Salvador Macías Hernández
Rodolfo Gómez Vega

MÍNIMOS CUADRADOS

27/10/2003 **Tarea No.9**

INTRODUCCIÓN. ¿Qué es regresión (regresión lineal)? La técnica tiene que ver con los métodos para volver numérica cualquiera de estas nociones:

- Asociación. Es el concepto más simple de los tres; se refiere a si los valores de dos variables parecen aparearse de una manera consistente (valores grandes de una variable con valores grandes de la otra o viceversa).
- Dependencia. Puede querer decir, y éste es su significado común en matemática y en física, dependencia exclusiva es decir si se da el valor ``x" a una variable se sigue que la variable ``y" toma tal otro valor. También puede querer decir ausencia de independencia probabilística como cuando se dice ``la temperatura del agua depende de la distancia desde el grifo". (también depende de otras cosas).
- Causalidad. Es el concepto más fuerte y requiere que la relación posea varios atributos.
  - Consistencia en la relación. Es decir que la relación se presente en diferentes niveles de otras variables y siempre con la misma intensidad y en el mismo sentido.
  - Repetibilidad. Que cuando se intervenga y se cambie el valor de x, inmediatamente cambie de una forma acorde el de y.
  - Un mecanismo. Un mecanismo inteligible que nos lleve paso a paso de x a y.

Significados de la regresión

Dos significados de regresión: esperanza condicional y ajuste de una función.

Propósitos de la regresión.

Como resumen. Dice con una fórmula rápidamente la relación numérica entre x e y. Lo cual sirve para enfocar un análisis causal: Para medir el tamaño de un efecto. Como al decir en cuánto cambia y al cambiar una unidad a x. Un caso extremo de causal es cuando queremos establecer una ley empírica. Para predicción. Como cuando nos preguntamos el valor de y cuando x tenga tal otro valor. Para quitar el efecto de una variable y seguir estudiando qué otras variables afectan (están relacionadas) con la y.

Dependiendo del propósito que tengamos en mente para la regresión nos interesará un buen ajuste o conocer los parámetros con precisión o sólo saber si las variables que usamos son ``buenas''. Revisamos esto para cada caso.

Regresión como esperanza condicional

Llamamos esperanza condicional de y dado un valor fijo para x al valor promedio de y cuando x toma un valor fijo.

Para calcular esta esperanza condicional veamos varios planes:

Desde el punto de vista empírico, idealmente, tendríamos grandes cantidades de observaciones de las variables. Partiríamos el eje x, en intervalos pequeños y para cada intervalo calcularíamos la media. La gráfica resultante de ese proceso sería la regresión.

Note que todavía sería mejor tener una regresión para varios percentiles, no sólo para la media, de modo que tuvieramos una idea mejor de la distribución condicional de y dado x.

En la práctica es muy difícil que tengamos número suficientemente grande de observaciones para hacer realidad el plan guajiro delineado aquí arribita. Usualmente tenemos (o tomamos) escasas observaciones (quizás una para cada x con las equis muy separadas). A veces se hace un suavizado de los datos de la variable ``y'' obteniendo una curva que puede ser suave pero de forma matemática complicada.

Otra forma de calcular requiere encontrar una expresión matemática sencilla que aproxime la curva resultante de la esperanza condicional.

La última forma de proceder en esta situación, en mayor detalle es así: después de haber visto que varios suavizados sugieren la misma forma funcional para la regresión (la gráfica de la esperanza condicional vs x), decidimos que hay una forma para la curva de regresión (recta, parábola, exponencial o lo que sea) y ajustamos una tal curva a los datos. Usualmente la curva tiene pocos parámetros y seleccionamos los valores de éstos con maña tal que se satisfaga algún criterio (mínimos cuadrados, mínimos valores absolutos, o cualquier otro).

Ajuste de una función (recta)

La última de las formas de calcular mencionadas en el párrafo anterior, tiene las siguientes notas: muchas veces se escoge la forma de la regresión, no por un conocimiento genuino sino por un desconocimiento, más o menos, criminal. a veces esa forma funcional viene de alguna consideración de la tesoría económica. a veces se usa un argumento de diferenciabilidad de la curva de la regresión o algo parecido. en esta forma de resolver el problema, nadie puede creer que la forma funcional resultante sea la ``correcta'', simplemente se quiere que sea útil. puede haber varias formas funcionales que ajusten bien.

Para hacer un ajuste de un modelo lineal, partimos de unos datos acomodados en una tabla:
una columna de n valores de y.
una columna de unos.
una (o más) columnas de x.
Un ajuste es una combinación lineal de las columnas de x. Los coeficientes de la combinación lineal son los coeficientes de la ecuación. Qué tan bueno o malo es un ajuste se puede medir en términos de los residuos que deja. Una manera ``natural'' de medir estos residuos consiste en considerar el (cuadrado) de la norma euclideana del vector de residuos. Esto se llama mínimos cuadrados.

¿CÓMO HACEMOS LOS MÍNIMOS CUADRADOS?

Partimos de un modelo de regresión simple Y=f(X). Y es la variable dependiente o explicada (endógena) y X es la variable independiente o explicativa (exógena). Si el modelo es lineal se puede escribir en la forma Y= b0+b1X+u dónde u es una variable estocástica denominada perturbación, estableciéndose diversas hipótesis sobre su comportamiento. Se supone que u tiene una influencia puramente aleatoria sobre la variable explicada. A partir de una muestra determinada se puede obtener una estimación de los parámetros b0, b1
Y= ^b0+^b1X+ûi dónde û son los residuos, es decir la diferencia entre el valor observado de la variable que tratamos de explicar y el valor ajustado de la misma.
Se hacen las siguientes hipótesis con respecto a los términos de perturbación

- E(ui)=0
- E(ui)2=E(uj)2=s2 lo que es indicativo de homocedasticidad es decir, las observaciones de Y que corresponden a los diferentes valores de X tienen la misma varianza.
- E(ui,uj)=0 no existe autocorrelación en los residuos.
- E(ui,xi)=0 es decir la variable independiente se mide sin errores de observación o no es estocástica o es exógena.

La estimación es ^b0=Y-^b1X^b1=Sxy/S2x
Sea una colección de n datos, de los cuales sospechamos que al representarlos gráficamente , tratan de orientarse en línea recta. Y que cada dato difiere de la recta en δ. El error al cuadrado esta dado por:

$$(\delta y_i)^2 = \left[y_i - (mx_i + b)\right]^2$$

$$M = \sum_i (\delta y_i)^2 = \sum_i {y_i}^2 + m^2 \sum_i {x_i}^2 + nb^2 + 2mb \sum_i x_i - 2m \sum_i x_i y_i - 2b \sum_i y_i$$

y como función mínima se debe de hacer:

$$\frac{\partial M}{\partial m} = 0 \;\; ; \;\; \frac{\partial M}{\partial b} = 0$$

$$m = \frac{n\sum_i (x_i y_i) - \sum_i x_i \sum_i y_i}{n\sum_i x_i^{\,2} - \left(\sum_i x_i\right)^2}$$

$$b = \frac{\sum_i x_i^{\,2} \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{n\sum_i x_i^{\,2} - \left(\sum_i x_i\right)^2}$$

Y un método mejor es el siguiente:
Los puntos P(1, ½), Q(2,1), R(3, 3/2), están situados sobre la recta cuya ecuación es

$$y = ½ x$$

Por dicha razón, no es posible hallar una recta que pase por P(1, ½), Q(2,1) y R(3,2), ya que la recta que pasa por P y Q es la citada anteriormente y el único punto de dicha recta de abscisa 3 es R(3,3/2) y R(3,2).
Por supuesto, que al colocar valores a las variables x e y, en una ecuación de la forma y = mx + b, a partir de la tabla:

X	Y
1	1/2
2	1
3	2

Llegaremos al sistema inconsistente:

$$m + b = ½$$
$$2m + b = 1$$
$$3m + b = 2$$

La ecuación normal para tal sistema $\mathbf{A^T A x = A^T B}$, es:

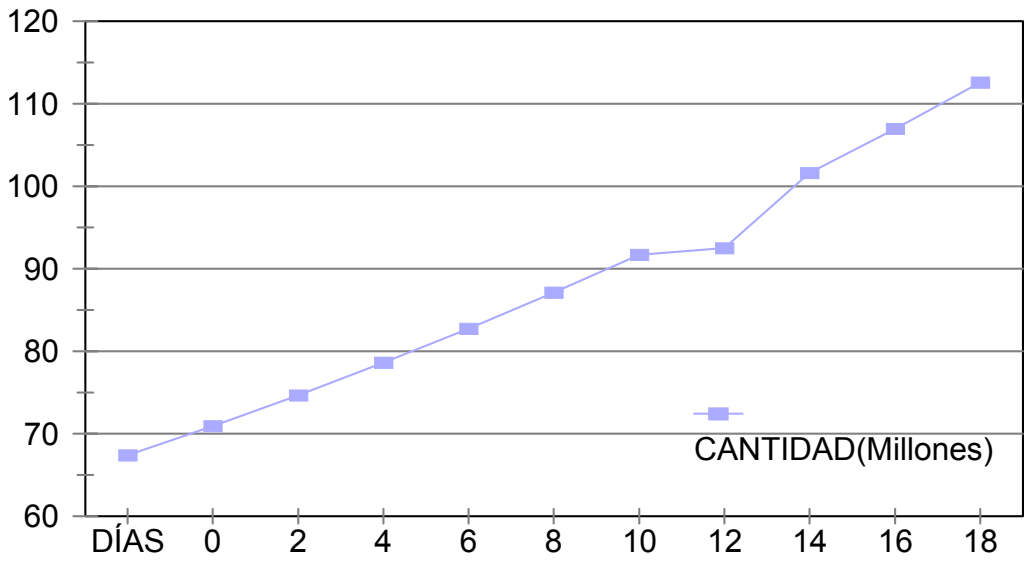
$$\begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} ½ \\ 1 \\ 2 \end{bmatrix}$$

tal sistema tiene solución única m = ¾ , b= -1/3

La ecuación de la recta sería: y = ¾ x - 1/3

PROBLEMA En un laboratorio farmacéutico que desarrolla antibióticos, es importante determinar el crecimiento de poblaciones de cepas de bacterias al transcurrir el tiempo. Los siguientes datos representan el crecimiento bacteriano en un caldo de cultivo, durante varios días:

DÍAS	CANTIDAD(Millones)
0	67.38
2	70.93
4	74.67
6	78.6
8	82.74
10	87.1
12	91.69
14	92.51
16	101.6
18	106.95
20	112.58



Aplicando el segundo método:

$$A = \begin{pmatrix} 0 & 1 \\ 2 & 1 \\ 4 & 1 \\ 6 & 1 \\ 8 & 1 \\ 10 & 1 \\ 12 & 1 \\ 14 & 1 \\ 16 & 1 \\ 18 & 1 \\ 20 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 67.38 \\ 70.93 \\ 74.67 \\ 78.6 \\ 82.74 \\ 87.1 \\ 91.69 \\ 92.51 \\ 101.6 \\ 106.95 \\ 112.58 \end{pmatrix}$$

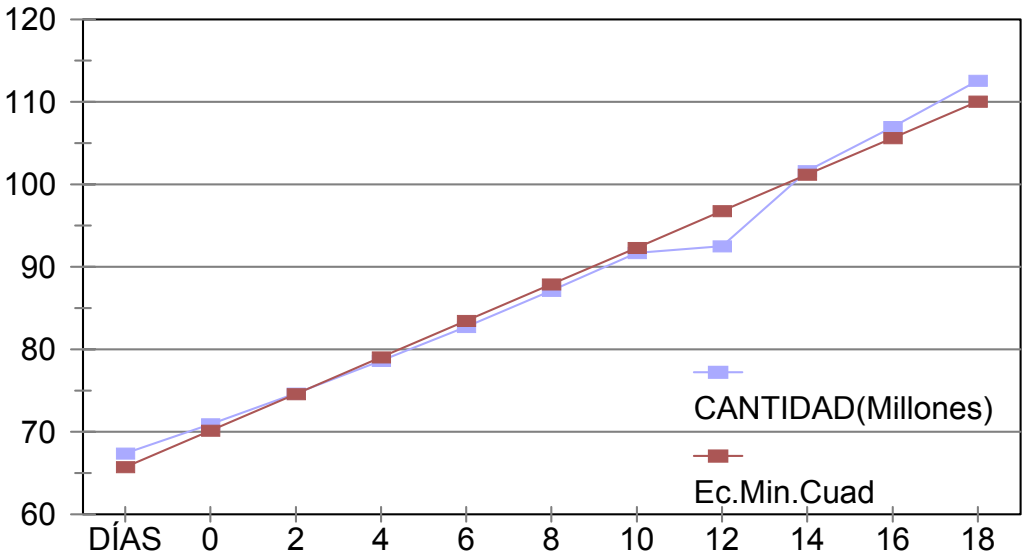
$$X = (A^T A)^{-1}(A^T B)$$

$$X = \begin{pmatrix} 2.21654545454545015` \\ 65.7209090909091475` \end{pmatrix}$$

por tanto la ecuación es:

$$y = 2.21655x + 65.7209$$

y la predicción para el tiempo 30 es de 132.2174



PROGRAMA

```
'------------------------------------------------------------
'PROGRAMA PARA RESOLVER PROBLEMA DE PAGINA DE CURSO  (MINIMOS CUADRADOS)
'------------------------------------------------------------
LET n = 0
LET Xf = 0
LET Yf = 0
LET XY = 0
LET X2 = 0

INPUT "Cuantos valores son? ", n
FOR MI = 1 TO n
 INPUT "Dame el valor de X ", X
 INPUT "Dame el valor de Y ", Y
 XY = XY + X * Y
 Xf = Xf + X
 Yf = Yf + Y
 X2 = X2 + X * X
NEXT

M = (n * XY - Xf * Yf) / (n * X2 - Xf * Xf)
B = (X2 * Yf - Xf * XY) / (n * X2 - Xf * Xf)
PRINT "M=", M
PRINT "B=", B
```

BIBLIOGRAFÍA

Análisis Numérico
Richard L. Buden; J. Douglas Faire
Numerical methods for scientists and engineers
R.W. Hamming
Metodos Numericos y Programacion Fortran
Mccraken, Daniel D
Qbasic, del menú de ayuda
Microsoft
http://www.geocities.com/josearturobarreto/capitulo9.htm