OCW-UPM Estadística para Ingeniería Civil y Medioambiental

Autores: E. M. García del Toro, C. Hermoso, E. J. Huertas

LABORATORIO DE ESTADÍSTICA CON MATLAB

PRÁCTICA 2 - Regresión lineal con MATLAB

Obtención de las rectas de regresión lineal Y/X e X/Y con MATLAB.

En esta práctica estudiaremos las distribuciones bidimensionales, es decir, aquellas en las que se observan a la vez dos características diferentes de un mismo fenómeno físico. Cada dato entonces consta de un par de valores o coordenadas (x_i, y_i). Usualmente, el conjunto de observaciones se disponen en dos filas o columnas, de modo que una contiene los datos de la variable X y la otra los de la variable Y. La importancia de las distribuciones bidimensionales radica en investigar como influye una variable sobre la otra. Esta puede ser una dependencia causa efecto, por ejemplo, la cantidad de lluvia (causa), da lugar a un aumento de la producción agrícola (efecto). O bien, el aumento de la tensión sobre un material, da lugar a una disminución del tiempo que tarda en fracturarse.

Si utilizamos un sistema de ejes cartesianos para representar los datos, obtendremos un conjunto de puntos conocido como **diagrama de dispersión**, cuyo análisis permite estudiar cualitativamente la relación entre ambas variables. El siguiente paso, es la determinación de la relación funcional entre las dos variables X e Y que mejor ajusta a los datos. Se denomina **regresión lineal** cuando dicha función es lineal, es decir, requiere la determinación de dos parámetros: la pendiente b y la ordenada en el origen a de la **recta de regresión**, $y = bx + a$.

La regresión nos permite además, determinar el grado de dependencia de las series de valores X e Y, estimando el valor y_* que se obtendría para un valor pedido x_* que no esté en la distribución original.

Para una abcisa dada x_i, se denomina **residuo** e_i a la diferencia entre el valor real observado y_i, y el valor estimado o predicho por la recta de regresión $\hat{y}_i = bx_i + a$, es decir

$$e_i = y_i - \hat{y}_i = y_i - (bx_i + a).$$

A modo de resumen, la **recta de regresión Y/X** (lo anterior se lee: "recta de regresión de Y sobre X") viene dada por la ecuación explícita $y = bx + a$, donde

$$a = \bar{y} - b\bar{x}, \quad b = \frac{S_{XY}}{S_X^2}. \tag{1}$$

Es aquella que cumple la propiedad de minimizar la suma de los cuadrados de los residuos e_i en la variable Y, y se utiliza para para estimar los valores de la Y a partir de los de la X.

Por otro lado, si lo que queremos es estimar los valores de la X a partir de los de la Y, necesitaremos una recta de regresión distinta (en principio) de la anterior, que verifique la propiedad de minimizar la suma al cuadrado de los residuos en la variable X, a saber, para una y_i dada, $e_{x,i} = x_i - \hat{x}_i = x_i - (\tilde{b}y + \tilde{a})$. Esta recta se denomina **recta de regresión X/Y** (es decir, "recta de regresión de X sobre Y"), y su ecuación explícita es $x = \tilde{b}y + \tilde{a}$, donde

$$\tilde{a} = \bar{x} - \tilde{b}\bar{y}, \quad \tilde{b} = \frac{S_{XY}}{S_Y^2}. \tag{2}$$

Ambas rectas se cortan siempre en el punto $(\bar{x}, \bar{y})$. ¿Existe la posibilidad de que ambas rectas se corten en más puntos? ¿Bajo qué condiciones? ¿Qué ocurre en este caso? **(Trabajo para el alumno)**

Ejercicio. Análisis de la calidad de las aguas de un embalse.

En un análisis de las aguas de cierto embalse, se obtienen los siguientes valores de *concentración de sólidos suspendidos* (medida en mg/l) y de *turbidez* del agua (medida en Unidades Nefelométricas de Turbidez, o Nephelometric Turbidity Unit (NTU))

turbidez (NTU)	95	100	102	104	100	98	96	100	110	99
sólidos suspendidos (mg/l)	85	94	84	88	85	92	76	90	102	89

Con ayuda de MATLAB, responda a las siguientes preguntas:

a) Obtenga la el diagrama de dispersión tomando y como ordenadas, y la ecuación de la recta de mínimos cuadrados que expresa la concentración de sólidos suspendidos en función de la turbidez del agua. Póngala en forma explícita $y = bx + a$ y en forma general $Ax + By = C$.

b) Obtenga el valor predicho de la concentración de sólidos suspendidos para una turbidez de 95 NTU ¿Cuánto vale el residuo de esta estimación?

c) Obtenga el valor predicho de la concentración de sólidos suspendidos para una turbidez de 103 NTU ¿Cuánto vale el residuo de esta estimación? ¿Tiene sentido plantearse esta pregunta? ¿Porqué?

d) Obtenga el valor predicho de la concentración de sólidos suspendidos para una turbidez de 20 NTU ¿Tiene sentido plantearse esta pregunta? ¿Porqué?

e) ¿En qué porcentaje la variación en la concentración de sólidos suspendidos es explicada por la variación en la turbidez del agua? ¿Cuál es el coeficiente de correlación de Pearson? Interprétalo.

f) Si en cierto punto del embalse la concentración de sólidos suspendidos presenta un valor de 86 mg/l, ¿qué valor de turbidez puede esperarse en esa zona? (*ver Nota abajo antes de realizar este apartado*). Póngala en forma explícita $x = \tilde{b}y + \tilde{a}$ y en forma general $\tilde{A}x + \tilde{B}y = \tilde{C}$.

g) Resuelva el sistema formado por las ecuaciones en forma general obtenidas en los apartados a) y f). ¿Qué valores espera obtener en la solución de este sistema lineal?

Nota: Observa que, para obtener la estimación pedida en el apartado d), se necesita una recta de regresión diferente a la calculada en el apartado a). Es decir, hay que obtener en primer lugar la recta que permite predecir los valores de turbidez del agua **en función de** la concentración de sólidos suspendidos en la misma.

Solución. Primero introducimos los pares de datos en el workspace o memoria de MATLAB

```
>> clear all;clc;
   x=[95 100 102 104 100 98 96 100 110 99];
   y=[85 94 84 88 85 92 76 90 102 89];
```

a) El diagrama de dispersión pedido se obtiene mediante los comandos

```
>> plot(x,y,'ro','markersize',5,'markerfacecolor','r')
   hold on
   xlabel('X:Turbidez del agua, en NTU');
   ylabel('Y:Concentracion de solidos suspendidos, en mg/l');
   title('Regresión lineal Y/X');
   grid on
```

Para sacar la gráfica del dibujo de la recta Y/X, en la ventana que contiene la imagen ir a Tools → Basic fitting, marcar la casilla *linear* y marcar la flecha grande para ver los detalles de la recta. Ésta aparecerá dibujada junto al diagrama de dispersión. Numéricamente, los coeficientes b y a de la recta de regresión Y/X vienen dados por

```
>> [r,b,a] = regression(x,y) %r es el coeficiente de correlacion de Pearson
r=0.6933
b=1.1192
a=-23.8698
```

Luego $y = 1.1192x - 23.8698$, y en forma general $1.1192x - y = 23.8698$.

Podemos calcular también lo mismo pero de otra forma, usando las expresiones (1) y el comando (**ejercicio para el alumno**)

```
>> cov(x,y,1) %es la matriz de varianzas-covarianzas.
```

Se tiene que

$$cov(x, y, 1) = \begin{pmatrix} S_X^2 & S_{XY} \\ S_{XY} & S_Y^2 \end{pmatrix} = \begin{pmatrix} 16.4400 & 18.4000 \\ 18.4000 & 42.8500 \end{pmatrix}$$

b) Sustituimos $x = 95$ en la recta anterior, y obtenemos $1.1192 \cdot 95 - 23.8698 = 82.454$. Como hay un punto con abcisa exactamente 95 NTU, podemos calcular el residuo correspondiente

$$\begin{aligned} e &= 85 - (1.1192 \cdot 95 - 23.8698) \\ &= 85 - 82.454 = 2.546. \end{aligned}$$

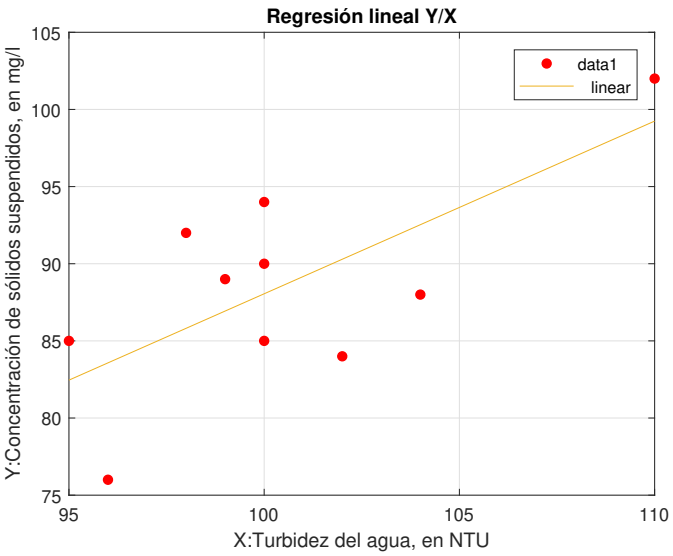

Figure 1: Recta de regresión Y/X entre turbidez del agua y concentración de sólidos suspendidos

c) El valor predicho de Y en este caso es $1.1192 \cdot 103 - 23.8698 = 91.408$, pero como no hay ningún dato que tenga de abcisa 103, no hay valor "real" de la Y en este caso, y no podemos calcular el residuo. No tiene sentido aquí preguntar por el residuo.

d) Aquí, el valor estimado de la concentración de sólidos en suspensión es de $1.1192 \cdot 20 - 23.8698 = -1.4858$ mg/l. **No** tiene sentido un valor **negativo** de la concentración de sólidos en suspensión. Esto es debido a que estamos pidiendo predicciones fuera del rango donde nuestra recta Y/X tiene validez, es decir, muy alejados del rango de valores de la variable X.

e) Nos piden el coeficiente de determinación (utilizar por ejemplo la fórmula siguiente, o la salida r del apartado a))

$$R^2 = r^2 = \frac{S_{XY}^2}{S_X^2 \cdot S_Y^2} = \frac{18.4000^2}{16.4400 \cdot 42.8500} = 0.4806.$$

Luego la variación en la concentración de sólidos suspendidos es explicada en un 48% por la variación en la turbidez del agua. Por otro lado, $r = \sqrt{0.4806} = 0.6933$ por lo que existe una relación directa (a mayores valores de X, valores crecientes de Y) y moderadamente fuerte entre las variables X e Y.

f) El nuevo diagrama de dispersión se obtiene mediante los comandos

```
>> plot(y,x,'b*','markersize',5,'markerfacecolor','b')
   hold on
   xlabel('Y:Concentracion de solidos suspendidos, en mg/l');
   ylabel('X:Turbidez del agua, en NTU');
   title('Regresion lineal X/Y');
   grid on
```

los coeficientes $\tilde{b}$ y $\tilde{a}$ de la recta de regresión Y/X vienen dados por

```
>> [rw,bw,aw] = regression(y,x) %rw debe ser el mismo que en el apartado a)
rw=0.6933
bw=0.4294
aw=62.3977
```

El coeficiente rw debe tener el mismo valor que el r en el apartado a) ¿Por qué?

Podemos calcular también los mismos coeficientes usando las expresiones (2) y la matriz de varianzas-covarianzas que obtuvimos en el apartado a) **(ejercicio para el alumno)**.

g) El sistema que nos piden es el formado por las rectas de regresión X/Y e Y/X, luego la solución de dicho sistema debe ser el punto $(\bar{x}, \bar{y})$, donde se cortan siempre estas rectas. Así tenemos

$$\begin{aligned} 1.1192x - y &= 23.8698 \\ x - 0.4294y &= 62.3977 \end{aligned}$$

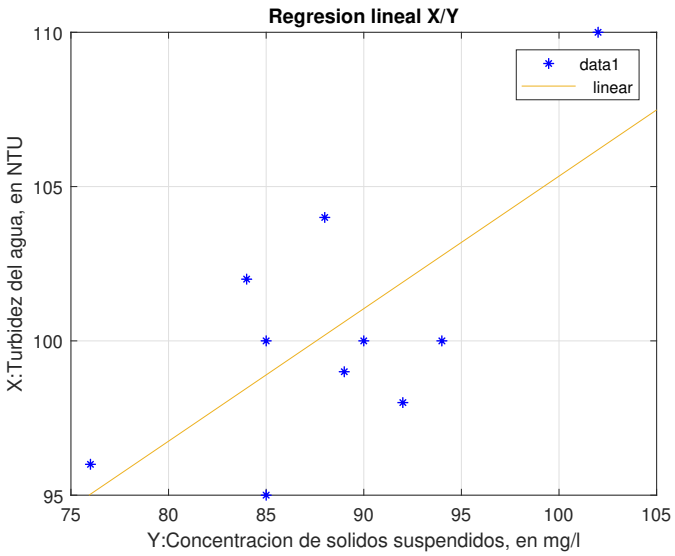

Figure 2: Recta de regresión X/Y entre concentración de sólidos suspendidos y turbidez del agua

para resolverlo en MATLAB escribimos la matriz de coeficientes y de terminos independientes de la forma

```
>> A=[1.1192,-1;1,-0.4294]
>> p=[23.8698;62.3977]
>> A\p
ans =100.3975
      88.4951
```

Comprobar lo anterior introduciendo

```
>> mean(x)
>> mean(y)
```

Trabajo del alumno 1. Resolver de forma semejante al ejercicio de la práctica

Se administra un larvicida volátil en una balsa de riego para cultivos destinados a consumo humano. A continuación se determinan las concentraciones de larvicida (en μg/ml) en relación al paso del tiempo (en horas) y los resultados son los siguientes:

tiempo (h)	1	1.5	2	3	6	15
concentración (μg/ml)	11.8	11.0	10.9	10.1	9.6	5.7

a) Dibuja el diagrama de dispersión (nube de puntos) de los datos anteriores.

b) Determina, a partir de la forma de la nube, si el modelo de regresión lineal es el adecuado. En caso afirmativo, obtén la expresión matemática que relaciona la concentración de larvicida en el agua en función del tiempo.

c) Estima el valor de la concentración a las 9 horas.

d) Calcula el coeficiente de correlación e interprétalo.

e) Se sabe que el agua tratada con este larvicida puede ser destinada a riego siempre que no contenga una concentración superior a 2.5 μg/ml. En base a la información de que dispones, ¿crees que sería adecuado regar con el agua de la balsa pasadas 24 horas desde la administración del larvicida?

Indicaciones y Comandos de MatLab

```
clear all;clc;
x=[95 100 102 104 100 98 96 100 110 99];
y=[85 94 84 88 85 92 76 90 102 89];

[r,b,a] = regression(x,y)

yr=b*x+a;
plot(x,y,'ro','markersize',5,'markerfacecolor','r')
hold on
plot(x,yr)
xlabel('X:Turbidez del agua, en NTU');
ylabel('Y:Concentración de sólidos suspendidos, en mg/l');
title('Regresión lineal Y/X');
grid on



[rw,bw,aw] = regression(y,x)

x=[95 100 102 104 100 98 96 100 110 99];
y=[85 94 84 88 85 92 76 90 102 89];

plot(y,x,'b*','markersize',5,'markerfacecolor','b')
grid on
xlabel('Y:Concentración de sólidos suspendidos, en mg/l');
ylabel('X:Turbidez del agua, en NTU');
title('Regresión lineal X/Y');


cov(x,y)
```