Teórica 01

Peredo Leonel

En muchas disciplinas científicas interesa saber cómo se relacionan distintas variables entre sí. Una de las herramientas principales que tiene la estadística para hacer eso es la regresión.
El modelo de regresión lineal es un método conceptualmente simple para investigar la relación entre dos o más variables. Esta relación se expresa en la forma de una ecuación o un modelo que conecta una variable respuesta o variable dependiente (continua) y una o muchas variables explicativas o covariables. Es una técnica clásica y muy utilizada.
La teoría de modelos lineales es un caso especial de la teoría más general que cubre modelos más flexibles y realistas. Precisamente porque es un caso tan especial, permite muchos atajos simplificadores, que pueden facilitar el aprendizaje, especialmente sin matemáticas avanzadas.
Debido a que los modelos lineales son tan simples, han sido y son tremendamente utilizados. Esto significa que muchas aplicaciones de la estadística se ha realizado sobre modelos lineales. También significa que muchos de los consumidores de estadística esperan modelos lineales o compararán los modelos obtenidos con modelos lineales. Por tanto, es importante entender a fondo tanto cómo funcionan, como cuáles son sus limitaciones.
La regresión lineal se ocupa de investigar la relación entre dos o más variables continuas.
Comenzaremos tratando de describir el vínculo entre dos variables aleatorias continuas. Medimos ambas variables en la misma unidad: puede tratarse de un individuo, un país, un animal, una escuela, etc.
Ejemplo: Se miden en el año 2015, para 187 países:
Y: Expectativa de vida: Número promedio de años que un niño recién nacido espera vivir, si los patrones de mortalidad no cambiaran (`life`)
X: Mortalidad infantil: Número de niños de 0 a 5 años que mueren en un año, por cada 1000 niños vivos (`child`)

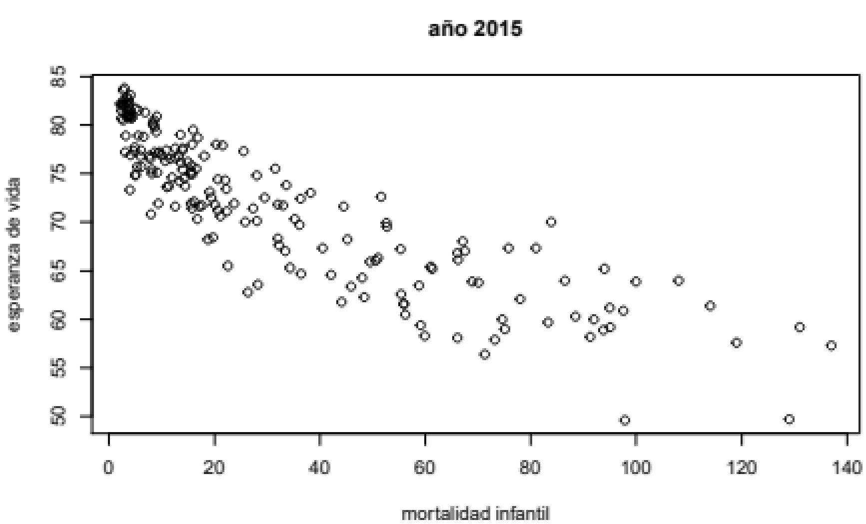


Figure 1: Expectativa de vida en función de Mortalidad en 187 países, año 2015

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, Yi). Se pierde la informaci´on del individuo (pa´ıs) Con este gr´afico podemos determinar si existe alg´un tipo de relaci´on entre X e Y . En este caso vemos que a medida que aumenta la mortalidad infantil, decrece la esperanza de vida. Queremos modelar la relaci´on entre ambas variables. El objetivo es tratar de explicar la esperanza de vida a partir de la mortalidad infantil. Otro ejemplo: Pearson-Lee data. Karl Pearson organiz´o la recolecci´on de datos de 1100 familias en Inglaterra en el per´ıodo 1893 a 1898. Este conjunto de datos en particular: Heights en el paquete alr4 de R da la altura de madres e hijas (en pulgadas), con hasta dos hijas por madre. Todas las hijas tienen 18 a~nos o m´as, y todas las madres son menores de 65 a~nos. En la fuente los datos aparecen redondeados a la pulgada m´as cercana. En la librer´ıa se les agrega un error de redondeo para que el gr´afico no sea discreto. Mostramos datos en cm.

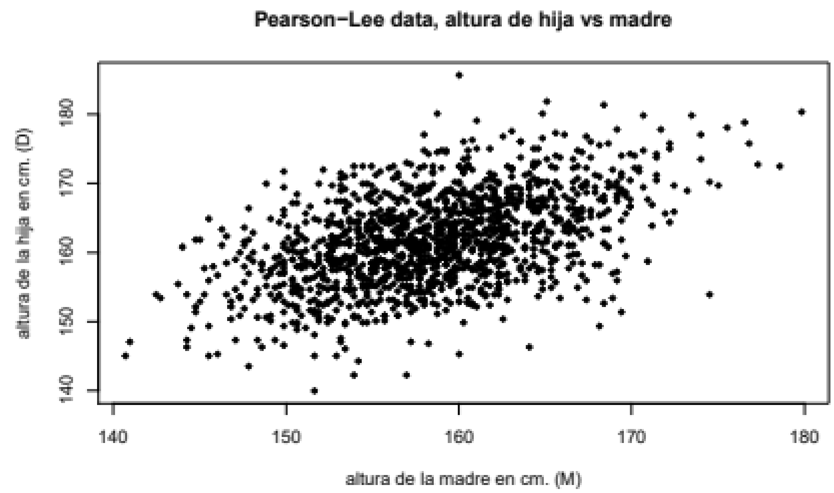


Figure 2: Scatterplot

Ejemplo con más variables: Para n = 193 pa´ıses, medimos en 2015 las siguientes variables: Y = life es la esperanza de vida al nacer (en a~nos) X1 = income: Producto Bruto Interno, per c´apita (en USD) X2 = child Tasa de Mortalidad de 0 a 5 a~nos, por cada mil ni~nos nacidos vivos en el a~no. X3 = dtp3: porcentaje de ni~nos de un a~no inmunizados con tres dosis de vacuna contra la difteria, t´etanos y pertussi (DTP3) X4 = school: n´umero de a~nos de escolaridad promedio en hombres de 25–34 a~nos. X5 = status: grado de desarrollo del pa´ıs (”developed” o ”not.developed”) El objetivo es explicar a Y ¿C´omo lo visualizamos?

Correlación de Pearson

La correlaci´on (poblacional) de un vector aleatorio (X, Y):

FORMULA

Correlación muestral:

FORMULA

FORMULAS DIAPO 14

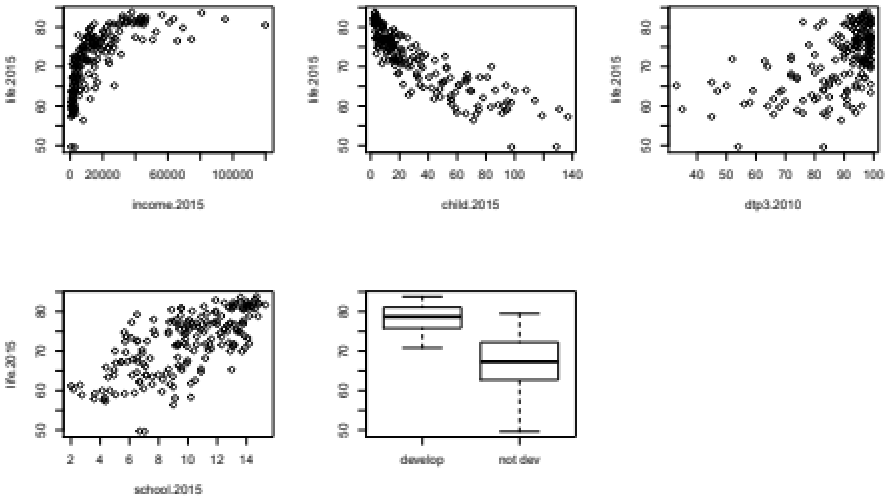
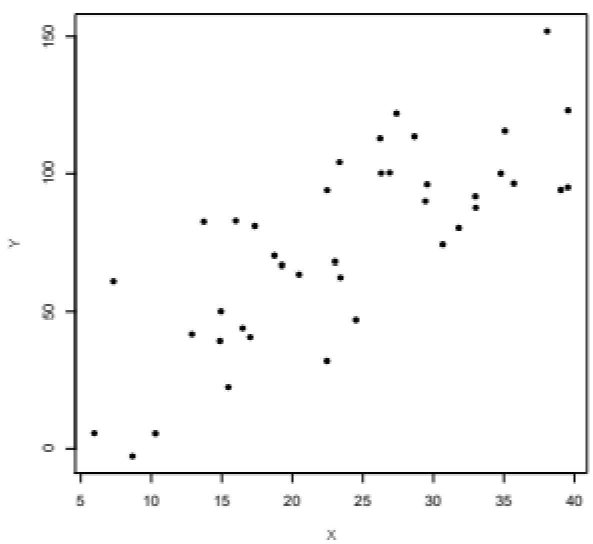
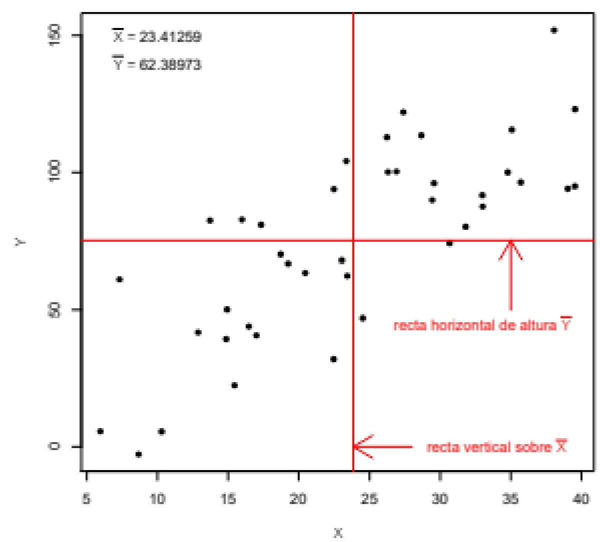
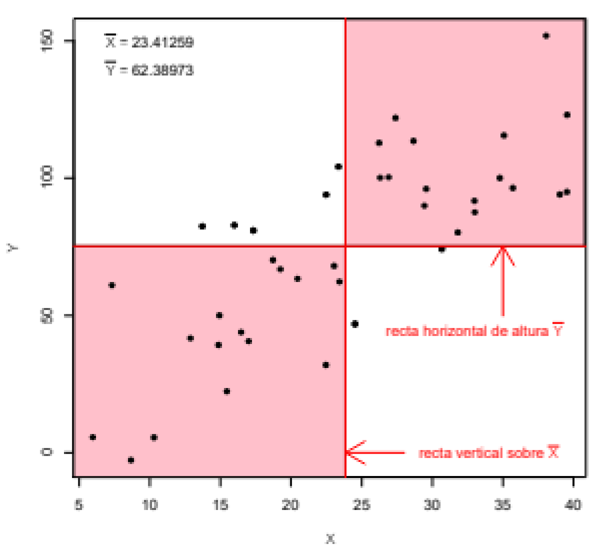


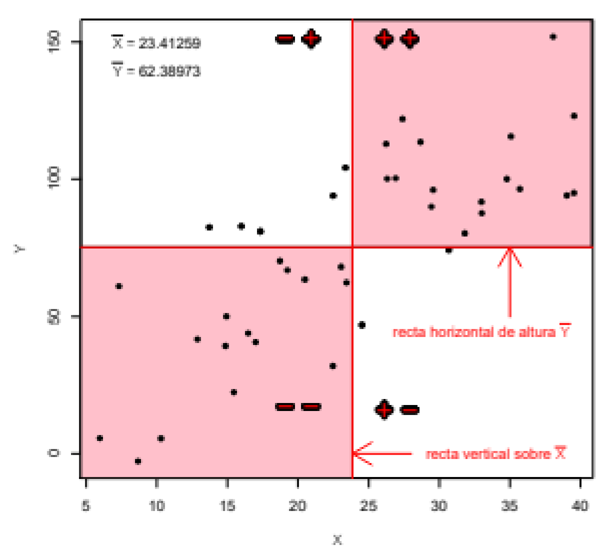
Figure 3: Scatterplot

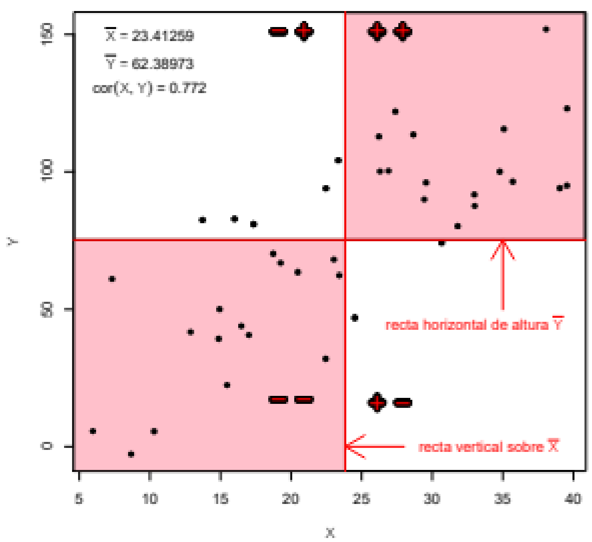
Interpretación de la correlación











Propiedades del coeficiente de correlación

1. -1 leq r leq 1.
2. El valor absoluto de r, |r| mide la fuerza de la asociaci´on lineal entre X e Y , a mayor valor absoluto, hay una asociaci´on lineal m´as fuerte entre X e Y .
3. El caso particular r = 0 indica que no hay asociaci´on lineal entre X e Y .
4. El caso r = 1 indica asociaci´on lineal perfecta. O sea que los puntos est´an ubicados sobre una recta de pendiente (o inclinaci´on) 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 asociaci´on positiva entre las variables (si r > 0); o asociaci´on negativa entre ellas (si r < 0).
7. r = 0,90 indica que los puntos est´an ubicados muy cerca de una recta creciente, r = 0,80 indica que los puntos est´an cerca, pero no tanto, de una recta creciente.
8. r no depende de las unidades en que son medidas las variables (mil´ımetros, cent´ımetros, metros o kil´ometros, por ejemplo) .
9. Los roles de X e Y son sim´etricos para el c´alculo de r.
10. Cuidado: el coeficiente de correlaci´on de Pearson es muy sensible a observaciones at´ıpicas. Hay que hacer siempre un scatter plot de los datos antes de resumirlos con r.

La temperatura corporal de mam´ıferos y p´ajaros tiende a fluctuar durante el d´ıa seg´un un ritmo circadiano regular. En un estudio 1 se registra la temperatura corporal de 10 ardillas ant´ılopes cada 6 minutos a lo largo de 10 d´ıas consecutivos en condiciones de laboratorio. Elegimos una ardilla y promediamos las temperaturas de los 10 d´ıas para obtener un conjunto de datos de 24 x 10 observaciones. Los autores trataban de contestar a la pregunta: ¿Hay una asociaci´on entre la hora del d´ıa y la temperatura corporal? Para contestarla, tenemos dos estrategias. Calcular la correlaci´on entre la hora del d´ıa y la temperatura corporal de la ardilla Graficar ambas variables: horario y temperatura en un scatter plot

Xi = horario de la i´esima medici´on Yi = temperatura promedio a lo largo de 10 d´ıas de las 10 mediciones realizadas en el horario X

Calculamos la correlaci´on muestral: `cor(horario, temperatura6) [1] -0.05863851` Parece no haber relaci´on entre ambas. ¿Eso mide la correlaci´on? Casi no hay relaci´on lineal entre ambas variables.

Una correlaci´on cercana a cero no significa (necesariamente) que las dos variables no est´an asociadas: la correlaci´on mide s´olo la fuerza de una relaci´on lineal

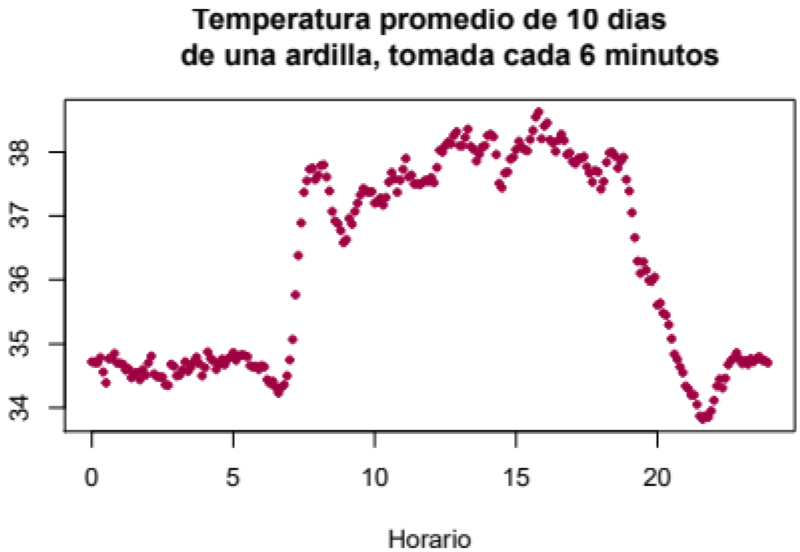
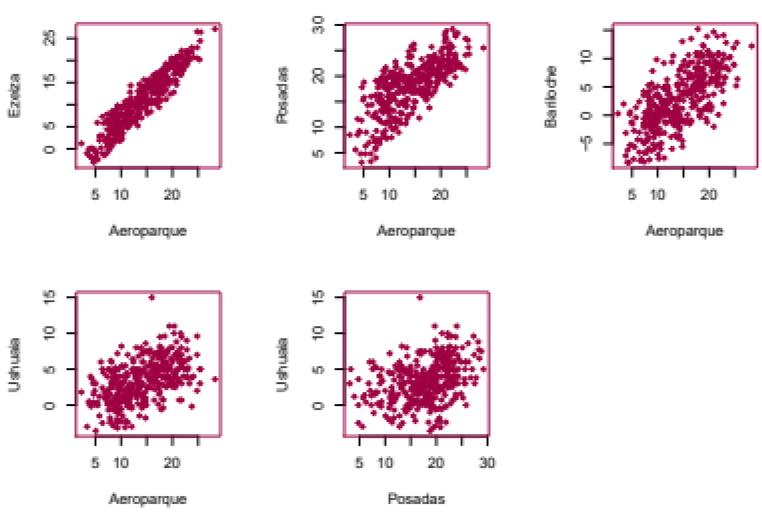
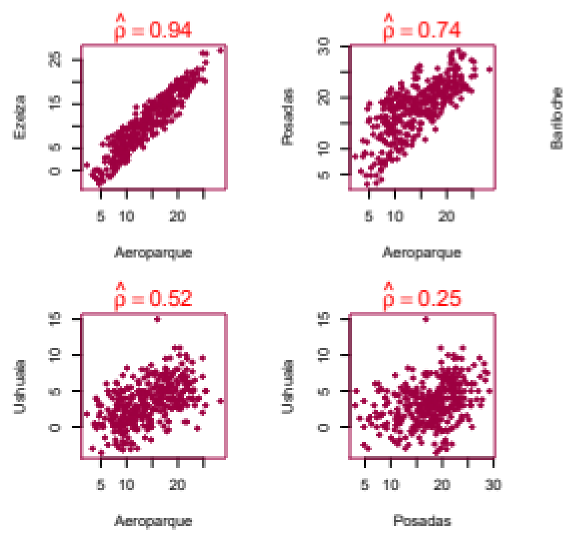
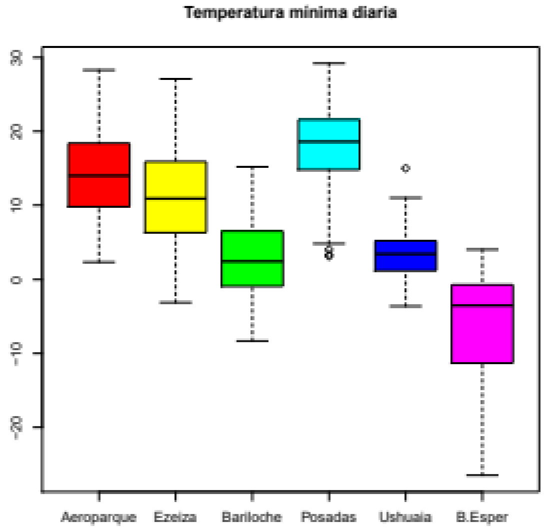


Figure 4: Scatterplot

En el sitio del Servicio Meteorol´ogico Nacional pueden bajarse los datos de temperaturas m´aximas y m´ınimas diarias de los distintos observatorios ubicados en el pa´ıs. 2 Elegimos 5 localidades, queremos ver c´omo se relacionan entre s´ı las temperaturas m´ınimas del mismo d´ıa. As´ı tenemos un vector aleatorio (Ai , Ei , Bi , Pi ,Ui), con 1 LEQ i LEQ n = 365 Ai = temperatura m´ınima del d´ıa i en Aeroparque Ei = temperatura m´ınima del d´ıa i en Ezeiza Bi = temperatura m´ınima del d´ıa i en Bariloche Pi = temperatura m´ınima del d´ıa i en Posadas Ui = temperatura m´ınima del d´ıa i en Ushuaia







La teoría de modelos lineales es un caso especial de la teoría más general que cubre modelos más flexibles y realistas. Precisamente porque es un caso tan especial, permite muchos atajos simplificadores, que pueden facilitan el aprendizaje, especialmente sin matemáticas avanzadas. 2 Debido a que los modelos lineales son tan simples, han sido y son tremendamente utilizados. Esto significa que muchas aplicaciones de la estadística se ha realizado sobre modelos lineales. También significa que muchos de los consumidores de estadística esperan modelos lineales o compararán los modelos obtenidos con modelos lineales. Por tanto, es importante entender a fondo tanto cómo funcionan como cuáles son sus limitaciones.

El modelo de regresión lineal es un modelo para el vínculo de dos variables aleatorias que denominaremos X = variable predictora o covariable e Y = variable dependiente o de respuesta. El modelo lineal (simple pues sólo vincula una variable predictora con Y) asume que 1 La distribución de X no está especificada, incluso puede ser determinística. 2 Proponemos el siguiente modelo para las variables: FORMULA donde E es el término del error. 3 Asumimos que la variable aleatoria error E tiene esperanza 0, varianza constante desconocida que llamaremos SIGMA 2 , no está correlacionado con X y no está correlacionado con los errores de otras observaciones. En el modelo (2) los números BETA0 y BETA1 son constantes desconocidas que se denominan parámetros del modelo, o coeficientes de la ecuación. Los parámetros se denominan BETA0 = ordenada al origen BETA1 = pendiente. El supuesto de la relación funcional entre X e Y sea lineal es no trivial, ya dijimos que muchas variables no lo cumplen. El requisito de que el error tenga varianza constante, se lo suele llamar homoscedasticidad, y tampoco es no trivial. Lo mismo pasa con las no correlaciones. Pero el supuesto de que los errores tengan esperanza cero sí es trivial. Verificar que los supuestos se cumplan para un conjunto de datos será uno de los objetivos que atacaremos más adelante en la materia.

FORMULAS DIAPO 32, 33 y 34