteriana*) en un cierto país y la correspondiente tasa de mortalidad infantil para niños menores a cinco años. El Fondo de las Naciones Unidas para la Infancia considera a la tasa de mortalidad infantil para niños menores a cinco años como uno de los indicadores más importantes del nivel de bienestar de una población infantil. Datos publicados en United Nations Children's Fund, *The State of the World's Children 1994*, New York: Oxford University Press. Y tratados en el libro Pagano, Gauvreau, y Pagano [2000], Capítulo 17.

Los datos para 20 países, del año 1992, se muestran en la Tabla 2. Si X representa el porcentaje de niños vacunados a la edad de un año, e Y representa la tasa de mortalidad infantil de niños menores de 5 años, tenemos una pareja de resultados (X_i, Y_i) para cada país en la muestra.

¿Cómo se lee la información desplegada en la Tabla 2? Por ejemplo, para Bolivia $X_1 = 77,0$, es decir, en el año 1992, un 77% de los niños menores de un año

Tabla 2: Datos para 20 países en los que se midieron dos variables, X : porcentaje de niños vacunados a la edad de un año en cada país, Y : es la tasa de mortalidad infantil de niños menores de 5 años en cada país. Archivo: `paises.txt`.

| País | Porcentaje vacunado | Tasa de mortalidad menor a 5 años |
|-----------------|------------------------|--------------------------------------|
| Bolivia | 77,0 | 118,0 |
| Brasil | 69,0 | 65,0 |
| Camboya | 32,0 | 184,0 |
| Canadá | 85,0 | 8,0 |
| China | 94,0 | 43,0 |
| República Checa | 99,0 | 12,0 |
| Egipto | 89,0 | 55,0 |
| Etiopía | 13,0 | 208,0 |
| Finlandia | 95,0 | 7,0 |
| Francia | 95,0 | 9,0 |
| Grecia | 54,0 | 9,0 |
| India | 89,0 | 124,0 |
| Italia | 95,0 | 10,0 |
| Japón | 87,0 | 6,0 |
| México | 91,0 | 33,0 |
| Polonia | 98,0 | 16,0 |
| Federación Rusa | 73,0 | 32,0 |
| Senegal | 47,0 | 145,0 |
| Turquía | 76,0 | 87,0 |
| Reino Unido | 90,0 | 9,0 |

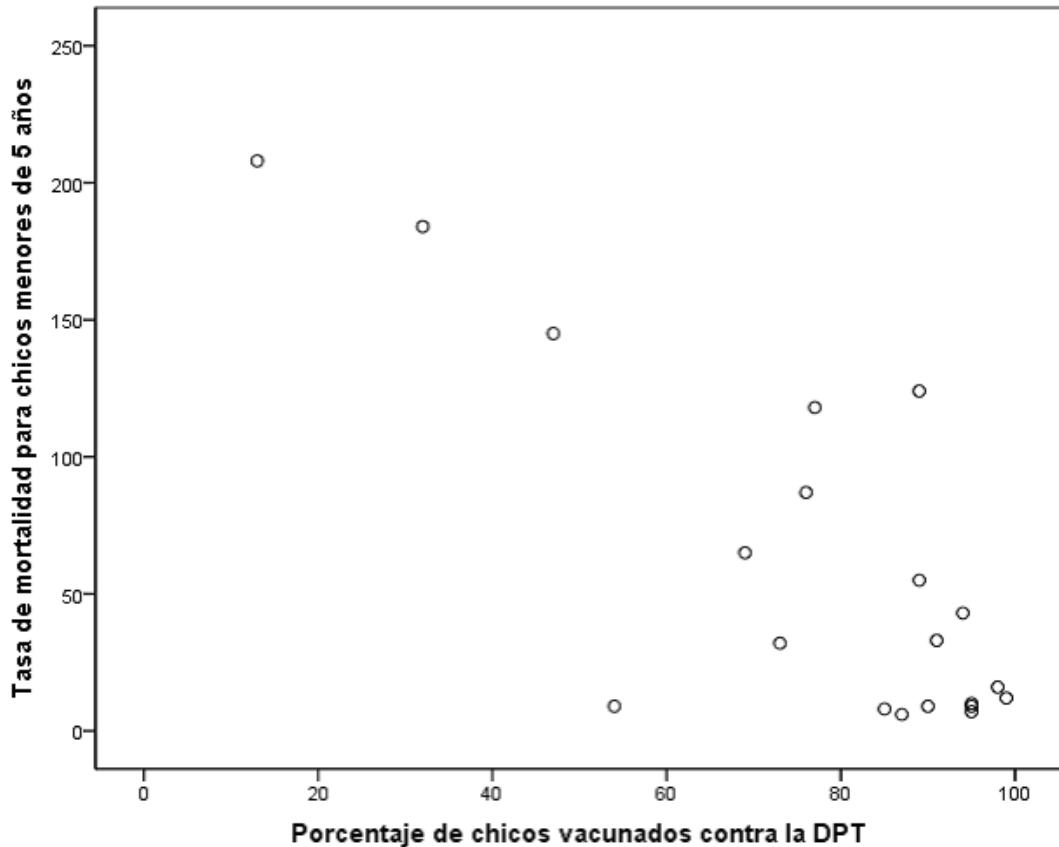
estaban vacunados contra la DPT y (en el año 1992) 118 niños menores de 5 años murieron por cada 1000 niños nacidos vivos.

¿Cómo puede visualizarse esta información? La forma más sencilla es mediante un gráfico de dispersión (o scatter plot). En un scatter plot se ubican los resultados de una variable (X) en el eje horizontal y los de la otra variable (Y) en el eje vertical. Cada punto en el gráfico representa una observación (X_i, Y_i) .

En este tipo de gráfico se pierde la información del individuo (paciente o país), y aunque si hubiera pocos puntos se los podrían rotular, esencialmente esta información no suele estar disponible en un scatter plot. El gráfico de dispersión de los datos de la Tabla 2 puede verse en la Figura 1. Ahí vemos que, por ejemplo, Bolivia está representada por el punto $(77, 118)$.

Usualmente con este gráfico podemos determinar si existe algún tipo de relación

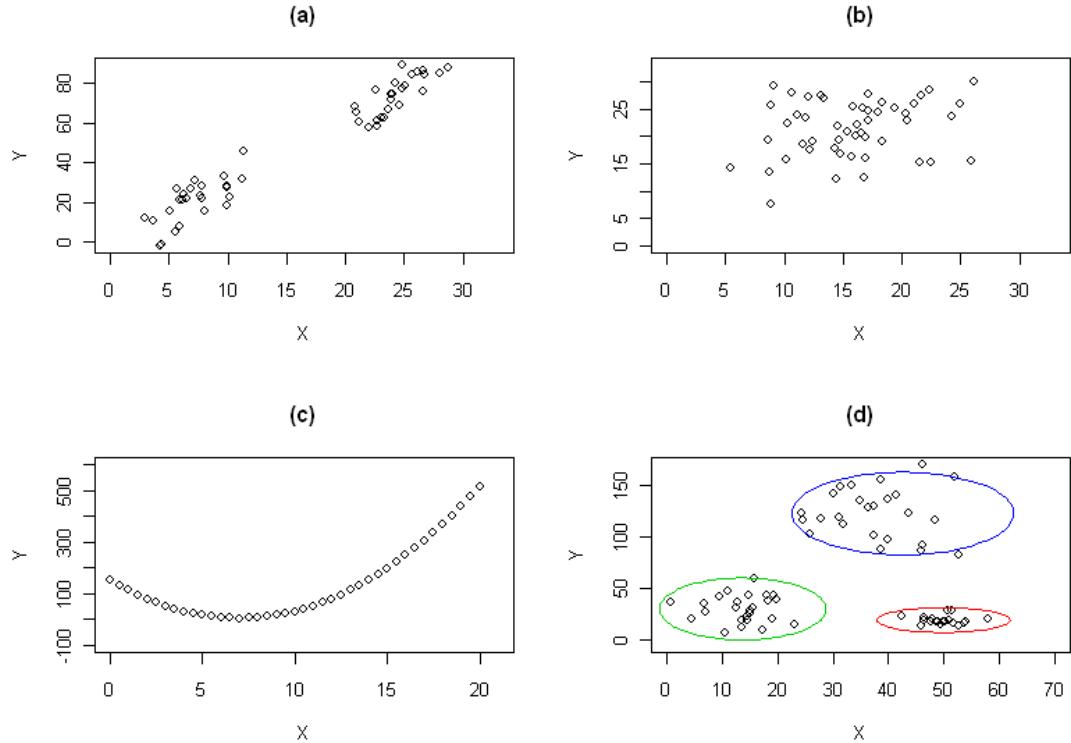
Figura 1: Scatter plot: tasa de mortalidad infantil (menor a 5 años) versus el porcentaje de chicos menores de un año vacunados contra la DPT.



entre X e Y . Para este caso vemos que a medida que aumenta el porcentaje de niños inmunizados, decrece la tasa de mortalidad. ¿Qué otras cosas podríamos observar? En la Figura 2 ilustramos algunas posibilidades, que describimos a continuación.

- **Ausencia de datos.** Puede ser que no hayamos medido ninguna observación cuya variable X se encuentre entre cierto rango de valores (en la Figura 2 (a) por ejemplo, no hay observaciones con coordenada X entre los valores 13 y 21). O que esta combinación entre X e Y no exista, o no se dé biológicamente. Esto indica que la relación que observamos entre las variables graficadas es solamente válida para algunos valores de las variables.
- **No asociación.** ¿Cómo luce un gráfico de dispersión de dos variables que no están asociadas? En la Figura 2 (b) hay un ejemplo. Luce como una nube de

Figura 2: Gráficos de dispersión de cuatro conjuntos de datos diferentes: (a) ausencia de datos; (b) no asociación; (c) vínculo curvilíneo; (d) agrupamientos.



puntos: los valores bajos de X pueden aparecer asociados tanto con valores altos de Y como con valores bajos de Y . Lo mismo para los valores altos de X . Lo mismo para los valores intermedios de X .

- **Vínculo curvilíneo.** Esto aparece cuando los valores de Y se vinculan a los de X por medio de una función. Por ejemplo, si en el eje X graficáramos los valores del tiempo medidos con un cronómetro a intervalos regulares y en el eje Y la posición de un objeto en cada instante de tiempo medido, y si este objeto se moviera siguiendo un movimiento rectilíneo uniformemente variado, observaríamos en el gráfico una función cuadrática, como aparece en la Figura 2 (c). A veces la curva no describe la ubicación de los puntos en la gráfica de manera exacta, sino en forma aproximada (hay errores de medición, por ejemplo, o una relación sólo aproximadamente cuadrática entre las variables).

- **Agrupamientos.** Cuando en el gráfico de dispersión se ven las observaciones separadas en grupos esto puede indicar que hay variables que están faltando incluir en el análisis. Por ejemplo, la Figura 2 (d) puede tratarse de mediciones del largo de pétalo y del sépalo de una flor, de un grupo de flores para las cuales no se ha registrado la variedad. Si habláramos con el biólogo que llevó a cabo las mediciones podríamos encontrar que se trató de tres variedades distintas de flores, que dieron origen a los tres grupos indicados con elipses de colores en el gráfico.

Esencialmente, en el scatter plot nos interesa evaluar

- la *forma* de la relación entre las dos variables
 - lineal
 - no lineal: cuadrática, exponencial, etc.
 - ausencia de relación
- el *sentido* de la asociación, que puede ser
 - asociación creciente: ambas variables aumentan simultáneamente
 - asociación decreciente: cuando una variable aumenta, la otra disminuye
 - no asociación
- la *fuerza* de la asociación, esto tiene que ver con la dispersión de los datos. Si el vínculo puede resumirse con una recta o una curva, cuán alejados de dicha recta (o curva) están los datos. Esto suele resumirse cualitativamente: diremos que la asociación es fuerte, moderada o débil, de acuerdo a si los puntos graficados presentan poca, moderada o mucha dispersión de la recta (o curva) que los describe.
- la identificación de unas pocas observaciones que no siguen el patrón general, o de otras características de los mismos como ausencia de datos en algunas regiones, agrupamientos, variabilidad de los datos que depende de una de las variables, etc.

1.1.1. Desventajas de los scatter plots

Los scatter plots son herramientas básicas del estudio de varias variables simultáneas. Sin embargo adolecen de dos problemas, esencialmente.

1. Si hay muchas observaciones todas iguales, en general no se las puede graficar a todas. En el gráfico de dispersión uno no puede notar si hay puntos repetidos en la muestra observada.
2. Sólo se pueden visualizar los vínculos entre dos variables. En gráficos tridimensionales se podrían graficar hasta tres variables, y luego habría que elegir con mucho cuidado el punto de vista del observador para exhibir las características más sobresalientes del gráfico. Cuando el interés está puesto en estudiar varias variables simultáneamente, pueden hacerse varios gráficos de dispersión simultáneos. Es decir, cuando tenemos las variables (X, Y, Z) haremos tres gráficos: Y versus X , Z versus X , y Z versus Y . Los haremos en la Sección 5.1.1.

1.2. Coeficiente de correlación de Pearson

Descriptivamente hablando, en estas notas estaremos interesados en las situaciones donde aparece una relación entre X e Y del estilo de las graficadas en la Figura 3, que pueden globalmente describirse bien a través de una relación lineal (puntos situados más o menos cerca del gráfico de una recta). Cuando los gráficos de dispersión son del estilo de los que aparecen en la Figura 2 (a), (c) ó (d) las técnicas estadísticas que mejor describen este tipo de vínculo entre las variables no se encuadran dentro de la regresión lineal. En la Figura 3 (a) se ve una *asociación positiva* entre las variables, esto quiere decir que a medida que crece X , esencialmente Y crece. En cambio, en la Figura 3 (b) las variables están *negativamente asociadas*: cuando X crece, Y decrece, en general.

1.2.1. Definición del coeficiente de correlación

Para cuantificar el grado de asociación entre X e Y se pueden describir coeficientes. Antes de hacerlo, repasemos los coeficientes poblacionales que asumimos conocidos, ya que se ven en cualquier curso introductorio de probabilidades y estadística

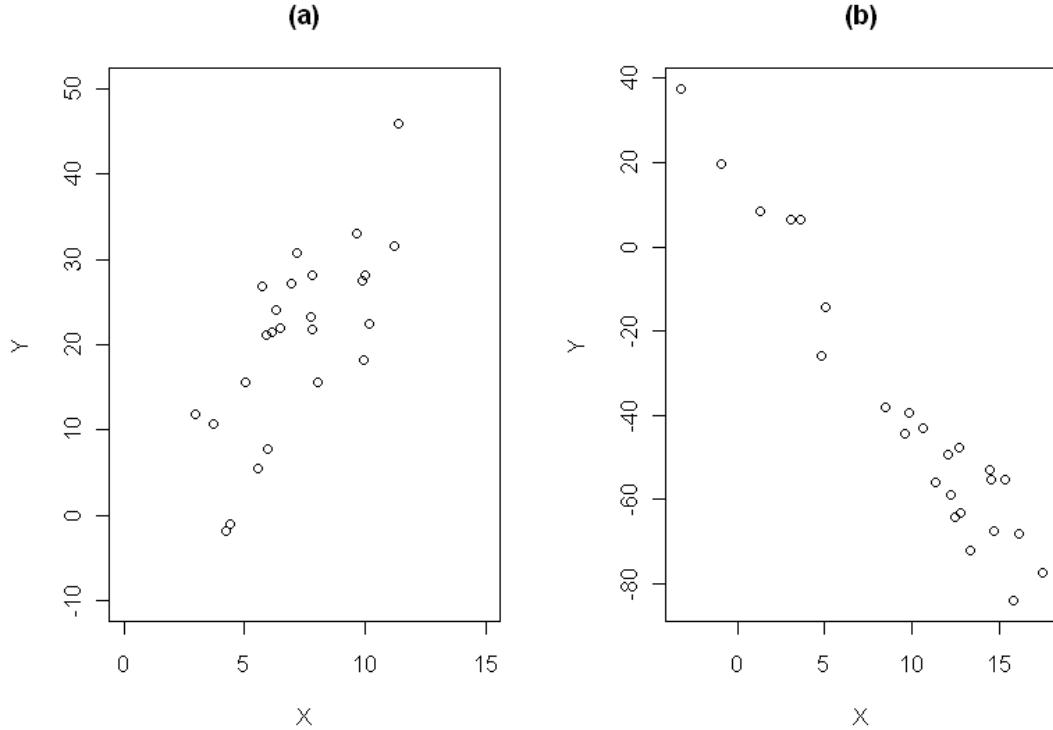
Para una sola variable numérica X podemos definir la **esperanza de X**

$$\mu_X = E(X)$$

como el valor poblacional que describe el centro de la variable. A su vez, tenemos también la **varianza poblacional de X** que es

$$\sigma_X^2 = E([X - E(X)]^2) = Var(X)$$

Figura 3: Dos conjuntos de datos con asociación lineal entre X e Y : el gráfico (a) muestra asociación lineal positiva, el (b) muestra asociación lineal negativa entre ambas.



que es una medida de la variación de la variable X respecto de su centro dado por $E(X)$. A partir de ella se define el **desvío estándar poblacional de X** por

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{Var(X)},$$

que es una medida de dispersión de la variable X .

¿Cómo estimamos estos valores poblacionales, en general desconocidos, a través de una muestra X_1, X_2, \dots, X_n de variables independientes con la misma distribución que la variable X ? A la media poblacional, μ_X la estimamos por el promedio de las n observaciones disponibles. Llamaremos $\hat{\mu}_X$ al estimador, es decir, a la función o cuenta que hacemos con las variables X_1, X_2, \dots, X_n observadas para estimar al número fijo μ_X (en este sentido, $\hat{\mu}_X$ en realidad es un $\hat{\mu}_X(X_1, X_2, \dots, X_n)$), y

escribimos

$$\hat{\mu}_X = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i,$$

\bar{X}_n o bien \bar{X} es el *promedio o media muestral*. A la varianza poblacional la estimamos por

$$\hat{\sigma}_X^2 = S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

que es la *varianza muestral*. Entonces, el desvío estándar poblacional queda estimado por el *desvío estándar muestral*, es decir,

$$\hat{\sigma}_X = S_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

Ahora estamos en condiciones de pensar en cómo definir un coeficiente que resuma el vínculo entre dos variables aleatorias X e Y medidas en el mismo individuo. El más utilizado de todos es el que se conoce como *coeficiente de correlación*, que se simboliza con una letra griega *rho*: ρ ó ρ_{XY} y se define por

$$\begin{aligned} \rho_{XY} &= E \left[\left(\frac{X - \mu_X}{\sigma_X} \right) \left(\frac{Y - \mu_Y}{\sigma_Y} \right) \right] \\ &= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}, \end{aligned}$$

o sea, el número promedio a nivel población del producto de X menos su media por Y menos su media divididos por el producto de los desvíos estándares. ¿Cómo estimamos a ρ ? A través de r el *coeficiente de correlación de Pearson, o coeficiente de correlación muestral*, dado por

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{S_X \cdot S_Y}.$$

Al numerador, se lo denomina covarianza muestral entre X e Y ,

$$\text{covarianza muestral} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

y el denominador es el producto de los desvíos muestrales de cada muestra por

separado

$$\begin{aligned} S_X &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \\ S_Y &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}. \end{aligned}$$

Otra forma de escribir a r es la siguiente

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 \right]}}.$$

Observemos que el numerador $\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ puede ser positivo o negativo, pero el denominador $\sqrt{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 \right]}$ siempre es positivo. Luego el signo de r está determinado por el del numerador. Veamos de qué depende.

$$\text{signo de } (X_i - \bar{X}) = \begin{cases} + & \text{si } X_i \text{ es más grande que } \bar{X} \\ - & \text{si } X_i \text{ es más chico que } \bar{X} \end{cases}$$

y también

$$\text{signo de } (Y_i - \bar{Y}) = \begin{cases} + & \text{si } Y_i \text{ es más grande que } \bar{Y} \\ - & \text{si } Y_i \text{ es más chico que } \bar{Y} \end{cases}$$

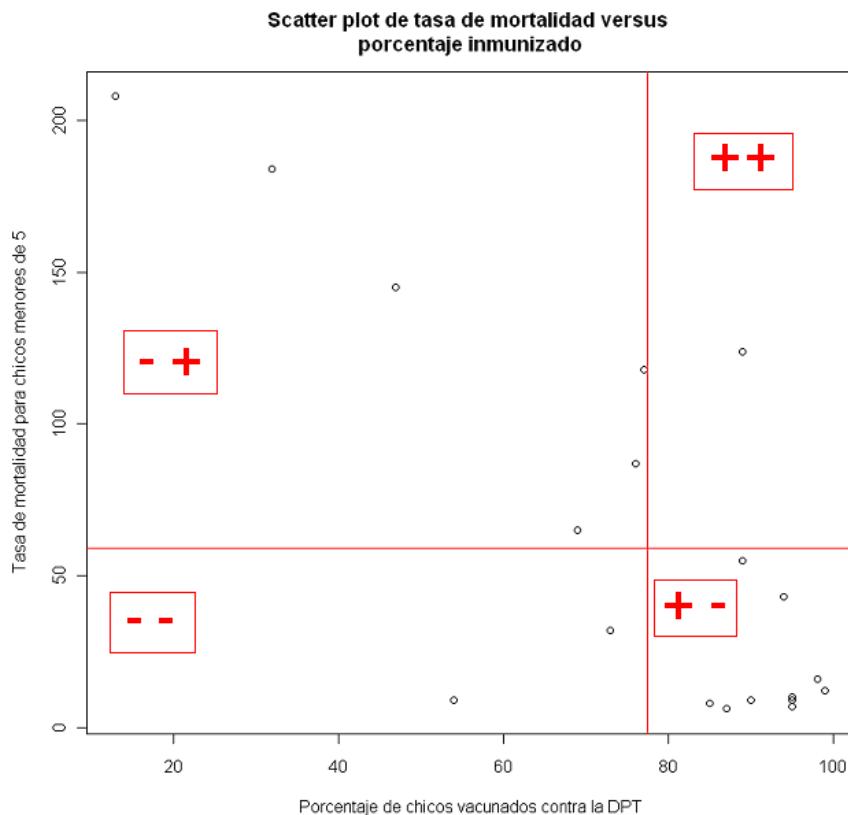
Luego, el

$$\text{signo de } (X_i - \bar{X})(Y_i - \bar{Y}) = \begin{cases} + & \text{si } ++ \text{ ó } -- \\ - & \text{si } +- \text{ ó } -+ \end{cases}$$

Hacemos un scatter plot de las observaciones. Luego ubicamos en el plano el punto (\bar{X}, \bar{Y}) . Trazamos una línea vertical que pase por \bar{X} y otra línea horizontal que pase a la altura de \bar{Y} . Esto divide al gráfico en cuatro cuadrantes, como puede verse en la Figura 4. Luego, el signo del sumando iésimo de r será positivo, si para el individuo iésimo tanto X_i como Y_i son mayores que su respectivo promedio (es decir, la observación cae en el cuadrante noreste, al que hemos denotado por $++$)

o bien ambos valores son simultáneamente menores que su promedio, es decir, la observación cae en el cuadrante suroeste, que hemos denotado por $- -$. En cambio, el sumando iésimo de r será negativo en el caso en el que la observación iésima tenga un valor X_i por encima de su promedio pero la Y_i sea menor que su promedio, o bien la X_i sea menor a su promedio y la Y_i sea mayor a su promedio.

Figura 4: Scatter plot de la tasa de mortalidad versus el porcentaje de niños menores a un año inmunizados, con la recta vertical y horizontal que pasan por (\bar{X}, \bar{Y}) , y los signos de cada sumando que interviene en el cálculo de r .



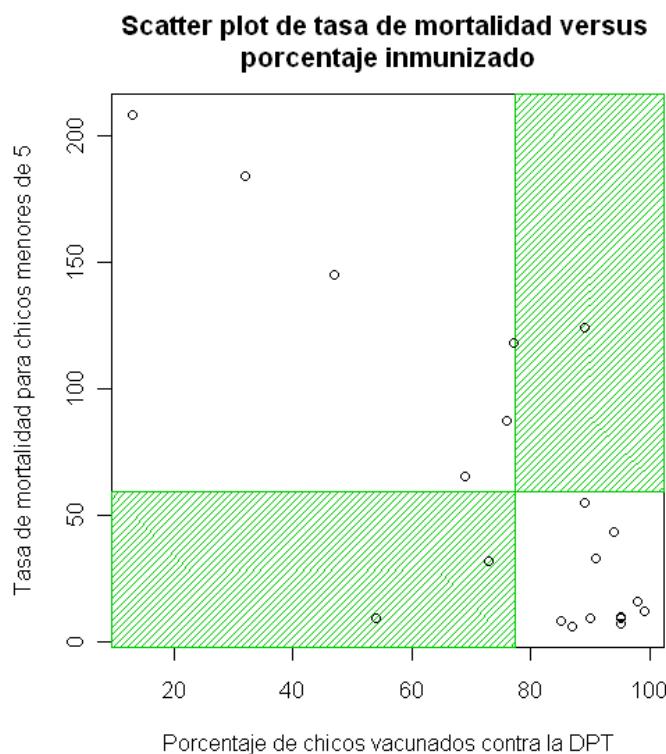
Esto en cuanto a cada sumando en particular. ¿Qué significará el signo de r ?

Si r da positivo, será indicio de que la mayoría de las observaciones caen en los cuadrantes noreste (NE) y suroeste (SO), marcados con color en la Figura 5. Es decir, que cuando los valores de las X suelen estar por encima del promedio ocurre, simultáneamente, que los valores de Y también están sobre su promedio. Análogamente, cuando en un individuo el valor de X está por debajo del promedio, lo mismo ocurre con su valor de Y . En general, un valor positivo de r indica que

hay una asociación positiva entre las variables (cuando una crece, la otra también lo hace).

Si r da negativo, en cambio, tenemos una indicación de mayor número de observaciones en los otros cuadrantes marcados con fondo blanco en la Figura 5, y se invierten las situaciones descriptas anteriormente. Es decir, que cuando los valores de las X suelen estar por encima del promedio ocurre, simultáneamente, que los valores de Y están por debajo de su promedio. Análogamente, cuando en un individuo el valor de X está por debajo del promedio, ocurre lo inverso con su valor de Y , que superará a su promedio. En general, un valor negativo de r es indicador de asociación negativa entre las variables (cuando una crece, la otra decrece).

Figura 5: Scatter plot de la tasa de mortalidad versus el porcentaje de niños menores a un año inmunizados, con los cuatro cuadrantes delimitados por (\bar{X}, \bar{Y}) . Las observaciones que caen en la región coloreada darán sumandos positivos del r .



Ejemplo 1.2 Veamos qué ocurre en nuestro ejemplo. Calculamos los promedios

de ambas variables, obtenemos

$$\begin{aligned}\bar{X} &= 77,4 \\ \bar{Y} &= 59\end{aligned}$$

y le superponemos al scatter plot dos líneas rectas, una vertical que corta al eje x en 77,4 y otra horizontal que corta al eje y en $Y = 59$. Las Figuras 4 y 5 muestran el gráfico de esta situación. Observamos que en los dos cuadrantes coloreados hay muy pocas observaciones (exactamente 3 de un total de 20).

El coeficiente de correlación muestral en este caso da $-0,791$, un valor negativo, lo cual hubiéramos podido anticipar ya que la mayoría de los términos involucrados en el cálculo de r (17 de los 20 sumandos) serán menores o iguales a cero.

1.2.2. Propiedades del coeficiente de correlación muestral (y también de ρ)

A continuación damos las propiedades del coeficiente de correlación muestral r , pero estas también son válidas para el coeficiente de correlación poblacional ρ .

1. $-1 \leq r \leq 1$. El valor del coeficiente r está entre 1 y menos 1 porque puede probarse que el denominador es más grande (a lo sumo igual) que el numerador.
2. El valor absoluto de r , $|r|$ mide la fuerza de la asociación lineal entre X e Y , a mayor valor absoluto, hay una asociación lineal más fuerte entre X e Y .
3. El caso particular $r = 0$ indica que no hay asociación lineal entre X e Y .
4. El caso $r = 1$ indica asociación lineal perfecta. O sea que los puntos están ubicados sobre una recta de pendiente (o inclinación) positiva.
5. En el caso $r = -1$ tenemos a los puntos ubicados sobre una recta de pendiente negativa (o sea, decreciente).
6. El signo de r indica que hay asociación positiva entre las variables (si $r > 0$); o asociación negativa entre ellas (si $r < 0$).
7. $r = 0,90$ indica que los puntos están ubicados muy cerca de una recta creciente.
8. $r = 0,80$ indica que los puntos están cerca, pero no tanto, de una recta creciente. En la Figura 6 se pueden ver distintos grados de correlación, que están comentados más abajo.

9. r no depende de las unidades en que son medidas las variables (milímetros, centímetros, metros o kilómetros, por ejemplo) .
10. Los roles de X e Y son simétricos para el cálculo de r .
11. **Cuidado:** el coeficiente de correlación de Pearson es muy sensible a observaciones atípicas. Hay que hacer **siempre** un scatter plot de los datos antes de resumirlos con r . La sola presencia de una observación atípica (o de unas pocas observaciones que siguen un patrón raro) puede hacer que el valor de r resulte, por ejemplo positivo, cuando en verdad las variables X e Y están negativamente asociadas. Si vemos el scatterplot de los datos, esta (o estas) observación atípica debiera destacarse en su patrón alejado del resto y seremos capaces de detectarla. Si en cambio sólo nos limitamos a calcular el r esta situación podría escapársenos y podríamos terminar infiriendo un vínculo entre X e Y que es espúreo.

Un ejemplo de fuerte correlación positiva se da entre el volumen espiratorio esforzado (VEF), una medida de la función pulmonar, y la altura. En la Figura 6 (a) se muestra un gráfico de dispersión de observaciones de estas variables, que tienen correlación $\rho = 0,90$. En la Figura 6 (b) se puede observar una correlación positiva más débil entre niveles séricos de colesterol y la ingesta diaria de colesterol, aquí $\rho = 0,3$. Una fuerte correlación negativa ($\rho = -0,8$) se da entre la frecuencia del pulso en reposo (o la frecuencia cardíaca) y la edad, medidas en niños menores a diez años. Ahí vemos que a medida que un chico crece, la frecuencia de su pulso desciende. Una correlación negativa más débil $\rho = -0,2$ existe entre VEF y el número de cigarrillos fumados por día, como se ve en la Figura 6 (d).

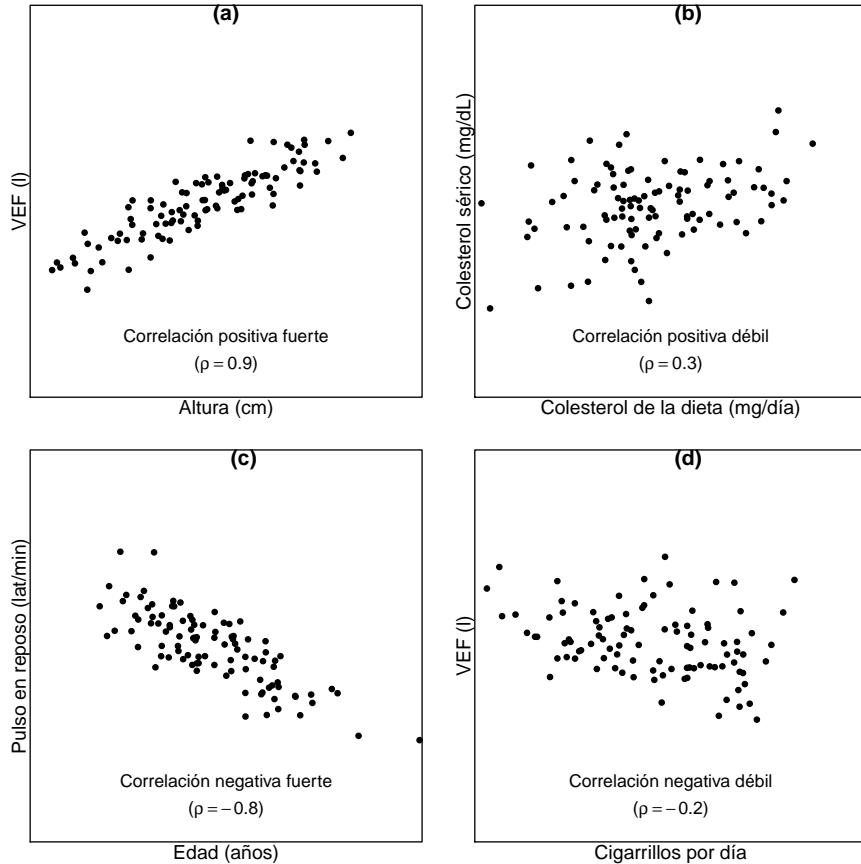
Cabe hacer un comentario respecto de la interpretación del coeficiente de correlación. Altos grados de asociación lineal entre X e Y no son señales de causalidad, es decir, una relación de causa y efecto entre ambas variables. Una alta correlación observada entre dos variables es compatible con la situación de que existan modelos que explican a Y por X , o bien a X por Y , o bien que exista una tercer variable que las determine a ambas simultáneamente.

1.2.3. Inferencia de ρ

La pregunta que nos hacemos en esta sección es la clásica pregunta de inferencia estadística, ¿qué podemos decir de ρ a partir de r ?

Queremos sacar conclusiones acerca del parámetro poblacional ρ a partir de la muestra de observaciones $(X_1, Y_1), \dots, (X_n, Y_n)$. En el ejemplo, la pregunta que podríamos hacer es ¿qué podemos decir del vínculo entre inmunización contra la DPT y la tasa de mortalidad infantil para menores a cinco años? Sólo contamos

Figura 6: Interpretación de distintos grados de correlación. Inspirado en: Rosner [2006], pág. 137.



con observaciones de 20 países en 1992. El test que más nos interesará es el que tiene las siguientes hipótesis

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0, \end{aligned}$$

ya que la afirmación de la hipótesis nula, $\rho = 0$, puede escribirse como “no hay asociación lineal entre X e Y a nivel poblacional”, mientras que la hipótesis alternativa postula que sí hay tal relación entre las variables. O sea, en el caso del ejemplo, sabemos que la correlación muestral observada entre ambas variables fue $r = -0,791$, y la pregunta ¿será que entre las dos variables consideradas no hay asociación lineal, y sólo por casualidad en la muestra obtenida vemos un valor de

$r = -0,791$? ¿O será que $\rho \neq 0$? Como el coeficiente de correlación muestral r es un estimador del valor poblacional ρ , a través de él podemos proponer un test para estas hipótesis.

Test para $\rho = 0$ Los supuestos para llevar a cabo el test son que los pares de observaciones $(X_1, Y_1), \dots, (X_n, Y_n)$ sean independientes entre sí, idénticamente distribuidos, y tengan distribución (conjunta) normal bivariada (ver la definición de esto en la Observación 1.1). En particular, esto implica que cada una de las muestras X_1, \dots, X_n e Y_1, \dots, Y_n tengan distribución normal. Si la hipótesis nula es verdadera, entonces el estadístico

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

que no es más que $\hat{\rho}$ dividido por un estimador de su desvío estándar, tiene distribución t de Student con $n - 2$ grados de libertad, lo cual notaremos

$$T \sim t_{n-2} \text{ bajo } H_0.$$

Si H_0 fuera cierto, ρ sería cero y su estimador $r = \hat{\rho}$ debería tomar valores muy cercanos a cero. Lo mismo debería pasar con T que es $\hat{\rho}$ estandarizado. Por lo tanto rechazaríamos la hipótesis nula cuando T tome valores muy alejados de 0, tanto positivos como negativos. El test rechaza H_0 cuando T toma valores muy grandes o muy pequeños, es decir, rechazamos la hipótesis nula con nivel $1 - \alpha$ cuando

$$T \geq t_{n-2, 1-\frac{\alpha}{2}} \text{ ó } T \leq -t_{n-2, 1-\frac{\alpha}{2}}$$

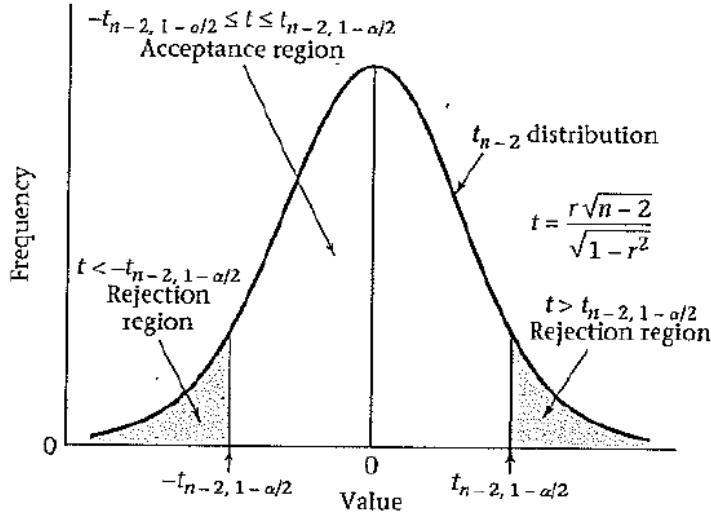
donde $t_{n-2, 1-\frac{\alpha}{2}}$ es el percentil $1 - \frac{\alpha}{2}$ de una distribución t_{n-2} , o sea el valor que deja a su izquierda un área de $1 - \frac{\alpha}{2}$. Es un test bilateral. La región de rechazo aparece dibujada en la Figura 7. El p-valor puede calcularse como

$$p\text{-valor} = P(|T| \geq |T_{obs}|),$$

donde $T \sim t_{n-2}$ y en general lo devuelve el paquete estadístico. Si el tamaño de la muestra fuera suficientemente grande, aplicando una versión del teorema central del límite no necesitaríamos que la muestra fuera normal bivariada.

Ejemplo 1.3 En la Tabla 3 aparece la salida del software libre R R Core Team [2015] para los datos del Ejemplo 1.1. Hemos llamado `immunized` al porcentaje de chicos vacunados contra la DPT y `under5` a la tasa de mortalidad para chicos menores a 5 años. Vemos que en este caso el p-valor del test resulta ser menor a 0,05, por lo que rechazamos la hipótesis nula y concluimos que el coeficiente de

Figura 7: Región de rechazo y aceptación para el test de t para una correlación. Fuente: Rosner [2006], pág. 457.



correlación poblacional ρ es no nulo, mostrando que la tasa de vacunación y la tasa de mortalidad infantil menor a 5 años están correlacionadas. Confiaremos en esta conclusión si somos capaces de creer que los datos de la muestra conjunta provienen de una distribución normal bivariada. En particular, debe cumplirse que ambas variables tengan distribución normal. Para validar al menos este último supuesto deberíamos realizar gráficos de probabilidad normal (qq-plots), histogramas o boxplots, y tests de normalidad (por ejemplo, el test de Shapiro-Wilks) y ver que ambos conjuntos de datos pasan las comprobaciones. Sin embargo, para estos conjuntos de datos no puede asumirse la distribución normal ya que ambos tienen distribución asimétrica: el porcentaje de niños vacunados con cola pesada a la derecha, la tasa de mortalidad con cola pesada a izquierda, como puede observarse en la Figura 8. Por lo tanto, no puede asumirse que conjuntamente tengan distribución normal bivariada, y no puede aplicarse el test de correlación antes descripto.

Observación 1.1 ¿Qué quiere decir que las observaciones $(X_1, Y_1), \dots, (X_n, Y_n)$ tengan distribución conjunta normal bivariada? Es un término técnico. Decir que un vector aleatorio (X, Y) tenga dicha distribución conjunta quiere decir que existen cinco números reales $\mu_1, \mu_2, \sigma_1 > 0, \sigma_2 > 0$ y $-1 < \rho < 1$ tales que la función

Tabla 3: Cálculo de la correlación entre el porcentaje de chicos vacunados contra la DPT y la tasa de mortalidad para chicos menores a 5 años, con el cálculo del p-valor para el test de las hipótesis $H_0 : \rho = 0$, versus $H_1 : \rho \neq 0$, e intervalo de confianza para ρ . Salida del R.

```
> cor.test(immunized,under5, method = "pearson")

Pearson's product-moment correlation

data: immunized and under5
t = -5.4864, df = 18, p-value = 3.281e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.9137250 -0.5362744
sample estimates:
cor
-0.7910654
```

de densidad conjunta para el vector (X, Y) está dada por

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - 2\rho \left(\frac{x-\mu_1}{\sigma_1} \right) \left(\frac{y-\mu_2}{\sigma_2} \right) \right] \right\}, \quad (1)$$

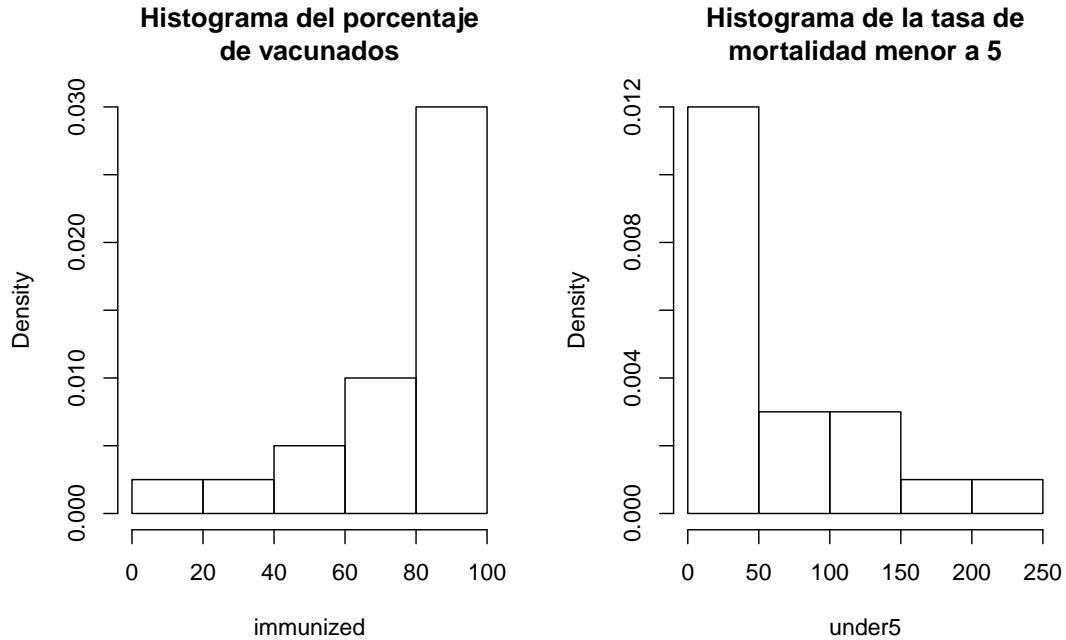
de modo que dada una determinada región A contenida en \mathbb{R}^2 , la probabilidad de que el vector (X, Y) pertenezca a dicha región está dada por

$$P((X, Y) \in A) = \int \int_A f_{XY}(x, y) dx dy.$$

Es decir, se calcula hallando el área bajo la función f_{XY} definida en (1) y sobre la región A . En la Figura 9 puede verse el gráfico de la densidad conjunta, que es una superficie en el espacio tridimensional. A los números $\mu_1, \mu_2, \sigma_1, \sigma_2$ y ρ se los denomina parámetros de la distribución normal bivariada (ya que una vez que se fija sus valores numéricos, la densidad queda determinada).

Como ya mencionamos, puede probarse que cuando (X, Y) tiene distribución normal bivariada, entonces cada una de las variables X e Y tienen distribución normal. Más aún, μ_1 es la media de X y σ_1^2 es su varianza, lo cual notamos $X \sim$

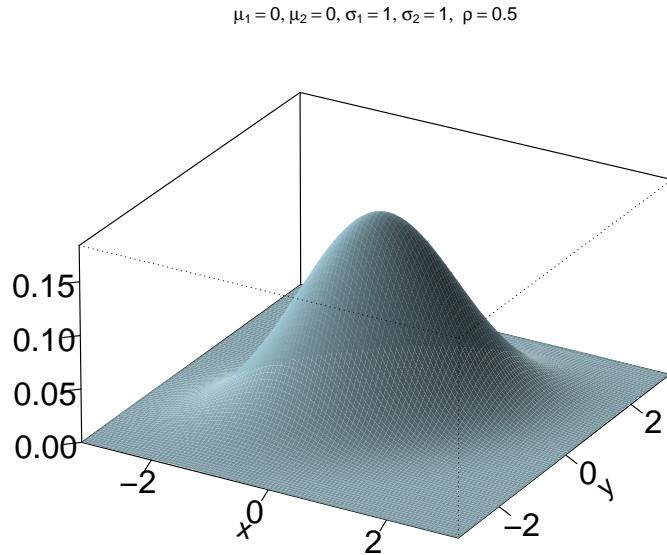
Figura 8: Histograma para los datos de porcentaje de niños vacunados y tasas de mortalidad infantil, para los datos del Ejemplo 1.1.



$N(\mu_1, \sigma_1^2)$ y también $Y \sim N(\mu_2, \sigma_2^2)$. Además ρ es el coeficiente de correlación entre X e Y . La recíproca no es siempre cierta, sin embargo: si sabemos que tanto X como Y tienen distribución normal, entonces no siempre la distribución conjunta del vector (X, Y) es la normal bivariada dada por (1).

Observación 1.2 ¿Cómo es un scatterplot de observaciones $(X_1, Y_1), \dots, (X_n, Y_n)$ independientes que tienen distribución normal bivariada? Por supuesto, el gráfico de dispersión dependerá de los valores de los parámetros. En general, los puntos yacerán en una zona que puede ser razonablemente bien descripta como una elipse con centro en (μ_1, μ_2) . En la Figura 10 se ven los gráficos de dispersión correspondientes a distintas combinaciones de parámetros. En ellos vemos que a medida que ρ se acerca a uno, los puntos se acercan más a una recta, cuya pendiente y ordenada al origen depende de los valores de los cinco parámetros. La pendiente será positiva si ρ es positiva, y será negativa en caso contrario.

Figura 9: Densidad conjunta normal bivariada, definida en (1.1) construida con los valores de parámetros especificados más abajo ($\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.5$).



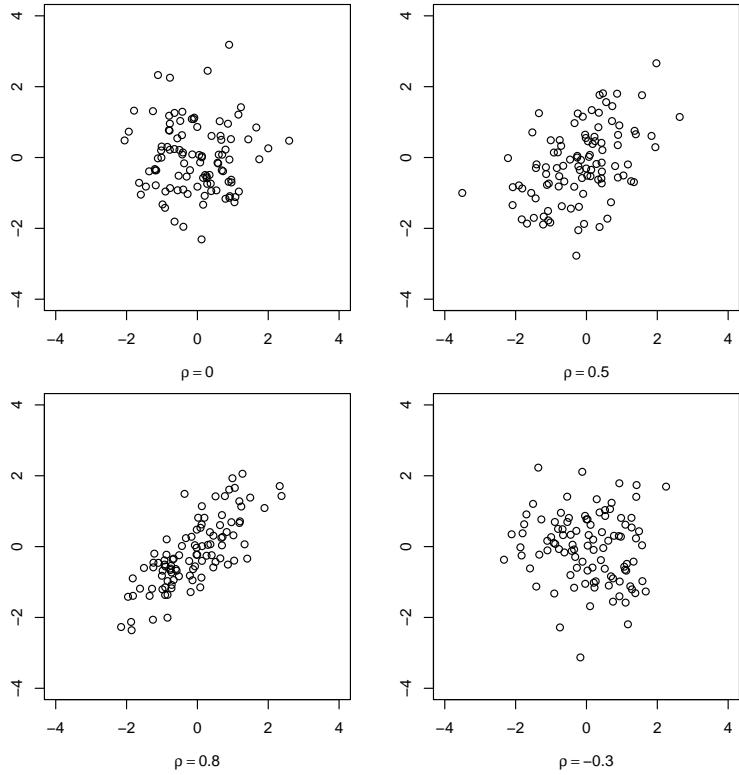
Test para $\rho = \rho_0$ A veces es de interés testear si la verdadera correlación poblacional es igual a un cierto valor ρ_0 predeterminado. Es decir, se quieren testear las hipótesis

$$\begin{aligned} H_0 &: \rho = \rho_0 \\ H_1 &: \rho \neq \rho_0. \end{aligned}$$

Por supuesto, esto no ocurre muy frecuentemente, pero puede surgir una pregunta de este tipo en algunas aplicaciones. La cuestión es que cuando $\rho = \rho_0$ el estadístico T descripto en la sección anterior no tiene distribución t de Student, sino que tiene una distribución sesgada.

Para testear las hipótesis recién propuestas, está el test basado en la transformación z de Fisher. Como en el anterior se requiere que las observaciones $(X_1, Y_1), \dots, (X_n, Y_n)$ sean independientes entre sí, idénticamente distribuidos y

Figura 10: Gráficos de dispersión de datos bivariados con distribución normal bivariada con parámetros: $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$ y distintos valores de ρ , que se indican en cada gráfico.



tengan distribución conjunta normal bivariada. El test se realiza de la siguiente forma. Primero se calcula la transformación z de Fisher sobre el coeficiente de correlación, que es

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right).$$

Bajo H_0 , puede probarse que la distribución de z es aproximadamente

$$N \left(\frac{1}{2} \ln \left(\frac{1+\rho_0}{1-\rho_0} \right), \frac{1}{n-3} \right).$$

Luego, esta distribución se utiliza para calcular el p-valor del test, o dar la región de rechazo de nivel α . El p-valor se obtendrá estandarizando el valor de z observado y calculando la probabilidad de obtener un valor tan alejado del cero o más alejado aún como el observado, usando la función de distribución acumulada

normal estándar, es decir

$$\begin{aligned} z_{\text{est}} &= \frac{\frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) - \frac{1}{2} \ln \left(\frac{1+\rho_0}{1-\rho_0} \right)}{\sqrt{\frac{1}{n-3}}} \\ p - \text{valor} &= P(|Z| \geq |z_{\text{est}}|). \end{aligned}$$

Esto lo realiza el paquete estadístico. En el ejemplo no puede aplicarse este test puesto que hemos visto ya que ninguna de las dos muestras es normal (por lo tanto los pares no pueden tener distribución conjunta normal bivariada), y este test es aún más sensible que el anterior al supuesto de normalidad.

Intervalo de confianza para ρ Puede resultar de interés disponer de un intervalo de confianza para el verdadero coeficiente de correlación poblacional, ρ , que nos dé indicios de qué parámetros poblacionales pueden describir apropiadamente a nuestros datos. Para construirlo se recurre a la transformación z presentada en la sección anterior. Luego se utiliza la distribución normal para encontrar los percentiles adecuados para describir el comportamiento del z estandarizado, y finalmente se aplica la inversa de la transformación z para obtener un intervalo de confianza para ρ . Los supuestos para llevar a cabo este procedimiento son los mismos que los presentados para ambos tests de las subsecciones anteriores. Finalmente el intervalo de confianza de nivel $1 - \alpha$ para ρ está dado por $[\rho_I, \rho_D]$ donde

$$\begin{aligned} z_{\text{obs}} &= \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \\ \rho_I &= \frac{e^{2(z_{\text{obs}} - z_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{n-3}})} - 1}{e^{2(z_{\text{obs}} - z_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{n-3}})} + 1} \\ \rho_D &= \frac{e^{2(z_{\text{obs}} + z_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{n-3}})} - 1}{e^{2(z_{\text{obs}} + z_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{n-3}})} + 1} \end{aligned}$$

y $z_{1-\frac{\alpha}{2}}$ es el percentil $1 - \frac{\alpha}{2}$ de la normal estándar. En el caso del ejemplo no tiene sentido mirarlo porque no se cumplen los supuestos, pero puede verse la salida del R en la Tabla 3 donde aparece calculado por la computadora, y da $[-0,91, -0,54]$.

1.3. Coeficiente de correlación de Spearman

Existen otras medidas de asociación entre dos variables que no son tan sensibles a observaciones atípicas como el coeficiente de correlación de Pearson, ni necesitan el supuesto de normalidad para testearse. La más difundida de ellas es

el coeficiente de correlación de Spearman, que presentamos en esta sección. El coeficiente de correlación de Spearman se encuadra entre las técnicas estadísticas no paramétricas, que resultan robustas bajo la presencia de outliers ya que reemplazan los valores observados por los rangos o rankings de las variables. Se calcula del siguiente modo.

1. Se ordena cada muestra por separado, de menor a mayor. A cada observación se le calcula el ranking que tiene (o rango, o número de observación en la muestra ordenada). De este modo, la observación más pequeña de las X' s recibe el número 1 como rango, la segunda recibe el número 2, etcétera, la más grande de todas las X' s recibirá el rango n . Si hubiera dos o más observaciones empataadas en algún puesto (por ejemplo, si las dos observaciones más pequeñas tomaran el mismo valor de X , entonces se promedian los rangos que les tocarián: cada una tendrá rango 1,5, en este ejemplo, ya que $\frac{1+2}{2} = 1,5$. En el caso en el que las tres primeras observaciones fueran empataadas, a las tres les tocaría el promedio entre 1, 2 y 3, que resultará ser $\frac{1+2+3}{3} = 2$). A este proceso se lo denomina *ranquear las observaciones X* . Llamemos $R(X_i)$ al rango obtenido por la i -ésima observación X .
2. Se reemplaza a cada observación X_i por su rango $R(X_i)$.
3. Se ranquean las observaciones Y , obteniéndose $R(Y_i)$ de la misma forma en que se hizo en el ítem 1 para las X' s.
4. Se reemplaza a cada observación Y_i por su rango $R(Y_i)$. Observemos que conocemos la suma de todos los rangos de ambas muestras (es la suma de $1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}$).
5. Se calcula el coeficiente de correlación de Pearson entre los pares $(R(X_i), R(Y_i))$. El valor obtenido es el coeficiente de correlación de Spearman, que denotaremos r_S .

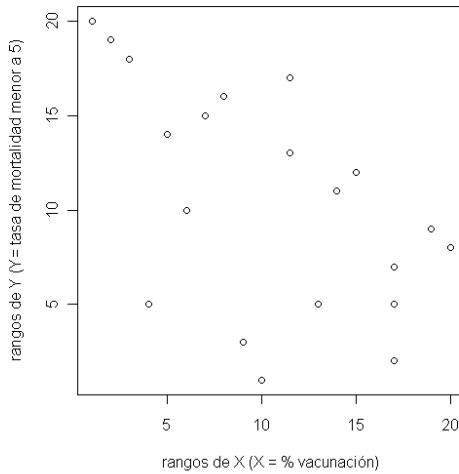
Ilustramos el procedimiento con los datos del Ejemplo 1.1, de la vacunación de DPT, en la Tabla 4. Allí figuran las originales X e Y en las columnas 1 y 3, y los rangos de cada muestra: los rangos de las X' s en la columna 2 y los rangos de las Y' s en la columna 4. Ahí vemos que Etiopía es el país de la muestra con menor tasa de vacunación, por eso su valor X recibe el rango 1. Lo sigue Camboya. Observamos que hay dos países cuyo porcentaje de vacunación es 89 %: Egipto e India. Ambos empatan en los puestos 11 y 12 de la muestra ordenada, por eso reciben el rango 11,5. Y también hay 3 países con un 95 % de bebés vacunados (Finlandia, Francia e Italia) que, a su vez, empatan en los puestos 16, 17 y 18 y reciben el rango promedio de esos tres valores, o sea, 17. Es interesante observar

que Etiopía recibe el rango 1 (el menor) para el porcentaje de vacunación, y el rango 20 (el mayor) para la tasa de mortalidad menor a 5 años, Camboya, a su vez, recibe el rango 2 (el segundo más chico) para el porcentaje de vacunación, y el rango 19 (el penúltimo) para la tasa de mortalidad. En ambos órdenes, lo sigue Senegal, esto muestra la asociación negativa entre ambas variables. Para evaluar si esto ocurre con el resto de los países, hacemos un scatter plot de los rangos de Y versus los rangos de X en la Figura 11. En ella se ve una asociación negativa entre los rangos de ambas variables, aunque no se trata de una asociación muy fuerte, sino más bien moderada. Los tres puntos con menores rangos de X mantienen una relación lineal perfecta, como habíamos observado. Sin embargo, ese ordenamiento se desdibuja en las demás observaciones.

Tabla 4: Datos para los 20 países, con las variables, X : porcentaje de niños vacunados a la edad de un año en cada país, rangos de la X : ranking que ocupa la observación en la muestra ordenada de las X 's, Y : tasa de mortalidad infantil de niños menores de 5 años en cada país, rangos de la Y : posición que ocupa la observación en la muestra ordenada de las Y 's.

| País | Porcentaje vacunado (X) | Rangos de X | Tasa de mortalidad menor a 5 años (Y) | Rangos de Y |
|-----------------|-----------------------------|---------------|---|---------------|
| Bolivia | 77,0 | 8 | 118,0 | 16 |
| Brasil | 69,0 | 5 | 65,0 | 14 |
| Camboya | 32,0 | 2 | 184,0 | 19 |
| Canadá | 85,0 | 9 | 8,0 | 3 |
| China | 94,0 | 15 | 43,0 | 12 |
| República Checa | 99,0 | 20 | 12,0 | 8 |
| Egipto | 89,0 | 11,5 | 55,0 | 13 |
| Etiopía | 13,0 | 1 | 208,0 | 20 |
| Finlandia | 95,0 | 17 | 7,0 | 2 |
| Francia | 95,0 | 17 | 9,0 | 5 |
| Grecia | 54,0 | 4 | 9,0 | 5 |
| India | 89,0 | 11,5 | 124,0 | 17 |
| Italia | 95,0 | 17 | 10,0 | 7 |
| Japón | 87,0 | 10 | 6,0 | 1 |
| México | 91,0 | 14 | 33,0 | 11 |
| Polonia | 98,0 | 19 | 16,0 | 9 |
| Federación Rusa | 73,0 | 6 | 32,0 | 10 |
| Senegal | 47,0 | 3 | 145,0 | 18 |
| Turquía | 76,0 | 7 | 87,0 | 15 |
| Reino Unido | 90,0 | 13 | 9,0 | 5 |

Figura 11: Gráfico de dispersión entre los rangos de Y (es decir, los rangos de la tasa de mortalidad menor a 5 años) y los rangos de X (es decir, del porcentaje de niños menores a un año vacunados contra la DPT). Se ve una asociación negativa, aunque no muy estrecha.



¿Cómo resumimos el grado de asociación observado entre los rangos? Con el cálculo del coeficiente de correlación entre ellos. En este caso da $r_S = -0,543$, como puede verse en la Tabla 5. Este número resulta menor en magnitud que el coeficiente de correlación de Pearson, pero sugiere una moderada relación entre las variables. Esta asociación es negativa.

Otro test de asociación entre variables. El coeficiente de correlación de Spearman puede usarse para testear las hipótesis

H_0 : No hay asociación entre X e Y

H_1 : Hay asociación entre X e Y : hay una tendencia de que los valores más grandes de X se aparezcan con los valores más grandes de Y , o bien la tendencia es que los valores más pequeños de X se aparezcan con los valores más grandes de Y

Como H_1 incluye la posibilidad de que la asociación sea positiva o negativa, este es un test de dos colas. El test rechazará para valores grandes de $|r_S|$ (valor absoluto de r_S). La distribución de r_S bajo H_0 no es difícil de obtener. Se basa en el hecho de

que, bajo H_0 , para cada rango de X_i , $R(X_i)$, todos los rangos de Y_i son igualmente probables, siempre que no haya asociación entre ambas variables. El p -valor puede calcularse de manera exacta si $n < 10$ y no hay empates en la muestra, y de manera aproximada para n mayores.

Si n es muy grande, se utiliza la misma distribución t de la Sección anterior, t_{n-2} . La ventaja de este test por sobre el test de Pearson es que requiere menos supuestos para llevarlo a cabo. Basta con que los pares de observaciones $(X_1, Y_1), \dots, (X_n, Y_n)$ sean independientes entre sí e idénticamente distribuidos. No es necesario asumir nada respecto de la distribución de cada muestra, de hecho basta que la escala de las observaciones sea ordinal para poder aplicarlo. Puede utilizarse si hay observaciones atípicas. La desventaja radica en la potencia del test. El test de Spearman tiene una potencia menor en el caso en el que ambas muestras son normales (en cualquier otro caso, el de Pearson no puede aplicarse). Pero, por supuesto que si con el test de Spearman se logra rechazar la hipótesis nula, ya no es necesario preocuparse por la potencia, ni utilizar el coeficiente de Pearson, que resulta más eficiente.

Tabla 5: Cálculo de la correlación de Spearman entre el porcentaje de chicos vacunados contra la DPT (`immunized`) y la tasa de mortalidad para chicos menores a 5 años (`under5`), con el cálculo del p-valor con el coeficiente de Spearman, para el test de las hipótesis H_0 : no hay asociación entre las variables, versus H_1 : las variables están positiva o negativamente asociadas. Salida del R.

```
> cor.test(immunized,under5, method = "spearman")

Spearman's rank correlation rho

data: immunized and under5
S = 2052.444, p-value = 0.01332
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
-0.5431913
```

En el ejemplo vemos que el p -valor del test de Spearman es 0,013 que al ser menor a 0,05 nos permite rechazar la hipótesis nula y concluir que la verdadera correlación poblacional entre el porcentaje de niños vacunados y la tasa de mortalidad menor a 5 años, es distinta de cero.

Otra medida no paramétrica de asociación entre dos variables está dada por el τ de Kendall. Resume la asociación a través de los rangos de las observaciones

de ambas muestras, pero de una manera diferente. Puede consultarse Kendall y Gibbons [1990] para una discusión completa del tema.

Un último comentario respecto de la correlación en el contexto del estudio de regresión lineal. En este contexto, no estaremos tan interesados en los tests presentados para probar si existe correlación entre las variables, sino más bien en el uso de la correlación a nivel descriptivo. Nos servirá en una primera etapa exploratoria de los datos para ver si las variables bajo nuestra consideración están asociadas con una variable Y que nos interesa explicar, y qué grado de fuerza tiene esa asociación lineal. Y también nos servirá para entender ciertos comportamientos extraños que presenta la regresión lineal múltiple cuando se ajusta para modelos con muchas covariables muy correlacionadas entre sí.

1.4. Ejercicios

Con R hacer scatterplots es muy sencillo. Además es tan útil lo que puede aprenderse de los datos que vale la pena entrenarse exponiéndose a muchos ejemplos. Con el tiempo se gana familiaridad con los tipos de patrones que se ven. De a poco uno aprende a reconocer cómo los diagramas de dispersión pueden revelar la naturaleza de la relación entre dos variables.

En esta ejercitación trabajaremos con algunos conjuntos de datos que están disponibles a través del paquete `openintro` de R. Brevemente:

`mammals`: El conjunto de datos de mamíferos contiene información sobre 62 especies diferentes de mamíferos, incluyendo su peso corporal, el peso del cerebro, el tiempo de gestación y algunas otras variables.

`bdims`: El conjunto de datos `bdims` contiene medidas de circunferencia del cuerpo y diámetro esquelético para 507 individuos físicamente activos.

`smoking`: El conjunto de datos `smoking` contiene información sobre los hábitos de fumar de 1.691 ciudadanos del Reino Unido.

`cars`: El conjunto de datos `cars` está compuesto por la información de 54 autos modelo 1993. Se relevan 6 variables de cada uno (tamaño, precio en dólares, rendimiento en ciudad (millas por galón), tipo de tracción, cantidad de pasajeros, peso).

Para ver una documentación más completa, utilice las funciones `? ó help()`, una vez cargado el paquete. Por ejemplo, `help(mammals)`. Esta práctica se resuelve con el `script_correlacion.R`

Ejercicio 1.1 Mamíferos, Parte I. Usando el conjunto de datos de `mammals`, crear un diagrama de dispersión que muestre cómo el peso del cerebro de un mamífero (`BrainWt`) varía en función de su peso corporal (`BodyWt`).

Ejercicio 1.2 *Medidas del cuerpo, Parte I.* Utilizando el conjunto de datos `bdims`, realizar un diagrama de dispersión que muestre cómo el peso de una persona (`wgt`) varía en función de su altura (`hgt`). Identifique el género de las observaciones en el scatterplot, para ello pinte de rojo a las mujeres y de azul a los hombres, use la instrucción `col` de R. Observar que en esta base de datos, `sex = 1` para los hombres y `sex = 0` para las mujeres.

Ejercicio 1.3 Utilizando el conjunto de datos `smoking`, realizar un diagrama de dispersión que ilustre cómo varía la cantidad de cigarrillos que fuma por día una persona durante el fin de semana (`amtWeekends`), en función de su edad (`age`).

Ejercicio 1.4 Utilizando el conjunto de datos `cars`, realizar un scatter plot del rendimiento del auto en la ciudad (`mpgCity`) en función del peso del auto (`weight`).

Ejercicio 1.5 Para cada uno de los cuatro scatterplots anteriores describa la forma, la dirección y la fuerza de la relación entre las dos variables involucradas. Respuestas posibles:

- forma: lineal, no lineal (cuadrática, exponencial, etc.)
- dirección: positiva, negativa
- fuerza de la relación: fuerte, moderada, débil, no asociación. Tiene que ver con cuán dispersos están las observaciones respecto del patrón descripto en la forma.

Ejercicio 1.6 ¿Para cuáles de los 4 conjuntos de datos tiene sentido resumir la relación entre ambas variables con el coeficiente de correlación muestral de Pearson? Para los casos en los cuales contestó que era apropiado,

- (a) calcúlelo usando R.
- (b) Testee las siguientes hipótesis

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

para cada uno de esos conjuntos. Antes de hacerlo defina a ρ en palabras. Observe que en el ítem 1.6 (a) calculó un estimador de esta cantidad, para cada conjunto. ¿En qué casos rechaza la hipótesis nula, a nivel 0.05?

Ejercicio 1.7 Mamíferos, Parte II. El conjunto de datos de *mammals* presenta un scatterplot que no es razonable resumir con el coeficiente de correlación muestral. El gráfico no es lindo por varios motivos, básicamente las observaciones parecen estar en escalas distintas, hay muchas observaciones superpuestas, necesitaríamos hacer un zoom del gráfico en la zona cercana al origen, a expensas de perder las dos observaciones con valores mucho más grandes que el resto. Podemos comparar lo que pasaría si no hubiéramos observado el diagrama de dispersión y quisieramos resumir los datos con el coeficiente de correlación.

- (a) Calcule el coeficiente de correlación muestral de Pearson para los 62 mamíferos.
- (b) Identifique las dos observaciones que tienen valores de peso corporal y cerebral más grandes que el resto. Realice un scatter plot de las restantes 60 variables. ¿Cómo podría describir este gráfico? Calcule el coeficiente de correlación muestral de Pearson para estas 60 observaciones.
- (c) El gráfico hecho en el ítem anterior no corrige el problema original del todo. La forma general podría describirse como un abanico: claramente las variables están asociadas, la asociación es positiva (ambas crecen simultáneamente) pero la dispersión de los datos parece aumentar a medida que ambas variables aumentan. Esta forma es frecuente en los conjuntos de datos, suelen corresponder a observaciones que están medidas en escalas que no son comparables entre sí y suele corregirse al tomar logaritmo en ambas variables. Para ver el efecto de transformar las variables, realice un scatterplot con todas las observaciones, del logaritmo (en base 10, o en base e) del peso del cerebro en función del logaritmo del peso corporal. Observe el gráfico. ¿Cómo lo describiría? Calcule la correlación de Pearson para los datos transformados.
- (d) Para ambos conjuntos de datos (transformados por el logaritmo y sin transformar) calcule la correlación de Spearman.

Ejercicio 1.8 ¿Con qué coeficiente de correlación, Pearson o Spearman, resumiría los datos de *cars*? (*weight*, *mpgCity*)